



Ss. Cyril and Methodius University in Skopje
**FACULTY OF COMPUTER
SCIENCE AND ENGINEERING**



2022

Proceedings of the 19th International Conference for Informatics and Information Technology

**Held at the Faculty of Computer Science and Engineering,
Skopje, North Macedonia
5-6th May, 2022**

Editors:

**Katarina Trojchanec Dineva
Vladimir Zdraveski**

ISBN 978-608-4699-15-6

Conference for Informatics and Information Technology 2022

Website: <http://ciit.finki.ukim.mk>

Email: ciit@finki.ukim.mk

Publisher:

Faculty of Computer Science and Engineering, Skopje, N. Macedonia,

Ss. Cyril and Methodius University in Skopje, N. Macedonia

Address: Rugjer Boshkovikj 16, P.O. Box 393, 1000 Skopje, N. Macedonia

Website: <http://www.finki.ukim.mk/>

Email: contact@finki.ukim.mk

Proceedings Editors:

Katarina Trojchanec Dineva

Vladimir Zdraveski

Technical editing: Stefan Andonov, Jana Kuzmanova, Vlatko Spasev, Dimitar Kitanovski,
Jovana Dobрева, Katarina Trojchanec Dineva and Vladimir Zdraveski

Cover page: Vangel Ajanovski

Total print run:150

Printed in Skopje, N. Macedonia, 2022

ISBN: 978-608-4699-15-6

CIP - Каталогизација во публикација

Национална и универзитетска библиотека "Св. Климент Охридски Скопје

004.7:621.39(062)

004(062)

CONFERENCE for Informatics and Information Technology (19 ; 2022 ; Skopje) Proceedings of the 19th Conference for Informatics and Information Technology, held at the Faculty of Computer Science and Engineering, Skopje, North Macedonia 5-6th May, 2022 / editors Katarina Trojchanec Dineva, Vladimir Zdraveski. - Skopje : Faculty of computer science and engineering, 2022

Начин на пристапување (URL):

<http://ciit.finki.ukim.mk/proceedings/ciit-2022-proceedings.pdf>.

- Текст во PDF формат, содржи XI, 163 стр., илустр. - Наслов преземен од екранот.

- Опис на изворот на ден 26.12.2022. - Библиографија кон трудовите

ISBN 978-608-4699-15-6

а) Информациско-комуникациски технологии – Собири

б) Компјутерски науки – Собири

COBISS.MK-ID 59035141

Preface

This volume contains the papers presented at the 19th International Conference for Informatics and Information Technology (CIIT 2022), which was held as a hybrid event on May 5-6, 2022. The conference was organized by the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Republic of North Macedonia.

As the COVID-19 pandemic was still active, CIIT 22 was organized as a hybrid conference. However, despite the chance to meet each other in person, this was also an opportunity for the conference to have increased attendance and for the authors to be able to present their work to a broader audience online. The conference was publicly available to everyone interested. Likewise, we had students and staff of the Faculty of Computer Science and Engineering, as well as attendees from other companies and institutions. Additionally, we had participants at the conference from neighboring countries. In general, we had a lot of excellent submissions from the country and the region.

The aim of the 19th edition of the CIIT conference, as in previous editions, was to provide an opportunity for young researchers from the country and abroad to present their work to a broader research community and get valuable feedback on their work. Regardless of all the limitations and circumstances that we faced due to the pandemic, we had a rigorous review process, and the quality, we believe, did not fall behind. For the 19th edition of the conference, we had 43 paper submissions, of which 17 were accepted as full papers, 2 short papers, and 14 student papers. During the conference, we had 32 presentations organized in 4 paper sessions and 2 student sessions. Additionally, we had 3 keynote lectures and 3 workshops as part of the conference.

This year we continued the tradition of this conference and awarded our hard-working students. However, the grading considered several aspects of each paper: technical quality, scientific contribution, and oral presentation. Each member of the evaluation commission graded each paper regarding those qualities, and we also considered the reviewer grades and comments. In the end, we had a difficult task, but we finally reached a consensus and decided to give two awards, one for the best student paper and one for the best full paper.

We tried to take advantage of both ways of holding the conference (physical and online) and hopefully we really managed to have more of the advantages and less of the disadvantages. The hybrid format of the conference allowed the participants to attend in a preferred way (with physical presence or online) all the talks covering a diverse range of topics. We are proud to have participants from almost all academic institutions in Macedonia and several from the state public institutions and business sector as well as from abroad. We had the pleasure of hosting three invited speakers. Our first keynote speaker was Sashko Ristov, PhD, a postdoctoral university assistant at the University of Innsbruck, Austria. His research interests include performance modeling and optimization of distributed systems, in particular workflow applications and serverless computing.

Our second keynote speaker was Tome Eftimov, PhD, and Gjorgjina Cenikj. Tome Eftimov is a senior researcher at the Computer Systems Department at the Jožef Stefan Institute. He is a visiting assistant professor at the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje. He was a postdoctoral research fellow at the Stanford University, USA, where he investigated biomedical relations outcomes by using AI methods. In addition, he was a research associate at the University of California, San Francisco, investigating AI methods for rheumatology concepts extraction from electronic health records. His research interests

include statistical data analysis, metaheuristics, natural language processing, representation learning, and machine learning.

Gjorgjina Cenikj is a student researcher at the Computer Systems Department at the Jožef Stefan Institute. She is currently pursuing a Masters degree in the area of Information and Communication Technologies, focusing on the development of Information Extraction methods for the domain of food and nutrition. Her main research interests include machine learning, natural language processing, representation learning and recommender systems.

The third keynote speaker was Drashko Nakikj, PhD. Dr. Nakikj's primary areas of expertise and investigation are individual, collective and collaborative sensemaking of health information for patients and physicians, as well as building tools that support those processes. He previously worked as a Postdoctoral Researcher at the Department of Biomedical Informatics at Harvard Medical School, USA. Prior to that, he obtained his PhD in Biomedical Informatics at the Department of Biomedical Informatics at Columbia University, USA. Dr. Nakikj holds a MS degree in Computer Science and Information Technology and BS in Computer Science, Information Technology and Automation, both from the Faculty of Engineering and Information Technologies at the Ss. Cyril and Methodius University in Skopje, North Macedonia.

Part of the conference success is owed to the support received from our partners and sponsors: Ss. Cyril and Methodius University in Skopje and ICT-ACT Association in Skopje.

November, 2022
Skopje

Katarina Trojachanec Dineva
Vladimir Zdraveski

Organization

Conference Chairs

Katarina Trojachanec Dineva	Assistant Professor - Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Vladimir Zdraveski	Assistant Professor - Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia

Organizing Committee

Stefan Andonov	Lab. Assistant - Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Jana Kuzmanova	Lab. Assistant - Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Vlatko Spasev	Lab. Assistant - Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Dimitar Kitanovski	Lab. Assistant - Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Jovana Dobрева	Lab. Assistant - Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia

Program Committee

Adrijan Božinovski	School of Computer Science and Information Technology, University American College Skopje, N. Macedonia
Aleksandra Dedinec	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Aleksandra Mileva	Faculty of Computer Science, University Goce Delcev, Shtip, N. Macedonia
Aleksandra Popovska-Mitrovikj	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Alexandru Nicolin	IFIN-HH, Romania
Ana Madevska Bogdanova	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia

Andrea Kulakov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Andreja Naumoski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Antun Balaz	Institute of Physics Belgrade, Serbia
Biljana Stojkoska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Biljana Tojtovska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Blagoj Ristevski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Boban Joksimoski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Bojana Koteska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Danco Davcev	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Dejan Gjorgjevikj	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Dejan Spasov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Dimitar Trajanov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Eftim Zdravevski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Emil Stankov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Galina Bogdanova	Institute of Mathematics and Informatics at Bulgarian Academy of Sciences, Bulgaria
Georgina Mirceva	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Gjorgji Madjarov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Goce Armenski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Goran Velinov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Hazim Kemal Ekenel	Istanbul Technical University (ITU), Turkey
Hristijan Gjoreski	Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Hristina Mihajloska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Igor Mishkovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Ilinka Ivanoska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia

Ivan Chorbev	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Ivan Kitanovski	Faculty of Computer Science and Engineering, "Ss. Cyril and Methodius" University in Skopje, Macedonia
Ivan Miguel Pires	Instituto de Telecomunicações in Covilhã, Portugal
Ivaylo Donchev	St Cyril and St Methodius University of Veliko Turnovo, Bulgaria
Ivica Dimitrovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Katerina Zdravkova	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Kire Trivodaliev	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Konstantin Delchev	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Lasko Basnarkov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Ljupcho Antovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Magdalena Kostoska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Marcin Paprzycki	Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
Marija Mihova	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Marjan Gusev	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Mile Jovanov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Milos Jovanovik	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Miroslav Mirchev	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Monika Simjanoska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Natasha Ilievska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Natasha Stojkovikj	University "Goce Delcev" - Stip, N. Macedonia
Nevena Ackovska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Nuno Pombo	Department of Informatics, University Beira Interior in Covilhã, Portugal
Pance Ribarski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Petre Lameski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Riste Stojanov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia

Sasho Gramatikov	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Sasko Ristov	Institute of Computer Science, University of Innsbruck, Austria
Simona Samardjiska	Digital Security Group at Radboud University, The Netherlands
Slobodan Kalajdziski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Smile Markovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Smilka Janeska-Sarkanjac	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Stela Zhelezova	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Suzana Loshkovska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Tome Eftimov	Computer Systems, Jožef Stefan Institute, Slovenia
Tsonka Baicheva	Institute of Mathematics and Informatics, Bulgaria
Valentina Janev	Institute "Mihajlo Pupin" Belgrade, Serbia
Vangel Ajanovski	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Verica Bakeva	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Vesna Dimitrievska Ristovska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Vesna Dimitrova	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Vladimir Trajkovik	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Yuri Borissov	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Zaneta Popeska	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia

Table of Contents

Full Papers

Language Agnostic Voice Recognition Model	2
<i>Tea Janeva, Kostadin Mishev and Monika Simjanoska</i>	
Cryptanalysis of Round-Reduced ASCON powered by ML	7
<i>Dushica Jankovikj, Hristina Mihajloska Trpceska and Vesna Dimitrova</i>	
Framework for a Multipurpose Remotely Accessible Laboratory for Education	13
<i>Vojdan Kjorveziroski</i>	
Understanding Public Perception of Waste Management in Western Balkans using Tweets and News Reports Analysis	19
<i>Ivana Gjorshoska, Angela Madjar, Aleksandra Dedinec and Jana Prodanova</i>	
Differences in cognitive ability assessment results between Millennial and Generation Z cohorts	25
<i>Ivica Pesovski, Andrea Kulakov and Vladimir Trajkovik</i>	
Investigating Public Awareness of Air Pollution in Western Balkans by analyzing Tweets and News Article Teasers	31
<i>Angela Madjar, Ivana Gjorshoska, Jana Prodanova and Aleksandra Dedinec</i>	
Model Hyper Parameter Tuning using Ant Colony Optimization	37
<i>Aleksandar Trajkovski and Gjorgji Madjarov</i>	
Blood pressure class estimation using CNN-GRU model	42
<i>Ivan Kuzmanov, Magdalena Kostoska and Ana Madevska Bogdanova</i>	
The influence of stock market indexes (S&P500 and Dow Jones) on cryptocurrencies prices	48
<i>Gorast Angelovski, Ana Todorovska, Ivan Rusevski, Jovana Marojevikj, Eva Spirovska, Hristijan Peshov, Irena Vodenska, Lubomir T. Chitkushev and Dimitar Trajanov</i>	
Analysis of write amplification on solid state drives in datacenters	54
<i>Sladana Đurašević, Uroš Pešović, Vanja Luković and Borislav Đorđević</i>	
Electricity price forecasting for the day-ahead market in Croatia	58
<i>Kristijan Koteski and Aleksandra Dedinec</i>	
Framework for Using BigchainDB in Software Application	64
<i>Avni Rustemi, Vladimir Atanasovski and Aleksandar Risteski</i>	
Missing Data in Longitudinal Image Retrieval for Alzheimer's Disease	69
<i>Katarina Trojchanec Dineva, Ivica Dimitrovski, Ivan Kitanovski and Suzana Loshkovska</i>	
Creating a distributed, privacy-aware e-learning system based on item response theory ..	75
<i>Bojan Kostadinov and Irena Stojmenovska</i>	
System for reproducible code environments - Dohrnii	80
<i>Dimitar Mileski and Marjan Gusev</i>	

logs2graphs: Data-driven graph representation and visualization of log data	86
<i>Stefan Andonov, Viktor Jovev, Aleksandar Kitanovski, Aleksandar Krsteski and Gjorgji Madjarov</i>	
Artificial Intelligence in Robotics	91
<i>Martin Stefanov, Nikola Stojanovski and Vlatko Spasev</i>	

Short Papers

Wearable Patch for Mass Casualty Screening with Graphene Sensors	99
<i>Teodora Vićentić, Bojana Koteska, Fedor Lehoccki, Milena Rašljić Rafajilović, Ana Madevska Bogdanova, Stefan Ilić, Igor Pašti and Marko Spasenović</i>	
Activation functions' impact on regularization of deep neural networks application in atomic simulations	103
<i>Ljubinka Sandjakoska and Ana Madevska Bogdanova</i>	

Student Papers

Virtualization technology and possibilities of its use in education	108
<i>Bojan Karkalashev and Marjan Gusev</i>	
Open source BI tools for SMEs: Comparative analysis of Pentaho CE and Knowage CE stacks	111
<i>Igor Jakimovski</i>	
Trends in Autonomous Vehicles	115
<i>Vesela Trajkoska</i>	
Anonymous Blockchain based model for e-Voting	119
<i>Marija Taneska and Fat Halimi</i>	
Perception of benefits and costs of computer games among high school students in Macedonia	123
<i>Marina Dodevska, Nikola Janev and Vladimir Trajkovikj</i>	
Survey on VoIP call security perception during COVID-19 pandemic	128
<i>Marija Taneska and Sasho Gramatikov</i>	
Forecasting the prices of the day-ahead electricity markets using real data from SEEPEX	132
<i>Pavle Petkovski and Aleksandra Dedinec</i>	
Digital Piracy in the Western Balkans with a special reference to North Macedonia	136
<i>Simon Atanasovski, Diellza Sherifi and Sasho Gramatikov</i>	
Privacy Risks and Security Threats in Online Social Networks	140
<i>Marija Bukalevska and Aleksandar Karadimce</i>	
Accelerating data compression using general purpose GPUs	144
<i>Kristijan Ristovski and Vladimir Zdraveski</i>	

Implementation Of The Viola Jones Algorithm For Face Detection And Tracking On Robots With Limited Computation Power.....	148
<i>Dimitar Bezhanovski and Aleksandar Stankovski</i>	
Running Semantic Search Over Complete English Wikipedia on a Local Computer	152
<i>Stojancho Tadjarski and Ana Madevska Bogdanova</i>	
An Overview of Robotic Process Automation.....	156
<i>Bojana Naumoska and Georgina Mirceva</i>	
Algorithms for Process Mining – An Overview	160
<i>Sandra Nikoloska, Biljana Sazdova and Georgina Mirceva</i>	

FULL PAPERS

Language Agnostic Voice Recognition Model

Tea Janeva

Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering, Skopje
 N. Macedonia
 tea.janeva@students.finki.ukim.mk

Kostadin Mishev

Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering, Skopje
 N. Macedonia
 kostadin.mishev@finki.ukim.mk

Monika Simjanoska

Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering, Skopje
 N. Macedonia
 monika.simjanoska@finki.ukim.mk

Abstract—Voice recognition is the ability of a machine to identify a person based on their unique voiceprint. As this task is becoming more important and dominant in everyday people's lives, this paper is testing different approaches for its implementation. Using a multilanguage database and working with the different frequencies' characteristics, five machine learning models such as Random Forest, XGBoost, MLP, SVM and Gradient Boosting, along with CNN deep learning model were implemented. The models were trained on three different tasks, gender prediction, age range prediction, and combined gender and age range prediction. These models were evaluated using accuracy, F1-score and MCC score. The results showed that Random Forest outperforms other models by achieving an accuracy of more than 0.9 for all the three classification tasks.

Index Terms—Voice recognition, Deep learning, Machine learning, Explainable Machine learning

I. INTRODUCTION

Voice recognition is slowly but surely becoming a key part of the future of communication. This task is the ability of a machine or a program to receive and understand spoken commands from a speaker and later to identify this person based on their unique voiceprint. On the other hand gender recognition and recognition of the age range group of the user are becoming essential information for the interactions of the users, and altogether for the social community. Both of these tasks are something that the human brain can automatically identify, and at the moment these are some of the most explored subjects in the world of voice recognition - the technology behind some of the most popular virtual assistants such as Amazon's Alexa, Apple's Siri, Google's Google Assistant and Microsoft's Cortana.

Most of the voice recognition applications work by analyzing the sound provided by the user and then implementing a variety of methods to understand it and identify the speaker. As this is a technology that first appeared 50 years ago, throughout the process of creating a working voice recognition application, different approaches were tried and were successful. The work that is currently available and working can be separated in 2 categories, text-dependent systems that require predetermined voice passphrases and text independent systems where the subject of analysis is conversational speech. The first category requires the speaker to provide audio of certain key-words and sentences. Some of the most used methods are Dynamic Time Warping (DTW)-based methods and hidden Markov model (HMM)-based methods. Text independent systems given their nature are more popular.

The ability to predict without constantly needing to be provided with audios of certain key-words gave the possibility for these systems to be done sequentially. Some of the most used methods are Long-Term-Statistics-Based Methods, VQ-Based Methods, Ergodic-HMM-Based Methods and Speech-Recognition-Based Methods. Both systems have a mutual serious weakness, their security systems can be easily bypassed. Trying to overcome this weakness, another category of voice recognition applications is available, a text-prompted speaker recognition where the idea is that the password sentences are changed every time. Looking more in depth behind the elemental job, voice recognition can be classified into 2 parts, speaker identification and speaker verification. Speaker identification is determining which speaker is the system currently working with, and speaker verification is the process of confirming or denying the identity that is calculated [3].

Analyzing the current work made in this field and for the selected subjects, the majority of speech and later on voice recognition is conducted on single language datasets, mostly working with English speakers. This study investigates different approaches, giving an importance in predicting the speaker on a model that is trained on multiple languages at once. Languages differ in many ways, including the way it is spoken, the way the words are formed and therefore the way the sentences are put together. The difference that is of importance for voice recognition are the different spoken ways, more specifically the features that different languages offer. Focusing on the frequency every language has, there can be seen a lot of differences as each language has its own, with the highest frequency range found in UK English, which peaks at 12000 Hz. Some of the languages with lowest frequency range used in this study are French and Spanish where on average the peak is at 2000 Hz. On the other hand, in general female frequencies are higher than male and correspondingly as the speaker ages the frequency of their speech lowers. [6]

In this paper we are proposing the use of both Machine Learning (ML) and Deep Learning (DL) models. Focusing on the different aspects that working with multiple languages at once offers, we implement and test the models on diverse information and features. Combining both types of models, more ways were explored and more results were acquired for determining the best approach.

The rest of the paper is organized as follows. In Section II different researches on the topic are explored and dis-

cussed. In the methodology Section III an accent is put on explaining the dataset and its creation, the process behind the data preprocessing and also all the ML and DL models. Explainable Machine Learning is introduced as a way to explain the models and the influence of the features on the final prediction. The results are presented and discussed in Section IV and a conclusion is derived in the final Section V.

II. RELATED WORK

Most often the traditional ML models are modified and implemented and later satisfying results are achieved. For example, a system using bootstrapping for audio based gender recognition implemented 5 models such as Naive Bayes, Nearest Neighbor, Decision Tree, Support Vector Machine (SVM) and Logistic Regression on a dataset based on 10 news broadcast excerpts. The classification accuracy for all the models was around 85% [8]. Another example of research in this field combining both gender and age recognition is shown in the models presented in [5]. There were 7 different proposed and tested methods using Gaussian mixture models (GMM) and Support vector machine (SVM) models based on different features such as mel-frequency cepstral coefficient (MFCC), GMM mean supervectors, acoustic, prosodic and voice quality, UBM weight posterior probability supervectors, GMM maximum likelihood linear regression (MLLR) matrix supervectors, polynomial expansion coefficients of the syllable level prosodic and weighted summation of the fusion of all the previously listed. The dataset used is the aGender database. The final results are around 88% weighted accuracy for gender prediction and around 50% for the age prediction. One of the best accuracy for this task was acquired by an approach trying to automatize the Bengali Voice based gender classification [1]. In this study Mel-frequency cepstral coefficient (MFCC) was used for feature extraction and for the ML models Logistic Regression, Random Forest and Gradient Boosting were implemented. Testing on a dataset with more than 250 speakers, an accuracy of 99.13% was obtained.

Using a Multilayer Perceptron on a dataset with more than 3000 samples of both female and male speakers, an accuracy of 96.76% was obtained [2]. The data was preprocessed extracting 22 acoustic parameters from the acoustic signals. Making a combination of acoustic and pitch features and implementing a set of Neural Networks is an example of some of the most accurate approaches for gender classification [4]. The system classifier contains individual experts with 4 MLPs each. Each MLP is separately trained using the Back Propagation algorithm. When testing this model on a Switchboard dataset with 19 male and 19 female speakers an accuracy of 98.5% is reached.

III. THE METHODOLOGY

A. Dataset

This study is conducted on multiple datasets containing audios from different languages. Primarily the dataset contained only English speakers from different backgrounds, but to benchmark the results it was updated and another 15 languages were implemented. The main goal for introducing new languages and working with more than one at a time was to explore the similarities and differences between the voices and their characteristics from different parts of the world.

In the final dataset the most dominant language is English represented by approximately 6000 speakers with more than 64000 audios, however, the total number of speakers in the dataset was 13399. The languages that are present in the final, combined dataset are: English, Italian, French, Spanish, Russian, Portuguese, Croatian, Ukrainian, Greek, Turkish, Catalan, Bulgarian, Dutch, Hebrew, Persian and Albanian. The distribution between languages can be seen in Table I. When collecting audios from languages different to English, the same format of the audios was preserved concerning length and number of audios per speaker. All of the data was collected from VoxForge [9] which offers open speech datasets.

TABLE I
DISTRIBUTION OF DIFFERENT LANGUAGES IN THE DATASET.

Language	Number of speakers
HR	13
FA	15
NB	23
SQ	33
UK	39
CA	41
BG	44
EL	112
TR	154
PT-BR	319
RU	576
NL	682
IT	922
ES	1986
FR	2101
EN	6339

The main focus of this study is towards distinguishing between 6 different classes combined from age and gender.

The numbers between female and male speakers are drastically different, as presented in Table II, implying imbalanced dataset. The first part of this study gives an accent specifically on the gender, and using all the features from the audios of the speakers trying to predict if the speaker in question is male or female.

TABLE II
REPRESENTATION OF GENDER SAMPLES.

Gender	Number
Male	10883
Female	1130

Focusing on the age range, there are 3 age groups implemented as youth, adult and senior. Each of these groups is represented by a different age range. Youth is characterized by 25 years and younger (kids under 18 years are excluded), adult group contains the ages between 25 and 55, and the senior group contains the older than 55. The majority of the dataset is represented by the adult group, then the young group, and the least amount of data is for seniors as can be seen in Table III.

Combining the 2 classes, we created 6 new possible classes: female youth, female adult, female senior, male youth, male adult, and male senior. Within these 6 classes, the

TABLE III
DISTRIBUTION OF AGE RANGE SAMPLES.

Age range	Number
Youth	1643
Adult	10027
Senior	182

dominant is adult males, and the group for which there are not representatives is senior females as can be seen in Table IV. Thus, this class is completely omitted in the further analysis.

The data downloaded to create the dataset is in a specific format, being separated in 2 parts. The first part is the audio of the speaker, split in multiple shorter audios, each of around 10 seconds in duration. The second part is the information about the speaker, such as gender, age range, pronunciation, language, etc. When adding new audios from different languages in the database we followed the same structuring format.

TABLE IV
DISTRIBUTION OF SAMPLES CONSIDERING COMBINED CLASSES.

Age range and Gender	Number
Youth + Male	1507
Adult + Male	8993
Senior + Male	182
Youth + Female	101
Adult + Female	997
Senior + Female	0

B. Preprocessing

Since we are training the classifiers by using both traditional Machine Learning and Deep Learning approach, the dataset was preprocessed in two different manners, accordingly.

For the Machine Learning, the audio files were preprocessed to extract the features describing the frequencies represented in the particular audio, and those are: nobs, mean, skew, kurtosis, median, mode, std, low, peak, q25, q72 and iqr.

For the DL models, raw audios were used as input in the CNN layer of the DL architecture.

The dataset was labeled to correspond to the three different classifiers we intend to create, the first to predict only the gender, the second to predict the age range and the third to predict a label that corresponds to the combination of both gender and age range as presented in Table IV, but without the class for which we do not have any representatives.

C. Machine Learning

Combining 16 languages, we have obtained dataset with 13399 speakers available. After preprocessing the dataset and extracting the features, it was used to build three types of intelligent models capable of predicting the gender, the age range, and the gender and age combined as one label, for a particular speaker.

Five different models were trained for each classification task by using Random Forest, Gradient Boosting, Support Vector Machines, Multilayer Perceptron, and XGBoost. We

experimented with different hyperparameters to obtain the best results shown in the following Section IV.

As evaluation metrics we used accuracy, F1 score, and MCC score. N-fold cross-validation was implemented, trying both 5-fold and 10-fold for all the models previously described in this section.

D. Deep Learning

The Deep Learning model used in this study is the Convolution Neural Network (CNN). Multiple different layers have been implemented such as Conv2D, BatchNormalization, MaxPool2D, Flatten, Dropout and Dense as shown in Figure 1. Adam was used as optimizer, and for the loss we used binary cross entropy. The final result was evaluated by using the same metrics as in the traditional ML approach.

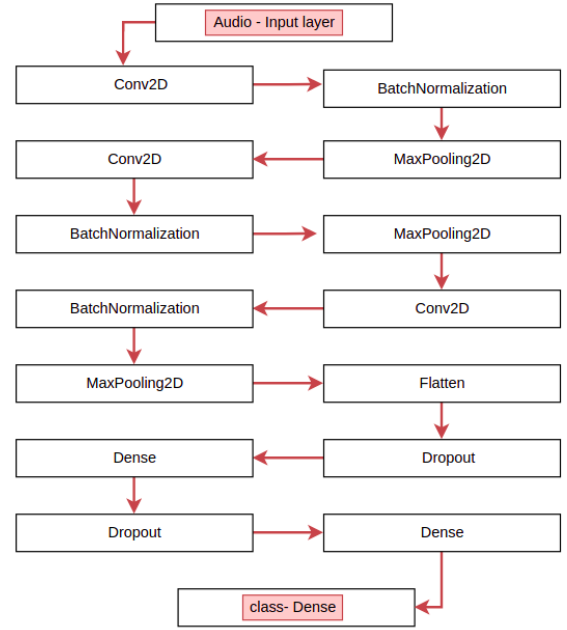


Fig. 1. CNN model- layers

E. Explainable Machine Learning

To understand how the features affect the gender, age and the combined labels classification, an explanation of contribution on each attribute from the dataset to each classification target was done by using SHAP method [7]. To be certain of the impact, Shapley values were calculated for each of the 10-fold round, and consequently averaged. The results are shown in the next Section IV.

IV. RESULTS

This section presents the evaluation metrics obtained from the three types of classifications we presented in the previous sections.

For the traditional ML case, five ML models were trained by using 10-fold cross-validation for all the three types of classifications.

Random Forest performed best at all classification tasks: gender, age range and combined (gender + age range) prediction. Tables V, VI and VII show the performance of all the classification models for the three tasks, respectively.

The MLP and SVM models gave poor results considering the F1-scores and the MCC scores.

TABLE V
RESULTS FROM 10-FOLD FOR ML MODELS FOR GENDER CLASSIFICATION.

Model	Accuracy	F1 score	MCC score
XGBoost	0.895	0.786	0.709
Random Forest	0.972	0.918	0.926
Gradient Boosting	0.949	0.947	0.870
Support Vector Machine	0.781	0.211	0.124
MLP	0.899	0.609	0.243

TABLE VI
RESULTS FROM 10-FOLD FOR ML MODELS FOR AGE RANGE CLASSIFICATION.

Model	Accuracy	F1 score	MCC score
XGBoost	0.907	0.743	0.623
Random Forest	0.977	0.922	0.906
Gradient Boosting	0.895	0.748	0.533
Support Vector Machine	0.848	0.306	0.0
MLP	0.829	0.395	0.188

TABLE VII
RESULTS FROM 10-FOLD FOR ML MODELS FOR COMBINED CLASSIFICATION.

Model	Accuracy	F1 score	MCC score
XGBoost	0.977	0.920	0.854
Random Forest	0.992	0.973	0.946
Gradient Boosting	0.988	0.963	0.931
Support Vector Machine	0.922	0.546	0.139
MLP	0.771	0.372	0.331

For the DL case, the presented CNN architecture in the previous section was used for the three classification tasks denoted as T1, T2 and T3 in Table VIII corresponding to gender, age range and combined classification.

TABLE VIII
CNN EVALUATION FOR THE THREE CLASSIFICATION TASKS.

Task	Accuracy	F1 score	MCC score
T1	0.832	0.767	0.617
T2	0.8	0.692	0.516
T3	0.838	0.359	0.579

Analyzing the results from the CNN evaluation it can be perceived that the DL approach showed worse performance than Random Forest, XGBoost and Gradient Boosting, regarding all the three evaluation metrics at all classification tasks. We assume the results would be better if more data were available for the DL approach.

Analyzing the importance of the features by using the SHAP method as described in the previous section, Figure 2 shows that the median feature has the biggest impact on the gender prediction. Closely behind are impacts from q25 and q75.

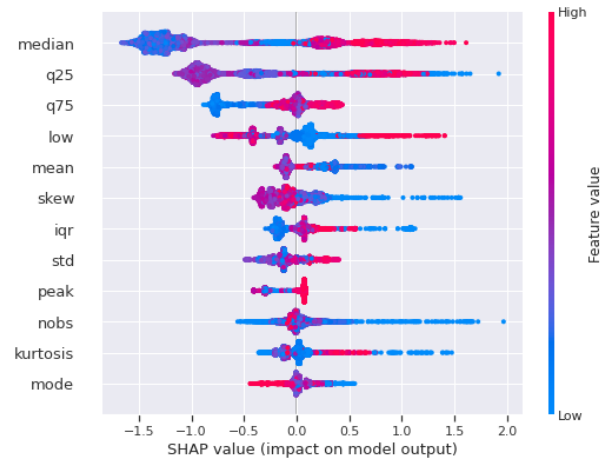


Fig. 2. The most important features for gender prediction.

Looking over the SHAP values for the age range prediction, from Figure 3 it can be seen that the values of skew affect the prediction the most along with Low (meaning the low frequency) and the feature q25.

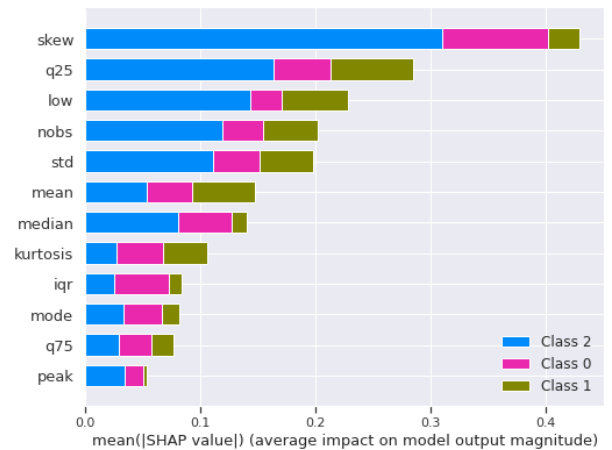


Fig. 3. The most important features for the age range prediction.

Analyzing the SHAP values for the combined gender + age range classification gave significant insight in the features importance overlapping more with the values for gender prediction than age range prediction. The most important feature can be seen as q75, closely followed by the median and q25 values. All the contributions can be observed in Figure 4.

V. CONCLUSION

In this paper, we have presented different approaches to voice recognition by using a multilingual database encompassing 13399 speakers. Using both machine learning and deep learning models, three different classifiers were trained to predict the gender, the age range, and combined label of both gender and age range. The results were evaluated by using accuracy, F1-score and MCC score. Random Forest showed significantly better results when compared to other ML models and DL network.

Additionally, Shapley values were used to explain the influence of the selected features on the model's ability to predict the three types of targets. The frequencies' median and q25 features have shown to have biggest influence on

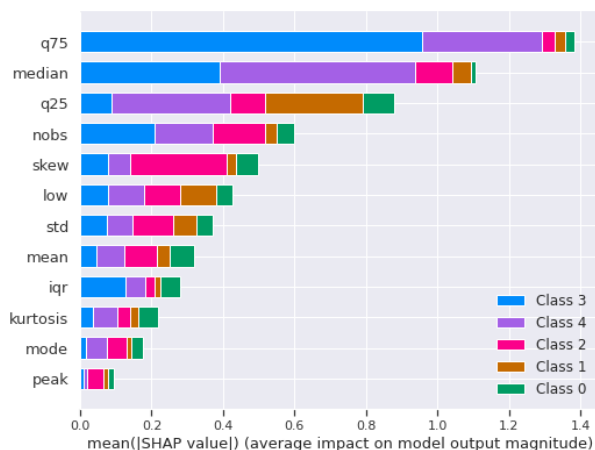


Fig. 4. The most important features for the gender and age range prediction.

the models ability to decide the gender and the age of the speaker.

As there is a disbalance related to gender, the next steps would be to include more speakers so that a more balanced dataset is available as this problem might influence the accuracy of the experiments. Future work would also include implementation of the Macedonian language as part of the dataset.

REFERENCES

- [1] S. S. I. Badhon, M. H. Rahaman, and F. R. Rupon. A machine learning approach to automating bengali voice based gender classification. In *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, pages 55–61. IEEE, 2019.
- [2] M. Buyukyilmaz and A. O. Cibikdiken. Voice gender recognition using deep learning. In *Proceedings of 2016 International Conference on Modeling, Simulation and Optimization Technologies and Applications (MSOTA2016)*. Paris, France: Atlantis Press, volume 10, 2016.
- [3] S. Furui. An overview of speaker recognition technology. *Automatic speech and speaker recognition*, pages 31–56, 1996.
- [4] H. Harb and L. Chen. Voice-based gender identification in multimedia applications. *Journal of intelligent information systems*, 24(2):179–198, 2005.
- [5] M. Li, K. J. Han, and S. Narayanan. Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer Speech & Language*, 27(1):151–167, 2013.
- [6] J. Lo. Sound frequencies of language. 2018.
- [7] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017.
- [8] G. Tzanetakis. Audio-based gender identification using bootstrapping. In *PACRIM. 2005 IEEE Pacific Rim Conference on Communications, Computers and signal Processing, 2005.*, pages 432–433. IEEE, 2005.
- [9] Voxforge.org. Free speech... recognition (linux, windows and mac) - voxforge.org. <http://www.voxforge.org/>, 2022.

Cryptanalysis of Round-Reduced ASCON powered by ML

Dushica Jankovikj

Faculty of Computer Science and
Engineering

Ss. Cyril and Methodius University
Skopje, Republic of North Macedonia
dushica.jankovikj@students.finki.ukim.mk

Hristina Mihajloska Trpeska

Faculty of Computer Science and
Engineering

Ss. Cyril and Methodius University
Skopje, Republic of North Macedonia
hristina.mihajloska@finki.ukim.mk

Vesna Dimitrova

Faculty of Computer Science and
Engineering

Ss. Cyril and Methodius University
Skopje, Republic of North Macedonia
vesna.dimitrova@finki.ukim.mk

Abstract— Our research focuses on attacking Ascon, a lightweight block cipher presented as a candidate in the NIST Lightweight Cryptography Standardization Process. This block cipher provides authenticated encryption with associated data functionalities. We propose a cryptanalysis model based on deep learning (DL), where the goal is to predict plaintext bits given knowledge of the ciphertext and other publicly known cipher input parameters. Our experiments show that such known-plaintext attacks can be successfully executed on a round reduced version of the cipher stripped of the finalization phase. This, in turn, validates the theoretical results. Cryptographic algorithms are complex for the purpose of security and cannot be easily broken by an ML model in their regular form (not reduced). We explore multiple dataset generation techniques, model design, and training hyperparameters.

Keywords—lightweight cryptography, cryptanalysis, known plaintext attack, machine learning, deep learning

I. INTRODUCTION

Cryptanalysis studies the inner workings of ciphers, the ciphertexts they produce, and cryptosystems in general. The goal is understanding the inner workings of information systems, uncovering hidden aspects of their operation, and then discovering or developing techniques toward breaking them or exposing their weaknesses. This discipline is not inherently malicious, it is often used to discover the use cases when a breach in a security system can be achieved so it can be prevented in the future by improving the system. Cipher standardization includes extensive security evaluations where any attacks the cipher might fall weak to are discussed. Attackers performing cryptanalysis hope to obtain access to encrypted data and to infer the raw data value without knowledge of the actual cryptographic key.

Cryptanalysis of block ciphers is not a new idea [1]. This has been persistently studied, and as a result many cryptanalytic techniques have been proposed. Most papers list differential, linear and integral cryptanalysis, attacks which exploit the algebraic degree, meet-in-the middle attack. Legacy cryptanalytic technologies are known to require a great deal of resources, whether that be time, memory, known plaintexts.

Due to the fact that machine learning (ML) and cryptanalysis share many of the same concepts and concerns, they have been described as “sister fields” [2]. This relationship originates from the similarity in the goal of both disciplines: to learn some unknown function given pairs of input and output values (in the cryptanalysis scenario, these values can be ciphertext and plaintext pairs). We propose a ML based approach to cryptanalysis, more precisely, within the deep learning (DL) realm.

Encryption methods are considered lightweight if they are characterized by small memory requirements and computational complexity. Their use is suitable in constrained

devices (low memory, power or computation resources) where performance of heavier cryptographic standards is not acceptable. Due to the lower complexity of operation as a requirement, these types of ciphers seem eligible for cryptanalysis using machine learning.

II. ASCON

A. Notation

This section defines some basic notation which will be used in the following sections. Here we define that $\oplus$ stands for XOR, right and left bit rotations will be denoted as $\gg$ and $\ll$ respectively, while $a||b$ represents the concatenation of two bit strings a and b .

B. Ascon

Ascon is a family of ciphers equipped to handle authenticated encryption with associated data and hashing [3]. Our focus falls on the encryption functionality of the cipher.

Ascon uses a duplex-sponge-based mode of operation for authenticated encryption. At the core of operation stands a lightweight permutation, which is used for all family members. The permutation is used to apply substitution-permutation network (SPN) based transformations iteratively using multiple rounds. Input parameters are key, tag and nonce where different modes support different bit lengths of these parameters. Namely Ascon branches out into two versions, Ascon128 and Ascon96, both of which have differences in the parameters and the achieved security levels. The recommended choice is said to be Ascon128 and it is the focus of the analysis in this paper. This mode operates on a plaintext block size of 64 bits.

The cipher operates using a 320 bit internal state which gets updated and transformed by the sponge function. The permutation gets applied in $a=12$ rounds (for initialization and finalization phases) and $b=8$ rounds (for encryption process). Encryption happens when the input plaintext block is XOR-ed with the first 64 bits of the state. The part of the state that “meets” the plaintext block (the first 64 bits of the state) are called the rate r bits, while the rest are called the capacity bits. Cipher operations can be divided in four phases:

1. Initialization: The bits of the internal state are populated using the key K (whose number of bits are represented by k) and nonce N . More precisely, the internal state is filled by the vector $IV||K||N$ where IV is some initialization vector and calculated as: $k||r||a||b||0160-k$. After the internal state registry of bits is no longer empty, initialization can begin. The 320 bit internal state is initialized using 12 rounds of the Ascon permutation p .
2. Associated Data processing: updates the state using associated data blocks A_i .

3. Plaintext Processing: injects plaintext blocks P_i into the state and extracts ciphertext blocks C_i . C_i can be calculated as $P_i \oplus S_r$, where S_r presents the r bits from an internal state. After this operation is performed, the internal bits in S_r accept the value of C_i themselves.
4. Finalization: injects the key K into the capacity bits of the state (S_c). After this, finalization is carried out similarly to initialization - in 12 permutation rounds. Upon completion, an authentication tag T is produced with authenticates both the associated data and the encrypted message.

As mentioned, the internal state S of the cipher consists of 320 bits. It is divided into 5 “words” of 64 bits each. We will denote them S_1, S_2, S_3, S_4, S_5 ; where S_1 is equivalent to S_r . The Ascon permutation module consists of three transformations that affect the state as follows:

1. Addition of Round Constants: XOR-s a round specific 1-byte constant to the state word S_3 . The round constants exist within the cipher stored in a lookup table [3].
2. Nonlinear Substitution Layer: applies a 5-bit S-box 64 times in parallel in a bit-sliced fashion (vertically, across words) [4]. The S-box mapping rules are presented in the official cipher specification [3].
3. Linear Diffusion Layer: provides diffusion within each state word by XOR-ing different rotated copies of each word (horizontally, within each word). The transformations for each state word are shown below.

$$\begin{aligned}
 S_1 &= S_1 \oplus (S_1 \gg 19) \oplus (S_1 \gg 28) \\
 S_2 &= S_2 \oplus (S_2 \gg 61) \oplus (S_2 \gg 39) \\
 S_3 &= S_3 \oplus (S_3 \gg 1) \oplus (S_3 \gg 6) \\
 S_4 &= S_4 \oplus (S_4 \gg 10) \oplus (S_4 \gg 17) \\
 S_5 &= S_5 \oplus (S_5 \gg 7) \oplus (S_5 \gg 41)
 \end{aligned}$$

The permutation in Ascon is used during the initialization, when it is applied in 12 rounds. During encryption, it is applied in 8 rounds in between each plaintext being processed.

C. Round reducing ciphers

In the world of ciphers, rounds refer to the number of times some internal cipher function is applied to the data. For Ascon, rounds are used when the cipher's permutation is applied to the cipher state in different stages of the cipher operation. As mentioned, initialization applies the permutation 12 times (variable a), while the encryption uses the same permutation 8 times (variable b). Multiple applications of some function that mix up and shift the data bits ensure better confusion of the bits - security through complexity of the transformation. Namely, by mixing the data repeatedly (multiple rounds), it gets harder to simply return the data to its internal state. For Ascon, multiple iterations of the SPN transformations lead to a state where the transformation becomes layered, and the result is hard to express in terms of a single formula where all the modifications applied to a specific bit through the rounds can be tracked.

Cipher specifications are created with a thick background of experimentation and proof that defines the specific number of rounds used for the cipher operation. The number of

rounds is chosen with security, but also with performance in mind. For example, Ascon with more than 8 rounds in the encryption permutation module might be more secure, but not secure to the point where the performance hindering is worth it. This is a lightweight cipher after all, so performance optimization is key.

Our cryptanalysis efforts focus on a round reduced Ascon, but only in the encryption phase. This means, the initialization carries out using all the rounds it is meant to, but for the encryption permutation module we experimented with using 1, up to 8 rounds.

D. Simplified Ascon

Aside from using a round reduced Ascon, which is common in cryptanalysis attempts, we additionally strip the cipher of the initialization phase. By using an empty associated data input, that section can also be considered avoided. The figure below shows the sections that aren't relevant to our work as grayed out. Beginning with an initialized state, the first plaintext block P_1 is injected in the first state word. This modified the state itself but outputs C_1 as a result as well. Then the state is processed using the encryption permutation p_b (where b is reduced). Next, P_2 is XOR-ed to the first state word to obtain C_2 .

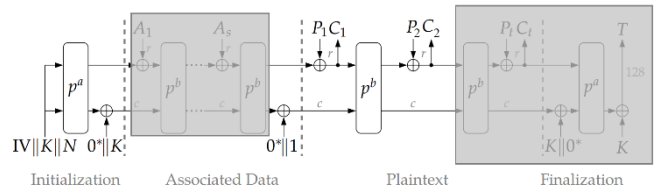


Fig. 1. Ascon encryption (simplified version omits greyed out processes)

We are focused on discovering the relationship between P_2 and C_2 , thus producing plaintext recovery attacks. Additionally, if knowledge of P_1 and C_1 is assumed, knowing the first state word after initialization is also assumed.

E. Traditional cryptanalysis of Ascon

The security analysis of Ascon [5] proposed by its creators gives a detailed inspection of the possible weaknesses when rounds are being reduced. In their work they apply cube-like, differential, and linear cryptanalysis to evaluate the security. The focus falls on practical key-recovery attacks on round-reduced versions of Ascon-128. The initialization phase was round reduced to 5 out of 12 rounds. Theoretical key recovery attacks were proven as feasible for up to 6 rounds of initialization. From the aspect of forgery, a practical forgery attack was presented for 3 rounds of the finalization, while a theoretical forgery attack was presented for 4 rounds finalization and zero-sum distinguishers for the full 12-round Ascon permutation.

They present attempts of linear cryptanalysis of Ascon which can be considered first in the related research efforts and prove the bounds on the minimum number of active s-boxes for the Ascon permutation.

III. MACHINE LEARNING FRAMEWORK

A. Machine learning concepts

Machine learning (ML) is a term that can be used to describe algorithms that can perform intelligent predictions based on a data set. The dataset part is necessary since ML seeks to automatically learn meaningful relationships and patterns from examples and observations.

The algorithms do multiple passes over the data which enables them to iteratively learn and uncover hidden relationships and complex patterns in the data points.

In literature, the algorithms are often referred to as models. Choosing the appropriate model for a learning task depends on the complexity of the function, and the dimensionality and volume of data.

Artificial neural networks represent a family of algorithms whose internal processing units are mathematical functions. These models take inspiration from the neurons in the human brain, and so, the processing units are, each connection transmitting signals like synapses in the brain. The strength of these signals depends on an internal weight for each model, which is updated and adjusted through the learning process. This affects subsequent processing in the network since the connected neurons only process signals if the signal strength exceeds a certain threshold. An activation function defines this threshold.

Neurons aren't chaotically distributed throughout the model; instead, they are organized into layers. The neurons within a layer do not communicate; instead, they are linked only to neurons in the previous and following layer. Depending on the type of neurons, the information (signal) flow can be one-directional (data flows from one layer to the next), bi-directional (data flows in different directions). Data enters the model through the input layer and exists in the form of predictions through the output layer. The layers in between are hidden and are tasked with learning a non-linear mapping between the input and output layers.

Deep learning (DL) is a machine learning technique which uses deep neural networks (DNN) that contain more than one hidden layer. These networks follow deeply nested architectures and can consist of advanced neurons with multiple activation functions and more complex internal functions. The simplest forms of ML can estimate and learn simple linear functions, while complex neural networks can be used to estimate complex functions. DL uses multiple layers to progressively extract higher level learnable features from the input data and as a result is capable of learning nonconvex and nonlinear functions [6].

B. Machine learning in cryptanalysis

The closeness of cryptanalysis and machine learning has been theoretically established for three decades [3], but even so, the research community has been scarce of practical efforts which exploit more possibilities this similarity offers. For the most part, practical work in ML driven side channel analysis (SCA) tasks has shown as very valuable. The earliest efforts utilized support vector machines (SVM) [7], but recently DL enhanced SCA has gained a lot of traction.

Ghor's work in [2] introduces a ML based cryptanalysis strategy which later inspires further work in the domain. His work uses DL mechanisms to produce a neural based distinguisher tasked to perform key recovery attacks on the lightweight block cipher Speck [8].

The authors in [9] proposes a detailed analysis of the functionality of the new neural distinguisher who was thus far known to work as a black box. Studying the classified sets led to the conclusion that the neural distinguisher relies on the differential distribution on the ciphertext pairs. Moreover, they propose a ML based distinguisher which performs similarly using simple standard machine learning tools.

The other authors, [6] proposes a DL based cryptanalysis model whose task is to predict the key of block ciphers given known (plaintext, ciphertext) pairs. Their experiments show that the DL based approach successfully recovers the key bits in a reduced key space scenario. When key space restriction is not applied, the attacks are not successful unless the ciphers are round reduced. Their work is focused on lightweight block ciphers: simplified DES, Simon, and Speck.

Perushevska et al. in [10] apply DL for the cryptanalysis of the DES algorithm. The model uses (ciphertext, plaintext) pairs during training to learn to predict the plaintext. DES works in 16 rounds; their work analyses the full implementation and a reduced-round implementation with only one round. Also, in [11] they implement a deep neural network to perform a known-plaintext attack on AES. The goal is to recover the bits of the plaintext. Their work encompasses a wide range of experiments using different key sizes and modes of operation on AES. The results show that the proposed approach can restore the bits in the whole data set with a probability of more than 98%, and more than half of the plaintext bytes with a probability of 99%.

IV. METHODOLOGY

A. Prediction task

Our aim is to construct a model which would be able to predict the plaintext P_i given the ciphertext C_i . This effectively is a known plaintext attack where using pairs (P_i , C_i) we aim to train a DL model. The model is then tested in order to evaluate the number of plaintext bits it can correctly guess when provided with the ciphertext bits as input.

B. Dataset generation

A cipher algorithm is a sequence of operations which can roughly be explained as a very complex mathematical formula where the end goal is to conceal and obscure the input data. However, in order to conduct any data transformation, a few other elements are needed. These elements come in the form of raw input data which introduces randomness to the process since the inner workings of a cipher are not secret. A key and nonce are necessary to initialize the state of Ascon, this is a necessary step before any data concealing can even begin.

Encryption is already a complex nonlinear data transformation and as such it is not straightforward to learn. We keep this in mind when generating a dataset to train the model with. By carefully crafting the dataset and placing some restrictions and rules upon it, we ensure that aside from

the complexity of the algorithm, the input data is not entirely random and all over the place. That can be achieved by using a dataset where the inputs have small variability: any input i does not differ from any other input j by more than x bits (where x is some small number). This way, the model can gradually learn how small differences in the input manifest in the output. The same key was used for the creation of one dataset, since we want to simulate it as an ingrained part of the cipher itself, and not as a randomizing input parameter.

Dataset generation begins with the generation of one random key which will be used throughout the generation process. For the dataset creation, we created 1000 low variability ciphertexts and 1000 low variability nonces and then we combined each ciphertext with all the nonces from the nonces dataset. This resulted in a dataset with 1 million pairs of type (ciphertext, nonce). For each pair, the nonce and the predefined fixed key were used to initialize the cipher state. After initialization, associated data comes into play (mentioned in the introduction). However, the associated data affects the internal state, making it go through more rounds of permutation. We decided to eliminate this step (since it is optional) in order to simplify the processing as much as possible, so we didn't use associated data input (left it empty). This still does a change to the state, and 1 is XOR-ed to the last bit of the state. After the initialization, one is able to see the existing state of the cipher.

Ascon encrypts data by XOR-ing the data block (64 bits) to the first word of the state (also 64 bits, so no need for padding): $P_i \oplus S_1$. Knowing the plaintext-ciphertext pair produced by this XOR reveals the first word of the state itself.

In our dataset generation we wish to focus on the encryptions produced when the cipher state has gone through encryption permutation rounds. This is possible by skipping the first plaintext block (since it only meets with the initialized state that doesn't get affected by the encryption permutation) and focusing on the second block of the plaintext. We set the first block of the plaintext to all 0's. Knowing this, it can be assumed that the first word of the state is revealed as well, since we controlled the first input block. This is a subject to a later discussion.

Now, moving on to a phase where the encryption permutation module has been executed and using b encryption rounds, the permutation was applied to the state. Ascon applies the permutation in $b=8$ rounds. We experimented with round reduced versions which use from 1 up to 8 rounds, and for each round option we generated the next cipher state. This is important since XOR-ing the state to the generated ciphertext in the pairs provides the plaintext in the equation. This might seem like a backwards way to generate (plaintext, ciphertext) pairs, but since the ciphertext is the input to the model, it has to be the starting point (generated first) so the variability across the dataset can be controlled and reduced.

The resulting dataset contains 1 million records, each with values: (ciphertext, nonce, plaintext, state). Since the nonces are repeated for the dataset, this can be considered as a simulation of a nonce misuse scenario. This is not a disruption of the rules for train test datasets - no input train data is repeated in the test data. Additionally, even if the

number of unique ciphertexts in the dataset is 1000, the number of unique (ciphertext, nonce) input pairs is 1.000.000.

C. Feature selection

Hinting plaintext bits. The problem can be simplified further, by tasking the model to predict only a subset of the plaintext bits, while the other bits of the plaintext are revealed to it and are used as input when training. This captures the scenarios when chunks of plain data can be sniffed and uncovered, and only some parts remain unknown. In our work, we have experimented with revealing the first x bits of the plaintext and predicting the remaining bits, where x varies from 0 to 32.

Previous state. As hinted earlier, the state of the first word of the cipher after initialization is a sequence of 64 bits which might be revealed to the model. Just as the section above, this leads to branching in the experimentation department. Namely, one can use the state as additional information to the ciphertext and nonce in the task of predicting the plaintext. This state can be considered as completely unknown, partially revealed (a subset of the bits used as input to the ML model) or completely known.

D. Applying machine learning

The design of the models is responsible for the model performance. Specifying the hyperparameters of a model refers to defining the number of layers and neurons, activation function.

The number of layers, the number of nodes and the type of neurons in each hidden layer are hyperparameters that control the topology of the network. In our work we used a deep neural network with fully connected (dense) layers whose inside neurons connect to every neuron in the preceding layer. At each node, the input data (signal) gets multiplied by the weights in a node. The resulting value is then transformed using an activation function to produce the output (or activation) of the node.

In order to use stochastic gradient descent with backpropagation of errors when training DNN's, the activation function must be chosen so that complex relationships between input and output data can be learnt all while avoiding oversaturation. The ReLU activation function outputs the input if it is positive, otherwise outputs zero. Neurons in all layers except the output layers were set to use this activation function.

The sigmoid activation function is nonlinear and transforms the input data into a value between 0.0 and 0.1. Inputs larger than 1.0 are mapped to the upper bound (1.0), while extremely small values converge to 0.0. This is the default function to use when a binary answer is required. This activation is applied in the output layer of the neural network; the final sigmoid activation function performs the binary classification, and it dictates the prediction for each output bit. The cutoff is done by a threshold in a way that if $\text{sigmoid}(x)$ is greater than the threshold then the result is 1 (the default value for the threshold is 0.5).

1) Optimizer

During the compilation of the model, the Adam optimizer, and binary cross entropy loss were used. Adam is an optimization algorithm created as an extension to stochastic

gradient descent whose task is to iteratively update network weights based on input data. The algorithm works with the gradient and the squared gradient variables to calculate an exponential moving average [12].

2) Loss

Loss functions measure the deviation between a model's estimation of a value and the actual value. The loss function maps model predictions to their associated costs. Classification tasks aim to produce a discrete class prediction as output. Binary classification tasks are even simpler since only two classes exist. Our research is in the realm of binary classification, in the sense that we are predicting the values of bits in the plaintext which.

Entropy can be defined as a measure of randomness of the data during processing. It is a way to express the uncertainty associated with a given probability distribution. Difference between the randomness of two random variables is binary cross entropy. As the predicted probability diverges increasingly from the expected label the cross-entropy loss increases.

3) Train/test partitions

The models were trained on 80/20 train test datasets derived from the generated dataset (1 million records). Data is randomly picked from the dataset to generate the train and test datasets, and they are used equally for each model. This means, all models see the same training data, and get evaluated for the same test data.

4) Training Hyperparameters

Machine learning models have hyperparameters which control the learning process. These parameters can in turn affect the time to train and test the model, as well as the model's performance as well. Performance variations of the models are likely caused by a subset of all the hyperparameters, and we will be focusing on two of them below.

The batch size controls the accuracy of the estimate of the error gradient during training. Using larger batch sizes allows for speedups fueled by parallel computation using a GPU. A batch refers to the number of training samples used in an iteration

An epoch encompasses a single complete pass of all the training data through the machine learning algorithm. The model learns by updating the weight of nodes at the end of each epoch. Defining the number of epochs means controlling the number of times the model gets the opportunity to learn from the data.

If the batch size is defined big enough to include the entire training data, then the number of epochs will be equal to the number of iterations. This is usually not the case in practice since most models require more than one epoch to familiarize themselves with the data and learn. Generally, if d is the size of the dataset, while the number of epochs is e and batch size is b , the number of iterations i can be expressed as $i = d * e / b$

Choosing the hyperparameter values requires a deep understanding of the data supported by a lot of experimentation. The following section provides our insights about these hyperparameters.

V. EXPERIMENTS AND RESULTS

A. Setup

The neural networks in our research were implemented using the TensorFlow 2, Keras module. Initially we experimented with different neural network architectures, but in the end we settled for a fully connected neural network with 5 hidden layers between the input and output layers. The number of nodes is discussed below:

- Input layer: 1024 neurons.
- Hidden layers: The number of neurons in the layers ordered from input to output: 2048, 2048, 2014, 512, 256 fully connected neurons.
- Output layer: number of bits the model aims to learn and predict.

We experimented with networks with more layers and different neuron configurations, but that didn't seem to affect performance in a noticeable way. However, this is not eliminated as a direction for work in the future.

As mentioned, we used ReLU activation for the input and hidden layers, and sigmoid activation for the output layer. The Adam optimizer was used with the binary crossentropy loss function.

Our experiments focused on estimating the model's performance when different options are used for the input data:

- Plaintext hinting: the number of bits of the plaintext revealed to the model on input, while the rest is used as target when classifying
- State word reveal: whether the state of the cipher from the previous step (right after initialization) is available as input.

The number of epochs significantly affected performance consistency; variable results were obtained from multiple runs using 50 epochs for training. Better stability is achieved when 500 epochs are used. We used 5000 for batch size.

B. Performance metric

The classification results are expressed in terms of "percentage of correctly guessed bits". This is an aggregate metric across all test prediction results. More precisely, the metric percentage of correctly guessed bits is calculated for each test sample and then averaged for all. We also include a minimum and maximum guessed bits within one test sample metric. The goal is to catch the instances with low accuracy even if their number is low and will not affect the global metric in a noticeable way.

C. Results

The experiments displayed in the tables below used 150 epochs for training using a batch size of 5000.

When hinting the first j plaintext bits to the model and using them as input we use the notation $P[:j]$. We experimented with different options for plaintext hinting:

- Revealing the first half of the plaintext P and using it as input during training, while the other half was the target for predictions.

- Revealing only the first 20 bits and predicting the remaining 44 bits.
- Revealing no plaintext bits and predicting the entire plaintext block.

State hinting: The experiments where the input is enriched with the state of the cipher immediately after initialization are denoted as CNS0. The ones where only the ciphertext bits and the nonce bits are used as input for the model are denoted CN.

In the table below, the column Rounds denotes the number of encryption rounds used when generating the dataset, while the column Accuracy displays the percentage of correctly guessed bits metric.

TABLE I. DL MODEL PERFORMANCE SUMMARY

Rounds	Plaintext hinting	Accuracy
1	P[:32]	0.9947
2	P[:32]	0.9858
3	P[:32]	0.9989
4	P[:32]	0.9989
5	P[:32]	0.99997
6	P[:32]	0.7736
7	P[:32]	0.5149
8	P[:32]	0.5339
1	P[:20]	0.9879
2	P[:20]	0.9756
3	P[:20]	0.9747
4	P[:20]	0.9818
5	P[:20]	0.8907
6	P[:20]	0.8532
7	P[:20]	0.5108
8	P[:20]	0.5339

The experiment results in the table above were obtained using full state hinting (all the bits of the previous state are revealed: CTNS0)

When using no plaintext hints and no state hints, the performance of the models drops quickly: the precision for predicting the plaintext bits for $b=1$ is somewhere around 0.7.

Our experimental results confirm that as the number of rounds increases, the harder it becomes to learn the inherent plaintext-ciphertext bit relationships. This aligns with the theoretical results, and hope lies in the future that deeper analysis and more advanced DL techniques will exceed the theoretical results.

VI. CONCLUSION

Our work marks an experimental approach to cryptanalysis of Ascon using DL models. We believe this to be the first ML-based cryptanalysis of the named cipher. While lightweight ciphers are not new to DL-based attacks, the previous attempts are focused on key recovery. We successfully predict the plaintext (with known plaintext attack) with an accuracy of 0.998 percent.

A significant drawback is that our efforts attack a very niche-specific dataset where the randomness and variability are brought to a minimum. The generated dataset has a nonce misuse situation, which does the job of reducing randomness but is not a likely real-world scenario. The cipher itself is stripped of the finalization phase, which introduces a lot of security and ensures data integrity.

In practice, the cipher assumes and enforces randomness by applying nonces correctly and functioning as a unit where all the phases are being executed. Such functions can be very complex, and ML would not be able to find meaningful relationships between the inputs and the outputs. In a non-restricted scenario, DL-based cryptanalysis fails to attack the block ciphers.

We plan to extend our work in the future by randomizing the datasets more and strengthening the model performance using techniques more focused on the model design and training. Hopefully, this inspires the application of more ML-based attacks in the domain of cryptanalysis in the future.

ACKNOWLEDGEMENT

This research was partially supported by Faculty of Computer Science and Engineering at “Ss Cyril and Methodius” University in Skopje.

REFERENCES

- [1] A. Gohr, “Improving Attacks on Round-Reduced Speck32/64 Using Deep Learning,” Bundesamt für Sicherheit in der Informationstechnik (BSI), Germany, 2019
- [2] R. L. Rivest, “Cryptography and Machine Learning,” Laboratory for Computer Science Massachusetts Institute of Technology Cambridge, MA 02139,
- [3] C. Dobraunig, M. Eichlseder, F. Mendel, and M. Schläffer, “Ascon v1.2,” May 2021
- [4] P. Grabher, J. Großschädl, and D. Page, “Light-Weight Instruction Set Extensions for Bit-Sliced Cryptography”, In: Oswald, E., Rohatgi, P. (eds) Cryptographic Hardware and Embedded Systems – CHES 2008, 2008
- [5] C. Dobraunig, M. Eichlseder, F. Mendel, and M. Schläffer, “Cryptanalysis of Ascon,” March 2021
- [6] U. M. Khokhar, “Deep Learning-Based Cryptanalysis of Lightweight Block Ciphers”, July 2020
- [7] G. Hospodar, B. Gierlichs, E. De Mulder, I. Verbauwhede, and J. Vandewalle, “Machine learning in side-channel analysis: a first study”, October 2011
- [8] R. Beaulieu, D. Shors, J. Smith, S. Treatman-Clark, B. Weeks, L. Wingers, “The Simon and Speck Families of Lightweight Block Ciphers”, June 2013
- [9] A. Benamira, D. Gerault, T. Peyrin, and Q. Q. Tan, “A Deeper Look at Machine Learning-Based Cryptanalysis”,
- [10] M. Gj. Perusheska, H. M. Trpceska, V. Dimitrova, “Deep Learning-Based Cryptanalysis of Different AES Modes of Operation”, March 2022
- [11] S. Andonov, J. Dobrova, L. Lumburovska, S. Pavlov, V. Dimitrova, and A. Popovska-Mitrovikj “Application of Machine Learning in DES Cryptanalysis”, September 2020
- [12] D. P. Kingma, J. Ba, “Adam: A Method for Stochastic Optimization”, 2014

Framework for a Multipurpose Remotely Accessible Laboratory for Education

Vojdan Kjorveziroski

Faculty of Computer Science and Engineering

Ss. Cyril and Methodius University

Skopje, North Macedonia

Email: vojdan.kjorveziroski@finki.ukim.mk

Abstract—Hands-on exercises are an important aspect of students' education and they have a positive effect on their overall knowledge retention rates. With the Covid-19 pandemic on one hand, and the ever-increasing number of new students on the other, in-person computer laboratories are no longer a feasible option for the implementation of such practical lessons. The introduction of remote learning laboratories is one potential solution to both of these problems. However, existing remote laboratory implementations only focus on either a single scientific field, are not versatile in terms of supported infrastructure, or support only limited runtime options for the software to be used as part of the exercises. To overcome these issues, we have devised a remote laboratory framework consisting of 8 fundamental feature requirements which would allow flexible use in different fields of study, while at the same time enabling easy extension with existing and new simulation software. We discuss different implementation routes and report on the progress that we have already made in terms of setting up such a remote laboratory for the needs of our courses. We conclude that while most features can be implemented with existing open source software, challenges arise in terms of providing a self-service portal from which students can on-demand deploy resources.

Index Terms—remote learning, remote laboratory, infrastructure orchestration, virtualization, containerization

I. INTRODUCTION

Hands-on experience is one of the pillars of natural sciences education, including computer science. The introduction of real-world practical examples in the curricula directly impacts the students' knowledge retention rates and has an overall positive impact on the education process as a whole [1]. However, in the majority of cases the design and organization of traditional hands-on exercises is a burdensome process for all parties involved, both for the educators and the students alike, no matter whether they are executed remotely or in-person. On one hand the required resources need to be provisioned and configured, while on the other hand it cannot be assumed that all students have the necessary equipment at their personal disposal for successfully taking part in the devised scenarios at home. An obvious solution to the latter problem would be to simply make available on-premise equipment which would be provided for in-person students' use, but even this might prove challenging. In light of recent events, strict lockdowns during the Covid-19 pandemic largely prevented the use of institutional computer science laboratories. Furthermore, the number of natural sciences students increases every year [2] and universities might have logistical difficulties scaling the available resources to students' demand.

The work presented in this paper has received funding from the Faculty of Computer Science and Engineering under the "SCAP" project.

One potential solution which would allow to overcome these problems is the introduction of remotely accessible laboratories where the necessary computing infrastructure can be provisioned, configured, and made available to students either by the system administration department of the university or by the course educators themselves. By leveraging either the private cloud infrastructure of the institution or by utilizing public cloud solutions, students can be provided with effortless remote access capabilities, while at the same time not requiring any high-performance infrastructure locally [3]. In such a scenario, the exercise setup would be completed beforehand, without the need for additional effort by the students, thus allowing them to focus on solving the challenges themselves instead of spending time on environment setup.

Should these remote laboratories be implemented in a scalable and easy to maintain manner, their advantages would be many-fold. Firstly, they would facilitate easy collaboration between students who all have access to the same centralized resources on a scale which is simply not feasible when each student completes the exercises on their own local machine or on-premise with a subset of the group. Furthermore, by following security best practices in terms of isolating the remote laboratory from the rest of the production systems, even potentially disruptive exercises can be organized, such as the showcasing of vulnerable software components or practicing with malicious exploits, concepts important for cybersecurity related courses [4]. Finally, by maintaining a pool of historical exercises, students can on-demand, and in their own time, use the infrastructure to revise important concepts in which they might be interested in.

The aim of this paper is to provide a framework for the implementation of a general purpose, infrastructure agnostic, versatile remotely accessible laboratory that can be utilized as part of various university courses requiring hands-on exercises, while not being strictly limited to the area of computer science. The main contributions of this work are:

- Presentation and discussion regarding the state of the art research in terms of implementing remote access laboratories in an education context.
- Definition of a framework for future remote access laboratories, not limited to a particular scientific field.
- Discussion of potential routes in which remote learning laboratories conformant to the framework can be implemented.
- Reporting on real-world experience acquired during efforts to implement such a remote learning laboratory in a number of computer science related courses.

The rest of this paper is organized as follows: in Section II we discuss related work to the area of remotely accessible

laboratories for education purposes, before moving forward with Section III where we outline the set of requirements that we have devised for the implementation of such a remote laboratory. In Section IV we discuss concrete ways in which the requirements outlined in Section III can be met, while also describing various approaches which we have already used during a number of courses at our faculty. We conclude the paper with Section V where we summarize the presented framework and discuss future planned work.

II. RELATED WORK

The need for remotely accessible laboratories that can augment the education process while allowing students to freely experiment and improve their understanding of novel concepts has already been recognized. Efforts have been made to improve the usability of such laboratories by both academia and industry. In the text that follows we outline the state of the art research related to this topic.

Remote learning laboratories have been traditionally very popular in a computer networking context, providing an easy and cost-effective way of gaining hands-on experience with enterprise level devices. Using different forms of virtualization, coupled with advancements in software defined networking (SDN), network devices such as routers or switches can be run on general purpose hardware, eliminating the need for students to buy expensive equipment, thus allowing them to easily acquire much needed experience using the latest revisions of both the software and hardware [5]. Unfortunately the network simulators that make all of this possible usually have very high system requirements in order to be able to virtualize these complex devices and are difficult to set up for inexperienced users taking their first steps in these fields. To alleviate these problems, centralized infrastructures where users can deploy a number of virtualized network devices for training purposes in an ad-hoc manner have been developed. Álvarez et al. [6] present one such implementation of a remote laboratory based on QEMU/KVM virtual machines that users can leverage as an alternative to commercial network training platforms offered by major vendors. Continuing this trend, Scazzariello et al. [7] describe a more modern architecture for deploying virtualized network devices, utilizing a container orchestrator to manage the containerized network devices and ensure the long term scalability of the platform. By developing a custom container network interface (CNI) plugin [8], [9], the authors have managed to adapt the current limitations of the Kubernetes orchestrator of assigning multiple independent interfaces per containers, thus enabling this new use-case.

Another area of particular interest where remote laboratories can be utilized is in the context of cybersecurity exercises. Cybersecurity courses are faced with the challenge of finding a solution to the ever-present problem of how to securely offer vulnerable environments where students can get hands on experience regarding topics that they have only theoretically learned during the curriculum. Sianipar describes Tele-Lab [10] which leverages virtual machines, and in certain cases containers, to offer vulnerable virtual environments which can then be customized using shell scripts after their deployment.

It should be noted that general purpose remote laboratories not targeted at a particular subject matter have been described in literature as well [11]–[15]. Unfortunately, the majority of them focus exclusively on a given runtime technology such

as virtual machines [14], [15] or containers [11], [12], thus limiting the number of possible scenarios and impacting the reusability of already created content by forcing reimplementations in the supported runtime environment.

Apart from academia, industry has also identified the benefits of remote laboratories, with many cloud providers offering virtual training for their own products in a sandbox environment [16], [17]. Even though most of these courses are free and accessible to the wider audience, they focus exclusively on the technology of the given cloud provider. Taking into account the low levels of compatibility between offerings of different cloud vendors, vendor lock-in is a major challenge in this area [18].

In conclusion, while there are existing efforts for implementing remotely accessible laboratories for education, they are focused on hosting only certain types of workloads relevant to a particular subject matter, offer limited support for different runtime environments, or can be deployed only on certain types of infrastructure. In our opinion a development of a general purpose remote laboratory framework is needed. Such a framework should be infrastructure agnostic, highly scalable, with support for different runtime environments allowing the deployment of both new and existing workloads. A later practical implementation of the framework would be beneficial not only for computer science education, but also for other scientific disciplines.

III. REQUIREMENTS EVALUATION FOR A REMOTE LEARNING LABORATORY

The design of a multipurpose remotely accessible laboratory for education demands careful consideration and evaluation of requirements. Different use-case scenarios might require distinct, often conflicting features. Careful balance must be stricken in terms of supported scenarios, environment isolation, and ease-of-use both in terms of administration and regular usage. On one hand, use-cases such as remote development in various programming languages require a large catalog, supporting different tools and programming languages. On the other hand, strong security measures combined with strict network isolation are essential for scenarios related to cybersecurity, guaranteeing the integrity of the underlying platform itself. Finally, various other fields already have a sizable catalog of simulation software which can only run in certain environments, for example on a given version of an operating system, or only within containers. Effort should be made to support different runtime environments as to facilitate easy incorporation of existing software into the remote laboratory.

To tackle these challenges, in the remainder of this section we outline 8 fundamental features (abbreviated as F1-F8) for the implementation of a remote laboratory.

- F1: Infrastructure agnostic – The remote laboratory should support deployment both in a virtualized environment, but also directly on bare-metal servers. Different deployment options should be interoperable with each other, allowing the virtual laboratory to be hosted both on physical and virtual servers at the same time. This would ensure both the performance and elasticity of the underlying platform. In times of high resource demand, virtualized resources either from on-premise institutional private clouds or public clouds can augment the per-

manent hardware resources where the platform itself is hosted.

- F2: Support for different runtime environments – Effort should be made to support both containers and virtual machines as runtime environments for the deployed scenarios and software on top of the platform. This would ensure greater compatibility with existing software that can currently be deployed in a standalone fashion either as containers or inside virtual machines, while saving time which would otherwise be dedicated to conversion. The support for different runtime environments would also future proof the platform, enabling novel technologies such as serverless functions to be utilized, should the need arise.
- F3: Free from vendor lock-in – Even though both virtualization and containerization are general concepts, unfortunately in practice their various implementations are not compatible with each other. This problem is even more pronounced when it comes to hypervisors, with different virtual hard disk formats and guest drivers required for interacting with the host operating system. In terms of containerization, the Open Container Initiative (OCI) [19] aims to create industry standards around container formats and runtimes, allowing container image reuse among different containerization solutions. Nevertheless, the definition of an abstraction layer which will support various virtual machine templates regardless of the hypervisor where they were created is paramount to enable easy migration of existing software solutions to the remote laboratory.
- F4: Integration with existing external systems – The remote laboratory cannot be seen as an isolated silo in terms of the existing and well-proven e-learning systems in use today. The majority of universities have already deployed learning management systems (LMS) for keeping track of students' records and for distributing learning materials. Any platform which would enable the implementation of a remote laboratory needs to support integration with such existing systems, aiding both the registration process of the new users, as well as syncing course participation information necessary for filtering the available scenarios. Furthermore, such an integration should be versatile as to allow both manual and automatic data exchange between the systems which can later also be analyzed to determine students' performance [14].
- F5: End-to-end connectivity between the students and the remote laboratory – Secure end-to-end connectivity between the remote laboratory and the students' computing devices would allow direct communication, no matter the communication protocol that the target software uses. Additionally, such an end-to-end connectivity, might also enable the incorporation of local resources available to the students within the context of an elaborate scenario, a strategy already discussed in [10].
- F6: Efficient use of compute resources – The remote laboratory should offer versatile options for supporting an influx of students. During the implementation process, scenarios where the requested computing capacity exceeds the available capacity should be taken into account. This should not be seen as an unlikely

occurrence, since the number of students increases every year, and courses with large number of students which would incorporate the use of the remote laboratory within their curriculum might outpace the rate at which new computing capacity is made available. Multiplexing strategies need to be implemented to deal with such scenarios.

- F7: Effortless collaboration between students – One of the major benefits of a centralized remotely accessible laboratory is that it enables easy collaboration between students no matter their group size. To facilitate this, despite supporting an isolated tenant model, where each student is given an independent remote working environment, shared access to a centralized instance should be possible as well.
- F8: Versatile interfacing options – To avoid requiring any prerequisite knowledge, students and educators alike should be able to choose the interface option most suited to them, either a graphical user interface (GUI) or a command line interface (CLI). The support of a CLI option would make bulk resource provisioning by educators easier, together with requiring less effort for integration with third party systems. On the other hand, a GUI would be necessary from the students perspective, providing easy way to browse the list of supported software and scenarios.

In the section that follows we proceed by outlining guidelines for meeting the stated feature requirements defined in the framework.

IV. DISCUSSION OF IMPLEMENTATION OPTIONS

The main pillar which would allow the development of a remote laboratory that satisfies all of the previously outlined features is the introduction of a layered design using multiple abstractions. The implementation of such abstraction layers would ensure the wide-ranging compatibility with different virtualization and containerization platforms while providing a unified view towards the end-users. Fortunately, the advancements made in terms of various configuration management, orchestration, and provisioning tools allow easy integration with diverse systems while relying on a single interface – that exposed by the chosen tool.

Figure 1 provides a visual overview of the implementation possibilities of the framework, outlining the relationship between the various feature requirements. In the remainder of this section, we proceed with the discussion of implementation options for the relevant features, grouped in different subsections.

A. Infrastructure and Runtime Versatility

The F1, F2, and F3 features can be satisfied with an elaborate use of popular and open source provisioning, orchestration, and configuration management tools.

Provisioning tools such as Terraform [20] can be used to deploy infrastructure such as virtual machines, firewalls, routing rules, and containers across different environments, as long as they expose a documented application programming interface (API). With the help of the wider open source community, such tools have implemented compatible modules for various solutions [21], allowing effortless deployment using a unified syntax no matter the underlying platform, thus solving both F1 and F3. With the promotion of the infrastructure as

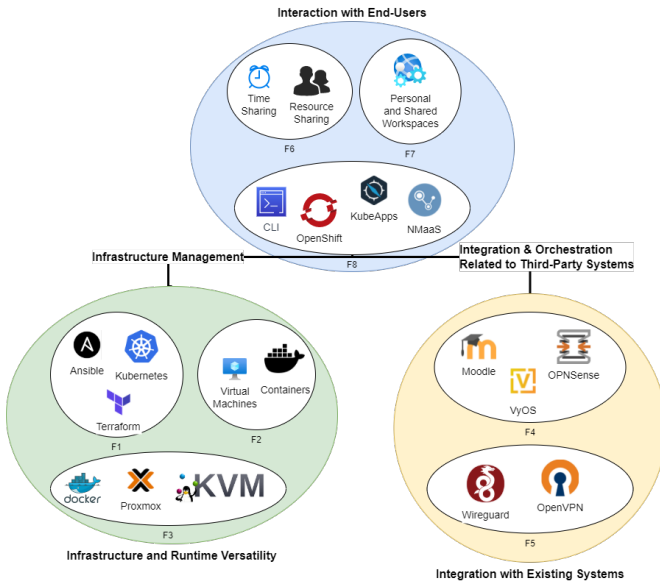


Fig. 1. Visual Representation of the Implementation Options for the Reference Framework

code (IaC) concept [22], infrastructure configuration can be written once and applied across different environments that are compatible with a given simulation software, without the need for any new tooling.

Once the infrastructure has been deployed using a provisioning tool it can be further customized with the help of configuration management solutions, such as Puppet [24] or Ansible [23]. In this manner, different lab scenarios can be described exactly once using the chosen configuration management tool and later shared between users with the aim of extending the number of supported scenarios. Such an approach would allow the creation of scenario catalogs based on IaC scripts written in different configuration management languages. This strategy of describing exercises using configuration management tools has been successfully tested and integrated in practice by the SecGen project [25], [26].

After provisioning and configuration of both the infrastructure and simulation software, the question of its management arises. For this task, orchestration tools can be used which aim to reconcile the actual state of the services with the recorded desired state specified by the administrator. While the use of container orchestrators for managing large fleets of containers has been extensively studied and implemented in practice even in remote learning laboratories [7], their shortcoming is that they are focused solely on a single runtime environment – containers. To satisfy F2, an orchestrator should support both containers and virtual machines at the same time, using the same interface. One possible solution to this problem is the Kubevirt project [27] which modifies the Kubernetes orchestrator to support QEMU/KVM based virtual machines as well. In this manner the Kubernetes scheduler can be used to schedule virtual machines across all cluster nodes, together with containers using a single seamless API, thus fully satisfying F2.

We have already evaluated the use of the SecGen tool to generate randomized virtual machines in the context of a network security course [28]. We have also employed the combination of Terraform and Kubevirt in a more recent instance of the course, deploying both containers and virtual machines on a large scale, for every student taking part. We

can report that with the help of these tools it is possible to implement a remote learning laboratory, but without any user-facing interface, students are not able to deploy resources on-demand, and instead the grunt of the work in terms of instance deployment is left to the educators themselves. Options which would allow direct end-user control are discussed in the subsection Interaction with End-Users below.

B. Integration with Existing Systems

To satisfy the requirements set out by F4 and F5 any remote learning laboratory will need to support integration and data exchange with existing systems which are already in use. Unfortunately, in practice, this might be a challenging task, since any third-party component would need to have a well-defined API through which communication can be established. Furthermore, in certain cases different people might be responsible for the maintenance of the remote laboratory on one hand, and the maintenance of the third party systems such as the LMS on another, a fact that would further complicate the testing and enablement of such integrations.

Perhaps the most important and common integration that will be required is that with an existing LMS. Fortunately, many popular learning management systems today expose an API which can be consumed for harvesting relevant student data for the remote laboratory, such as basic information, enrolled courses, and student performance [29]. In cases where such an API is not available, administrators will need to resort to programmatically executing web requests which would achieve the desired functions by simulating a real client. In our case, we have successfully used both options before; we have utilized the Moodle web service functionality as well as created Python wrappers for programmatically simulating requests, which allowed us to perform actions which are not natively supported by the web service implementation yet. It should be noted though that any large scale implementation of a remote learning laboratory will have to extend this integration towards the end-users, so that they can independently deploy resources directly from the user interface of their choice.

To facilitate remote connectivity to the environment, users will need to resort to the use of virtual private networks (VPNs). As was the case with the LMS, integration with existing systems will be required as well, since many institutions already have VPN concentrators in place both for staff and students. The only requirement is that granular firewall rules can be configured, targeted at each users of the remote laboratory individually. This would ensure network level isolation between the rest of the infrastructure and the users of the remote laboratory, as well as among the users themselves, ensuring that they can only access resources which have been personally deployed. In cases where the infrastructure is being set up from scratch, many popular open source solutions exist that can act as both a firewall and a VPN concentrator, such as PfSense, OPNSense, or VyOS. Care must be taken to first evaluate the quality of any API before making a selection. In our current remote learning environment we are using OPNSense, but since not all required features have been exposed via the API yet, additional tooling had to be created which programmatically executed web requests, simulating a client browser.

C. Interaction with End-Users

The last three feature requirements, F6, F7, and F8 all deal with the way that end-users interact with the remote learning laboratory. A common issue being faced is that the number of students attending a given course drastically varies per year, and as such it might not always be feasible to provision additional hardware resources to meet the new requirements in time. This problem can be solved by introducing some sort of a sharing concept in terms of the compute infrastructure. Perhaps the least intrusive solution for the lack of resources would be to implement a time sharing mechanism, where students can schedule when they would like to use the remote learning laboratory. This would limit the number of users at any single point in time, allowing the infrastructure to meet the demand of users.

Certain scenarios and exercises require team work and the remote learning laboratory together with the implemented security and isolation features should be flexible enough to support this. This can easily be solved with the introduction of the workspaces concept, where each user is given their own dedicated workspace, but should the need arise these workspaces can be shared between multiple users either perpetually or for a limited time. In this manner, resources deployed by a single student in a given resource would be accessible by all other teammates.

The last, and perhaps the most challenging feature to implement is a user facing application that would allow students to access, configure, and deploy resources within the virtual environment by themselves. Such an application should incorporate support for all of the features described previously, such as: allow users to deploy workloads from the catalog depending on the courses that they are enrolled in; allow users to schedule a desired timeslot for using the remote learning laboratory; dynamically generate firewall rules to reflect different sharing states of the workspaces; facilitate easy addition of new exercises by educators and administrators; interface with the different orchestration, configuration, and provisioning tools to support the on-demand deployment of applications within the catalog. To the best of our knowledge, no such tool exists that can satisfy all of these requirements by itself at present. However, the use of Kubevirt as a Kubernetes extension that allows the Kubernetes API to be used to manage virtual machines along with containers would allow the repurposing of Kubernetes focused applications that allow users to select and deploy workloads. Examples include NMaaS [30], OpenShift [31], and KubeApps [32]. However, it should be noted that none of these satisfy all of the features in their current state, so they can only serve as a starting point from where additional modifications will need to be made to their source code.

To solve this final piece of the puzzle, we plan to evaluate the various existing open source solutions and determine whether they can be modified to support all 8 feature requirements or whether a new platform will have to be created from scratch.

V. CONCLUSION

We have presented a framework consisting of 8 distinct feature requirements for the implementation of a universal and versatile remote learning laboratory, supporting both new and existing simulation software. Through the discussion of possible implementation routes, we conclude that the

majority of the requirements can be satisfied using open source software, thanks to the recent advancements made in terms of the robustness of configuration management, orchestration, and provisioning tools. We have also reported on our current progress in terms of implementing a conformant remote learning laboratory for use at our faculty during various courses. Even though this remote laboratory, in its current state, is still work in progress, students have already expressed satisfaction and they prefer assignments which make use of the centralized infrastructure instead of those that must be completed locally. This is understandable, since by making use of the remote laboratory students do not need to configure anything, and can instead focus directly on the task at hand, foregoing any manual setup procedures for the used software. However, attention should be paid to the level of abstraction introduced when deploying software to be used during the exercises, since too much abstraction would have a negative impact in terms of the students' understanding of the underlying architecture of the used simulation tools. This is especially true for computer science subjects, where manually dealing with the complete installation process can prove beneficial and can be used to teach valuable troubleshooting lessons. Nevertheless, using the presented framework a balanced approach can be implemented, where students are tasked with installing the software manually within their assigned virtual workspaces.

One open issue remains before the conceptual framework can be realized in its entirety. To support the desired level of versatility in terms of runtime environments, supporting both virtual machines and containers, either some of the existing open source platforms will have to be adapted or a new one built from scratch. No matter the route taken, the resulting platform will have to support direct access by students, allowing them to on-demand deploy exercises while providing integration with external systems for acquiring information regarding course enrolment, grades, and dynamically provisioning firewall rules for isolation. We plan to evaluate how the existing open source candidates can be adapted to meet the requirements in a future work, and if the required effort is proven to be infeasible, implement such a platform from scratch, thus concluding all of the 8 outlined features in the presented framework.

REFERENCES

- [1] P. M. Stohr-Hunt, 'An analysis of frequency of hands-on experience and science achievement', *Journal of Research in Science Teaching*, vol. 33, no. 1, pp. 101–109, 1996, doi: [https://doi.org/10.1002/\(SICI\)1098-2736\(199601\)33:1<101::AID-TEA6>3.0.CO;2-Z](https://doi.org/10.1002/(SICI)1098-2736(199601)33:1<101::AID-TEA6>3.0.CO;2-Z).
- [2] 'Tertiary education statistics', Eurostat. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Tertiary_education_statistics (accessed Apr. 03, 2022).
- [3] N. Petrovic, V. Nejkovic, and M. Tasic, 'Dealing with Scalability of Laboratory Sessions in Computer Science University Courses', in 2018 26th Telecommunications Forum (TELFOR), Nov. 2018, pp. 1–4. doi: <https://doi.org/10.1109/TELFOR.2018.8612090>.
- [4] D. C. Rowe, B. M. Lunt, and J. J. Ekstrom, 'The role of cyber-security in information technology education', in *Proceedings of the 2011 conference on Information technology education*, New York, NY, USA, Oct. 2011, pp. 113–122. doi: <https://doi.org/10.1145/2047594.2047628>.
- [5] R. K. CV and H. Goyal, 'IPv4 to IPv6 Migration and Performance Analysis using GNS3 and Wireshark', in 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), Mar. 2019, pp. 1–6. doi: <https://doi.org/10.1109/ViTECoN.2019.8899746>.
- [6] C. Álvarez, 'Developing a Flexible Virtual Networking Laboratory Platform for Education', in *Proceedings of the IV School on Systems and Networks*, SSN 2018, Valdivia, Chile, October 29–31, 2018, 2018,

- vol. 2178, pp. 35–38. [Online]. Available: http://ceur-ws.org/Vol-2178/SSN2018_paper_9.pdf.
- [7] M. Scazzariello, L. Ariemma, G. Di Battista, and M. Patrignani, 'Megalos: A Scalable Architecture for the Virtualization of Large Network Scenarios', *Future Internet*, vol. 13, no. 9, Art. no. 9, Sep. 2021, doi: <https://doi.org/10.3390/fi13090227>.
- [8] R. Kumar and M. C. Trivedi, 'Networking Analysis and Performance Comparison of Kubernetes CNI Plugins', in *Advances in Computer, Communication and Computational Sciences*, Singapore, 2021, pp. 99–109. doi: https://doi.org/10.1007/978-981-15-4409-5_9.
- [9] Megalos CNI. Kathará, 2022. Accessed: Apr. 03, 2022. [Online]. Available: <https://github.com/KatharaFramework/Megalos-CNI>.
- [10] J. H. Sianipar, 'Towards scalable and secure virtual laboratory for cybersecurity e-learning', *Universität Potsdam*, 2020. doi: <https://doi.org/10.25932/publishup-50279>.
- [11] G. Chen, 'CvLabs: A Container Based Interactive Virtual Lab for IT Education', *Georgia Institute of Technology, Technical Report*, Apr. 2020. Accessed: Mar. 29, 2022. [Online]. Available: <https://smartech.gatech.edu/handle/1853/64898>.
- [12] L. Tobarra, A. Robles-Gómez, R. Pastor, R. Hernández, A. Duque, and J. Cano, 'Students' Acceptance and Tracking of a New Container-Based Virtual Laboratory', *Applied Sciences*, vol. 10, no. 3, Art. no. 3, Jan. 2020, doi: <https://doi.org/10.3390/app10031091>.
- [13] T. Budai and M. Kuczmann, 'Towards a Modern, Integrated Virtual Laboratory System', *Acta Polytechnica Hungarica*, vol. 15, no. 3, p. 14, 2018.
- [14] Y. Deng, D. Lu, C.-J. Chung, D. Huang, and Z. Zeng, 'Personalized Learning in a Virtual Hands-on Lab Platform for Computer Science Education', in *2018 IEEE Frontiers in Education Conference (FIE)*, Oct. 2018, pp. 1–8. doi: <https://doi.org/10.1109/FIE.2018.8659291>.
- [15] M. Perales, L. Pedraza, and P. Moreno-Ger, 'Work-In-Progress: Improving Online Higher Education with Virtual and Remote Labs', in *2019 IEEE Global Engineering Education Conference (EDUCON)*, Apr. 2019, pp. 1136–1139. doi: <https://doi.org/10.1109/EDUCON.2019.8725272>.
- [16] 'AWS Educate', Amazon Web Services, Inc. <https://aws.amazon.com/education/awseducate/> (accessed Mar. 29, 2022).
- [17] 'How to get started in cloud computing — Google Cloud Training', Google Cloud. <https://cloud.google.com/training/getstarted> (accessed Apr. 03, 2022).
- [18] M. Armbrust et al., 'Above the Clouds: A Berkeley View of Cloud Computing', undefined, 2009, Accessed: Apr. 03, 2022. [Online]. Available: <https://www.semanticscholar.org/paper/Above-the-Clouds%3A-A-Berkeley-View-of-Cloud-Armbrust-Fox/1f7190fc294246f83f1f331cc51e3264851d0d36>
- [19] 'Open Container Initiative - Open Container Initiative'. <https://opencontainers.org/> (accessed Apr. 03, 2022).
- [20] 'Terraform by HashiCorp', Terraform by HashiCorp. <https://www.terraform.io/> (accessed Apr. 03, 2022).
- [21] 'Terraform Registry'. <https://registry.terraform.io/> (accessed Apr. 03, 2022).
- [22] T. A. Limoncelli, 'GitOps: A Path to More Self-service IT: IaC + PR = GitOps', *Queue*, vol. 16, no. 3, pp. 13–26, Jun. 2018, doi: <https://doi.org/10.1145/3236386.3237207>.
- [23] 'Ansible is Simple IT Automation'. <https://www.ansible.com> (accessed Apr. 03, 2022).
- [24] 'Powerful infrastructure automation and delivery — Puppet'. <https://www.puppet.com/> (accessed Apr. 03, 2022).
- [25] Z. C. Schreuders, T. Shaw, M. Shan-A-Khuda, G. Ravichandran, J. Keighley, and M. Ordean, 'Security Scenario Generator (SecGen): A Framework for Generating Randomly Vulnerable Rich-scenario VMs for Learning Computer Security and Hosting CTF Events', presented at the *2017 USENIX Workshop on Advances in Security Education (ASE 17)*, 2017. Accessed: Mar. 05, 2022. [Online]. Available: <https://www.usenix.org/conference/ase17/workshop-program/presentation/schreuders>
- [26] Cliffe, Security Scenario Generator (SecGen). 2022. Accessed: Apr. 03, 2022. [Online]. Available: <https://github.com/cliffe/SecGen>
- [27] 'KubeVirt.io', KubeVirt.io. <https://kubevirt.io/> (accessed Jan. 14, 2022).
- [28] V. Kjorveziroski, 'Engaging Students with Personalized and Remotely Orchestrated Cybersecurity Training Exercises', in *Proceedings of the 18th International Conference for Informatics and Information Technology*, May 2021, pp. 48–53. [Online]. Available: <http://ciit.finki.ukim.mk/proceedings/ciit-2021-proceedings.pdf>.
- [29] 'Web service API functions - MoodleDocs'. https://docs.moodle.org/dev/Web_service_API_functions (accessed Apr. 03, 2022).
- [30] 'NMaaS Home - Network Management as a Service - GEANT'. https://www.geant.org/Services/Connectivity_and_network/NMAaS (accessed Jan. 14, 2022).
- [31] 'OpenShift Virtualization'. <http://content.cloud.redhat.com/learn/topics/virtualization/> (accessed Apr. 03, 2022).
- [32] 'KubeApps, deploy your applications in Kubernetes'. <https://kubernetes.com/> (accessed Jan. 14, 2022).

Understanding Public Perception of Waste Management in Western Balkans using Tweets and News Reports Analysis

Ivana Gjorshoska

Macedonian Academy of Sciences and Arts
Skopje, Republic of North Macedonia
ivana.gjorshoska@students.finki.ukim.mk

Aleksandra Dedinec

Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of North Macedonia
aleksandra.kanevche@finki.ukim.mk

Angela Madjar

Macedonian Academy of Sciences and Arts
Skopje, Republic of North Macedonia
angela.madzhar@students.finki.ukim.mk

Jana Prodanova

Macedonian Academy of Sciences and Arts
Skopje, Republic of North Macedonia
jprodanova@manu.edu.mk

Abstract—The rapid urbanization and economic growth greatly contribute to the increasing amount of produced solid waste worldwide. As part of their pro-European policies, many countries from the Western Balkans have started to accept and apply the EU Waste Management Law in their legal systems as well as in practice in order to create adequate waste management schemes. The aim of this paper is to analyze the public response regarding waste management regulations and practices using Twitter and News reports data analysis within a period of four months, between November 1st, 2021 and February 28th, 2022. Sentiment analysis and Dynamic Topic Modeling were used to assess the public attitude in both news reports and Twitter responses, in addition to bigrams similarity using soft cosine measure to determine the semantic correlation between the two sources. The results suggest that there is a predominantly negative attitudes towards waste practices in both media outlets. Additionally, there is a strong semantic correlation between the discussed topics in news reports and on social media, leading to the conclusion that news outlets play a major role in promoting pro-environmental behavior in the general public in Western Balkans.

Keywords—waste management, Western Balkans, social media, twitter analysis, news reports, sentiment analysis, dynamic topic modeling, semantic similarity

I. INTRODUCTION

With the acceleration of urban development, the amount of generated waste in the world increased dramatically and this trend is only said to continue in the following years. According to the World Bank, rapid urbanization, population growth and economic development will increase annual waste generation by 70 percent over the next 30 years, from 2.01 billion tonnes being generated globally in 2016 to 3.40 billion tonnes in 2050 [1]. Around 37 percent of the global waste is disposed in some type of landfill and 33 percent is mismanaged through open dumping or burning, whereas adequate waste disposal is almost exclusively done in high-income countries. Inadequate waste management in urban areas negatively impacts the environment and the climate, public health and ultimately the economy. Environmental pollution of waste dumping affects health through short-term effects such as congenital anomalies, asthma and respiratory infection, while long-term exposure to waste may cause chronic respiratory and cardiovascular diseases, cancer and even brain, nerves, liver or kidneys diseases [2], [3]. To reduce the negative impacts of waste, different countries have

adopted various actions. Since 2000, Japan began implementing the Containers and Packaging Recycling Law [4]. The United States enforced the Resource Protection and Recycling Act as well as different cities establishing their own regulations [5].

In the European Union, 5.2 tonnes of waste were generated per EU inhabitant in 2018, 38.5 percent of which was landfilled while 37.9 percent was recycled [6]. Due to greater economic development and EU legislation, many of the higher recycling and collection rates happen in Western Europe. However, many countries, especially countries from Western Balkans aim to meet the EU Waste Management Law as part of their pro-European policies. The governments of Serbia and Montenegro reported 338 kg and 548 kg of municipal solid waste per inhabitant in 2019 respectively, while North Macedonia reported consistent increase over the last 10 years - from 351 kg in 2010 to 452 kg per inhabitant in 2020 [7]–[9]. Additionally, the amount of waste generated annually per capita is expected to increase linearly until it reaches the corresponding waste projections of EU28 in 2035 [11]. Latest reports in the aforementioned countries also suggest high collection rates, between 70 percent and 90 percent of collected waste, however only up to 3 percent of the collected waste is being recycled and the rest is either landfilled, in both regulated and unregulated landfills, or openly burned [10]. The current waste treatment and increasing trend of waste generation plays a major role in greenhouse gas emissions, where up to 80 percent of emissions caused by the waste management sector come from solid waste disposal in landfills [10]. The main struggle for waste management in Western Balkans can be attributed to outdated infrastructure for collection and transport, illegal dumpsites and the lack of facilities to sort, clean and aggregate recyclable materials [7]–[9]. Lack of adequate policies, inefficient allocation of external resources and lack of awareness among people and companies have been identified as barriers for transitioning to green economy [12].

There is significant research done in the field of public perception for waste management and other environment-related topics based on interactions on social media platforms. Some of the exemplary research includes analyzing attitude towards municipal waste policy as well as promotion of

campaigns to encourage consumers' awareness [13], [14]. However, there is a lack of research about public perception towards waste management in the Western Balkan region. The aim of this paper is to assess the public perception about waste management via Twitter posts, as well as to research the possible correlation between tweets related to waste and news reports in the Western Balkans. The paper is organized as follows. Section II describes the data collection process, storing and preprocessing. Section III presents and explains the methods used for sentiment analysis, topic modeling and similarity assessment between tweets and news reports. Section IV presents and discusses the results obtained from the analysis and finally, Section V provides conclusions of this study and presents possibilities for future work.

II. DATA

For the purpose of the study, we collected tweets related to waste using Twitter's stream API with an authenticated application that connects to a public stream consisting of a sample of tweets that are posted on Twitter [15]. Although many studies for sentiment analysis rely on using hashtags to filter tweets, the filtering of relevant tweets in our case was done using keywords instead of hashtags. This decision was done due to a lack of hashtags relevant to this topic used in the Western Balkan region. The tweets were filtered based on keywords related to waste in Macedonian and Serbo-Croatian i.e. "отпад" (waste), "отпад" (waste), "ѓуѓре" (garbage) and "рециклира" (recycle). Waste-related tweets were collected on a weekly basis over a period of 4 months starting November 1st, 2021 until February 28th, 2022.

After the initial collection of the tweets, they were saved as an Excel spreadsheet for further data cleansing which included checking and filtering the appropriate tweets about waste-related topics. This secondary filtering step was necessary for this study due to language ambiguity. Namely, words such as "trash" and "garbage" tend to be used as derogatory terms for people, objects or situations and this usage of the aforementioned keywords used for filtering can be seen in the Western Balkan region as well. The collected tweets were translated into English using an online document translation tool and parts of the tweets such as "RT" standing for retweet and usernames were removed from the tweets in order to focus on the tweets' content [16].

The decision to analyze news reports was done due to the fact that journalism on Western Balkans plays a major role in enhancing societal transitions [17]. In order to obtain news reports data, a web scraping method was used for data collection. We used the BeautifulSoup python library to scrape news reports from the websites time.mk and time.rs, which are cluster-based news websites that collect and aggregate news from 120 distinct sources [18]–[20]. Similarly as the tweets, the news reports were filtered using the keywords "otpad" (waste) and "отпад" (waste) and were collected weekly over the same time period as the Twitter data. During this process, we collected both the news title and the news preview to get a better picture about the contents of the news reports for the respective week. The news reports datasets were stored and translated using the same techniques as the waste-related tweets.

III. METHODOLOGY

A. Sentiment Analysis

Sentiment analysis is the practice of using algorithms to classify various samples of related text into positive and negative categories. This method is widely applied in analyzing customer reviews and responses, however it has also been used in online social media platforms for analyzing online opinion. Sentiment analysis within microblogging has verified Twitter as a valid online indicator of political sentiment reflecting the offline political landscape [21]. It has also been used for analyzing tweets for social events and health care tweets [22], [23].

There are several models and tools that can be used for sentiment analysis of text, however none of the tools available right now can do sentiment analysis on Macedonian or Serbo-Croatian text, resulting in the previously explained English translation of the collected data in Section II. For the purpose of sentiment analysis in our research, we used a pre-built Sentiment Intensity Analyzer [24], [25]. The analyzer is built using VADER, a human-centered sentiment lexicon that combines lexical features from well-established sentiment word-banks and features that are common in microblogging, including acronyms, slang and emojis. Every candidate feature of this lexicon has been assessed using a Wisdom-of-the-Crowd approach as well as using identified heuristics humans use to assess sentiment intensity of text, which lead to creating a human-validated sentiment intensity on corpora within four domains including social media text, movie reviews, technical product reviews and opinion news articles. The Sentiment Intensity Analyzer from the NLTK library uses this lexicon to assess the sentiment intensity on a piece of text by verifying whether any of the words in the content are available in the lexicon and producing measurements that assess the sentiment intensity in the text from the identified words [24], [25]. Using this tool, the tweets and news reports for each week were classified in three groups, i.e. tweets and news reports containing positive, negative or neutral sentiment.

B. Dynamic Topic Modeling

Topic modeling is a method for unsupervised classification of documents which finds natural groups of items within a collection and is used for discovering hidden semantic structures in a body of text. Within our analysis, we used a BERTopic model that leverages BERT embeddings and c-TF-IDF to create dense clusters allowing for easily interpretable topics [26]. BERTopic uses document embedding representation with pre-trained Sentence-BERT framework which allows converting sentences and paragraphs into vector representations using the well-known BERT language model. These sentence embeddings are then clustered using Hierarchical DBSCAN, an extension of the DBSCAN algorithm that is able to find clusters of varying density, and Uniform manifold approximation and projection (UPAM) for vector dimensionality reduction. Lastly, topic representations are extracted from the clusters using class-based TF-IDF:

$$W_{t,c} = t f_{t,c} \cdot \log\left(1 + \frac{A}{t f_t}\right) \quad (1)$$

where term frequency models the frequency of the term t in a class c and the inverse document frequency is replaced by

an inverse class frequency to measure how much information a term provides to a class calculated as the logarithm of the average number of words per class A divided by the frequency of term t across all classes.

For the purpose of our analysis, fine-tuning on a pre-trained BERTopic model was done using minimum topic range of 50 and monograms, bigrams and trigrams which were derived from the translated tweets and news reports during the fine-tuning process. Additionally, because the BERTopic model is used to extract topics from large documents, a count vectorizer model was used to ensure relevant topic extraction [27]. The topic over time modeling was done using the fine-tuned BERTopic model with weekly timestamps and global and evolution tuning.

C. Bigrams similarity

In order to evaluate the similarity between the tweets and news reports, we used Soft Cosine Similarity as a method of assessing semantic similarity between two documents. This type of measure considers two vectors as pairs of features and considers their similarity in a vector space model [28]. Given two N-dimension vectors a and b , the soft cosine similarity is calculated as follows:

$$\text{soft_cosine}_1(a, b) = \frac{\sum_{i,j}^N s_{ij} a_i b_j}{\sqrt{\sum_{i,j}^N s_{ij} a_i a_j} \sqrt{\sum_{i,j}^N s_{ij} b_i b_j}} \quad (2)$$

where $s_{ij} = \text{similarity}(\text{feature}_i, \text{feature}_j)$.

To measure the soft cosine similarity between the most common terms found in tweets and news reports for each week, the top 10 most common bigrams for tweets and news reports were determined using a count vectorizer model [27]. All bigrams taken from tweets and news reports were respectively sorted in ascending order and a similarity matrix was calculated using GloVe vector representations [29]. Soft cosine similarity was calculated on the respective vector representations of the bigrams for each week.

IV. RESULTS

Table I and Figure 1 show statistics for the collected data over the 4 months of data collection period. Sentiment analysis of the data from both sources shows predominantly negative sentiment over positive and neutral content. Additionally, the percentage of texts classified as negative is much higher in the news reports compared to the percentage of negative sentiment in tweets. In Figure 1, a stable trend of news reports over time can be detected, however the trend of collected tweets shows a few bigger peaks. Namely, the first peak around December 6th, 2021 is connected to the ban of single-use plastic bags and additional rise of the cost of biodegradable plastic bags from 2 Macedonian Denars to 15 Macedonian Denars [30]. On the other hand, the second peak of tweets around the end of January 2022, happened after the news that over 300.000 expired COVID vaccines would be landfilled by the end of February due to lack of interest for COVID vaccination. Other smaller peaks include topics connected to generated plastic waste during the COVID-19 pandemic (between November 15th and November 29th), total generated municipal waste (between the end of December 2021 and the beginning of January 2022), as well as electronic waste and landfills regulations (around January 16th, 2022).

TABLE I. Number of Collected Tweets/News Reports

Source	Total	Unique	Positive(%)	Negative(%)	Neutral(%)
Twitter	21124	5287	22.55	68.10	9.35
News	1405	1137	15.23	81.35	3.42

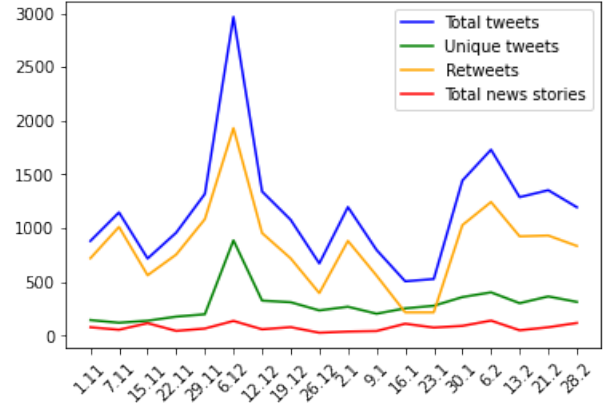


Fig. 1. Number of Tweets and News Reports over time

The sentiment analysis of tweets and news reports for each week follows similar trends as the overall collected data, as shown in Figure 2 and Figure 3. The number of tweets with negative sentiment is significantly bigger than the number of tweets with positive and neutral sentiment and the news report trends adhere to these trend patterns as well. The dynamic topic modeling analysis also shows similarities within the tweets and the news reports over the 4 month period, as displayed in Figure 4. Figure 4a shows the trends of the 10 most common topics identified in tweets. Some of these topics include: people inadequately throwing waste, plastic bags, nuclear waste as a result of nuclear power plants, medical waste and its management by the municipalities, burning wood and other materials for the purpose of heating, disposal of electronic waste such as lithium batteries, composting bio waste and vehicle recycling. Over the period of 4 months, all of the identified topics have similar frequency rates with the exception of the topic about plastic bags, that is connected to the North Macedonia's ban for single-use plastic bags [30]. Figure 4b shows the trends of the most common topics identified in news reports. There are fewer identified topic within the news reports data compared to the twitter data. The topics include: waste landfills and waste management in Skopje, the "plastic waste pandemic", medical waste and its management by the municipalities, plastics bags for packaging and space debris. The trends for the previously mentioned topics show bigger peaks during the beginning of December and toward the end of January, corresponding to the two aforementioned news events about the plastic bags ban and vaccines disposal.

Lastly, Figure 5 shows the semantic similarity between the most common bigrams found in tweets and news reports over time. The figure shows trends for bigram similarity for all collected data during the week, as well as similarity between bigrams of tweets and news reports classified as positive or negative accordingly. The obtained results show that there is a strong semantic similarity between the most common topics over time, since the threshold for strong soft cosine similarity is 0.5 [28]. The biggest values for cosine similarity for the data while not taking into consideration the sentiment is 0.86, 0.81 and 0.79, accordingly corresponding

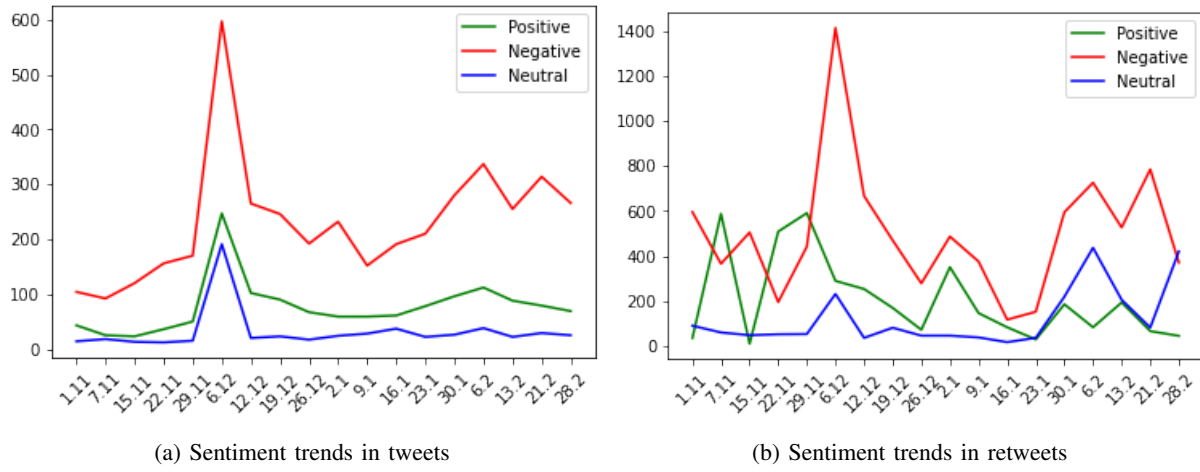


Fig. 2. Sentiment analysis

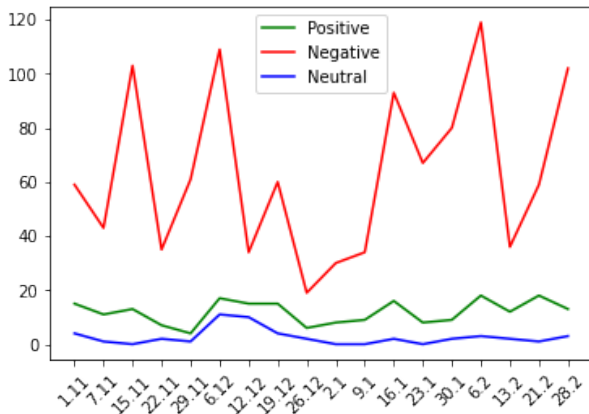


Fig. 3. Sentiment Analysis - News Reports

to the periods of end of February 2022, beginning of February 2022 and beginning of December 2021. The last two events are in correspondence with the previously mentioned news about plastic bags ban and vaccine disposal, whereas the first event with highest similarity value is connected to the war conflict between Russia and Ukraine, specifically the Russian occupation of Chernobyl and possible risks of major nuclear disaster and radioactive contamination across Europe [32]. Similar scores can be observed for cosine similarity of topics with negative sentiments, where the corresponding cosine similarity scores for the three major news events are 0.84, 0.83 and 0.78.

V. DISCUSSION AND CONCLUSION

The results obtained from all three analytic methods suggest strong correlation between the news reports and waste-related tweets which means that journalism has a major impact on the overall public response connected to waste management policies. Three major events were identified to have incited larger public response during the time frame between November 1st, 2021 and February 28th, 2022. The first event was the ban of single-use plastic bags in North Macedonia that began to be enforced as of December 1st, 2021 [30]. The main goal of these legal changes is to promote a new habit of using reusable bags made of eco-friendly materials. Despite the good intentions of the legal change, the results from the sentiment analysis show negative sentiment across both data sources. The main reason for this result is due to people's scepticism towards the governing bodies

for waste management. Namely, the vast majority of Twitter users during this period commented on the ineffectiveness of the ban due to the common practice of disposing mixed household waste in single-use plastic bags. Additionally, topics such as low recycling and waste selection rates as well as active burning of toxic waste were discussed during this event.

Another event that posed as a major initiator of public response was the great-scale COVID vaccines disposal due to expired dates in North Macedonia. In this instance, the majority of the Twitter users showed concerns about the management of medical waste, not only for the event of vaccine disposal, but also for the increased amount of medical waste generated over the last 2 years as a result of the COVID-19 pandemic [31]. The Twitter response about this issue is centered around the distrust of the waste managing facilities in the ability to adequately manage medical waste, since poor management of this type of waste can greatly threaten human and environmental health. The last identified provoking event was the armed conflict happening in Ukraine. The Russian invasion of Ukraine began on February 24th, 2022 and shortly after, the towns of Chernobyl and Prypyat were under Russian control. The major concern about nuclear disaster happened a few days after the initial invasion following the reports that the radiation level in Chernobyl drastically increased due to failure to maintain the nuclear reactor facility [32]. The Twitter response in the Western Balkans follow these concerns, mainly because of the geographical proximity of the affected region.

Over the last few years, the governments in the Western Balkan nations have made significant efforts in plotting a nation-wide plan to tackle the growing waste management issue. Last year, the North Macedonian Ministry of Environment and Physical Planning released a long-term strategy on climate action, which includes reports about waste generation trends as well as future plans for waste management [11]. Among the plans for building more efficient waste management systems, the North Macedonian government also has the intention of further incorporating the topics of climate change and sustainable development in the overall national education system, in addition to enforcing measures to raise climate awareness among the general public. The ban of single-use plastic is just one of these measures whose main intend is to limit the amount of municipal plastic waste and to raise the public awareness about waste management. Reducing

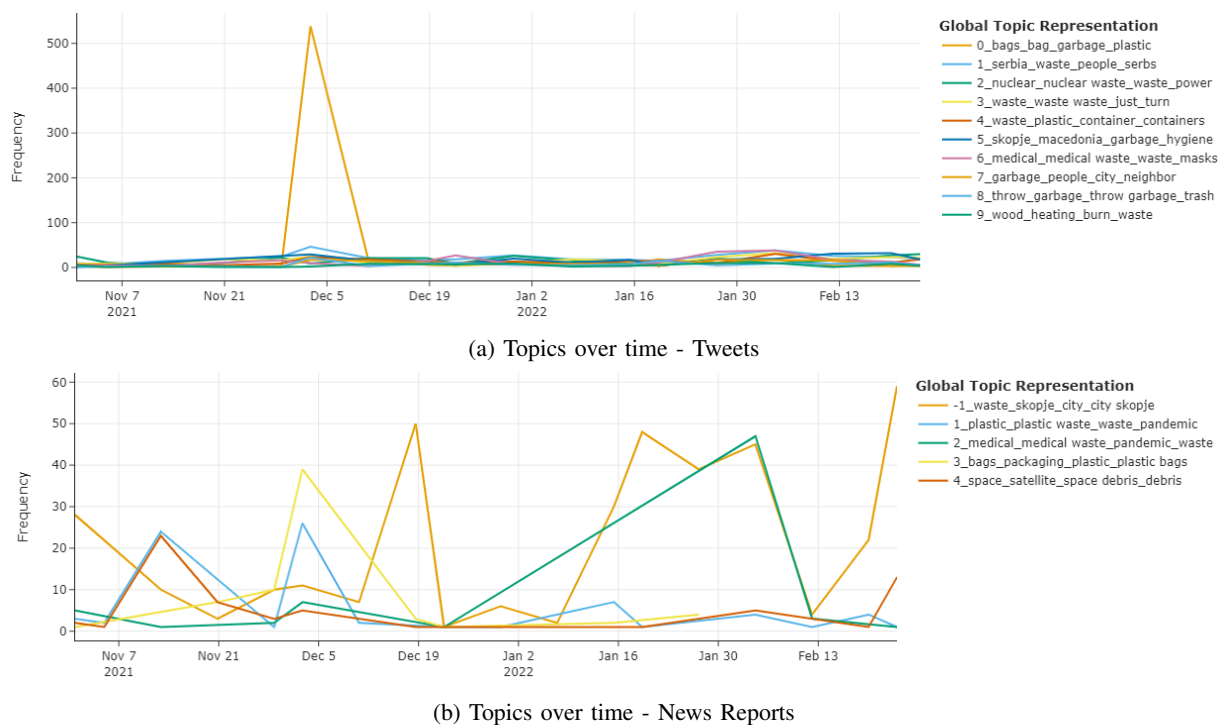


Fig. 4. Sentiment analysis - Tweets

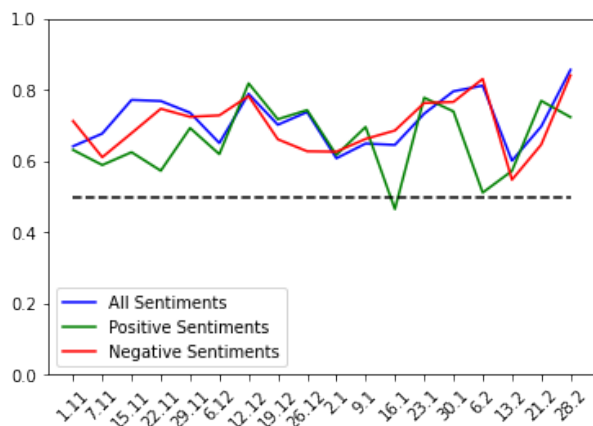


Fig. 5. Soft Cosine Similarity between Tweets and News Reports

the amount of waste through waste selection and recycling are one of the main goals of the laws regulating waste treatment, therefore it would be convenient to promote the positive aspect of these measures initially by the news media outlets. Since there is a strong positive correlation between the news reports and the social media response concerning waste management, the positive news reports may play a major role in promoting pro-environmental attitude on social media, and consequently increase pro-environmental behavior in the general public. Additionally, analysis of social media attitudes regarding waste management could be used for assessing and improving the existing waste regulations as well as developing new approaches to promote pro-environmental behaviour.

All things considered, this paper presented methods for data collection and analysis of social media posts and news reports in order to evaluate public response related to topics about waste management. The results show that news reports greatly influence the public response on Twitter. Moreover, events associated with political and health-related topics are

the main reason for the increased responses and discussions on social media in the Western Balkan region. Some possibilities for future work include analyzing public response of waste management schemes and events for a specific area of waste management, conducting a public survey to assess the level of awareness about waste management in Western Balkans, as well as collecting and analyzing data about public attitudes for waste-related topics over a longer period of time.

ACKNOWLEDGEMENT

This research was conducted with support of the Macedonian Academy of Sciences and Arts as part of an internship.

REFERENCES

- [1] D. Hoornweg and P. Bhada-Tata, "What a Waste : A Global Review of Solid Waste Management," Worldbank.org, 2012, doi: <http://documents.worldbank.org/curated/en/2012/03/16537275/waste-global-review-solid-waste-management>.
- [2] M. Kah, L. Levy and C. Brown (2012) Potential for Effects of Land Contamination on Human Health. 2. The Case of Waste Disposal Sites, Journal of Toxicology and Environmental Health, Part B, 15:7, 441-467, DOI: 10.1080/10937404.2012.736855.
- [3] D. Porta, S. Milani, A. I. Lazzarino, C. A. Perucci, and F. Forastiere. "Systematic review of epidemiological studies on health effects associated with management of solid waste." Environmental health 8, no. 1 (2009): 1-14.
- [4] M. Fujii, T. Fujita, X. Chen, S. Ohnishi, and N. Yamaguchi. "Smart recycling of organic solid wastes in an environmentally sustainable society." Resources, Conservation and Recycling 63 (2012): 1-8.
- [5] G. E. Louis, "A historical context of municipal solid waste management in the United States." Waste management & research 22, no. 4 (2004): 306-322.
- [6] "Waste statistics - Statistics Explained," Europa.eu, 2022. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Waste_statistics#Total_waste_generation [Accessed: 15-Mar-2022].
- [7] Vonschoenberg.info. Waste Management in Montenegro, 2022. [online] Available at: https://vonschoenberg.info/avscfiles/uploads/2021/06/CMS_Montenegro_Fact_Sheet_final.pdf [Accessed: 15-Mar-2022].
- [8] Vonschoenberg.info. Waste management in the Republic of Serbia, 2022. [online] Available at: https://vonschoenberg.info/avscfiles/uploads/2021/06/CMS_Serbia_Fact_Sheet_final.pdf [Accessed: 15-Mar-2022].
- [9] Vonschoenberg.info. Waste management in North Macedonia, 2022. [online] Available at: https://vonschoenberg.info/avscfiles/uploads/2021/06/CMS_North_Macedonia_Fact_Sheet_final.pdf [Accessed: 15-Mar-2022].

- [10] Р. Македонија, “Четврт национален план за климатски промени извештај за националниот инвентар на стакленички гасови” [Online]. Available: <https://api.klimatskipromeni.mk/data/rest/file/download/3260224e5a2745c8597aa095f5f7f30f56c5e11b9b2085c5b68258a87c6f93d.pdf> [Accessed: 15-Mar-2022].
- [11] “Long-term strategy on climate action and action plan” [Online]. Available: https://unfccc.int/sites/default/files/resource/MKD_LTS_Nov2021.pdf. [Accessed: 15-Mar-2022].
- [12] A. Licastro, and S. S. Bruno. "Drivers and barriers to a green economy. A review of selected balkan countries." *Cleaner Engineering and Technology* 4 (2021): 100228.
- [13] Z. Wu, Y. Zhang, Q. Chen, and H. Wang. "Attitude of Chinese public towards municipal solid waste sorting policy: A text mining study." *Science of the Total Environment* 756 (2021): 142674.
- [14] L.M. Heidbreder, M. Lange, and G. Reese. "# PlasticFree-July-Analyzing a Worldwide Campaign to Reduce Single-use Plastic Consumption with Twitter." *Environmental Communication* 15, no. 7 (2021): 937-953.
- [15] J. Roesslein, Tweepy: Twitter for Python! URL: <https://github.com/tweepy/tweepy>. (2020).
- [16] Online Doc Translator, 2021. [Online]. Available: <https://www.onlinedoctranslator.com/en/>. [Accessed: 1-Nov-2021].
- [17] K. Andresen, A. Hoxha, and J. Godole. "New roles for media in the Western Balkans: A study of transitional journalism." *Journalism Studies* 18, no. 5 (2017): 614-628.
- [18] L. Richardson. Beautiful soup documentation. April (2007).
- [19] “Вестѝ,” Time.mk, 2021. [Online]. Available: <https://time.mk/st/vesti>. [Accessed: 1-Nov-2021].
- [20] “Vesti,” Time.rs, 2021. [Online]. Available: <https://time.rs/>. [Accessed: 1-Nov-2021].
- [21] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welp. "Predicting elections with twitter: What 140 characters reveal about political sentiment." In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 4, no. 1, pp. 178-185. 2010.
- [22] X. Zhou, X. Tao, J. Yong, and Z. Yang. "Sentiment analysis on tweets for social events." In *Proceedings of the 2013 IEEE 17th international conference on computer supported cooperative work in design (CSCWD)*, pp. 557-562. IEEE, 2013.
- [23] S. Gohil, S. Vuik, and A. Darzi. "Sentiment analysis of health care tweets: review of the methods used." *JMIR public health and surveillance* 4, no. 2 (2018): e5789.
- [24] S. Bird, E. Klein, E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. "; Reilly Media, Inc." (2009).
- [25] C. J. Hutto, & E. E. Gilbert. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text , (2014). Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.
- [26] M. Grootendorst. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics (Zenodo v0.9.4). Zenodo v0.9.4. (2020) doi:10.5281/zenodo.4381785
- [27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel et al. "Scikit-learn: Machine learning in Python." *the Journal of machine Learning research* 12 (2011): 2825-2830.
- [28] G. Sidorov, A. Gelbukh, H. Gómez-Adorno, and D. Pinto. "Soft similarity and soft cosine measure: Similarity of features in vector space model." *Computación y Sistemas* 18, no. 3 (2014): 491-504.
- [29] J. Pennington, R. Socher, and C. D. Manning. "Glove: Global vectors for word representation." In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-1543. 2014.
- [30] N. Georgievski, “Single-use plastic bags banned in North Macedonia, biodegradable ones sell at €0.25,” Meta.mk, Dec-2021. [Online]. Available: <https://meta.mk/en/single-use-plastic-bags-banned-in-north-macedonia-biodegradable-ones-sell-at-e0-25/> [Accessed: 23-Mar-2022].
- [31] World, “Tonnes of COVID-19 health care waste expose urgent need to improve waste management systems,” Who.int, Feb-2022. [Online]. Available: <https://www.who.int/news/item/01-02-2022-tonnes-of-covid-19-health-care-waste-expose-urgent-need-to-improve-waste-management-systems#:~:text=Tens%20of%20thousands%20of%20tonnes,to%20a%20new%20WHO%20report> [Accessed: 23-Mar-2022].
- [32] P. Polityuk, F. Crellin, “Ukraine reports higher Chernobyl radiation after Russians capture plant,” Reuters, 25-Feb-2022. [Online]. Available: <https://www.reuters.com/world/europe/ukraine-nuclear-agency-reports-higher-chernobyl-radiation-levels-due-heavy-2022-02-25/>. [Accessed: 23-Mar-2022]

Differences in cognitive ability assessment results between Millennial and Generation Z cohorts

Ivica Pesovski

*Computer Science and Innovation
Brainster Next College
Skopje, RN Macedonia
ivica@next.edu.mk*

Andrea Kulakov

*Faculty of Computer Science & Engineering
Ss. Cyril and Methodius University
Skopje, RN Macedonia
andrea.kulakov@finki.ukim.mk*

Vladimir Trajkovic

*Faculty of Computer Science & Engineering
Ss. Cyril and Methodius University
Skopje, RN Macedonia
trvlado@finki.ukim.mk*

Abstract—The Flynn effect is a well-known phenomenon that has been documented by researchers in a variety of developed countries around the world. This phenomenon refers to the fact that people's cognitive test scores are increasing by about 3 points every decade. Many researchers have documented this spanning several decades since it was first defined by James Flynn in the 1980s. According to recent studies, this upward trend in cognitive assessment scores has reached a halt in certain nations and may possibly be reversing, a phenomenon known as the negative Flynn effect. The goal of this study is to see if the Flynn effect can be detected across generations in a developing country like RN Macedonia. The data for this study came from a competitive test for obtaining a scholarship for work-related courses. The test consisted of 40 questions that evaluated the participants in a variety of areas. The test was completed by 2700 participants. The key research groups include two generations: millennial and generation-z. Multiple cognitive areas were evaluated, including arithmetic reasoning, logical thinking, visual reasoning, attention to detail, abstract thinking, spatial and verbal reasoning, and more. The findings prove that cognitive test scores of people in RN Macedonia are improving, with a notable difference in average scores between the two generation groups in every field studied and a 1.79 points difference in collective intelligence results.

Index Terms—Flynn effect, intelligence, IQ, cognitive tests, millennial, generation-z

I. INTRODUCTION

If one is about to search any digital resource about IQ tests, how they work and whether they are a precise representation of a person's intelligence, there is a big chance one will run into and get familiar with a phenomenon known as the Flynn effect. It is named after James Flynn, a psychologist who first documented the fact that every decade the average IQ score increases by 3 points [1]. These claims are not clearly evident and were easily overseen by other researchers because of the way IQ scores are calculated. If we compare the results from IQ tests from one generation with the results of same tests from another generation we will not see evident difference because the tests are regularly being calibrated to keep the mean at 100. Flynn compared the results on the same scale and found an obvious 3 IQ points difference which can not be neglected. This means that person with IQ of 100 in 2020s, would have had IQ of around 118 back in the 1960s and could have been considered intelligently gifted, while vice-versa, person with IQ of 100 in 1960s would have had around 82 IQ points today, which would have put that person on the verge

of mental retardation. The occurrence of this effect sparked many discussion whether IQ tests measure intelligence which is innate or one that is taught. It can also justify the lower cognitive scores of elderly people as not being a result of aging, but rather general differences among generations.

There is available research from the past 10 years that that the Flynn effect is already stalling and even reversing - a phenomena known as negative Flynn effect [2] [3] [4]. Such research is usually conducted in developed countries, like Sweden [5], and the main reasons usually pointed out for this are the lower fertility rates [6], the decrease in quality of living, like increased air pollution [7], and even the increased migration since immigrants usually score lower on cognitive test etc [8]. Some radical and controversial reasons for this negative phenomenon were also pointed out through the years, like genetic deterioration also known as dysgenics, which will be discussed more in the related work section.

The goal of this paper is to analyze the trends of cognitive test scores in a developing country like RN Macedonia. This research represents a modified version of the Flynn effect occurrence model, since the scores from the tests were not gathered throughout the years, but rather a test was assigned at the same time to people born in different generations. Main focus is to show whether there is clear difference in the scores from people belonging to the millennial generation (people born between 1977 and 1995) and the people belonging to the Z generation (born between 1996-2015). The data gathered and presented in this paper can serve as starting point for tracking trends in the upcoming years.

The data that is analyzed was collected by organizing a competition which awarded the top 10 participants with a full scholarship for a year-long course for learning digital skills like computer programming, digital marketing, data science, graphic design and similar. The lower-ranked participants were given a smaller scholarship which they could use to co-finance their re-qualification. The incentive is high, mostly due to the increased digital transformation in many business domains which sky-rocketed the demand for IT professionals in 2022. Because of this incentive and the valuable award system, a total of 2700 people successfully completed the test and entered the competition. Another 2366 students attempted to solve the test but didn't made it to the end or their time expired.

This paper is organized in 5 sections. Section I is the introductory section, where the research question is proposed and an introduction to the topic is given. Section II provides an overview of the available literature for the Flynn effect in general and the most common reasoning why it may be happening. Section III outlines the methodology used, how the data was gathered and how it was analyzed. Section IV sums up the results from the research offering both visual and numerical representation of the findings as well as the limitations of the study. Section V wraps up the research paper with a conclusion from the reported findings and presents the future work that becomes available as a result of the findings and the data collected from this research.

II. RELATED WORK

A lot time and money has been invested into demystifying intelligence [9] and although many fields of study have seen ground-breaking advancements, intelligence still remains extremely vague. This can be witnessed in computer science as well - computers have mimicked a lot of our human skills, but the field of artificial intelligence is struggling with its progress [10].

The term Flynn effect by which this study was inspired was coined after James Flynn, who first documented this effect back in the 1980s. Flynn found out a rise in mean IQ of 13.8 points between Americans from 1932 to 1978 [11]. Other researchers easily omitted this evident trend because IQ scores are calculated relatively to the average score of the group, so they are regularly being calibrated, and the results of the cohorts usually oscillate around the mean of 100. However, when results from different cohorts are compared on the same scale, the spread of values is evident.

Francis Heylighen had an interesting viewpoint on the implications from this effect [12]. He uses the Flynn effect as justification for the lower IQ test scores of older generations compared to the younger generations. Before, it was believed that intelligence diminishes with age, but this finding now proves that older people were simply raised in a period when the general intelligence was lower. This is supported by a study on US adults from 2019 which finds that cognitive function is improving over time for US citizens aged over 60 years as a result of the greater educational attainment [13]. This is contrary to the findings of a more recent study from 2021 which analyzed the cognitive abilities of people older than 50 years in Germany, Sweden and Spain [5]. In this study Julia et al. tested different age cohorts using Wechsler tests to conclude that cognitive abilities diminish as people age. They did the research between 2004 and 2013 and they also confirmed that later generations have higher cognitive abilities, although these abilities still start diminishing after 50 years of age. These higher cognitive abilities represent the Flynn effect. They confirmed this finding happening in Germany and Spain, while in Sweden the gap between the generations was not notably big, which led them to believe that maybe there is an upper limit up to which these gains in IQ scores can go. It is unclear whether improvements in countries like Sweden are not happening because a ceiling effect has occurred or further improvements in the educational system and society in general can lead to further improvements.

There are various interpretations of the fact that IQ scores are increasing year by year. James Flynn himself argues that

IQ test don't actually measure inborn intelligence [14], but rather one that can be and has been taught. So, as education becomes more available and of higher quality, people get better at standardized test. Others, like Richard Lynn argue that this is the act of dysgenics, meaning that the quality of species deteriorates with time because people with poor genes, usually from lower socioeconomic classes are outbreeding people with good genes because of their higher fertility rate, something which we deem as a non-sense elitism on a verge of racism.

Mitchum and Fox in their work from 2013 [15], analyzed the scores that people in the 1960s got on a Raven's Progressive Matrices test and the scores that people get today. They confirmed the Flynn effect occurring and also concluded that the average scores from 50 years ago would certainly be below average now. However, they reason that these phenomena isn't bound to changes in intelligence, but rather in the environment. People today are faced with math algebra at a very early age and the tech-driven world which is inevitable part of modern life is a great beneficial factor toward one's abstract thinking skills. In their research they point out that even things we take for granted today, like computer's file system, with folders and files is a very advanced abstraction compared with the things people 50 years ago had access to. They see this rise in abstraction in everyday life too - the humour in comedy shows today is far more abstract than the one decades ago.

Flynn himself approves this theory as he concludes that in the 1900s only 3% of people in the USA worked jobs that were cognitively demanding (like doctors or teachers) [16], while in 2006 there, a huge 35% of the working population in the US works cognitively demanding professions [17].

There are evidences that intelligence is also influenced by the environment. Among others, Carmi points out that intelligence is not only biological but the culture, social structure, and the environment have effects on it [18]. This is backed up by another research from Nisbett who has shown a 12-point to 18-point increase in IQ when children are adopted from lower socioeconomic families to middle-class homes [19]. Similar findings were publicized in 1970 by Scarr who found out that poor performance of black children in IQ tests tremendously increased once they were adopted by white families [20], which back then was also a matter of socioeconomic effects.

When it comes to defining generations, a work by Michael Dimock from 2019 has been one of the most cited ones [21]. He defines Generation Z as people born starting 1997 and later, and Millennial generation as people born between 1981 and 1996. Our research primarily relies on a classification proposed by Jason Dorsey [22], one of the most popular and influential researchers and speaker on the topic of generations and future-proofing, especially on Millennial and Z generations [23]. His cut-off year between these two generations is 1995-1996.

We believe that for developing countries like RN Macedonia, the cutoff should happen later, because generations are defined by technological, economic, political and social shifts which usually come into effect later in developing countries. For this reason, besides the classification proposed by Jason, we will do a comparison of the results when year 1999-2000 is taken as year separating Millennial and Generation Z people. Because the time-span of a generation has historically

been and is analytically meaningful at around 16 to 20 years, we will increase the start year to 1980, so that Millennial are people born between 1980 to 1999 while Generation Z falls between 2000 and 2017.

III. METHODOLOGY

A competitive test consisting of 40 questions was organized to collect the data that is part of this research. The test was open for exactly one month and rewarded the best participants with full and partial scholarships for attending workplace related courses.

TABLE I
TEST PARTICIPANTS BY GENERATION

Generation	Birth year	# tests	# completed tests
baby-boomers	1946 - 1964	9	7
gen-x	1965 - 1976	203	79
millennials	1977 - 1995	2927	1490
gen-z	1996 - 2015	1927	1124
millennials*	1980 - 1999	3775	1997
gen-z*	2000 - 2015	892	529

* Generations if 1999/2000 is taken as a cutoff year

The test was supposed to be solved mostly by millennials, people born between 1977 and 1995, since they were most likely to be interested in professional re-qualification. Surprisingly, together with the millennials there were a lot gen-z people (people born between 1996 and 2015) among the users that took the test. The exact numbers of participants by generation are shown in Table I. Because there are not enough representatives from the other two present generations, baby-boomers and gen-x, these will be left-out from this research and the focus will be on millennial and gen-z participants.

The condition to gain a full scholarship for a year long course was to score best on a test that consisted of 40 questions which evaluated different skills among the contestants. The categories in which the participants contested, total number of questions from each category for all participants combined and total number of correct answers for each category is given in Table II.

Worth noting is that the contestants were not aware that different cognitive skills are measured and this did not even play a role in the final scholarship standings. The rules were simple, each participant gets 40 random questions with 25 minutes to solve. Each correct answer brings 1 point, each wrong question is discarded. A maximum of 40 points can be acquired. If more participants have same test scores, then

TABLE II
ALL CATEGORIES, TOTAL NUMBER OF POLLED QUESTION AND NUMBER OF CORRECT ANSWERS IN EACH CATEGORY

Category	Total questions		Correct answers	
	Millennial	Gen-z	Millennial	Gen-z
Basic Numeracy	2628	1819	1239	887
Numerical reasoning	4107	2810	1810	1257
Logical reasoning	3855	2793	1946	1474
Inductive reasoning	3710	2686	1613	1242
Spatial reasoning	3496	2962	1902	1632
Short logic puzzles	3856	2861	2161	1638
Verbal reasoning	6043	4742	2997	2368
Attention to details	5563	4313	3217	2539
Abstract thinking	3992	3114	2271	1815
General knowledge	22350	16860	10961	8274

they are ordered by the time left for completing the test, so the faster are the better.

The distribution of participants in the quiz by birth year is shown in Fig. 1. We can see that the most applications came from people just around the discussed cutoff years (late 1990s), so changing the year where Millennial ends and Generation Z begins by even a single year, will have a big impact on the results.

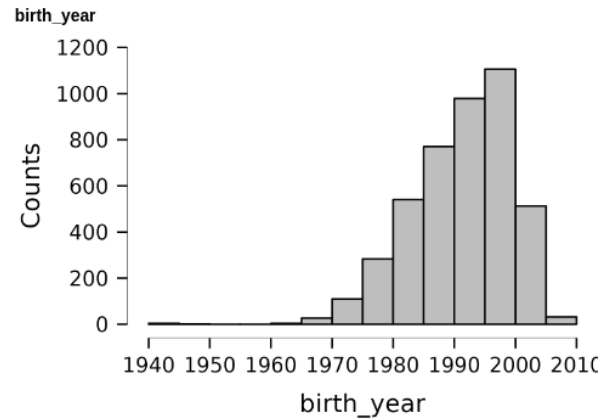


Fig. 1. Age distribution of quiz participants

Because the test consists of questions divided up into categories, the first analysis is a comparison of the scores different generations got in different categories. The results are calculated as percentage of correct answers out of the total possible answers. If a participant gave 20 correct answers, then the score will be 50 because 40 is the total number of questions.

Before analyzing the trends in cognitive scores between the 2 generations, the results are calibrated in order to show whether there are any precursors indicating the Flynn effect. The mean average of all 2700 participants is calculated, and everyone's score is calculated as percentage of this mean. For example, if the mean of all correct answers is 20, then 20 correct answers mean a score of 100, 30 correct answers equal score of 150, 10 correct answers equal score of 50 and so on. The average of these scores is then calculated for the 2 generations in order to prove or discard existence of Flynn effect.

IV. RESULTS AND DISCUSSION

The first result of this paper shows that the participants representing generation-Z scored higher in all categories, except in general knowledge where the difference is unnoticeable. This was to be expected, as the Flynn effect affects only the abstract thinking skills and general knowledge has already proven to be deteriorating [14]. This result can be extracted from the data available in Table II. The differences in each category between the two generations are given in Table III and graphical representation of this finding is given in Fig. 2.

The more interesting analysis is the analysis of combined cognitive skills. After dividing the participants in two groups and calculating these groups' average means, the result shows that the average number of correct answers given by millennial people is 20.21, while the average number of correct answers given by gen-z participants is 20.57. Indeed, this is not a huge difference, until we see how this reflects when presented in a way IQ scores are calculated.

TABLE III
SCORES FOR EACH GENERATION BY CATEGORY
(MILLENNIALS: 1977-1995, GEN-Z: 1996-2015)

Category	Millennial score	Gen-z score	Difference*
Basic Numeracy	47.1461	48.7631	1.62
Numerical reasoning	44.0711	44.7331	0.66
Logical reasoning	50.4799	52.7748	2.29
Inductive reasoning	43.4771	46.2398	2.76
Spatial reasoning	54.4050	55.0979	0.69
Short logic puzzles	56.0425	57.2527	1.21
Verbal reasoning	49.5946	49.9367	0.34
Attention to details	57.8285	58.8685	1.04
Abstract thinking	56.8888	58.2852	1.40
General knowledge	49.0425	49.0747	0.03

* Difference = Generation-z score - Millennial score

For that purpose we first calculate the mean of correct answers given by the whole population (both millennial and generation-z participants). This number is 20.37. Then each participant's score is calculated relatively to this number. So if someone gave 20 correct answers his score will be $(20/20.37) \times 100\%$, or score of 98.18. The formula for calculating participants score is shown below "(1)", where $\overline{score}$ is the mean average of all participants scores.

$$f(score_i) = \frac{score_i}{\overline{score}} \times 100 \quad (1)$$

This calculation replicates the way scores from IQ tests are measured and helps to evaluate each user relative to the average result of the whole group. This is why the effect of increasing IQ points was not noticed for many years, until Flynn calculated each person's score relative to the complete population scores. After the test scores are calculated for every participant, an average of these scores is calculated for the 2 focus groups, shown in Table IV.

TABLE IV
AVERAGE SCORES BY GENERATION

Generation	Average score	Generation	Average score
millennials	99.24	millennials*	100.03
gen-z	101.03	gen-z*	99.87

* Generations if 1999/2000 is taken as a cutoff year

From Table IV we can see that the average score of the millennial groups is 99.24 and the average score of the gen-z group is 101.03, the difference being 1.79 points. This result confirms the findings from other available research which confirm the presence of the Flynn effect, or at least some increase in cognitive ability test score for newer generations.

If we compare the results of the two generations after shifting the year separating the two generations of interest to year 2000, as proposed in Section II due to the latency in technological and social progress in developing countries, the results are presented in Fig. 3 and in Table V accordingly. The differences are even more expressed in most categories. The Millennial generation still scores better when evaluating general knowledge but also beats the Z generation when it comes to extroverted skills like verbal reasoning. We can see that differences in inductive reasoning and abstract thinking are by far most evident when considering this classification, which comes at no surprise.

In Table IV, we can see that the if proposed cutoff year is considered 2000, then Millennial generation performs better

TABLE V
SCORES FOR EACH GENERATION BY CATEGORY
(MILLENNIALS: 1980-1999, GEN-Z: 2000-2017)

Category	Millennial score	Gen-z score	Difference*
Basic Numeracy	48.0434	46.6334	-1.41
Numerical reasoning	44.5750	44.6241	0.05
Logical reasoning	51.2556	51.8634	0.6
Inductive reasoning	43.8376	47.3726	3.53
Spatial reasoning	54.3469	55.8625	1.51
Short logic puzzles	56.6712	57.1747	0.5
Verbal reasoning	50.1158	48.1230	-2
Attention to details	58.4374	59.0691	0.63
Abstract thinking	56.9372	59.6179	2.68
General knowledge	49.5610	47.7505	-1.81

* Difference = Generation-z score - Millennial score

TABLE VI
AVERAGE SCORES BY YEAR

Birth year	Participants	Score	Std. Deviation
1995	130	99.42	26.26
1996	141	99.34	26.89
1997	156	104.81	25.82
1998	164	99.46	28.82
1999	134	98.88	27.74
2000	147	97.66	28.12

at collective level than Generation Z. For that reason, guided by the density of participants born in late 1990s seen in Figure 1 we analyzed the scores year by year, for the years where the cutoff can be made and found that 1997 has the biggest impact on the final results as shown in Table VI.

A. Limitations

A known limitation of this research is that the data was gathered at once from different generations. This means that the members of the two generations are at different age at the time of testing which means that one generation had acquired longer life knowledge as preparation for the test. As such, this research is not exactly representing the differences in cognitive ability test scores between the generations because the test wasn't conducted in each generation's prime time.

Another limitation is that the participants in the study do not exactly replicate the general population of RN Macedonia in demographic characteristics and should be carefully selected to do so for the next iterations, so that clear conclusions concerning the whole countries can be made.

The findings are still confirming the findings of Francis Heylighen discussed in Section II. As he used the Flynn effect to conclude that intelligence is not diminishing as people age, but rather newer generations get a head-start for cognitive tests as the general intelligence levels are rising with time. This is evident in the results of this research. Although millennial and generation-z people are not that far apart in age, the constant difference in test scores, confirms that the more recent, Generation Z, already experiences gains in general intelligence.

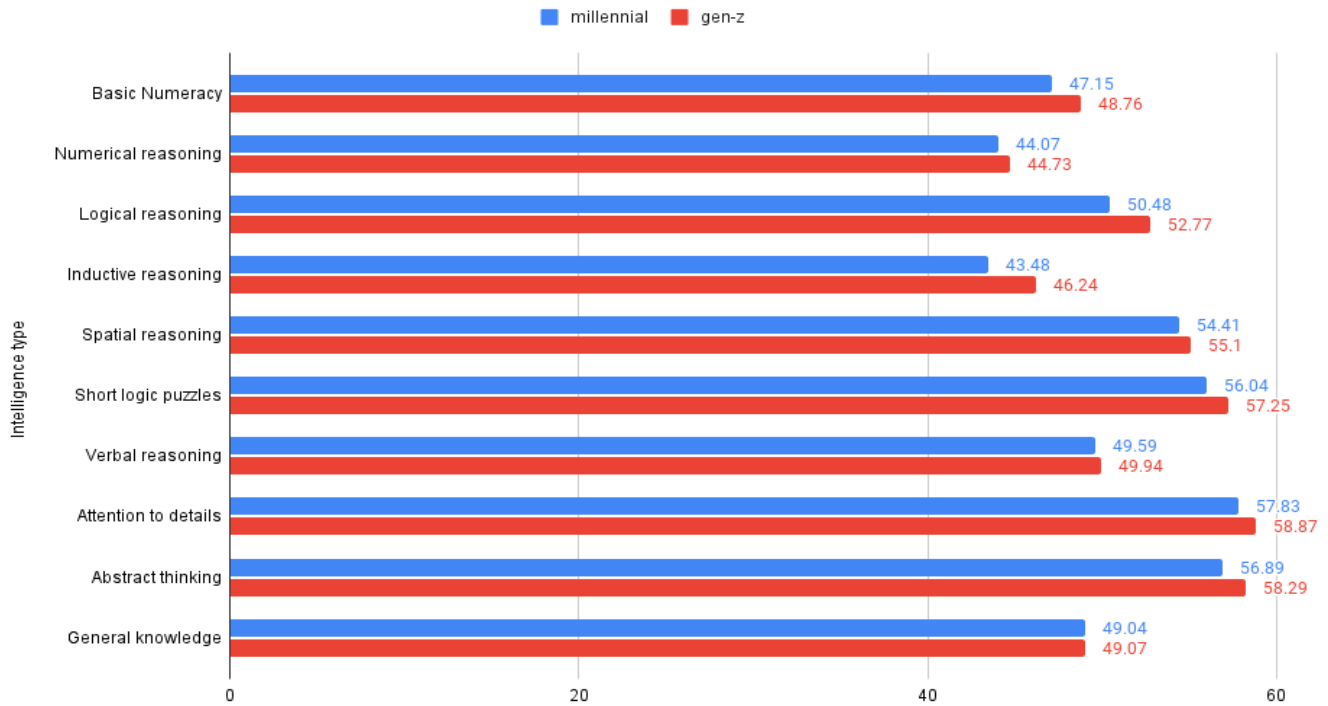


Fig. 2. Scores by category compared between millennial and generation-z participants (millennials: 1977-1995, gen-z: 1996-2015)

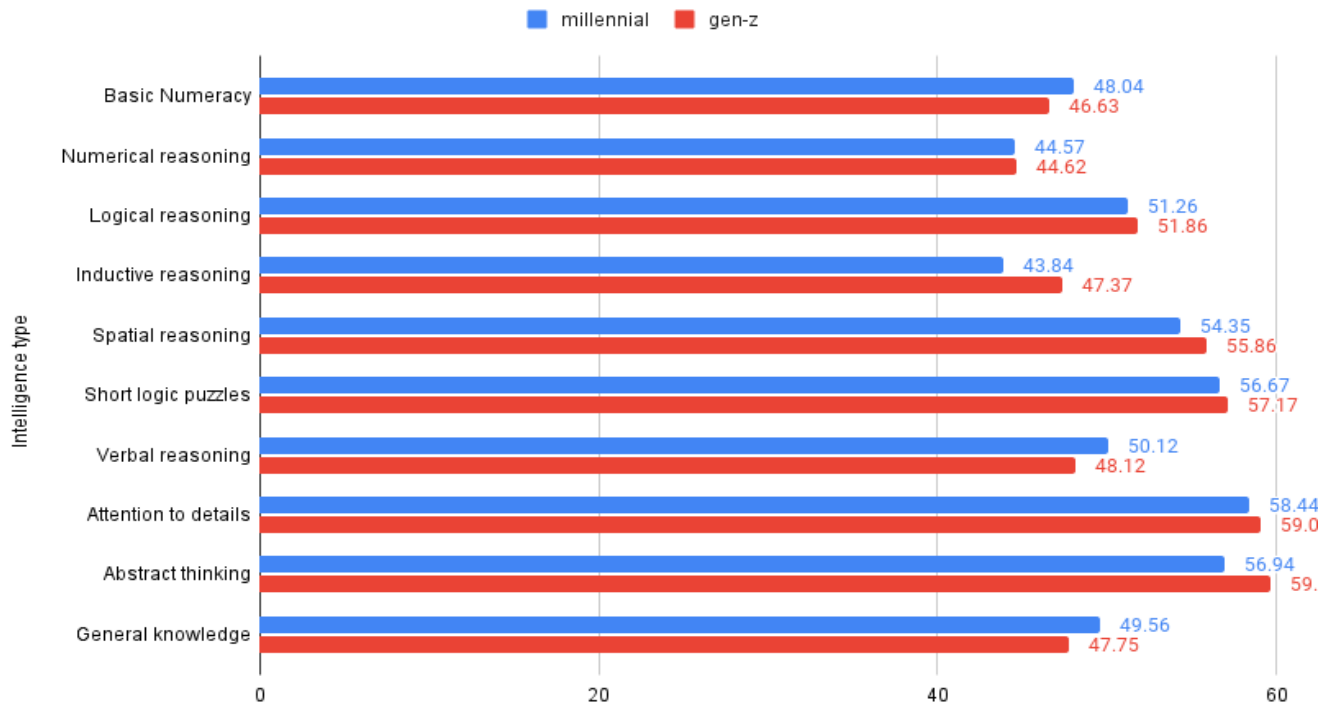


Fig. 3. Scores by category compared between millennial and generation-z participants (millennials: 1980-1999, gen-z: 2000-2017)

V. CONCLUSION AND FUTURE WORK

The findings in this research give a clearer perspective as to where RN Macedonia stands regarding having positive or negative Flynn effect. The fact that both generations solved the test in the same time can only diminish the effect of the findings, since the people from the senior generation had more time to develop their cognitive skills. Even with such

hypothesis, the participants from the newer generation (gen-z) outsourced their predecessors by almost 2 IQ points.

Comparing the scores on each category separately confirms these results - in every category except general knowledge (which is expected), newer generation scored clearly better results than its preceding generation. This result did not change even when shifting the year that defines when Millennial

generation ends and Generation Z begins in the range between 1995 and 2000.

These findings certainly contribute toward concluding that the Flynn effect is present among the people in developing countries like RN Macedonia, and these results will undoubtedly serve as starting point for more exact documentations of the Flynn effect phenomenon in the future.

The results from this study provide a preliminary insight in how generations in RN Macedonia are developing their cognitive skills. The collected data in this research presents a firm starting ground for future research in this field. The same data gathering process will be performed in the upcoming years with the focus on same age groups and in couple of years a true evaluation of the Flynn effect will be available.

It would be of great importance to compare the results of the upcoming, generation alpha, now that the results for generation-z are available.

REFERENCES

- [1] J. R. Flynn, "Massive iq gains in 14 nations: What iq tests really measure," *Psychological bulletin*, vol. 101, no. 2, p. 171, 1987.
- [2] T. W. Teasdale and D. R. Owen, "A long-term rise and recent decline in intelligence test performance: The flynn effect in reverse," *Personality and Individual Differences*, vol. 39, no. 4, pp. 837–843, 2005.
- [3] E. Dutton, D. van der Linden, and R. Lynn, "The negative flynn effect: A systematic literature review," *Intelligence*, vol. 59, pp. 163–169, 2016.
- [4] R. L. Williams, "Overview of the flynn effect," *Intelligence*, vol. 41, no. 6, pp. 753–764, 2013.
- [5] J. Grasshoff, J. Beller, B. G. Kuhlmann, and S. Geyer, "Increasingly capable at the ripe old age? cognitive abilities from 2004 to 2013 in germany, spain, and sweden," *Plos one*, vol. 16, no. 7, p. e0254038, 2021.
- [6] H. Rindermann, D. Becker, and T. R. Coyle, "Survey of expert opinion on intelligence: The flynn effect and the future of intelligence," *Personality and Individual Differences*, vol. 106, pp. 242–247, 2017.
- [7] B. Bratsberg and O. Rogeberg, "Flynn effect and its reversal are both environmentally caused," *Proceedings of the National Academy of Sciences*, vol. 115, no. 26, pp. 6674–6678, 2018.
- [8] T. W. Teasdale and D. R. Owen, "Secular declines in cognitive test scores: A reversal of the flynn effect," *Intelligence*, vol. 36, no. 2, pp. 121–126, 2008.
- [9] K. Richardson, *Understanding intelligence*. Cambridge University Press, 2022.
- [10] D. Partridge, *What Makes You Clever: The Puzzle of Intelligence*. World Scientific, 2014.
- [11] J. R. Flynn, "The mean iq of americans: Massive gains 1932 to 1978," *Psychological bulletin*, vol. 95, no. 1, p. 29, 1984.
- [12] F. Heylighen, "Increasing intelligence: the flynn effect," *Principia Cybernetica Web. Principia Cybernetica*, Brussels, 2000.
- [13] M. Bancks, A. Alonso, N. Allen, K. Yaffe, and M. Carnethon, "Temporal trends in cognitive function of older us adults associated with population changes in demographic and cardiovascular profiles," *J Epidemiol Community Health*, vol. 73, no. 7, pp. 612–618, 2019.
- [14] J. R. Flynn and J. R. Flynn, *Are we getting smarter?: Rising IQ in the twenty-first century*. Cambridge University Press, 2012.
- [15] M. C. Fox and A. L. Mitchum, "A knowledge-based theory of rising scores on "culture-free" tests," *Journal of Experimental Psychology: General*, vol. 142, no. 3, p. 979, 2013.
- [16] J. W. CNN, "What your iq score doesn't tell you," 2014. [Online]. Available: <https://edition.cnn.com/2014/02/19/health/iq-score-meaning/index.html>
- [17] R. W. Johnson, G. B. Mermin, and M. Resseger, "Job demands and work ability at older ages," *Journal of aging & social policy*, vol. 23, no. 2, pp. 101–118, 2011.
- [18] C. Schooler, "Environmental complexity and the flynn effect," 1998.
- [19] R. E. Nisbett, J. Aronson, C. Blair, W. Dickens, J. Flynn, D. F. Halpern, and E. Turkheimer, "Intelligence: new findings and theoretical developments," *American psychologist*, vol. 67, no. 2, p. 130, 2012.
- [20] S. Scarr and R. A. Weinberg, "Iq test performance of black children adopted by white families," *American Psychologist*, vol. 31, no. 10, p. 726, 1976.
- [21] M. Dimock, "Defining generations: Where millennials end and generation z begins," *Pew Research Center*, vol. 17, no. 1, pp. 1–7, 2019.
- [22] J. Dorsey, "Generations birth years - gen z, millennials, gen x, and baby boomers," 2022. [Online]. Available: <https://jasondorsey.com/about-generations/generations-birth-years/>
- [23] J. R. Dorsey and D. Villa, *Zconomy: how Gen Z will change the future of business—and what to do about it*. HarperCollins, 2020.

Investigating Public Awareness of Air Pollution in Western Balkans by analyzing Tweets and News Article Teasers

Angela Madjar

Macedonian Academy of Sciences and Arts
Skopje, Republic of Macedonia
angela.madzhar@students.finki.ukim.mk

Jana Prodanova

Macedonian Academy of Sciences and Arts
Skopje, Republic of Macedonia
jprodanova@manu.edu.mk

Ivana Gjorshoska

Macedonian Academy of Sciences and Arts
Skopje, Republic of Macedonia
ivana.gjorshoska@students.finki.ukim.mk

Aleksandra Dedinec

Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
aleksandra.kanevche@finki.ukim.mk

Abstract—Air pollution is a serious threat to the health of people living in Western Balkans, with the biomass being one of the main pollutants. This study is based on the assumption that air pollution escalation in Western Balkan countries during winter will provoke a more intense activity on Twitter, while acknowledging that people's opinions and feelings can often be influenced by content presented in news articles. The objective of this study is to investigate public awareness of air pollution in Macedonia, Serbia, Bosnia and Herzegovina and Montenegro. Natural Language Processing techniques such as Sentiment Analysis and Topic Modeling, as well as Statistical Analysis are employed to determine whether Twitter discussions regarding air pollution reflect the PM₁₀ levels measured by official air monitoring stations in these countries. Such analyses are also performed on news article teasers, attempting to determine whether mass media portrays a realistic ambient condition and aims to promote pro-environmental behavior. The results of this study suggest that using Correlation Analysis to check for resemblance between the frequency of sentiments detected in social media discussions and the temporal changes in the PM₁₀ concentration in the air, can serve as a measure of public awareness of air pollution. Analyzing the content of the tweets can reveal issues in the public opinion and thus, contribute to tackling them down.

Keywords—air pollution, pm10, western balkans, social media, tweets analysis, news media, sentiment analysis, topic modelling, cross correlation

I. INTRODUCTION

Air pollution is a major problem in all Western Balkan countries. People living in this region are estimated to lose up to 1.3 years of life due to the prominent levels of pollutants in the air [1]. According to the Air Quality Report (2018), the crude mortality rate attributable to air pollution in Macedonia was 154.7 per 100 000 inhabitants. In comparable neighboring countries such as Serbia, it was 200.7 per 100 000; in Montenegro was 110.9 per 100 000 and in Bosnia and Herzegovina it was 105.1 per 100 000 [2]. PM₁₀ is among the air pollutants whose levels are most frequently above the legislation limits in this region and are mainly emitted by human activities such as industry, household heating and transport [3][4]. Uncontrolled urbanization, illegal construction, poor planning and design also endanger the environment by contributing to such ambient condition [5]. Additionally, transboundary pollution from within and outside the region considerably bring about the observed

concentrations [3]. However, the Health and Environment Alliance (HEAL) has reported that the main sources of air pollution in the Western Balkans are the production of electricity from coal and using wood for heating households [6]. Consequently, the daily PM₁₀ limit of 40µg/m³ set out under national legislation is exceeded between 120 and 180 days a year, mostly during wintertime [1]. Exposure to such particles can affect the heart and lungs and cause serious health effects [7].

For this study, we supposed that the air pollution escalation in Western Balkans during winter will provoke a more intense activity on social media, especially on Twitter, where people share their opinions and feelings. However, the way people feel and think about a certain topic can often be influenced by sentiments expressed in news articles. Knowing that exposure to PM₁₀ particles harms the human health, our initial hypotheses were that the public has justified negative sentiments towards air pollution in their country and that news media reflects the real air pollution ambient, aiming to promote pro-environmental behavior.

To test our hypotheses and to investigate public perception of air pollution in several Western Balkan countries including Macedonia, Serbia, Bosnia and Herzegovina and Montenegro, we employ Natural Language Processing techniques such as Sentiment Analysis and Topic Modeling, as well as Statistical Analyses.

II. RELATED WORK

Analyses of micro-blogging posts referring air pollution has been an appealing research field for many researchers over the recent years. Such analyses were primarily done using Weibo (Chinese Twitter) content by separating the positive and negative sentiments with a manual qualitative classification method and using their frequency to improve correlation with the daily Air Quality Index. The study demonstrates that filtered social media messages can be used to monitor dynamics of air pollution to some extent. Such messages can reveal insights about public perceptions and concerns of air pollution [8]. Another study demonstrates the feasibility of using social media to monitor PM_{2.5} levels as an alternative method in areas without air pollution monitoring systems, showing that citizen-led monitoring can be used to better understand public's interaction with air quality issues and use Twitter discussions to promote pro-environmental behavior. The authors minimized media influence by filtering

tweets containing URLs and compared the frequency of different sentiment groups against the official PM_{2.5} data in Greater London [9]. Moreover, public response to a particular event, such as wildfire, can be captured by using a structural topic model to extract topics from tweets posted during the event, as demonstrated in [10]. As indicated in [11], experimenting with a wide range of Supervised and Unsupervised learning methods can provide information about the evolution of topics over time and determine similarities and differences in public response to air quality information.

III. DATA

A. Twitter Data

For the purpose of this study, we coded a searching application using the Python library Tweepy [12] to collect air quality-related tweets from November 1st, 2021 to February 28th, 2022. We used Twitter's Standard API [13], which allows searching for tweets posted within the last 7 days prior to the search time. Thus, tweets were collected on weekly basis using the keywords: "aerozagaduvanje" (air pollution), "аерозагадување" (air pollution), "zagaduvanje" (pollution), "загадување" (pollution), "pm10" (pm10), "дишеме" (we breathe) to capture tweets written in Macedonian language. To collect Western Balkan tweets (excluding tweets written in Albanian language), we used the terms: "zagadjenje" (pollution), "загађење ваздуха" (air pollution) and "zagadjenje vazduha" (air pollution).

Although "pollution" is a broad concept, we empirically concluded that twitter users most frequently refer to the terms "pollution" and "air pollution" interchangeably. The very few tweets discussing different kinds of pollution were excluded manually.

B. News Media data

Parallel to the collection of tweets, the powerful web-crawler tool Octoparse [14] was employed to weekly assemble teaser texts of news articles containing the above-mentioned keywords. A teaser is an illustrative short reading suggestion for an article that entices potential readers to read particular news items [15]. We crawled the news web sites Time.mk [16] and Time.rs [17], which are cluster-based news aggregators that analyze 15 000 news daily, collected from 120 distinct sources.

C. Official Air pollution data

Air pollution data was acquired in order to investigate the frequency of tweets and news articles during the peaks and falls of PM₁₀ particles measured by official measuring stations. The PM₁₀ data in Macedonia [18], Serbia [19], Montenegro [20] and Bosnia and Herzegovina [21] was collected from 21, 37, 6 and 16 monitoring sites respectively, owned and funded by local authorities. The hourly data measured from November 1st, 2021 to February 28th, 2022 was aggregated by week, so that the air pollution data is adjusted to the frequency of collection of Tweets and teasers. For each country, data from all of its monitoring stations was aggregated to encompass the entire country area. PM₁₀ data measured in Serbia was available from December 2nd, 2021 to February 28th, 2022.

IV. METHODOLOGY

A. Sentiment Analysis

To analyze the data in this study, we used VADER (Valence Aware Dictionary and sEntiment Reasoner) - a lexicon and rule-based sentiment analysis instrument optimized to find semantics in micro blog texts, such as tweets [22]. VADER relies on an English dictionary that maps lexical features to emotion intensities called sentiment or Valence scores. Valence scores of each word are measured on a scale from - 4 (most negative) to + 4 (most positive), with 0 indicating a neutral sentiment. The compound score of the whole text is obtained by summing the valence scores of each word in the lexicon, normalized to be between -1 (most extreme negative) and +1 (most extreme positive) by using the following normalization:

$$x = \frac{x}{\sqrt{x^2 + \alpha}} \quad (1)$$

In (1), x is the sum of Valence scores of constituent words and α is a normalization constant with a default value equal to 15. For severalizing the tweets into positive, negative, and neutral sentiment groups, the default threshold value of - 0.05 and + 0.05 was used.

Before using the built-in NLTK VADER Sentiment Analyzer, we first used an automatic document translator [23] to translate the collected data into English language.

B. Time Series Statistical Analysis

In terms of time series analysis, measuring similarity is important to assess the casual relationship between two signals in time. To compare the tweets and teasers against the PM₁₀ data, we used cross-correlation. The Cross Correlation Function (CCF) is the correlation between the observations of two time series x_t and y_t , separated by k time units (the correlation between y_{t+k} and x_t), where k is called a lag. The confidence interval is calculated with $\pm \frac{2}{\sqrt{n-|k|}}$, where n is the number of observations and k is the lag. The correlation is significant if its absolute value is greater than $\frac{2}{\sqrt{n-|k|}}$ [24]. The CCF is based on the assumption that the data is stationary, meaning that the mean and variance are constant and independent of time. If a time series has an upward or downward trend, it is commonly made stationary by differencing [25].

To test whether the obtained sentiment groups of tweets and teasers and the PM₁₀ data are stationary, we used the nonparametric Mann-Kendall test. The null hypothesis is that no monotonic trend is present in the data. The test indicated a decreasing trend in the teasers obtained from Time.rs. Prior to performing the CCF test, we conducted first order differencing to make this data stationary. For the rest of the data, the Mann-Kendall test did not indicate any trends and thus, the null hypothesis was accepted.

C. Topic Modelling

To understand what contributes to the air pollution and who is accountable according to the public opinion, we experimented with two topic modelling approaches: Latent Dirichlet Allocation (LDA) [26] and Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM) [27].

LDA assumes that a single document (tweet in our study) covers a small set of concise topics and calculates the contribution of each topic to the document. Topics are

identified based on the likelihood of co-occurrences of words contained in them [28]. Ranked lists of words associated with a given word w_n are obtained by calculating the sum of the weight of each topic generated by LDA multiplied by the weight of each word w_n contained in that topic. The ranking weight of the word i is computed as follows:

$$w_i = \sum_{j=1..N} w_{ij} * w_{nj} \quad (2)$$

In (2), N is the number of topics and w_{ij} denotes the weight of the word i in topic j .

Contrary to the LDA's assumption, GSDMM assumes that there is only one topic per document, making it suitable for detecting topics in smaller documents, such as tweets. Gibbs sampling describes the method of iterating through and reassigning clusters based on a conditional distribution. In the same manner the Naïve Bayes Classifier works, documents (tweets) are assigned to clusters based on the highest conditional probability.

To evaluate the performance of these approaches, we employed the Topic Coherence score. The coherence of a topic, used as a proxy for topic quality, is based on the distributional hypothesis that states that words with similar meaning tend to co-occur within a similar context [29]. Topics are considered to be coherent when all or most of the words in the topic are related. We consider the differences between each model as we learn an increasing number of topics, starting from 2. Prior to this analysis, we preprocessed the negative tweets to remove retweet symbols, special characters, URLs, emojis and extra spaces. To give more focus to the important information, we also removed stopwords, tokenized the tweets and reduced the words to their lemma.

V. RESULTS

A. Obtained sentiments

The statistics about the tweets and the news article teasers collected during the course of this 17-week study, as well as the sentiments obtained with the Sentiment Analysis for all of the datasets, are displayed in Table I. It is noticeable that the percentage of negative sentiments prevails in every dataset, while the percentage of neutral sentiments is the lowest. The relatively high number of retweets indicates a sense of agreement and approval among the users [30]. An important point is that the total number of tweets does not equal the sum of retweets and unique tweets (including replies). Sometimes, the original tweet that is being retweeted is not captured since it dates long before the time scope of this study. Moreover, it can happen for several original tweets to be identical. Thus, tweets with distinct preprocessed content (without retweet symbols and special characters) are considered as unique. Similarly, media teasers with distinct content count as unique.

TABLE I. STATISTICS OF THE COLLECTED AIR POLLUTION TWEETS AND NEWS ARTICLE TEASERS

Name	Total number	Retweets (%)	Unique (%)	Negative (%)	Positive (%)	Neutral (%)
Macedonian Tweets	1018	33.99	53.93	64.3	19.4	16.2
Time.mk teasers	994		48.89	55.2	39.8	4.9
Western Balkan Tweets	2664	47.56	44.52	54.3	29.09	16.61
Time.rs teasers	709		73.77	64.2	27.5	8.3

B. Cross-correlation

The severalized tweets and teasers obtained through the Sentiment Analysis were plotted weakly against the official PM₁₀ data to determine the resemblance between the sentiment groups and the actual air pollution data. The different sentiment groups of the Macedonian tweets were compared against the PM₁₀ data obtained from measuring stations in Macedonia. Sentiment groups of the rest of the tweets were compared against PM₁₀ data in Serbia, Montenegro and Bosnia and Herzegovina to check for correspondence.

The maximum cross-correlation between all categories of Macedonian air pollution tweets and PM₁₀ particles in the country was at lag 0, indicating that there is no lag or lead-time between levels of PM₁₀ particles and tweet frequency. A cross-correlation analysis between the number of negative tweets and PM₁₀ particles showed a coefficient of 0.62 ($p = 0.001$) and a coefficient of 0.54 ($p < 0.001$) between the number of neutral tweets and PM₁₀ particles (Fig. 1). The cross-correlation between the positive tweets and PM₁₀ particles was insignificant. Regarding the sentiment groups of Time.mk teasers, the maximum cross-correlation between the frequency of negative teasers and PM₁₀ data was 0.53 ($p = 0.002$), between the positive teasers and PM₁₀ data was 0.52 ($p < 0.001$) (Fig. 2) both at lag 0; while between the neutral teasers and PM₁₀ data it was insignificant.

There was no significant cross-correlation between the Balkan air pollution tweets and any of the PM₁₀ data measured in Serbia, Bosnia and Herzegovina and Montenegro. As for the Time.rs teasers, there was significant cross-correlation at lag 0 between the negative teasers and PM₁₀ data measured in Serbia, with a coefficient of 0.66 ($p < 0.001$) (Fig. 3) and at lag 3 between the neutral teasers and PM₁₀ data measured in Montenegro with a coefficient of 0.65 ($p < 0.001$) (Fig. 4).

C. Obtained topics

Topic Modelling is a challenging NLP task and choosing the “best model” is not an exact science. Context, and translation in our study, make for a wide range of possible meanings that can be best interpreted by human eye.

Despite visual inspection of the topic models, we also depended on the Topic Coherence to assess the quality of each model. We experimented with topic numbers ranging from 2 to 20, and with different hyperparameters. With a small number of topics, we only extracted broad topics, achieving low Topic Coherence score. On the other hand, with a large number of topics it was difficult to distinguish between them and the Topic Coherence score was low again, indicating that the chosen number of topics is probably wrong.

According to the analysis, for GSDMM on the Macedonian negative tweets, 9 topics were selected with a coherence score of 0.51, while for LDA 15 topics (coherence score = 0.41).

In addition, for GSDMM on the Western Balkan negative tweets, 7 topics were chosen (coherence score = 0.57), while for LDA 15 topics (coherence score = 0.38).

In other studies, it has been shown that traditional topic models such as LDA, experience large performance degradation over short texts [29] [31], which is in accordance with our results. Therefore, we focus on the topics obtained with GSDMM and present some of the most important ones in Table II, along with the most important words and their frequency of occurrence.

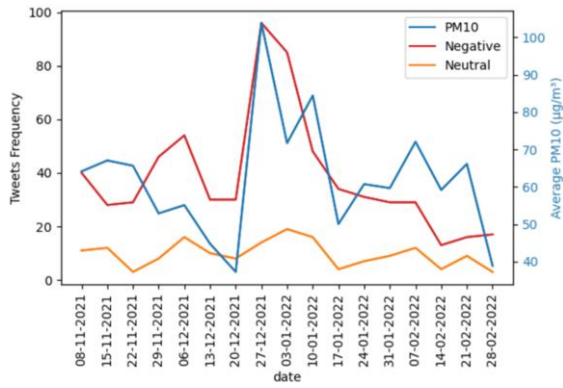


Fig. 1. Weekly comparison of official PM₁₀ data in Macedonia and frequency of Macedonian Tweets

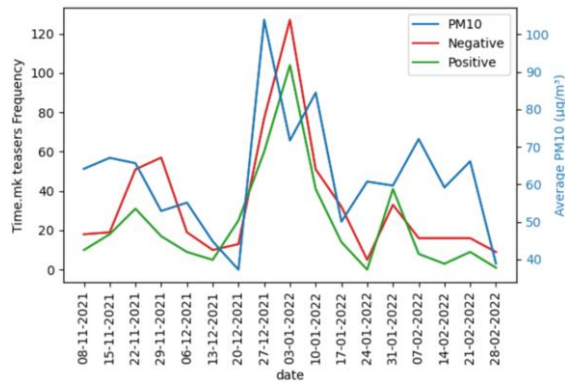


Fig. 2. Weekly comparison of official PM₁₀ data in Macedonia and frequency of Time.mk teasers

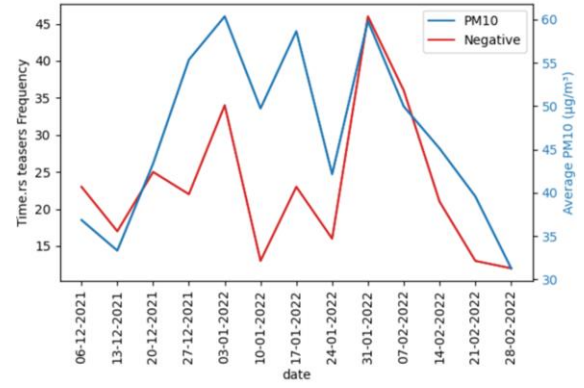


Fig. 3. Weekly comparison of official PM₁₀ data in Serbia and frequency of Time.rs teasers

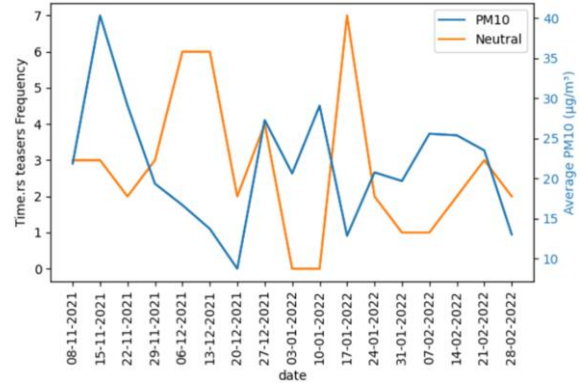


Fig. 4. Weekly comparison of official PM₁₀ data in Montenegro and frequency of Time.rs teasers

VI. DISCUSSION AND CONCLUSION

The analyses showed that negative Macedonian tweets were most predictive of PM₁₀ levels in the country, whereas positive tweets do not have comparable peaks and falls with the measured PM₁₀ particles. This suggests that during times of escalated air pollution, the public expresses negative sentiments and concerns on Twitter, which is in accordance with our first hypothesis that Twitter discussions about air pollution reflect measurements of PM₁₀ particles.

As for the News Media in Macedonia, the maximum cross-correlation was detected between negative teasers and PM₁₀ data, which again, is consistent with our second hypothesis that News Media provides truthful information on air pollution. However, the high cross correlation of positive teasers and PM₁₀ data should not be neglected, since it implies the possibility that news media tries to suppress public discussions about air pollution hazards.

The topic modelling indicated that the Macedonian public thinks that electricity production and transport lead to great air pollution, and thus, to health issues. They expect the government to take action and address the problem. However, studies conducted in Macedonia show that the ratio of air pollution caused by households and transport in the country is about 95% to 5% [32]. The experimenting with different numbers of topics revealed that some discussions about the negative effects of biomass are present on Twitter, suggesting awareness of this pollutant to some extent. Even so, the size of such clusters was significantly smaller compared to the size of clusters regarding transportation and a lower

coherence score was achieved. This leads to a conclusion that the use of biomass in households is not among the topics obtained with a highest coherence score, while the ambient condition is often attributed to the transport. These results suggest high public unawareness of the main air pollutants in the country, emphasizing the need to raise public consciousness of the dangers posed by using biomass and to promote pro-environmental behavior on this topic.

Contrary to the first hypothesis about Twitter discussions reflecting PM₁₀ concentration in the air, there was no resemblance between any of the sentiment groups of Western Balkan Tweets and the actual PM₁₀ data collected from Serbia, Bosnia and Herzegovina and Montenegro. However, most of the tweets were classified as negative, which suggests that people express negative sentiments on Twitter even when the PM₁₀ levels are moderate or low.

Although the number of articles regarding air pollution decreases over time, a high cross-correlation was detected between negative Time.rs teasers and Serbian PM₁₀ data, which again, confirms our second hypothesis about mass media portraying a realistic ambient condition.

Throughout the topic modeling, we discovered that many air pollution tweets refer to Serbia and cities in Serbia (Belgrade, Sabac, etc.). When tweeting on air pollution, the public generally expresses worries about health, climate change, industrial air pollution and pollution caused by power plants. Again, this indicates public unawareness of the biomass being a major air pollutant in the Western Balkans [6].

TABLE II. TOPICS OBTAINED THROUGH GSDMM TOPIC MODELLING

Topic (GSDMM)	Negative Macedonian Tweets	Topic (GSDMM)	Negative Balkan Tweets
Electricity	pollution (52), crisis (16), increase (16), price (14), air (11), electricity (11), cut (10), energy (9), terrible (9)	Power plant	pollution (407), air (188), cities (62), Belgrade (53), Serbia (48), terrible (45), plant (34), power (31), problem (24)
Transport	pollution (38), air (12), less (11), problem (9), car (7), brt (7), cities (6), need (6), tram (4)	Climate change	pollution (293), air (97), Serbia (83), problem (45), climate (30), change (30), increase (21), death (18), environment (17)
Government	pollution (12), people (11), anything (9), pm10 (7), immediately (6), mayor (6), don't (6), vmro (4), municipality (4)	Industrial air pollution	pollution (49), Sabac (25), miner (19), poison (17), plant (8), industry (7), factories (7), burn (7), suffer (7)
Health	die (8), pollution (6), breathe (6), people (6), lose (6), children (6), poison (4), burn (4), poor (4)	Health	pollution (7), Lazarevac (6), children (5), problem (3), nausea (3), dizziness (3), caught (3), heart (3), symptom (3)

Future research directions for this study include a more supervised translation of the collected data to improve VADER resistance to slang and sarcasm. Additionally, using geo-localized tweets in the future would eliminate the possibility of capturing Croatian tweets when using keywords for collecting Western Balkan tweets. Finally, separate analysis of the data obtained from each measuring station could provide insights on air pollution in smaller regions in every country, since levels of PM₁₀ vary largely over space and time.

Collectively, this study confirms the convenience of NLP techniques such as Sentiment Analysis and Topic Modelling for analyzing public thoughts and feelings on important topics such as air pollution, as well as the truthfulness of the information transmitted by news media. The cross-correlation between sentiments detected in social media discussions and real air pollution measurements in a country can serve as a measure of public awareness of air pollution. Topic Modelling techniques can reveal issues in the public opinion and thus, contribute to tackling down such problems.

ACKNOWLEDGMENT

The research presented in this article was conducted during an internship in the Macedonian Academy of Sciences and Arts.

REFERENCES

- [1] Colovic Daul, M., M. Kryzanowski, and O. Kujundzic. "Air Pollution and Human Health: The Case of the Western Balkans." UN Environ (2019).
- [2] Helotonio, Carvalho. "Air pollution-related deaths in Europe – time for action." Journal of Global Health 9.2 (2019).
- [3] Banja, M., G. Đukanović, and C. A. Belis. "Status of air pollutants and greenhouse gases in the Western Balkans." Publications Office of the European Union, EUR 30113 (2020): 1-53.
- [4] Meisner, Craig, Dragan Gjorgjev, and Fimka Tozija. "Estimating health impacts and economic costs of air pollution in the Republic of Macedonia." South Eastern European Journal of Public Health (SEEJPH) (2015).
- [5] Jovanovic, Mica. "Environmental Impact of Illegal Construction, Poor Planning and Design in Western Balkans: A Review."
- [6] I. Todorović, "HEAL: Biomass is one of main sources of air pollution in Western Balkans", Balkan Green Energy Group, 27-Jan-2022. Available at: <https://balkangreenenergynews.com/heal-biomass-is-one-of-main-sources-of-air-pollution-in-western-balkans/#:~:text=The%20European%20Environmental%20Agency%20estimated,challenge%20around%20improving%20woodburning%20technology.> [Accessed: 13-Mar-2022]
- [7] Environmental Health, "Particulate matter (PM10 and PM2.5)", Environmental Health, 25-Nov-2020. Available at: <https://www.health.nsw.gov.au/environment/air/Pages/particulate-matter.aspx>. [Accessed: 13-Mar-2022]
- [8] Jiang, Wei, et al. "Using social media to detect outdoor air pollution and monitor air quality index (AQI): a geo-targeted spatiotemporal analysis framework with Sina Weibo (Chinese Twitter)." PLoS one 10.10 (2015): e0141185.
- [9] Hsuen, Yulin, et al. "Feasibility of using social media to monitor outdoor air pollution in London, England." Preventive Medicine 121 (2019): 86-93.
- [10] Sachdeva, Sonya, and Sarah McCaffrey. "Using social media to predict air pollution during California wildfires." Proceedings of the 9th International Conference on Social Media and Society. 2018.
- [11] Gurajala, Supraja, Suresh Dhaniyala, and Jeanna N. Matthews. "Understanding public response to air quality using tweet analysis." Social Media+ Society 5.3 (2019): 2056305119867656.
- [12] J. Roesslein, "Tweepy: Twitter for Python!", 2020. Available at: <https://github.com/tweepy/tweepy>. [Accessed: 10-Oct-2021]
- [13] Twitter, "Standard v1.1", Available at: <https://developer.twitter.com/en/docs/twitter-api/v1>. [Accessed: 10-Oct-2021]
- [14] Almaqbali, Iqtibas Salim Hilal, et al. "Web Scrapping: Data Extraction from Websites." Journal of Student Research (2019).
- [15] Karn, Sanjeev Kumar, et al. "News Article Teaser Tweets and How to Generate Them." arXiv preprint arXiv:1807.11535 (2018).
- [16] I. Trajkovski, "How does TIME.mk work?", Time.mk, 2008. Available at: <https://time.mk/info/site>, [Accessed: 01-Nov-2021]
- [17] I. Trajkovski, Time.rs, 2008. Available at: <https://time.rs/>, [Accessed: 01-Nov-2021]
- [18] Ministry of environment and physical planning – Republic of North Macedonia, "Air Quality Portal". Available at: https://air.moepp.gov.mk/?page_id=175, [Accessed: 01-Mar-2022]
- [19] Republic of Serbia, Open Data Portal, "Air quality - unverified real-time clock data". Available at: <https://data.gov.rs/sr/datasets/kvalitet-vazduha/>, [Accessed: 01-Mar-2022]
- [20] Environmental Protection Agency of Montenegro, "Measurement data archive". Available at: <http://www.epa.org.me/vazduh/arhiv/7>, [Accessed: 01-Mar-2022]
- [21] Discomap EEA, "Download of air quality data". Available at: <https://discomap.eea.europa.eu/map/fme/AirQualityExport.htm>, [Accessed: 01-Mar-2022]
- [22] Hutto, Clayton, and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." Proceedings of the international AAAI conference on web and social media. Vol. 8. No. 1. 2014.

- [23] Online Doc Translator, 2021. Available at: <https://www.onlinedoctranslator.com/en/>, [Accessed: 01-Nov-2021].
- [24] Minitab, "Interpret all statistics and graphs for Cross Correlation", 2022. Available at: <https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/time-series/how-to/cross-correlation/interpret-the-results/all-statistics-and-graphs/>, [Accessed: 15-Mar-2022]
- [25] Dean, Roger T., and William Dunsmuir. "Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: The importance of constructing transfer function autoregressive models." *Behavior research methods* 48.2 (2016): 783-802.
- [26] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3 Jan (2003): 993-1022.
- [27] Yin, Jianhua, and Jianyong Wang. "A dirichlet multinomial mixture model-based approach for short text clustering." *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2014.
- [28] Korzycki, Michał, Izabela Gatkowska, and Wiesław Lubaszewski. "Can the human association norm evaluate machine-made association lists?." *Cognitive Approach to Natural Language Processing*. Elsevier, 2017. 21-40.
- [29] Syed, Shaheen, and Marco Spruit. "Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation." *2017 IEEE International conference on data science and advanced analytics (DSAA)*. IEEE, 2017.
- [30] Sharifi, Zahra, and Sajjad Shokouhyar. "Promoting consumer's attitude toward refurbished mobile phones: A social media analytics approach." *Resources, Conservation and Recycling* 167 (2021): 105398.
- [31] Qiang, Jipeng, et al. "Short text topic modeling techniques, applications, and performance: a survey." *IEEE Transactions on Knowledge and Data Engineering* (2020).
- [32] G. Kanevce, A. Dedinec, V. Taseska-Gjorgievska, A. Dedinec. "Transport in Skopje – realities and challenges, Path to green transport", Third Biennial Update Report on Climate Change, 2017. Available at: <https://api.klimatskipromeni.mk/data/rest/file/download/71dea57f28b54b8c5f35e41a364a586d58c97edef47aacf36268b5d4296667ec.pdf>, [Accessed: 28-Mar-2022]

Model Hyper Parameter Tuning using Ant Colony Optimization

Aleksandar Trajkovski

FCSE, Ss Cyril and Methodius University, Skopje
Elevate Global LLC

Skopje, North Macedonia

aleksandar.trajkovski.2@students.finki.ukim.mk

Gjorgji Madjarov

FCSE, Ss Cyril and Methodius University, Skopje
Elevate Global LLC

Skopje, North Macedonia

gjorgji.madjarov@finki.ukim.mk

Abstract—The process of adjusting the best hyper-parameters for machine learning models is a complex optimization problem over the models that in reality are "black-boxes". This process entails several challenges such as the time complexity of finding optimal results, different types of hyper-parameters (continuous and categorical hyper-parameters) and model over-fitting. The main benefits of the hyper-parameter adjustment process are the improvement of the efficiency and quality of the models. This paper presents an implementation for adjusting hyper-parameters of time series prediction regression models, which is based on pheromone paths of ant colonies used as a way of finding food in nature. This approach of leaving pheromone pathways has been used to quickly find values for the hyper-parameters of the regression models for which they give the most optimal results.

The proposed implementation for Ant Colony optimization is compared with three other parameter optimization approaches (Grid search, Bayesian optimization and Tree Parzen Estimators) on ten different regression datasets according to the time required to perform the optimization and the quality of the model prediction. The experimental evaluation shows that the proposed method needs minimal time to optimize the hyper-parameters, while preserving very good predictive performance in comparison to the competing approaches.

Index Terms—hyper-parameter optimization, hyper-parameters, Ant colony, Grid search, HPO, time series

A. Acronyms

MSE - Mean Squared Error;
ACO - Ant Colony Optimizer;
SVR - Support Vector Regressor;
HGB - Histogram Gradient Boosting;
GS - Grid search;
RS - Random Search;
TPE - Tree Parzen Estimators;
HPO - Hyper Parameter Optimization;
BS - Bayesian Search;

I. INTRODUCTION

Machine learning models use parameters that cannot be directly deduced from the data. These types of parameters are called hyper-parameters of a model [1]. The overall work and behavior of the model in the learning and prediction phases greatly depend on these parameters. Their optimization is a difficult and tedious task, which depends on the expertise of the engineer and the domain of the problem the engineers are trying to solve with the help of data modeling. The hyper-parameter tuning is a process of choosing optimal values of the corresponding hyper-parameters, for which the model will have the best results for a predefined metric used for assessing the quality of the forecast. The most well-known and widely used technique for hyper-parameter tuning

of models is the Grid Search technique, which gives the best results when searching the entire hyper-parameter space. However this technique has a high level of time complexity. To avoid the most common problems with selecting hyper-parameters values (manual adjustment by the engineer, use of "recommended" values, use of inadequate metrics to assess the quality of the forecast and the occurrence of model over fitting), automatic tools are made that significantly aid in the tuning process and minimize the prediction error [2]. The purpose of this paper is to describe, review and experimentally compare approaches for hyper-parameter tuning on regression models which use time series as input data. In this paper the tools AutoSklearn, Hyper-Opt will be considered as representatives of Bayesian optimization and Tree Parzen Estimators respectively. Moreover as third and fourth method of hyper-parameter tuning evaluation, the method Grid search and the proposed implementation using the principle of pheromone paths of ant colonies will be used. The proposed implementation for the Ant Colony Optimizer can be easily adapted to any type of problem and model. Furthermore the same implementation can be used for parallel execution and skipping of combinations through the pheromone path system, which significantly increases the speed of parameter tuning.

II. RELATED WORK

Hyper-parameter optimization (HPO) is the final step in model design and the initial step in model training [3]. There are two types of hyper-parameter optimization:

- **Manual:**
For models that are widely used, the adjustment of these parameters by the scientist is possible and depends on the experience of previously solved problems. With the increasing dimensionality of hyper-parameter space and its values, this approach becomes very difficult. To overcome this problem, a second type of optimization has been proposed.
- **Automatic:**
As the search space and parameter settings expand, the demand for computing power grows, as does the necessity for continuous trial-and-error access. As a result of that new types of optimizers emerged such as Grid search (GS), random search (RS), bayesian search (BS) and ant colony optimization (ACO). For this kind of HPO, the scientist is expected to set a starting configuration for the process of optimization. Afterwards, the best hyper-parameters are automatically

chosen after certain time. With this method, the procedure of manually specifying the hyper-parameters and analyzing the outcomes is no longer necessary [4].

A. Automatic hyper-parameter optimization approaches

There are two primary types of automatic hyper-parameter tuning approaches, depending on whether they can be adapted to iterations [5]:

- non-iterative and non-adaptive
- iterative and adaptable

Non-iterative and non-adaptive approaches are represented by the GS (Grid search) method.

Grid search is one of the most utilized hyper-parameter tuning approaches. This fundamental approach of parameter tuning begins by defining a hyper-parameter search space that must be explored in order to achieve a certain global optimum. It is the most commonly utilized approach due to its mathematical simplicity and the ability to adjust parameters in parallel. The most significant disadvantages of this strategy are its high time complexity and the so-called curse of dimensionality. The number of combinations, and thus the resources required to calculate them, grows exponentially as the number of parameters to be adjusted grows [6].

Furthermore, Bayesian search (BS) and Tree Parzen Estimators (TPE) are representatives for the iterative and adaptable type of approach.

Most of the time, the functions in the models that require parameters are "black-box" functions. These functions are difficult to evaluate and demand a substantial amount of computer resources to identify the global extreme. To optimize these types of functions, Bayesian optimization is most typically utilized. With each evaluation and new sample, it grows more and more accurate over time. The basic idea behind this strategy is to establish a balance between exploration and exploitation. By balancing these two characteristics, the problem of becoming trapped in a local optimum is avoided. When compared to GS, the main advantage is that it can be utilized for a wide range of functions, including convex, non-convex, and stochastic functions. Bayesian optimization is comprised of two parts: a surrogate model for modeling a function and a posterior probability-dependent activation function that seeks to choose the next sample of the distribution. The surrogate model is often a Gaussian process that assumes similar inputs will result in similar outcomes. Unlike GS method, this approach for hyper-parameter tuning explicitly determines the error function [4], [6]–[8]. Finally, the Tree Parzen Estimator (TPE) is utilized as an alternative to a suitable surrogate Bayesian model source. This approach works with a variety of search spaces, including values from a uniform distribution, category spaces, and normally distributed real values represented by a tree structure. The inability to represent interactions between hyper-parameters is a significant shortcoming of this method. The basic method of the Gaussian process outperforms TPE for hyper-parameters with no interaction.

B. Related work for Ant Colony Optimization

In the literature, the ant colony optimization algorithm is most commonly used to solve the traveling salesman problem. This technique has recently been introduced to hyper-parameter tuning for hyper-parameters in neural network layers and the construction of complete neural architectures,

as well as for improving predictions using fuzzy logic [9], [12]. The next paragraphs will go over some of the more noteworthy publications that utilize this concept. The Deep Swarm system, which is based on flock intelligence, is presented in [9]. To discover the optimum neural network architectures, the authors apply the Ant Colony Optimizer (ACO) approach. They have discovered a balance in the process of exploitation and exploration by using this concept. The authors are using ACO and cluster intelligence for the process of NAS (Neuron Architecture Search) due to some benefits such as decentralization, scalability and the ability to share already acquired knowledge.

Deep Swarm begins by generating a graph with a single starting vertex. After a certain number of ants have been generated, the ACO tuning process begins. Each of these ants aims to determine what the next layer in neural architecture will be based on certain selection rules. Following this selection, each ant makes additional adjustments to the internal hyper-parameters of each layer, thereby resolving the HPO problem (Hyper parameter optimization). To a certain depth, this procedure is repeated for each ant. Following the completion of this procedure, the pheromone matrix is locally updated. Only the best ant is capable of updating the global pheromone matrix. The authors emphasize as an important point in the evaluation process that by limiting the number of ants to one, the research and exploitation of the search space does not show its full potential, but by increasing the number of ants from four to eight, the time required for evaluation increases significantly, and model improvement is minimal (less than 0.13%). Finally, the authors report error rates of 0.46%, 1.79%, and 1.68% when comparing the new system, RS search, and GS search, respectively.

The authors of [10] present the OpenNAS system, which generates CNN (convolutional neural networks) for data sets that contain black and white or colored images. PSO (partial cluster optimization) and ACO are used to select the layers for neural networks. The idea for such a system emerged from the brute force algorithm's impracticality and the problem's complexity. For ACO tuning, this before mentioned system OpenNAS utilizes an implementation from the previous Deep Swarm system. According to the authors, as the number of ants increases, so does the search time. For the configuration chosen in that paper, PSO outperforms ACO by 3% on one data set and 1% on the second data set. The authors point out that the networks generated by ACO are much simpler, but with relatively well-adjusted parameters, based on their configuration. The paper concludes that increasing the depth of an ACO network results in more complex and better networks, whereas choosing the number of layers in a PSO network is completely stochastic. Then, in [11], ACO is used to configure a modular neural architecture for image pattern recognition. The process of fine-tuning the modular neural architecture for image pattern recognition begins with the establishment of a colony of ants in various positions. For each of these ants, an adjustable graph is created. The offered implementation makes use of three modules, each with seven vertices. The final result is 98.82 percent pattern recognition accuracy. Finally, [12] uses a new heuristic to find a suitable region in the universe of discourse, which is a region in which values for certain observable variables are found. ACO, a widely accepted solution to the graph search problem, has been used to find such a suitable region. The paper utilizes

time series with a high level of data noise. These are known as fuzzy series. Forecasting is based on TAIEX data and a special model for predicting this type of series, which is based on fuzzy logic.

III. ANT COLONY OPTIMIZER

A. Pseudo algorithm for Ant colony optimization

Ant colony optimizer (ACO) is based on and inspired by ant behavior in the wild in search of food, as well as their communication via chemical pheromone pathways that lead from their nest to the food source. With this in mind, the optimizer's implementation is based on marking specific parts of the search space of the algorithm's hyper-parameters which will have the greatest contribution to reducing the prediction error. Initially, this algorithm was designed to solve the travel salesman problem, but later, specific variants of it were identified and used to optimize models using categorical, discrete, or continuous hyper-parameter values.

Algorithm 1 Ant colony optimization pseudo-code

Input: (N No. of ants, E function, max No. of iteration n_{max})

Configuration selection from the hyper-parameter space P

Creation of matrix $N \times P$

Evaluation of solution matrix S with function E

Saving results into $R = E(S)$

for $i = 1, 2, 3, n_{max}$ **do**

Selection of new parameter configuration

Creation of new solution matrix S'

Evaluation of the newly created matrix S'

Saving results into $R' = E(S')$

Updating S with the best solutions from S and S'

Output: S, R

B. Proposed implementation for Ant Colony Optimization

The algorithm begins by creating a matrix of size $N \times P$, where N is the number of selected ants for tuning and P is the number of hyper-parameters that need to be adjusted. This matrix is initially initialized with values chosen at random for the hyper-parameters. Following this step, the matrix is evaluated, and the results are placed in a new matrix of size $N \times 1$. The RMSE metric is used to calculate how good the solution is. After calculating the metric value and storing the result in a matrix of size $N \times 1$, the solutions of the matrix $N \times P$ are arranged in order of best to worst. The next step in the algorithm is to assign weights to each of the solutions. For this purpose, the following formula is used:

$$\frac{1}{qn * 2 * \pi} e^{\frac{-i^2}{2 * q^2 * N^2}} \quad (1)$$

where $qn = N * q$ and q is a parameter of the algorithm that is variable and can be experimented with, and i represents the ordinal number of each of the solutions after prior ordering from best to worst.

The Gaussian distribution is determined by this formula, which has a mean of 1 and a standard deviation of qn . The values of these weights are unchanged because they use the ordinal numbers of the solutions from 1 to N (the number of ants in the initial configuration). The algorithm then calculates the standard deviation for each of the parameters. A matrix of size $P \times 1$ is created for this purpose. This

matrix will aid us in the creation of the Gauss kernel and the determination of hyper-parameter values. Furthermore, the pheromone evaporation parameter is introduced as a parameter that controls the solution's convergence.

Following the creation of the matrix with standard deviation, a new matrix of size $N \times P$ is created. To fill the hyper-parameter values, the previously generated standard deviation matrix is used in such a way that the value for the standard deviation corresponding to the parameter for which they are currently taken is multiplied by a randomly generated variable from the half-open interval $[0,1)$, and this value is added to the value for this parameter from the best found solution so far. To avoid generating parameter values that are outside the parameter limits, all values that cross the upper or lower limit for a parameter are assigned the maximum or minimum value of the limits. The newly generated matrix is sent to the optimization process in this manner. This procedure runs all models on a predetermined number of processor cores. After completing the optimization and obtaining a matrix with metric values, the new solutions are combined with the best solutions thus far and sorted from best to worst. Only N of these solutions are saved, and the procedure is repeated z times until the maximum number of iterations is reached or the change in improvement does not exceed the threshold limit defined at the start. With the selection of N best solutions from $2 * N$ solutions a balance between research and exploitation is introduced.

IV. EXPERIMENTAL EVALUATION

A. Model representatives and expectations

For the purpose of experimental evaluation as regression model representatives SVR (Support Vector Regression) and HGB (Histogram Gradient Boosting) models were chosen. SVR, or Support Vector Regression, is a supervised machine learning method. The primary goal of this type of prediction method is to locate a hyper plane in space that contains the greatest number of points used in training. If the problem cannot be solved linearly, the kernel function is used to find this plane, which aims to transcend the inputs to a higher dimension where they can be solved linearly. In the literature, RBF is most commonly used as a kernel function to generalize models. During the experiments, the RBF (Radial basis function) kernel function was used in each model. SVR differs from other types of regression because it seeks the best hyperplane with a predetermined limit, rather than focusing on implicitly minimizing the error between the actual and predicted values. When there are a large number of samples, SVR has a significant disadvantage in terms of complexity and training time. This model has the advantage of being resistant to anomalous points. In contrast, histogram gradient boosting is a decision tree ensemble model. It is distinct from other types of Boosting algorithms in that it discretizes continuous input values into a set of distinct values. This accelerates the training of the trees added to the ensembles. The disadvantage is that if the number of iterations is very large, over-fitting can occur. Furthermore, this model is not robust in terms of anomalous points, which are usually the result of a minor over-fitting. This is due to the fact that each new tree is updated based on the residuals and prediction errors of previous trees.

The training time of the SVR model is expected to be significantly faster for a smaller data set, while the training

time of the HGB model is expected to be faster as the number of training data increases. The experimental evaluation in this paper consists of setting parameters for the two types of regressors (SVR and HGB) mentioned above using four methods: Grid Search, HyperOpt, Auto-Sklearn, and the proposed Ant Colony optimizer.

B. Data sets

Ten data sets from the UCI Machine learning repository [13] with varying sampling intervals and observations ranging from 61 to 69681 samples were chosen for a thorough evaluation of the proposed optimization algorithm.

Summary results for both types of models will be presented for each set, sorted by the chosen Mean Squared Error (MSE) comparison metric and the time required to find a solution.

TABLE I
DATASET NAME AND REFERENCE

Dataset Name	Reference
ai4i2020	[14]
Istanbul Stock Exchange	[15]
Garments worker productivity	[16]
Air Quality UCI	[17]
Energy Data	[18]
PRSA	[19]
Daily Demand Forecasting Orders	[20]
Metro Interstate Traffic Volume	[21]
ETT	[22]
Household Power Consumption	[23]

C. Metrics and hyper-parameters

For the purpose of model comparison Mean Squared Error (MSE) and time required to find a solution are used. Mean Squared Error (MSE) is the mean root value of the difference between the actual and the predicted value. A lower value with this metric determines a better model. The formula for this metric is:

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (2)$$

where MSE is *MeanSquaredError*, n is the number of samples, Y are the true values and $\hat{Y}$ are the predicted values.

MSE error minimization is accomplished by calculating the MSE metric in regards to 80% of the total data without shuffling from each data set (due to the problem of continuous data in time series) and determining the mean value of the two trainings obtained through 2-fold cross validation. Following the tuning process on the training data sets, the value for the previously mentioned metric MSE, is calculated on the leftover 20% of the data sets.

SVR parameters C (regulatory parameter) and epsilon (a parameter that determines the extent to which no error is provided in the training function for observations provided in the interval from epsilon to the real value for that observation) are used as hyper-parameters for tuning. In all experiments for SVR, the kernel function is set to RBF (radial basis function). In contrast, the following parameters are chosen for Histogram Gradient Boosting (HGB): learning rate (parameter that controls the model's learning), maximum leaf nodes (minimum number of sheets in each tree), minimum sample leaf (minimum number of samples per sheet in wood), and max depth (maximum depth of each tree).

V. RESULTS AND DISCUSSION

A. Results

According to Table II, the proposed implementation ACO method is the most accurate in six of the ten data sets. On the other hand, the proposed implementation ACO for the HGB model (Histogram Gradient boosting) is the best when compared by MSE in three out of ten cases, equating with Grid search and Auto-Sklearn.

According to Table III, the proposed implementation ACO method is the fastest in six of the ten data sets. On the other hand, the proposed implementation ACO for the HGB model (Histogram Gradient boosting) is fastest when compared by time needed for tuning in nine out of ten cases.

Finally, whether using the SVR or HGB model, the proposed ACO implementation is always among the top three best methods.

B. Discussion

After evaluating the approaches made over the ten data sets that differ in the number of samples and the number of features they possess, as well as the size of the values in the predicted variable, it can be concluded that in seven of the ten SVR models the fastest solution is the proposed architecture. For the MSE metric ranking in many cases where ACO is in second place, it is preceded by GS, which is an expected result because the implementation of ACO generates approximate values for the parameters that in the experimental phase were set to have the above limit of 1000 for parameter C and 1 for epsilon parameter in SVR models. With that in mind, ACO is most accurate by MSE in six out of ten experiments, and Grid Search is most accurate in two out of ten cases. In one case, because no data preprocessing was performed, only Auto-Sklearn produced good results with a budget of 100 seconds and utilization of the frame's internal preprocessor. In the final experiment, the HyperOpt framework produces an MSE that differs from the ACO solution by one decimal place.

In nine out of ten experiments, ACO produces significantly faster results for the second type of model, HGB (Histogram Gradient Boosting). The results for this model are very close to MSE (differ by a third decimal place).

Another finding is that GS is the slowest of all models. When experimenting with Auto-Sklearn and HyperOpt, it is necessary to set a budget, which in Auto-Sklearn is fully utilized, whereas in HyperOpt it is possible but not always fully utilized. Determining the exact budget is difficult, which is why in this experiment, it is set to 500 and 100 seconds in Auto-Sklearn and 500 seconds in HyperOpt, respectively, and these numbers are chosen arbitrarily.

VI. CONCLUSION AND FURTHER WORK

In this paper we presented an implementation for Ant Colony Optimizer for hyper-parameter tuning. This implementation was compared with three different benchmark methods (Grid Search, Bayesian Optimization and Tree Parzen Estimators) on ten different datasets and on two different types of models (SVR and HGB).

The ant colonies optimization strategy for hyper-parameter tuning showed minimal time to tune the hyper parameters while providing competitive results for prediction accuracy compared with the other method mentioned above. It also has the potential for a lot of parallelism as well. Its accuracy is

TABLE II
RESULTS FOR MSE METRIC FOR MODELS SVR AND HGB FOR EACH OF THE TEN DATASETS

Dataset	SVR MSE				HGB MSE			
Method	ACO	Grid search	AutoSklearn	HyperOpt	ACO	Grid search	AutoSklearn	HyperOpt
Data set 1	84.41	67.75	71.95	94.66	2.17	1.55	2.76	3.94
Data set 2	0.06	0.07	0.06	0.069	0.05	0.051	0.05	0.05
Data set 3	0.02	0.43	0.13	0.43	0.01	0.01	0.01	0.01
Data set 4	0.24	0.32	0.23	0.22	0.29	0.29	0.28	0.29
Data set 5	2.95	2.17	0.27	3.50	0.37	0.36	0.48	0.35
Data set 6	6612.90	6612.86	8889.93	6616.26	6183.95	6133.22	6202.17	6212.68
Data set 7	555.61	574.62	989.49	573.82	1622.55	1669.85	1069.75	1669.85
Data set 8	3658869.40	3660087.39	3871753.23	3660049.28	3543492.39	3536480.03	3533824.67	3547464.79
Data set 9	61.13	61.37	60.90*	2225.12	67.94	63.97	68.66	67.94
Data set 10	564.97	565.82	72533.10	612.22	2111.98	2862.77	1163.81*	3079.90

* DummyRegressor used leading to not valid results.

TABLE III
RESULTS FOR TIME NEEDED FOR HYPER-PARAMETER TUNING FOR MODELS SVR AND HGB FOR EACH OF THE TEN DATASETS

Dataset	SVR Time				HGB Time			
Method	ACO	Grid search	AutoSklearn	HyperOpt	ACO	Grid search	AutoSklearn	HyperOpt
Data set 1	88.12726	263.16185	500	350	4.67375	524.975	100	90
Data set 2	0.46472	4.19407	100	340	2.204	268.064	100	100
Data set 3	0.68046	0.89757	100	340	2.83316	328.047	500	100
Data set 4	16.147	57.844	100	344	4.37914	506.48	100	100
Data set 5	357.07076	2049.26036	500	400	60.26691	858.13465	100	100
Data set 6	898.712	3348.097	100	357	6.81851	649.3429	100	100
Data set 7	1.63201	2.24134	100	340	5.48762	144.80978	100	100
Data set 8	553.03885	4077.37868	500	354	5.16121	396.8439	100	101
Data set 9	3353.06297	17669.359	500	409*	6.21469	955.67905	100	102
Data set 10	1218.08134	2804.524	500	340	384.21429	592.46702	500	100

* Ends on first iteration due to the big number of samples in the dataset.

heavily reliant on the parameter boundaries. These boundaries are arbitrarily chosen and set in all experiments without taking into account problem modeling knowledge. In practice, these boundaries would be precisely defined based on the set, model, and the domain knowledge.

Moreover in this paper only the concept of parameter tuning is applied to two previously selected models in the domain of time series regression. In addition to this paper, it is possible to implement an AutoML (Automated Machine Learning) system, which is a system for searching algorithms and setting hyper-parameters for them without any human intervention [24]. Such systems consist of several processes, such as preparing the dataset, generating key features from that set, generating machine learning models, optimizing them, and finally evaluating them. In the literature, the part of the selection of machine learning models and their optimization is known as the Combined Algorithm Selection and Hyperparameter Optimization, or CASH for short.

REFERENCES

- [1] M. Kuhn and K. Johnson, Applied Predictive Modeling. Springer Science Business Media, 2013.
- [2] "objective function." <https://xlinux.nist.gov/dads/HTML/objective.html>
- [3] T. Yu and H. Zhu, "Hyper-Parameter Optimization: A Review of Algorithms and Applications," arXiv.org, Mar. 12, 2020.
- [4] "Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization," Journal of Electronic Science and Technology, vol. 17, no. 1, pp. 26–40, doi: 10.11989/JEST.1674-862X.80904120.
- [5] "Best Practices for Hyperparameter Tuning with MLflow," Databricks, May 06, 2019. <https://databricks.com/session/best-practices-for-hyperparameter-tuning-with-mlflow>
- [6] "Hyperparameter Optimization With Random Search and Grid Search," Machine Learning Mastery, Sep. 13, 2020.
- [7] P. I. Frazier, "A Tutorial on Bayesian Optimization," arXiv.org, Jul. 08, 2018. <https://arxiv.org/abs/1807.02811.pdf>
- [8] A. Agnihotri and N. Batra, "Exploring Bayesian Optimization," Distill, vol. 5, no. 5, May 2020, doi: 10.23915/distill.00026
- [9] E. Byla and W. Pang, "DeepSwarm: Optimising Convolutional Neural Networks using Swarm Intelligence," arXiv.org, May 17, 2019. <https://arxiv.org/abs/1905.07350.pdf>
- [10] S. Lankford and D. Grimes, "[PDF] Neural Architecture Search using Particle Swarm and Ant Colony Optimization".
- [11] "Ant colony optimization for the design of Modular Neural Networks in pattern recognition," IEEE Xplore.<https://ieeexplore.ieee.org/document/7727194>
- [12] Q. Cai, D. Zhang, W. Zheng, and S. C. H. Leung, "A new fuzzy time series forecasting model combined with ant colony optimization and auto-regression," Knowledge-Based Systems, vol. 74, pp. 61–68, Jan. 2015, doi: 10.1016/j.knosys.2014.11.003.
- [13] "UCI Machine Learning Repository." <https://archive.ics.uci.edu/ml/index.php>
- [14] "UCI Machine Learning Repository: AI4I 2020 Predictive Maintenance Dataset Data Set." <https://archive.ics.uci.edu/ml/datasets/AI4I+2020+Predictive+Maintenance+Dataset>
- [15] "UCI Machine Learning Repository: ISTANBUL STOCK EXCHANGE Data Set." <https://archive.ics.uci.edu/ml/datasets/ISTANBUL+STOCK+EXCHANGE>
- [16] "UCI MLR: Productivity Prediction of Garment Employees Data Set." <https://archive.ics.uci.edu/ml/datasets/Productivity+Prediction+of+Garment+Employees>
- [17] "UCI Machine Learning Repository: Air Quality Data Set." <https://archive.ics.uci.edu/ml/datasets/Air+Quality>
- [18] "Appliances energy prediction," data.world, <https://data.world/uci/appliances-energy-prediction>
- [19] "UCI Machine Learning Repository: Beijing PM2.5 Data Data Set." <https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data>
- [20] "UCI Machine Learning Repository: Daily Demand Forecasting Orders Data Set." <https://archive.ics.uci.edu/ml/datasets/Daily+Demand+Forecasting+Orders>
- [21] "UCI Machine Learning Repository: Metro Interstate Traffic Volume Data Set." <https://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume>
- [22] zhouhaoyi, "GitHub - zhouhaoyi/ETDataset:" GitHub. <https://github.com/zhouhaoyi/ETDataset>
- [23] "Household power consumption." <https://archive.ics.uci.edu/ml/machine-learning-databases/00235/>
- [24] I. Guyon et al., "A Brief Review of the ChaLearn AutoML Challenge: Any-time Any-dataset Learning Without Human Intervention," PMLR, Dec. 04, 2016. https://proceedings.mlr.press/v64/guyon_review_2016.html <https://arxiv.org/abs/2003.05689.pdf>

Blood pressure class estimation using CNN-GRU model

Ivan Kuzmanov

Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering
 Skopje, North Macedonia
 ivan.kuzmanov@students.finki.ukim.mk

Magdalena Kostoska

Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering
 Skopje, North Macedonia
 magdalena.kostoska@finki.ukim.mk

Ana Madevska Bogdanova

Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering
 Skopje, North Macedonia
 ana.madevska.bogdanova@finki.ukim.mk

Abstract—Blood pressure (BP) estimation can add on great value in emergency medicine, especially in case of mass casualty situations. The presented research aims to create a model for BP class estimation using electrocardiogram (ECG) and photoplethysmogram (PPG) waveforms. We focus on developing a BP classification model as a convolutional neural network (CNN) - gated recurrent unit (GRU) hybrid model, containing both CNN and GRU layers. The used dataset is the publicly available UCI Machine Learning Repository dataset. We have achieved f1 score of 0.83, 0.73 and 0.74 respectively according to the BP classes and 78% overall accuracy.

Keywords—blood pressure (BP) estimation, triage, electrocardiogram (ECG), photoplethysmogram (PPG), gated recurrent unit (GRU), artificial neural network, deep learning, CNN-GRU hybrid model

I. INTRODUCTION

Blood pressure (BP) is an important metric to determine a patient's hemodynamic stability. While primary triage (ex. START) does not include blood pressure in the decision making process, this metric is important in the secondary triage. Blood pressure is usually measured noninvasively by using a cuff-based measuring device or invasively, in a specialized hospital setting. Using the first method (cuff-based), the values are obtained on demand or in regular intervals thus are non-continuous. The second method requires specific conditions, equipment and highly-trained staff. Furthermore in case of mass casualty event it may not be possible to repeatedly measure blood pressure with both of the methods.

The goal of our research is to estimate BP in real-time given the embedded electrocardiogram (ECG) and photoplethysmogram (PPG) signals' values. ECG represents the electrical activity of the heart, while PPG shows the changes of blood volume in the microvascular tissue. The ECG and PPG values will be obtained by using patch-like combined multi-sensor attached to a human's chest. Both of them are low cost sensors that can be used as integrated sensors on a patch. The sensors will obtain basic signals and vital parameters necessary for triage process [1]. The correlation among BP, ECG and PPG has been researched in other studies [2] [3] [4] [5].

The ECG and PPG signals are given in a time series. Recurrent neural networks are a type of neural networks that specialize in processing sequential information that are

specifically built to be able to follow long term dependencies. This type of network are typically slow and difficult to train. Because of the nature of the problem the classification needs to be continuous. To further increase the efficiency of the model it's combined with a CNN layer which essentially performs feature selection. The model our paper proposes is a CNN-GRU neural network.

The rest of the paper is organized as follows. Section 2 presents an overview on similar papers or related researches. The methodology is described in Section 3. The results of the experiments and the discussion are presented in Section 4, including the introduction of the dataset, the preprocessing part and the used methods. Section 5 presents the conclusion.

II. RELATED WORK

In this section, an exploratory analysis of related work is done. Estimation of blood pressure from ECG and PPG is a theme that has vastly been researched. There are mainly two approaches used in similar papers: the first one is extracting relevant features from the signals and the other one is building neural networks that can process raw signals. In this paper it's opted to go with the second approach because of lack of assurance that the PPG and ECG signals in this dataset have been recorded in sync.

The most commonly used features with the highest information gain are pulse transit time (PTT) and pulse arrival time (PAT). PTT is more prominent in recent work. The features are similar, however a study done on this topic shows that PAT is not unable to detect subtle BP changes as well as PTT [6]. PTT forms the basis for continuous BP estimation in many works [4] [7]. Often time the features are combined with others in an attempt to get better results. One such features is photoplethysmogram intensity ratio (PIR), a measure for the arterial diameter change [8]. The strong correlation between the ratio and BP is explored in a different study that also states that patient with hypertension have a higher mean PIR with higher variance for this value [9]. Another study performs a complexity analysis to extract features from ECG. The machine learning model of that study achieves mean absolute error (MAE) of 8.64 mmHg for SBP, 18.20 mmHg for DBP, and 13.52 mmHg for MAP [10]. These results are further improved with model calibration. The features are

not always limited to what can be extracted from the PPG and/or ECG signal. For example a study focusing on features extracted from PPG and combining them with demographic characteristics achieves root mean square error (RMSE) of 6.74 mmHg for SBP and 3.59 mmHg for DBP.

The bio signals that a body emits are depended on many factors such as: medications, circulatory system diseases and interaction between the different physiological systems. This contributes to patient-specific morphological contours for the signals. This is the main problem with using the approach with feature extraction. Other studies focus on building model capable of estimating blood pressure from raw signals. One such study's proposed solution is a hierarchical ANN-LSTM neural networks, where the lower level extract the features and the higher ones track time distributed dependencies. The proposed network's results are RMSE of 1.56 mmHg for SBP and 0.85 mmHg for DBP [11]. A hybrid CNN-LSTM network that works with a similar approach with the previously mentioned one achieves MAE of 4.43 ± 6.09 mmHg for SBP and 3.23 ± 4.75 mmHg for DBP [12]. A study attempts to predict the blood pressure in hemodialysis patients using a CNN-GRU neural network [13]. Another study attempts to utilize the difference between PPG and ECG to feed to a neural network with a CNN-LSTM architecture [14]. This difference is used to calculate PAT and it allows the network to learn PAT information.

III. MATERIALS AND METHODS

In this section we describe the dataset, data preprocessing procedures, the developed CNN-GRU model and the used evaluation metrics.

A. Dataset

Most of BP studies utilise the MIMIC database or some subset of it. However because of a current memory storage constraints, in this study the "Cuff-Less Blood Pressure Estimation Data Set" published in the UCI Machine Learning Repository [15] is used. This dataset is consisted of 12000 instances, each having three recorded signals of variable length. The signals correspond to PPG, ABP, and ECG recorded with a 125 Hz frequency. There is no other information in the dataset, nor patient identifiers, so it's impossible to identify whether two separate instance are signals from the same patient.

1) *PPG*: Photoplethysmograph, PPG, measures the blood volume changes in microvascular tissue by detecting changes of light absorption on the skin. It's a technology that has gotten prominence in the medical field in recent years because of it's low-cost and it's usability in understanding the cardiovascular system [3].

2) *ABP*: Arterial blood pressure, ABP is a waveform representing blood pressure measured from within an artery. Measurements of this kind can only be taken in a hospital setting. Blood pressure is the pressure created in the blood vessels of a cardiovascular system as the heart pumps the blood. The positive peaks of the waveform represent the moment when the heart contracts and the blood pressure is at it's highest and the negative peaks represent the moment when the heart relaxes after a contraction and the blood pressure is at it's lowest. The blood pressure in these moments are known as the systolic blood pressure, SBP, and diastolic blood pressure, DBP.

3) *ECG*: The heart plays a vital role in the circulation of blood. Electrocardiogram, ECG, measure the electrical activity of the heart. While it's measured no electricity is send to the body. It only tracks the polarization wave caused by the heart as it beats. The ECG in this study is a 1-channel one. With each beat an electrical impulses coordinate the heart's contraction. An ECG simply records these impulses as they move through different parts of the heart.

On Figure 1 a 8 second electrocardiogram is visualized and the peaks characteristic to ECG are marked.

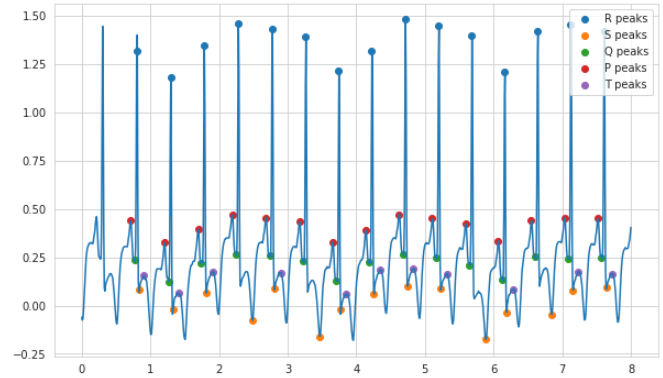


Fig. 1. 8s ECG waveform with annotated peaks

B. Preprocessing

1) *Data Validation*: The data is validated according to the two following criteria: 1. There aren't missing values in any of the three signals, PPG, ABP and ECG; 2. The pulse pressure is above 20 and less than 80. No data instance is taken into consideration when building the models if both of the criteria weren't met.

The average pulse pressure, i.e. the difference between SBP and DBP is between 40 and 60. Too low pulse pressure indicates poor heart function, while too high pulse pressure indicates a health problem. Valid data are considered to be in the range (20 , 80). It is also worth noting that the signal quality is mediocre. The values calculated from signals with extremely low or high pulse pressure may be a result of a sensor noise. This problem can be remedied with a larger dataset of higher quality signals in future work.

2) *Data Segmentation*: The signals per patient in the dataset are with variable length, within the range [1000, 74000] data points. The signals are recorded with frequency of 125 Hz, concluding that the shortest signal length in the dataset is 8 seconds and the longest - 9.9 minutes long.

Different studies suggest and work with segments of different length, usually within the range of 5-30 seconds [10] [12]. In this study it is decided to use 8 seconds (8s) segment length. This decision was influenced by the aforementioned fact that the shortest signal has a length of 8 seconds and if we were to work with longer signal per instance, we would have had to exclude the shorter signals, reducing the already small dataset. To further increase the number of samples, multiple segments are taken from the longer signals. The segments are taken sequentially with no overlap. The decision to use this type of segmentation was influenced by a similar study [16], where the same approach was used. It should be noted that by taking multiple samples from the same patient we introduce bias due to the patients specific morphological

contours. However this approach was deemed as necessary since this dataset is quite limited.

3) *Normalization*: The PPG and ECG signals are normalized by scaling the values into decimals between 0 and 1, with 0 being the minimum value of the signal and 1 being the maximum. Normalizing the data before training the models is a common practice and generally leads to better convergence in machine and deep learning models.

4) *Filtering*: Noise is a common problem with sensors. It can appear in a signal for a variety of reasons such as - noise caused by the electrical activity of the muscles, breathing, loss of contact with the skin, etc. Often, the noise has the same frequency as the signal and thus it's difficult to completely remove it without distorting the signal.

Filtering is the process of removing unwanted component from the signal. For the filtering of biological signals the most commonly used filters are notch and bandwidth filters. The decision to use this type of filters was made based on the results presented in previous works in this field which show that bandwidth filters are suitable for both PPG and ECG [11].

However, the PPG is filtered by a filter of the fourth order and the ECG by a filter of the fifth order.

5) *BP categorization*: For the purposes of this study it's necessary to transform the ABP waveform into 2 scalars, SBP and DBP, used for the blood pressure categorisation. To avoid possible outliers it's decided that instead of taking the min and max value in the segment, it's better to calculate the average of the local extremes, maxima for SBP and minima for DBP [17].

An illustration of the extraction process is presented on Figure 2.

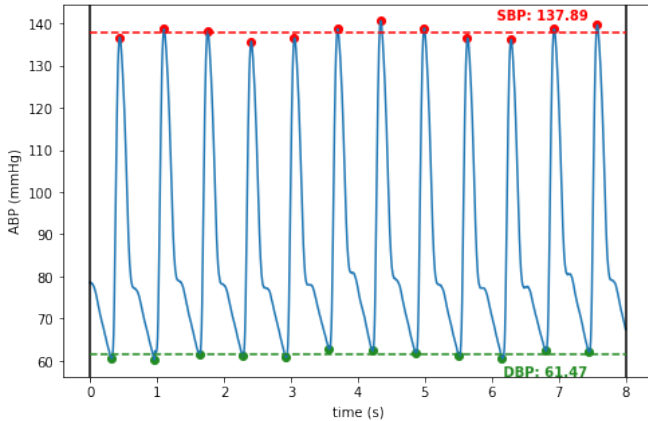


Fig. 2. Illustration of SBP and DBP extraction from ABP waveform

The European Society of Hypertension and the American College of Cardiology have different classifications schemes for blood pressure categorization. In this study it's decided to use the latter. Following the scheme there are four blood pressure: normal, elevated, stage 1 hypertension and stage 2 hypertension [18]. Since the dataset is small and the hypertension classes are under represented, it is decided to merge them into a single class. The elevated blood pressure class in this study is referred to as prehypertension. In Table I can be seen the conditions for division and the number of samples per class.

As shown in Table I, there is a class imbalance in the dataset used in this study. The results of models trained on

TABLE I
BLOOD PRESSURE CATEGORIZATION

Category	SBP	DBP	Number of Samples
Normal	< 120	< 80	118644
Prehypertension	120 – 139	80 – 89	102927
Hypertension	≥ 140	≥ 90	49948

an imbalanced dataset often favor the more common classes. Many solutions have been developed to counteract this problem. The most common and widely used are oversampling of the minority classes and undersampling of the majority classes. The most popular type of oversampling, SMOTE is not applicable to this problem, since the input vectors are time distributed sequences. Undersampling methods were also considered. These methods are useful in some specific cases, but usually achieve lower classification performance because the loss of data negatively impacts the models ability to learn. A problem with their adaptation for the purpose in this research, is the problem of measuring the similarity between two signals. There is an undersampling method that removes samples at random, however this method is unlikely to yield good results because of the above mentioned problem.

A simple solution of the problem of dealing with an imbalanced dataset is the method of assigning weight to the classes. The weights are assigned so that the training samples of the majority classes have a lesser impact and the minority classes have greater impact while training. The less represented a class is, the higher weight value is appointed and vice versa. The formula used to calculate the weights is $weight_class_i = \frac{total_num_samples}{num_classes * num_samples_i}$. The calculated weights for the classes are: 0.76 for Normal, 0.88 for Prehypertension and 1.81 for Hypertension.

C. Model Structure

Recurrent Neural Network (RNN) is a type of neural networks designed to be able to follow temporal dependencies through temporal sequence. A main problem with these networks is the vanishing gradient problem. The gradient disappears or explodes after a few timesteps, thus preventing learning long term dependencies. To overcome this problem several solutions have been developed. The most popular one is the Long Short Term Memory, LSTM. The cells in a LSTM layer have an internal state and the flow of data is strictly controlled through the use of three gates: input gate, output gate, and forget gate. Gated recurrent units, GRU are recurrent network layers that are similar to LSTM, however they have fewer parameters and lack a forget gate. Even with this simplification they perform comparably on a variety of problems [19]. The gates in a GRU cell are called an update gate and a reset gate. They control the flow of data. A GRU cell has the capacity to keep information in the form of a cell state for a long time without overwriting it.

A simple way to improve a recurrent neural networks is the use of bidirectional layers. A bidirectional layer is made up of two different layers linked to the same output layer. The time-distributed sequence the network is processing is passed once forwards and once backwards. By passing the sequence backwards, the network attempts to use future context to explain the present by looking at the previous data points. The same is true for GRU layers. Networks that utilize bidirectional layers are more computationally expensive than

networks of the same size that utilize only single directional recurrent layers [20].

Convolutional neural network, CNN, is a type of neural network that uses the mathematical concept of convolution, an operation on two functions that produces another function. It's a popular network that's most often used in problems involving computer vision and picture analysis. It should be noted that ability to extract patterns from data also makes applicable for solving other types of problems than those mentioned above such as processing natural languages, recommending systems, brain-computer interfaces and any other where patterns can be found.

The input vector consist of 2000 real numbers and it enters the network in the format (100, 10, 2), where the values correspond to (length of subsequences, number of timesteps, number of features). There are 2 features: PPG and ECG. The size of the subsequences and the size of the timesteps are inversely proportional. If the subsequences are too small the CNN won't be able to find patterns in them, while if there are too few timesteps the ability to track longterm dependencies granted by the GRU won't be fully utilized. Keeping that in mind it's empirically decided on previously mentioned values.

We use a multivariate CNN-GRU architecture. A CNN-GRU is a hybrid neural network that contains both CNN and GRU layers. The first layer is a time distributed 1d convolutional layer that receives the input vector. This layer processes the subsequences attempting to find usable features. Convolutional layers are typically followed by max pooling ones, such is the case in our model. Max pooling layers perform downsampling of the inputs from the previous layer retaining only the most salient features. A batch normalization layer normalizes the inputs and stabilizes the learning process. After this step, the data is passed to a bidirectional GRU layer. This layer attempts to follow the temporal dependencies of the features generated by the convolution layer. The network ends with a dense layer with three neurons, one for each class, with a Softmax activation. In the proposed model, each of the classes (Normal, Prehypertension, Hypertension) is represented as 0, 1, 2, respectively.

The model is somewhat similar to CNN-LSTM models given in related work, with the main difference being that most of those models are geared to solving regression, while ours works with classification, as well as the use of a GRU instead of an LSTM layer. The goal of this study is to aid in the development of a lightweight algorithm for blood pressure category classification. Since GRU layers are simpler and therefore faster yet comparable to LSTM, they are used as a part in this architecture.

D. Evaluation Metrics

In this section, the metrics that evaluate the performance of this model are explained. This model attempts to classify the category of blood pressure, hence the classification evaluation metrics are used:

- 1) Accuracy is the ratio of correctly predicted samples against the total number of samples;
- 2) Precision is the ratio of the number of correctly classified samples and the total number of samples classified as such;
- 3) Recall is the ratio of the number of correctly classified samples and the total number of samples of that class regardless of whether they were accurately classified;

- 4) F1 score is the harmonic mean of precision and recall.

After the aforementioned metrics were calculated, averaging schemes were used obtaining a single performance value. The schemes used in this paper are:

- 1) Macro average is a schema often used for multi-class classification that first calculates the above mentioned metrics independently and averages their results;
- 2) Weighted average is a schema that averages the other metrics by assigning them weight with regard to their class representation. The less represented a classes is in the testing dataset the higher its weight.

The support is just the number of testing samples for each class from which the other metrics are calculated. The metrics are presented via a classification report from the sklearn python package.

IV. RESULTS AND DISCUSSION

In this section, the achieved results are presented and discussed. The proposed model has a CNN-GRU architecture that predicts the BP category using ECG and PPG signals.

A. Model Training

The dataset is split into training and testing sets with a ratio of 85:15. A further 10% from the training set is designated for a validation dataset. By using a keras tuner, it is concluded that ADAM optimizer with a learning rate of 0.001 is suitable for this specific problem. The most successful model structure so far was trained with a batch size of 512 for 22 epochs.

B. Model Evaluation

In Figure 3 a classification report of the model's performance is shown. The testing dataset has 40728 samples divided in three categories - 17673 as normal BP, 15466 have prehypertension and 7589 have hypertension. The model has overall accuracy of 0.78. The most distinguishable class for the model is the normal class. It has a precision of 88%. The prehypertension class has the lowest f1-score, that leads us to a conclusion that the model struggles to learn the class boundaries between the normal and the prehypertension class, and the prehypertension and the hypertension class.

	precision	recall	f1-score	support
0	0.88	0.79	0.83	17673
1	0.72	0.75	0.73	15466
2	0.69	0.81	0.74	7589
accuracy			0.78	40728
macro avg	0.76	0.78	0.77	40728
weighted avg	0.78	0.78	0.78	40728

Fig. 3. Classification report

In Figure 4 the training and validation accuracy per epoch are shown and in Figure 5 the training and validation loss per epoch are shown.

C. Discussion

The proposed model obtains overall accuracy of 78%, which is higher than our previous best results with different LSTM models. While there is an improvement, the model would still require further development in order to be applicable.

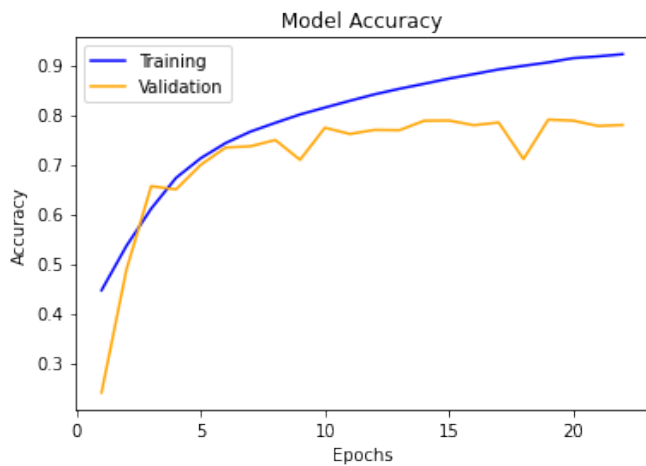


Fig. 4. Results - Training and validation accuracy

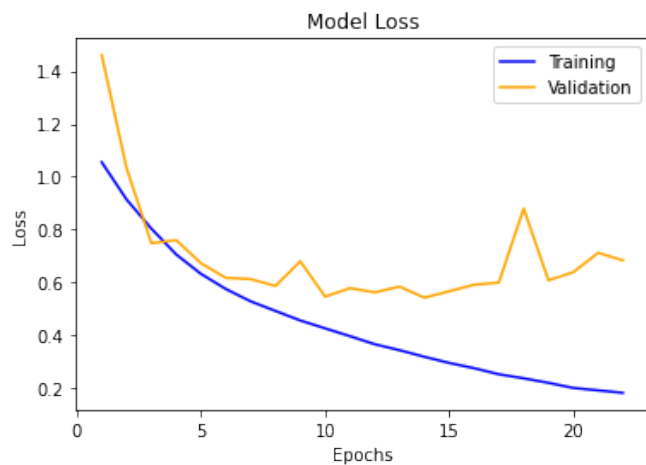


Fig. 5. Training and validation loss

Most of the studies listed in related work utilize regression models. Thus making direct comparison with our classification model impossible. It should also be taken in to account that the results the model achieves are somewhat influenced by the bias introduces during the segmentation phase. In future work a larger dataset will be used to reduce the bias and enhance the development of a generalized algorithm for blood pressure classification, as well as extending the model's pulse pressure limits to include abnormal cases.

One important aspect that is not considered in the current model are the demographic characteristics, since the information gain from these characteristics is significant. Arteries stiffen with age and as a result the blood pressure needed to deliver the blood rises, meaning older people have on average higher blood pressure.

Other aspect that may impact the model is the short segment size (8 sec). Our further research will focus on both of these aspects. The results show that the proposed model can distinguish between the different categories. This is important trait, since they indicate that using a larger dataset can improve the results. Hence we assume that using an much larger dataset from different sources will help to overcome the pointed aspects.

V. CONCLUSION

Our research is focused on blood pressure category estimation, given ECG and PPG signals. The idea is to use the embedded sensors placed on a patch in order to obtain the aforementioned signals, and using a tablet or a remote server to utilise our developed model. The whole system should be used in emergency and mass casualty situations. In these situations it is not practical to measure BP values manually with cuff-based devices for each subject in order to estimate the subject's hemodynamic state in conjunction with the other vital parameters obtained by the sensor. Other important aspect is that BP varies and thus the BP measurement should be continuous to enhance the second triage process. Given large number of subjects in situations with high number of casualties, it would not be possible to regularly manually measure BP and effectively trace changes in subject's health state. Hence BP category estimation can be an important future in the triage process and increase the survival rate.

In this paper we propose building and training of CNN-GRU model. The input form the model are sequences of ECG and PPG signals and output is the BP category. An important aspect is the preprocessing stage, given that we have raw signals. We normalize and filter the ECG and PPG signals from the dataset we use. Since the dataset contains ABP signals, we extract SBP and DBP from these signals. The model learns by adjusting its weights.

For imminent future work we plan on refining our model using much larger dataset, as MIMIC III, and including other futures and characteristics. Later we also plan to work on refining of the results by developing regression models for BP estimation with deep learning approach, using Big Data for selection of different features from the ECG and PPG signals.

ACKNOWLEDGMENT

This paper has been written thanks to the support of the "Smart Patch for Life Support Systems" - NATO project G5825 SP4LIFE and by the National project IBS4LIFE of Faculty of Computer Science and Engineering, at Ss. Cyril and Methodius University in Skopje.

REFERENCES

- [1] F. Lehocki, A. M. Bogdanova, M. Tysler, B. Ondrusova, M. Simjanoska, B. Koteska, M. Kostoska, M. Majak, and M. Macura, "Smart-patch for victims management in emergency telemedicine," in *2021 13th International Conference on Measurement*, 2021, pp. 146–149.
- [2] Y.-H. Li, L. N. Harfiya, K. Purwandari, and Y.-D. Lin, "Real-time cuffless continuous blood pressure estimation using deep learning model," *Sensors*, vol. 20, no. 19, 2020. [Online]. Available: <https://www.mdpi.com/1424-8220/20/19/5606>
- [3] Y.-C. Hsu, Y.-H. Li, C.-C. Chang, and L. N. Harfiya, "Generalized deep neural network model for cuffless blood pressure estimation with photoplethysmogram signal only," *Sensors*, vol. 20, no. 19, 2020. [Online]. Available: <https://www.mdpi.com/1424-8220/20/19/5668>
- [4] Y. Choi, Q. Zhang, and S. Ko, "Noninvasive cuffless blood pressure estimation using pulse transit time and hilbert–huang transform," *Computers & Electrical Engineering*, vol. 39, no. 1, pp. 103–111, 2013, special issue on Recent Advanced Technologies and Theories for Grid and Cloud Computing and Bio-engineering. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0045790612001711>
- [5] C. Poon and Y. Zhang, "Cuff-less and noninvasive measurements of arterial blood pressure by pulse transit time," in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, 2005, pp. 5877–5880.
- [6] G. Zhang, M. Gao, D. Xu, N. B. Olivier, and R. Mukkamala, "Pulse arrival time is not an adequate surrogate for pulse transit time as a marker of blood pressure," *Journal of Applied Physiology*, vol. 111, no. 6, pp. 1681–1686, 2011, pMID: 21960657. [Online]. Available: <https://doi.org/10.1152/jappphysiol.00980.2011>

- [7] S. Goli and Jayanthi, "Cuff less continuous non-invasive blood pressure measurement using pulse transit time measurement," *International Journal of Recent Development in Engineering and Technology*, vol. 2, no. 1, pp. 86–91, 2014.
- [8] X.-R. Ding and Y.-T. Zhang, "Photoplethysmogram intensity ratio: A potential indicator for improving the accuracy of ptt-based cuffless blood pressure estimation," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 398–401.
- [9] Y. Chen, Y. Zhu, H. T. Ma, and H. Huang, "A study of photoplethysmography intensity ratio in hypertension," in *2016 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, 2016, pp. 317–320.
- [10] M. Simjanoska, M. Gjoreski, M. Gams, and A. Madevska Bogdanova, "Non-invasive blood pressure estimation from ecg using machine learning techniques," *Sensors (Basel, Switzerland)*, vol. 18, no. 4, p. 1160, Apr 2018, 29641430[pmid]. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/29641430>
- [11] M. S. Tanveer and M. K. Hasan, "Cuffless blood pressure estimation from electrocardiogram and photoplethysmogram using waveform based ann-lstm network," *Biomedical Signal Processing and Control*, vol. 51, pp. 382–392, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1746809419300722>
- [12] S. Yang, Y. Zhang, S.-Y. Cho, R. Correia, and S. P. Morgan, "Non-invasive cuff-less blood pressure estimation using a hybrid deep learning model," *Optical and Quantum Electronics*, vol. 53, no. 2, p. 93, Jan 2021. [Online]. Available: <https://doi.org/10.1007/s11082-020-02667-0>
- [13] H.-W. Hu, W.-L. Jheng, E.-H. Huang, K.-Y. Chen, C.-H. Liu, H.-Y. Chi, W. Chin-Yu, Y.-C. Chen, J.-A. Wang, H.-M. Lin, S. Ghose, C.-W. Chen, and J.-F. Wang, "Blood pressure value prediction in hemodialysis patients using the cnn+gru network architecture," in *2021 9th International Conference on Orange Technology (ICOT)*, 2021, pp. 1–4.
- [14] D. U. Jeong and K. M. Lim, "Combined deep cnn-lstm network-based multitasking learning architecture for noninvasive continuous blood pressure estimation using difference in ecg-ppg features," *Scientific Reports*, vol. 11, no. 1, p. 13539, Jun 2021. [Online]. Available: <https://doi.org/10.1038/s41598-021-92997-0>
- [15] M. Kachuee, M. M. Kiani, H. Mohammadzade, and M. Shabany, "Cuff-less high-accuracy calibration-free blood pressure estimation using pulse transit time," in *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2015, pp. 1006–1009.
- [16] S. Baker, W. Xiang, and I. Atkinson, "A hybrid neural network for continuous and non-invasive estimation of blood pressure from raw electrocardiogram and photoplethysmogram waveforms," *Computer Methods and Programs in Biomedicine*, vol. 207, p. 106191, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0169260721002650>
- [17] T. Athaya and S. Choi, "An estimation method of continuous non-invasive arterial blood pressure waveform using photoplethysmography: A u-net architecture-based approach," *Sensors*, vol. 21, no. 5, 2021. [Online]. Available: <https://www.mdpi.com/1424-8220/21/5/1867>
- [18] A. V. Chobanian, G. L. Bakris, H. R. Black, W. C.ushman, L. A. Green, J. L. Izzo, D. W. Jones, B. J. Materson, S. Oparil, J. T. Wright, E. J. Roccella, and null null, "Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure," *Hypertension*, vol. 42, no. 6, pp. 1206–1252, 2003. [Online]. Available: <https://www.ahajournals.org/doi/abs/10.1161/01.HYP.0000107251.49515.c2>
- [19] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014.
- [20] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "The performance of lstm and bilstm in forecasting time series," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 3285–3292.

The influence of stock market indexes (S&P500 and Dow Jones) on cryptocurrencies prices

Gorast Angelovski

*Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
gorast.angelovski@students.finki.ukim.mk*

Ana Todorovska

*Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
ana.todorovska@finki.ukim.mk*

Ivan Rusevski

*Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
ivan.rusevski@students.finki.ukim.mk*

Jovana Marojevikj

*Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
jovana.marojevikj@students.finki.ukim.mk*

Eva Spirovska

*Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
eva.spirovska@students.finki.ukim.mk*

Hristijan Peshov

*Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
hristijan.peshov@students.finki.ukim.mk*

Irena Vodenska

*Laboratory for Interdisciplinary Finance and Economics
(LIFE) Research, Administrative Sciences Department,
Financial Management
Metropolitan College, Boston University
Boston, USA
vodenska@bu.edu*

Lubomir T. Chitkushev

*Laboratory for Interdisciplinary Finance and Economics
(LIFE) Research, Computer Science Department
Metropolitan College, Boston University
Boston, USA
ltc@bu.edu*

Dimitar Trajanov

*Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
dimitar.trajanov@finki.ukim.mk*

Abstract—In this paper we analyze openly available time series data for the prices of 18 cryptocurrencies and 2 stock market indexes (S&P500 and Dow Jones).

First, we calculate the correlation values between the cryptocurrencies and indexes datasets.

Then, we use a state of the art time series prediction library (XGBoost) in order to make prediction models for the daily prices of all the cryptocurrencies, using the stock market index datasets as input features in the training model.

We calculate metrics for the difference between the actual prices and the prices predicted using our models.

Finally, we show the feature importance score that our model attributed to each prediction model, and compare the score between the three input features (S&P500 dataset, Dow Jones dataset, and the actual cryptocurrency dataset).

Index Terms—cryptocurrencies, stock market indexes, correlations, RMSE, feature importance

I. INTRODUCTION

Cryptocurrencies have become a global phenomenon. With the rise of digitalisation in recent years, cryptocurrencies introduced a novel medium, and their market is growing rapidly and breaking records by the day. However, this market is also categorized as highly volatile, meaning that it is

subjected to rapid changes in price and demand, sometimes overnight. Some of these changes are attributed to isolated events which have the ability to trigger an effect of mass speculation, usually through social media such as Reddit and Twitter. This presents investors and traders with a major challenge when it comes to analyzing and predicting the cryptocurrency market.

On the other hand, there are the major stock indexes, such as the S&P500 and DowJones, which are the most popular index funds and are the so called 'standard' for stock investing since the 1900s. In the field of economics these indexes, among others, are considered as classical economic indicators - implying that through them we can track and monitor the overall status of the economy.

In this paper we are investigating the relationship between these two US stock market indexes and the price of cryptocurrencies. We are using time series analysis in order to try and predict the fluctuations in the prices of cryptocurrencies, in correlation with the price levels of the S&P500 and DowJones. Our idea is to find a correlation between the price movements of these two indexes and some of the most popular cryptocurrencies, and thus predict the general trend

in which the cryptocurrency market is going.

II. S&P500 AND DOW JONES INDEXES

The S&P 500 Index, or Standard & Poor's 500 Index, is a market-capitalization-weighted index of 500 leading publicly traded companies in the U.S. It is not an exact list of the top 500 U.S. companies by market cap because there are other criteria that the index includes. Still, the S&P 500 index is regarded as one of the best gauges of prominent American equities' performance, and by extension, that of the stock market overall.

The Dow Jones Industrial Average (DJIA), also known as the Dow 30, is a stock market index that tracks 30 large, publicly-owned blue-chip companies trading on the New York Stock Exchange (NYSE) and the Nasdaq. The Dow Jones is named after Charles Dow, who created the index in 1896 along with his business partner Edward Jones.

III. DATA

In this study the structured data we use for time series forecasting is publicly available.

A. Cryptocurrency prices

We collect historical prices for eighteen cryptocurrencies from Yahoo Finance for the period between March 2019 and March 2021.

For each cryptocurrency, the historical prices dataset contains a date, open price, high and low prices, close price, adjusted closing price, and volume.

The eighteen cryptocurrencies we picked based on popularity are the following: Bitcoin, Monero, Cardano, Stellar, Ripple, Iota, Chainlink, Litecoin, Ethereum, Neo, Dogecoin, Celsius, Dash, EOS, Nano, NEM, Maker, and VeChain.

B. Stock Market Indexes

We collect historical prices for the two biggest stock market indexes in the US, The S&P500 and The Dow Jones Industrial Average from Yahoo Finance for the period between March 2019 and March 2021.

For each stock market index, the historical prices dataset contains a date, open price, high and low prices, close price, adjusted closing price, and volume.

IV. METHODOLOGY

The methodology of this research consists of four phases, which are: data preprocessing, finding correlation between the indexes (S&P500 and DowJones) and cryptocurrencies, prediction of cryptocurrency prices, and analyzing the feature importance of the resulting time series.

The steps will be further analyzed in the text below.

A. Data preprocessing

The data used in this research includes the prices of the 18 most popular cryptocurrencies and 2 US stock market indexes (DowJones and S&P 500) in the date range from 01/03/2019 to 01/03/2021. Since the US stock market does not operate during weekends and holidays, these days are filtered out.

B. Correlation between stock market indexes and cryptocurrency prices

For the purpose of finding relations and patterns between the prices of the cryptocurrencies and indexes, it is relevant to measure the pairwise correlation of the time series data for the prices.

The two cases observed are: Correlation between S&P500 and each cryptocurrency, and correlation between DowJones and each cryptocurrency.

The pairwise correlation of DowJones daily prices and S&P500 daily prices is 0.937, meaning the prices are following the same trend.

This implies that their corresponding correlations to the cryptocurrencies will also be similar.

C. Time series predictions

For the purpose of predicting the daily prices, firstly the data is split such that 80 percent of it is used for training, and the remaining 20 is used for testing. Then, the data is transformed using a logarithmic function, as it is an effective method of removing exponential variance with time series data.

XGBoost is used as a model for prediction, as it is a state of the art model that we found shows the best results for time series forecasting in our previous research.

In order to get the optimal parameters for training the model, GridSearchCV is used. After the predictions are made, the data is transformed back to the original format using inverse log. As a accuracy metric, RMSE is used.

D. Feature importance

As for the feature importance, we calculate the F-score function, in order to quantify the extent to which a feature has contributed to the model's output.

V. RESULTS

A. Correlations

All of the resulting correlations fall in the range between 0.10 and 0.24, except for one outlier with a correlation of 0.00 with both indexes - Nem. This indicates a weak correlation between the prices of the cryptocurrencies and the stock market indexes.

The S&P 500 has an average correlation of 0.175, and the Dow Jones has an average correlation of 1.40

B. RMSE between predicted and actual prices

The RMSE represents the difference between predicted and actual values measured in USD.

The results vary greatly due to the inconsistent pricing of cryptocurrencies.

C. Feature importance using F-score

The highest feature importance is most often attributed to the price of the cryptocurrency itself - this occurs in 9 out of 18 cases.

The next highest feature importance is attributed to the price of the S&P 500 - in 8 out of 18 cases.

In 16 out of 18 cases, the S&P 500 has a higher feature

TABLE I
CORRELATIONS

Cryptocurrency	S&P500_corr	DowJones_corr
Bitcoin	0.17	0.13
Monero	0.19	0.14
Cardano	0.21	0.17
Stellar	0.21	0.18
Ripple	0.14	0.11
Iota	0.23	0.19
Chainlink	0.20	0.17
Litecoin	0.18	0.14
Ethereum	0.22	0.18
Neo	0.19	0.18
Dogecoin	0.12	0.10
Celsius	0.13	0.09
Dash	0.14	0.11
Eos	0.17	0.13
Nano	0.23	0.19
Nem	0.00	0.00
Maker	0.20	0.14
VeChain	0.22	0.16

TABLE II
RMSE BETWEEN PREDICTED AND ACTUAL PRICES

Cryptocurrency	RMSE
Bitcoin	1248.4452
Monero	7.8988
Cardano	0.0184
Stellar	0.0214
Ripple	0.0478
Iota	0.0203
Chainlink	1.4736
Litecoin	7.7578
Ethereum	62.0107
Neo	1.2886
Dogecoin	0.0038
Celsius	0.3219
Dash	10.2230
Eos	0.1604
Nano	0.3105
Nem	0.0550
Maker	86.4395
VeChain	0.0018

importance compared to the Dow Jones.

The Dow Jones index has the lowest feature importance scores, the lowest F-score of the three features in 14 out of 18 cases.

VI. CONCLUSION

This research was conducted as an extension of a larger research topic that we are working on.

Our research presents an opportunity to understand the relationship between these different markets and utilize this knowledge as an investment asset.

Based on the results obtained, we conclude that these indexes have an insignificant impact on the price of cryptocurrencies. It could be implied that the impact of the S&P500 on the price of cryptocurrencies is arguably larger than that of the Dow Jones. However, more research is needed in order to justify this conclusion.

In the future we intend to integrate the findings of this research with our other ongoing research, as well as answering some of the peculiarities of the model like why does the previous day price of some cryptocurrencies have the lowest F-score?

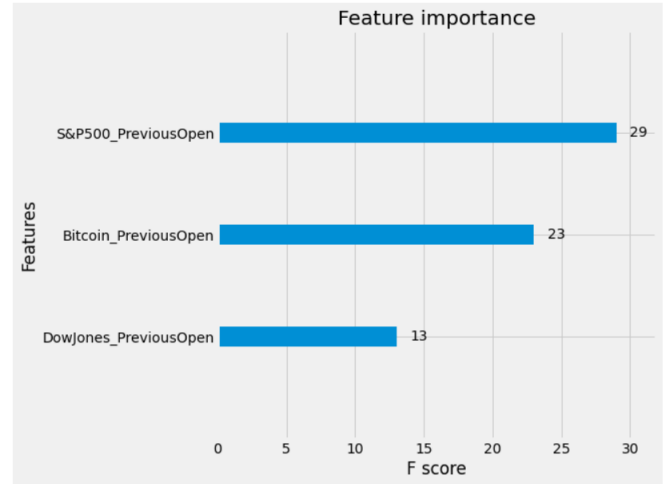


Fig. 1. Feature importance using F-score (Bitcoin)

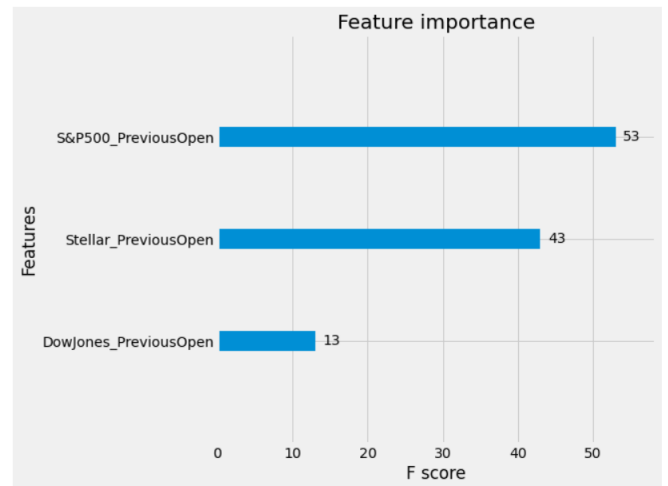


Fig. 2. Feature importance using F-score (Stellar)

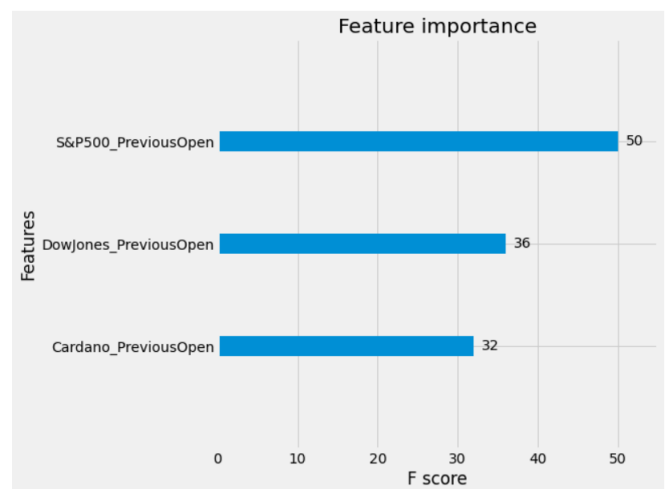


Fig. 3. Feature importance using F-score (Cardano)

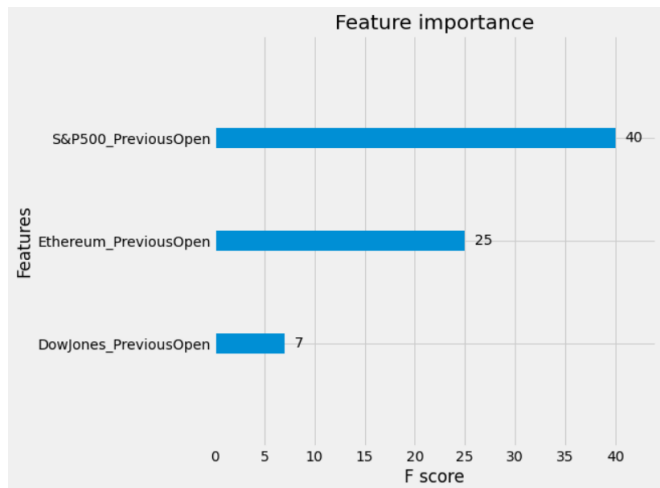


Fig. 4. Feature importance using F-score (Ethereum)

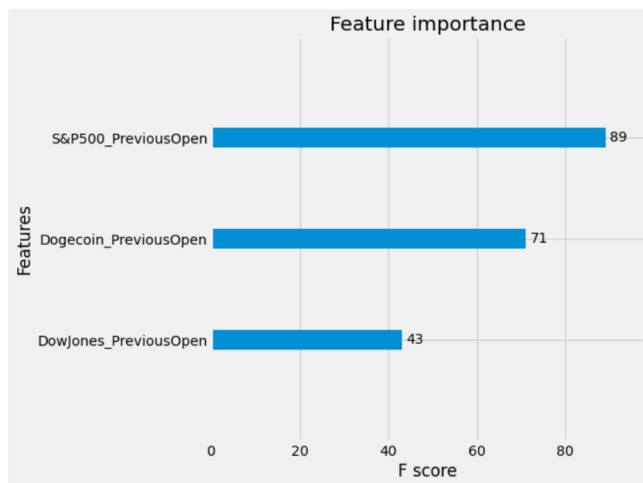


Fig. 7. Feature importance using F-score (Dogecoin)

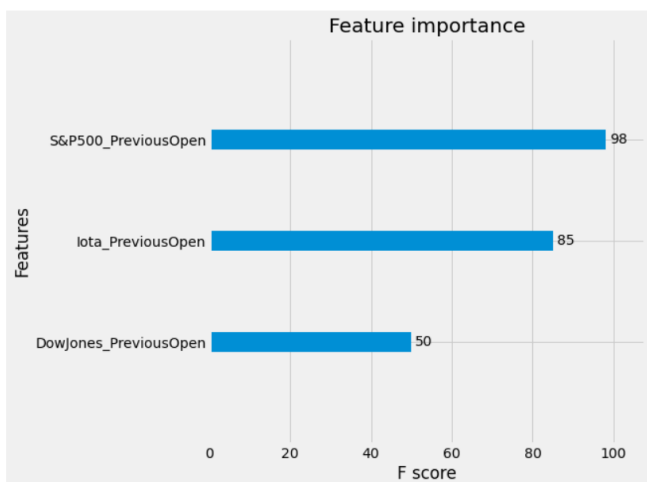


Fig. 5. Feature importance using F-score (Iota)

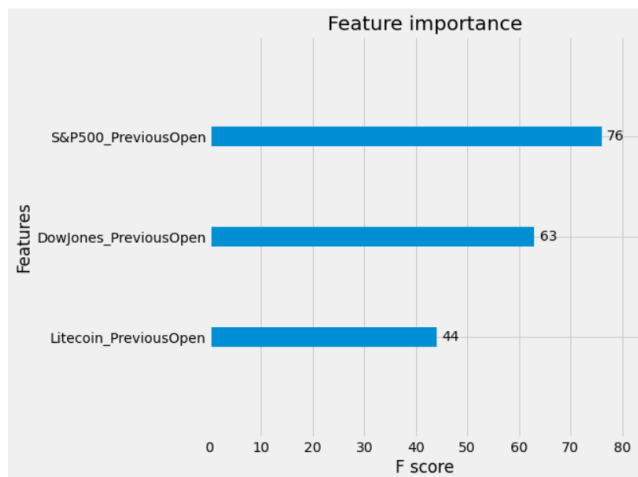


Fig. 8. Feature importance using F-score (Litecoin)

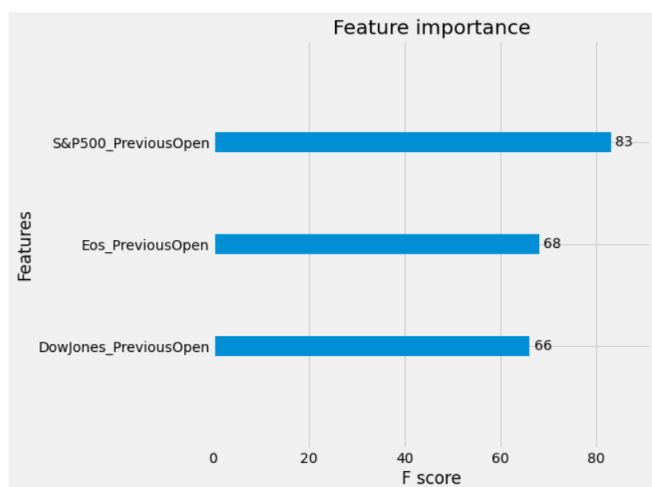


Fig. 6. Feature importance using F-score (Eos)

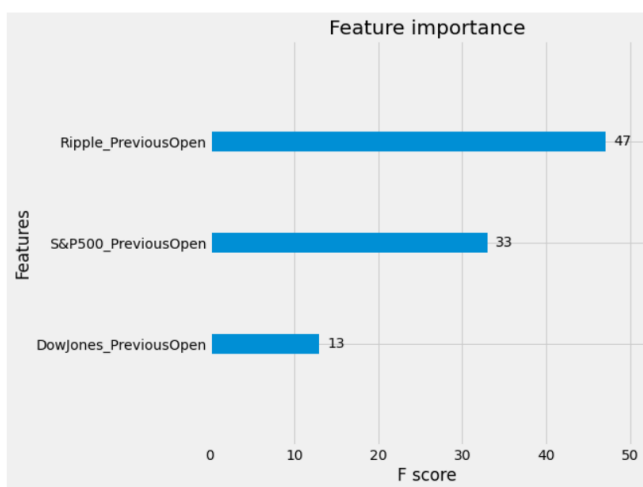


Fig. 9. Feature importance using F-score (Ripple)

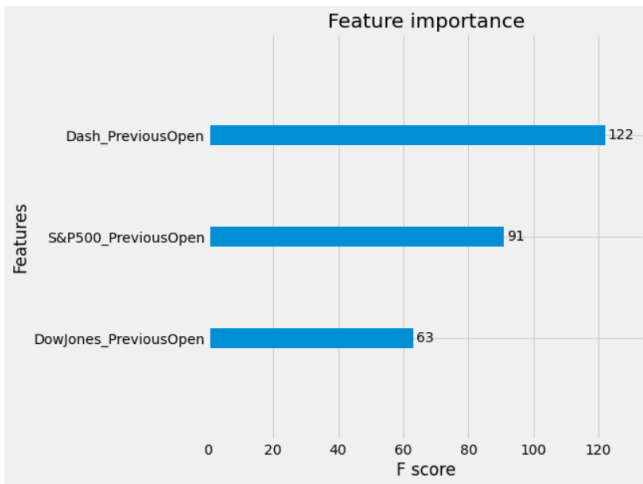


Fig. 10. Feature importance using F-score (Dash)

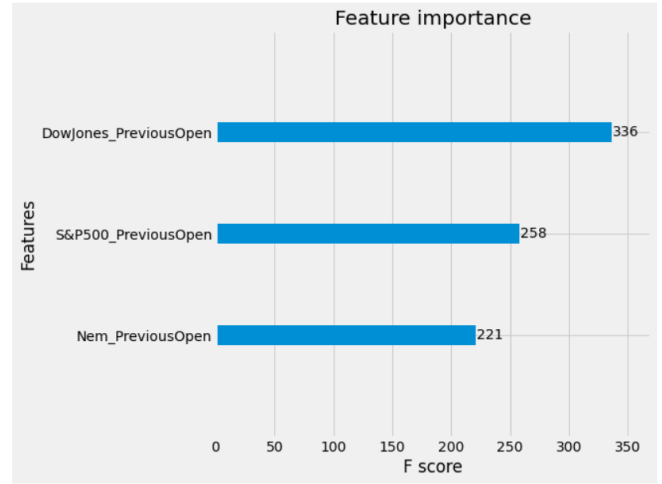


Fig. 13. Feature importance using F-score (Nem)

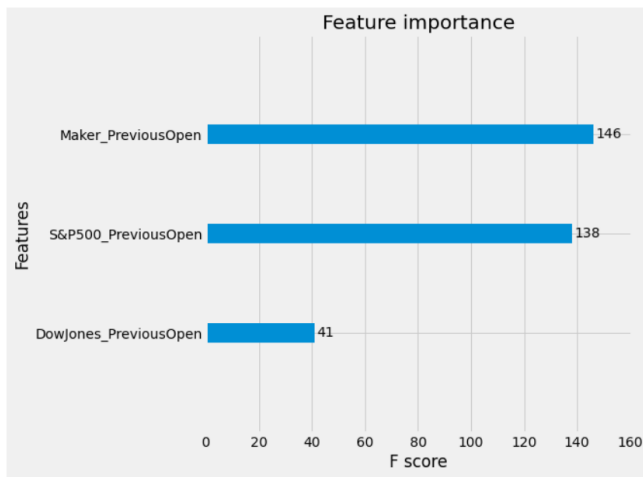


Fig. 11. Feature importance using F-score (Maker)

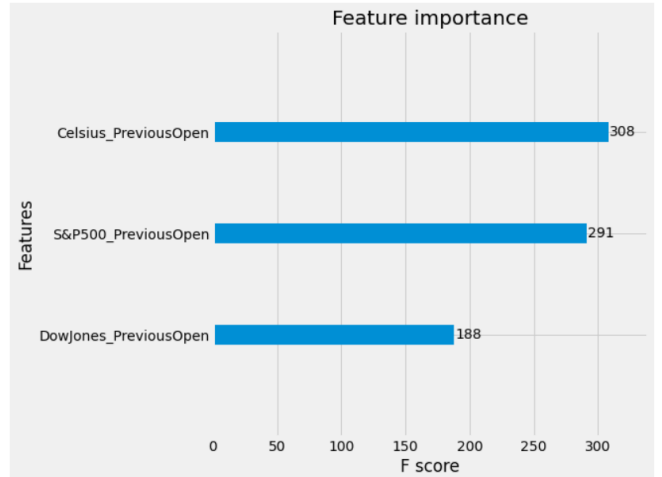


Fig. 14. Feature importance using F-score (Celsius)

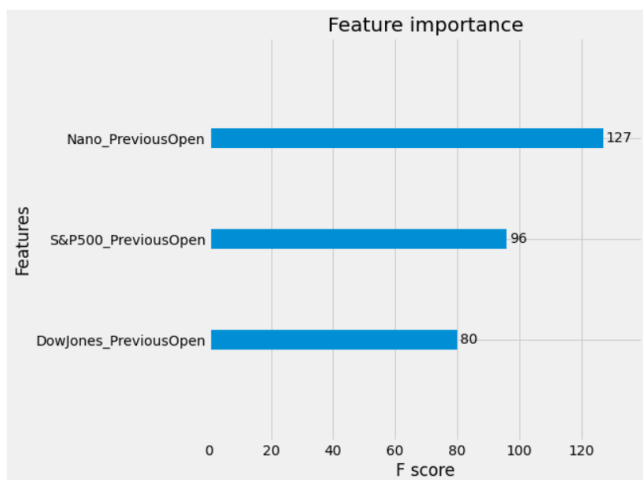


Fig. 12. Feature importance using F-score (Nano)

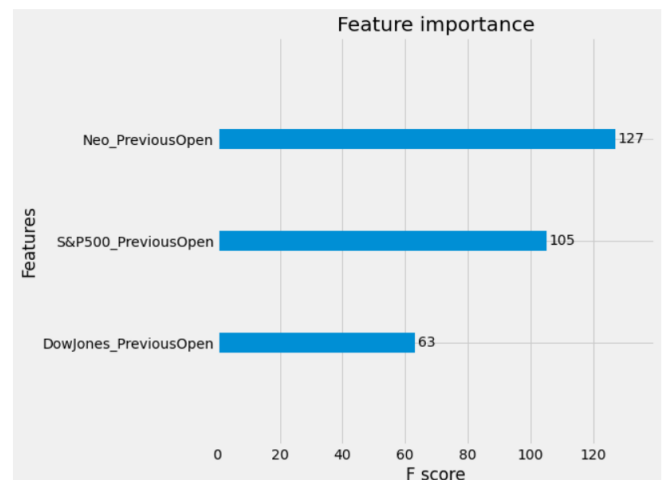


Fig. 15. Feature importance using F-score (Neo)

REFERENCES

- [1] Nakamoto Satoshi, "Bitcoin: A peer-to-peer electronic cash system." Manubot, 2019.
- [2] Chen, Tianqi and Guestrin, Carlos. "Xgboost: A scalable tree boosting system" Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp.785-794, 2016.
- [3] Sokolova, Marina, Nathalie Japkowicz, and Stan Szpakowicz. "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation." Australasian joint conference on artificial intelligence. Springer, Berlin, Heidelberg, 2006.
- [4] Todorovska, Ana, Eva Spirovska, Gorast Angelovski, Hristijan Peshov, Ivan Rusevski, Jovana Marojevikj, Irena Vodenska, Lubomir T. Chitkushev, and Dimitar Trajanov, "Analysis of cryptocurrency interdependencies." In Proceedings of Blockchain in Kyoto 2021 (BCK21) (p. 011004).
- [5] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).
- [6] Jovan Davchev, Kostadin Mishev, Irena Vodenska, Lou Chitkushev, and Dimitar Trajanov. Bitcoin price prediction using transfer learning on financial micro-blogs. In The 16th Annual International Conference on Computer Science and Education in Computer Science, September 5, Bulgaria (virtual), 2020.
- [7] Y. Jiao and J. Jakubowicz, "Predicting stock movement direction with machine learning: An extensive study on S&P 500 stocks," 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 4705-4713, doi: 10.1109/BigData.2017.8258518.
- [8] Cortez, Alexandra, and Ana Tulcanaza. "Bitcoin: its influence on the global World and its relationship with the stock exchange." Revista Chakiñan de Ciencias Sociales y Humanidades 5 (2018): 54-72.

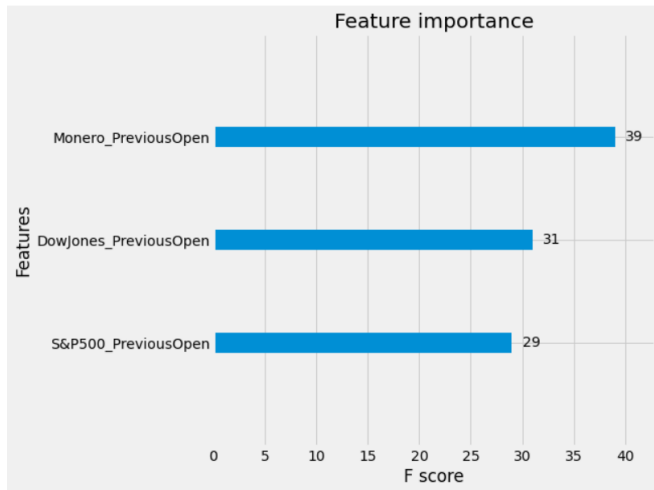


Fig. 16. Feature importance using F-score (Monero)

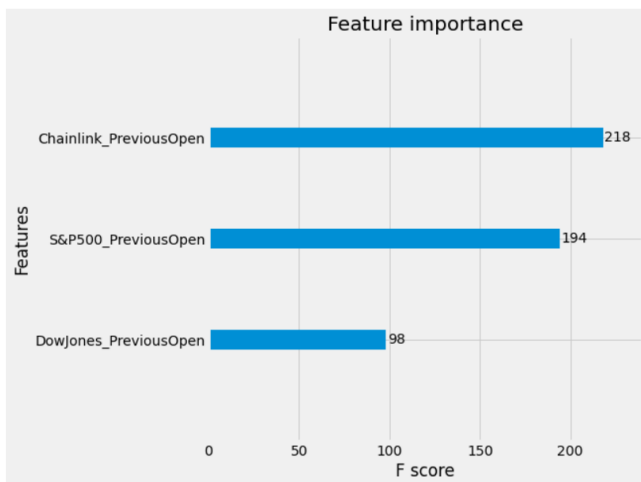


Fig. 17. Feature importance using F-score (Chainlink)

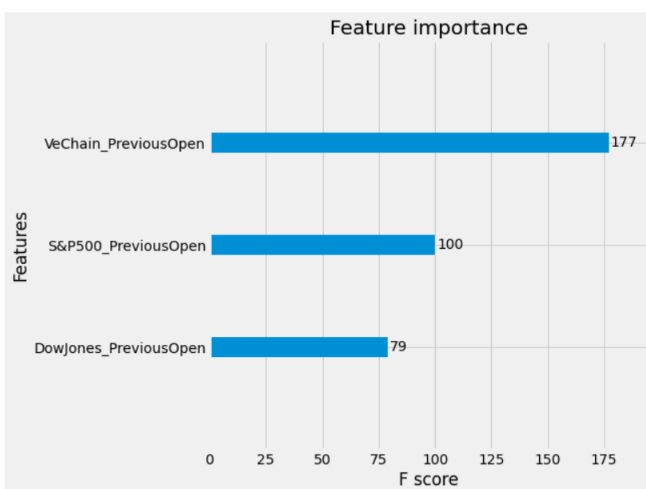


Fig. 18. Feature importance using F-score (VeChain)

Analysis of write amplification on solid state drives in datacenters

Sladana Đurašević

Faculty of technical sciences Čačak
University of Kragujevac
Čačak, Serbia
ORCID: 0000-0002-3598-5781

Vanja Luković

Faculty of technical sciences Čačak
University of Kragujevac
Čačak, Serbia
ORCID: 0000-0002-1887-6102

Uroš Pešović

Faculty of technical sciences Čačak
University of Kragujevac
Čačak, Serbia
ORCID: 0000-0001-8722-6544

Borislav Đorđević

Institute Mihajlo Pupin
Belgrade, Serbia
ORCID: 0000-0002-6145-4490

Abstract — Solid state drives (SSD) becomes the dominant type of secondary memory in personal computer systems, thanks to their superior read and write performance when compared to traditional hard disk drives (HDD). Due to the limitations of NAND flash memory technology SSD has several disadvantages that make them unfavorable for permanent data storage in data centers. One of the critical factors in SSD operation is write amplification which introduces additional system writes that must be performed for each user write, thus increasing cell wear and reducing drive performance. In this paper, we analyzed the write amplification of SSD drives operating in the Backblaze data center.

Keywords—Solid state drive, NAND FLASH, SMART, Write amplification

I. INTRODUCTION

Data storage becomes one of the fastest rising field in computing due to the exponential increase in the amount of data produced by almost every field of human activity. Starting in 2012, mankind entered so-called the Zettabyte Era, the period when the amount of digital data in the world first exceeded one zettabyte (ZB). According to the Digital Age 2025 study, it is expected that by the end of 2025 global amount of data will exceed 175 ZB. Digital data can be stored on different storage mediums such as hard disk drives, magnetic tape, optical medium, or semiconductor flash medium. Today most dominant mediums for data storage are hard disk drives and flash, which constitute almost 85% of global storage share, 60% share for hard disk drives and 25% share for flash storage. It is expected that by 2025 share of the flash storage will continue to increase to 35% while the share of HDD will decrease to 50% [1].

Hard disk drives are used for computer storage for more than 65 years thanks to the constant evolution of HDD technology in improving the storage density. The basic principle of HDD electromechanical design has not changed thus it represents a major drawback in further increasing the HDD performance in terms of read and write speeds. Flash technology appeared in the late eighties and was primarily used for storing small quantities of data in embedded systems. Thanks to the improvements in miniaturization, flash technology became competitive with hard disk drives in terms of data density and read and write performance. SSDs are capable to achieve much faster read and write speeds when compared to HDDs, due to the lack of moving parts thus SSDs are becoming the dominant type of secondary memory in personal computers. It is to be expected that

SSDs are more durable than hard drives, because they do not have mechanical parts, but it is not the case. Although they seem to be a superior solution for the average user compared to hard drives, SSDs have several disadvantages that make them unfavorable for permanent data storage in data centers that rely on HDDs.

The main SSD disadvantage is write mechanism which requires the movement of data to new memory cells which results in a phenomenon known as write amplification. Due to the limited write endurance of the flash medium, the occurrence of write amplification can significantly reduce SSD lifetime. The main design challenge in SSDs is to ensure equal medium wear while minimizing write amplification. In this paper, we will analyze the write amplification of SSDs operating in the Backblaze [2] data center. The primary factor we observe was the average erase count, which expectedly increased with the amount of data written to drive.

II. NAND FLASH TECHNOLOGY

The essence of the data storage in the flash cell is represented by the amount of electric charge, which is trapped between the layers of insulator inside the cell. Based on the number of discrete charge levels used, several types of flash cells can be used as shown in Fig 1. In the case of the SLC (Single Level Cell), a cell single binary bit is represented as the charged state for logical zero or the uncharged state as a logical one [3].

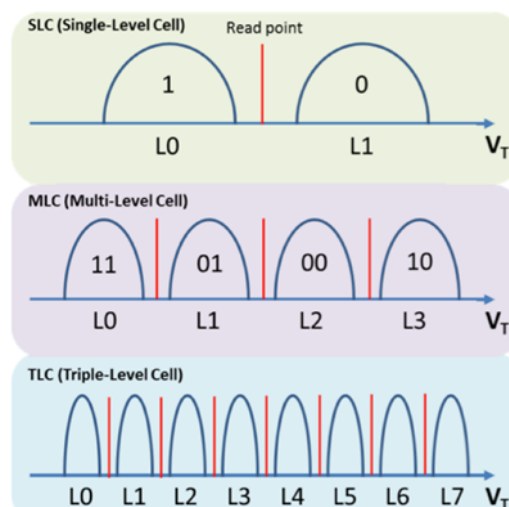


Fig. 1. FLASH cell technologies

This type of cell has low data density per area and could not compete with HDD which provides comparable data densities at fraction of the cost. Therefore, more advanced types of flash cells are in use, such as MLC (Multi-Level Cell) cells that store two bits using four charge levels or TLC (Triple Level Cell) cells that store three bits using eight charge levels. Furthermore, these technologies rely on vertical stacking of flash cells, known as 3D NAND which significantly increases storage density. Currently, QLC (Quad-Level Cell) flash cells, which store 4-bit value per cell, and PLC (Penta-Level Cell) flash cells, which store 5-bit value per cell, are being developed.

Placing more bits per one flash cell increases the capacity of the SSD without increasing the chip size, but also reduces storage reliability due to small differences between charge levels which makes it difficult to write and read bits. Also, a flash memory cell can only be written a certain number of times during its lifetime, with SLC cells allowing 100,000 erase cycles, MLC cells allowing 10,000 erase cycles, and TLC cells about 3,000 erase cycles [4]. Newer QLC and PLC cells allow just from 100 to 1000 erase cycles. Depending on the type of memory cells used in the SSD, the performance and reliability (lifespan) of the SSD are determined. Unlike a hard disk where data can be written to any location at any time and data can be easily overwritten, flash cells in an SSD must first be erased before new content can be written to them. Erase process discharges the flash cells to be able to store new data. Since flash cells have a limited number of erase cycles, it is necessary to ensure that data is cyclically written to all cells to wear them evenly. Flash controller maps logical data addresses to physical addresses on a flash using an FTL (Flash Translation Layer) table, in a process known as wear leveling. When writing data in a logical location, the physical location where the data was located will be erased, and the controller will write the data in a new physical location that has fewer erase cycles than the previous one and will perform mapping of the logical address in the FTL table to a new physical address.

Due to the organization of NAND flash memory, it is not possible to access cells individually to perform read and write operations. A page is the smallest unit of data that can be read or written to memory. The write operation can only be performed on pages that have been previously erased. In case it is necessary to change some data on the page, the content of the page is read to the internal register, the data is updated in register, and the updated version of the page is stored on a free page and this operation is called "read-modify-write". Unlike HDDs, the data is not updated on the spot, because the free page is at a different address from the page that originally contained the data, to ensure even wear. When data is saved on a new page, the original page is marked as obsolete and will remain so until it is deleted. Since an individual page cannot be deleted, the entire block to which the page belongs must be deleted, and the other pages with the correct data are moved to a new free block. By copying data from one block to another block, one lifecycle of that page or block is used [5]. This operation, called the garbage collection, is performed in the background by the SSD controller and is shown in Fig 2. Garbage collection is of great importance for SSDs, as it allows the drive to mitigate the impact of the cycle of erasing and writing data to flash memory. In that case, the SSD controller will have to perform additional data migration when the user wants to write new data, which results in additional data

writing to flash memory, which leads to greater memory wear and shortens the life of the SSD. A measure of this phenomenon is defined by WA (Write Amplification) parameter which represents the ratio between an actual amount of data written to the flash memory and amount of data to be written issued by host [6].

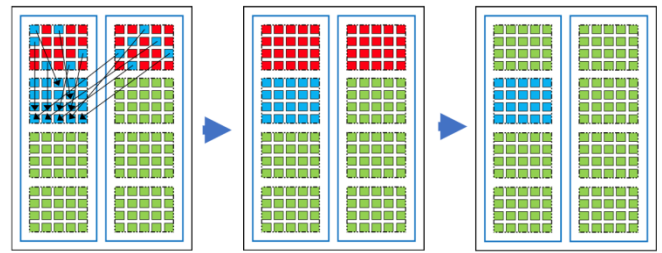


Fig. 2. NAND erase process: left - copy/move valid data (blue pages), middle - invalid blocks (red blocks), right erased block (green blocks).

With a blank SSD, all pages in blocks are free and data can be written to them immediately, so in this case the WA parameter is equal to one, which is the ideal value. However, as the SSD fills, there are fewer and fewer blank pages, and the controller must free up space in blocks, which affects the appearance of increased writing. The higher the WA parameter, the more additional data will be written to flash memory leading to increased flash memory wear, which will reduce SSD life. Also, movement of additional data will take up the bandwidth of flash memory, which affects the performance of the SSD.

The main design challenge in SSDs is focused on efficient flash controllers, which can improve SSD performance and extend the service life. There are many ways how flash controller can reduce WA, such as data compression, over-provisioning and use of sequential writing. Data compression requires high performance hardware, while over-provisioning limits the available free space.

III. RESULTS

The SSD disks analyzed in this paper were used as system disks on the Backblaze data storage servers to store the server operating system. Storage servers contain up to 60 HDDs that permanently store user data, while the one SSD. is used to boot the storage server operating system, and to store log files and temporary files produced by the storage server. Each day, the SSD will read, write and delete files depending on the activity of the storage server itself. SSDs began to be used in the Backblaze data center from 2018, and by the end of 2021, the number of SSDs reached 2,200 on data warehouse servers. In this paper, we analyze two models of SSDs from Seagate, which are the most common in the data center ZA2000CM10002 and ZA250CM10003.

TABLE I. SSD MODELS USED IN RESEARCH

Model	ZA250CM10002	ZA250CM10003
Capacity	250 GB	250 GB
No. drives	562	1090
Average age (months)	21.7	11.1
Operating days	204 287	276 281
Failed drives	2	8
Annular failure rate	0.36 %	1.06 %

The Backblaze records S.M.A.R.T. attributes for each HDD and SSD every day and publishes this data quarterly in form of the open data set [5]. S.M.A.R.T. is an abbreviation for Self-Monitoring, Analysis, and Reporting Technology and was developed to report on various indicators related to the reliability of HDD drives to predict failures. This technology is also applied to SSDs, but due to differences in technology, different parameters are available. The monitored parameters of the SSD include, operating time and number of power-on cycles, temperature, the degree of wear of the cells, the total amount of data read and written to SSD. The SMART parameters available for monitoring vary from manufacturer to manufacturer, so in addition to the above, there may be parameters that are specific to a particular disk model.

The Backblaze data set of SMART parameters is published quarterly in the form of csv files for each calendar day. Each line of the csv file contains a set SMART parameter for a specific model of SSD or HDD and disk serial number. This data was imported into the SQL database using SQLite3 after which the several SMART parameters for the target disk model were extracted using the appropriate SQL query.

The first parameter, SMART 241, is the total amount of data written by the host in GB and it represents the load under which SSD is subjected during its lifetime. In the first part of the research, the influence of host writing on the number of erasure cycles for both drive models is being analyzed. SMART 173 parameter contains an average, minimum and maximum count of erased SSD blocks for that particular day. Using this parameter, it is possible to predict the remaining life of the disk where the manufacturer specifies a lifetime of at least 1500 erase cycles. Obtained results for SSD model ZA250CM10002, are shown in Fig 3. as gray lines for every drive model, where the mean value for the number of erase cycles for all drives is shown by the blue solid line, while blue dashed lines represent one standard deviation from the mean value.

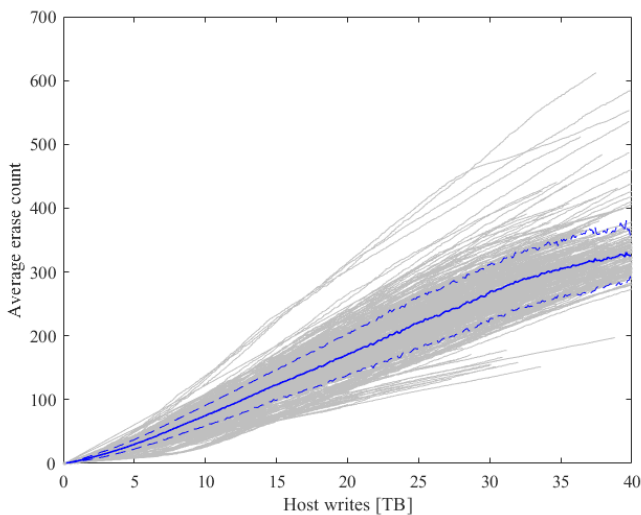


Fig. 3. Average erase count for ZA250CM10002 SSD

Results show that the average erase count is equal to 300, after 40 TB of data are written to SSD, this drive model has a wear level of around 20%. Given that these drives on average operate for almost two years, their useful life will be almost ten years. Obtained results for SSD model ZA250CM10003,

are shown in Fig 4. and this drive has a much higher number of erased cycles from the previous drive model, 600 erased cycles after 30 TB of written data. This reveals that the wear level for drive model ZA250CM10003 is equal to 40% and since these drives on average operate for one year, their useful life will be just two and a half years, four times shorter than for previous model ZA250CM10002.

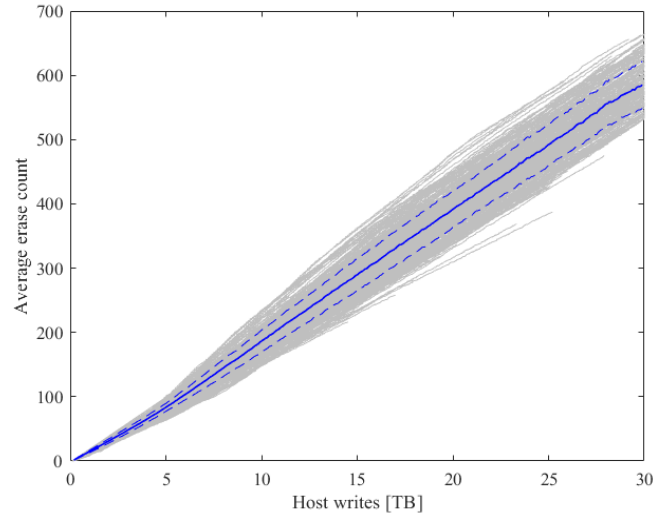


Fig. 4. Average erase count for ZA250CM10003 SSD

The SMART 233 parameter represents the total amount of data written to disk in GB. The ratio between this parameter and total host writes represents the WA parameter as shown by Eq. 1.

$$WA = \frac{SMART\ 233}{SMART\ 241} \quad (1)$$

In the second part of research influence of host writes on WA is analyzed for both drive models. Results are shown in Fig 7. and Fig 8. as gray lines for every drive model, where the mean value for the number of erase cycles for all drives is shown by the blue solid line, while blue dashed lines represent one standard deviation from the mean value. In case of the ZA250CM10002 model, result shown in Fig 5. suggest that this drive might use compression since some of the drives have WA factor lower than one. Later WA factor increases to a stable value which is less than two.

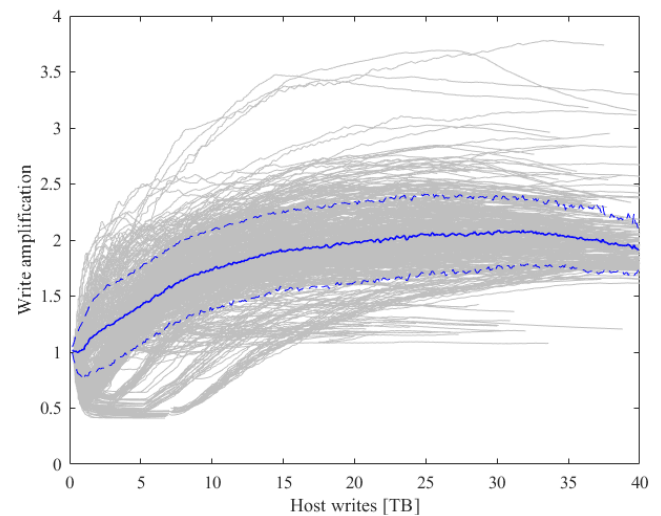


Fig. 5. Write amplification for ZA250CM10002 SSD

In case of the ZA250CM10003 model, result shown in Fig 6. suggest that this drive does not use compression like previous drive model and WA factor increases to a stable value which is around two and a half.

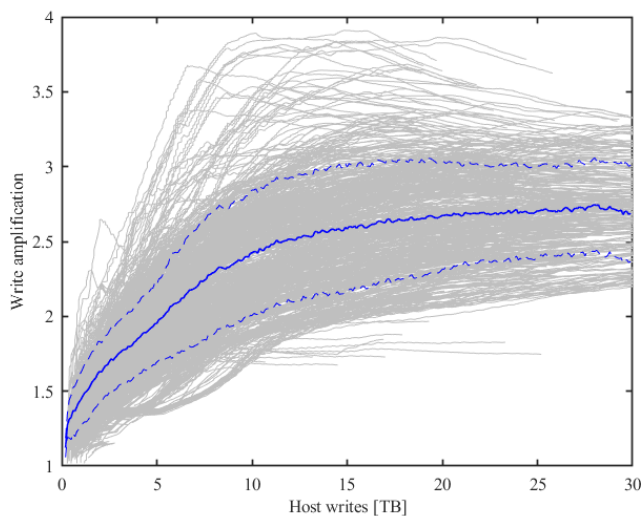


Fig. 6. Write amplification for ZA250CM10003 SSD

IV. CONCLUSION

In this research, we performed the analysis of the write amplification on the SSD drive operating in the Backblaze data center. The result showed that for two similar SSD models from the same manufacturer there is a significant difference in write amplification which suggest that the latter SSD model will probably reach the end of its service life much sooner than the older model. The main conclusion is that the drive ZA250CM10002 might use compression since some of the drives have WA lower than one at the beginning of their lifetime. Further research will be focused on the

analysis of data which will be published by the Backblaze in following year to check these claims.

ACKNOWLEDGMENT

The research in this paper is part of the project 451-03-68/2022-14/200132 funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia.

REFERENCES

- [1] D. Reinsel, J. Gantz, J. Rydning, "Data Age 2025: The Evolution of Data to Life-Critical Don't Focus on Big Data; Focus on the Data That's Big", IDC White Paper sponsored by Seagate, 2017, Framingham, MA, USA
- [2] A. Klein, The SSD Edition: 2021 Drive Stats, <https://www.backblaze.com/blog/ssd-edition-2021-drive-stats-review/>, accessed 10th April 2022.
- [3] Y. Cai, S. Ghose, E. Haratsch, Y. Luo, O. Mutlu, Errors in Flash-Memory-Based Solid-State Drives: Analysis, Mitigation, and Recovery, DOI: <https://doi.org/10.48550/arXiv.1711.11427>
- [4] M. K. Jibbe, B Chan, Analysis of SSD health and Prediction of SSD life, Storage Developer Conference, September 21, 2016, Santa Clara, USA
- [5] F. Chen, D. Koufaty, X. Zhang, Understanding Intrinsic Characteristics and System Implications of Flash Memory based Solid State Drives, SIGMETRICS/Performance'09, June 15–19, 2009, Seattle, WA, USA.
- [6] X. Hu, E. Eleftheriou, R. Haas, I. Iliadis, R. Pletka, Write Amplification Analysis in Flash-Based Solid State Drives, SYSTOR '09: Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference, Haifa, Israel, May 2009, pp. 1–9, <https://doi.org/10.1145/1534530.1534544>
- [7] BackBlaze SMART data set, <https://www.backblaze.com/b2/hard-drive-test-data.html>, accessed 10th April 2022.

Electricity price forecasting for the day-ahead market in Croatia

Kristijan Koteski

Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Skopje, Macedonia
kristijan.koteski@students.finki.ukim.mk

Aleksandra Dedinec

Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University in Skopje
Skopje, Macedonia
aleksandra.kanevche@finki.ukim.mk

Abstract—This paper presents a Machine Learning model, specifically a Neural Network, to forecast the prices of the Croatian day ahead power exchange. The price forecasting is of great importance for building and maintaining the smart energy grids that are a part of Smart Cities around the globe. Knowing the price of electricity can make a switch to renewable energy sources easier to predict and maintain so that it is available to the citizens of the city.

The day ahead market requests bids for the next day for all of the 24 hours in that day. From all of the participants' bids, a price is calculated where supply meets the demand and this is the price that the models try to predict without using the bids from the participants. Predicting the price of electricity for markets that produce a big part of its electricity from renewable sources can be unpredictable and cause a lot of problems for researchers. Croatia is trying to increase production of electricity from renewable sources in the following years, which will make the prices more unpredictable.

The data used is collected for the period from 11.02.2016 to 24.08.2021. Multiple Neural Network structures were explored, but the best one was a Neural Network with 2 hidden layers with 11 neurons each. The testing period was the last 20 % of the data as the training and testing data was split 80:20 and some years were dropped for better results.

The results for the model were evaluated using multiple evaluation metrics, most of which being error rates.

Keywords—Artificial Intelligence, Machine Learning, Neural Networks, time-series forecasting, electricity price forecasting, day-ahead market forecasting for Croatia, Smart City, Smart energy grids

I. INTRODUCTION

Forecasting the prices of the day-ahead market is a field of research that holds great importance for the maintenance and implementation of smart energy grids as a key part of Smart Cities around the globe. The predicted prices can help an electricity company optimize its bidding strategy and its own production or consumption schedule in order to reduce the risk or maximize the profits in day-ahead trading [1]. For this to be effective, a decent accuracy has to be achieved so the company can have some confidence in the predicted prices and act accordingly. Unfortunately, the prices can be difficult to predict due to various unforeseen circumstances, especially renewable energy sources [3-6]. As the world is trying to switch to renewable energy sources to achieve The Sustainable Development Goals (SDGs), the need to accurately predict the prices on the day-ahead electricity market becomes an important task.

The problem is formally presented as predicting the prices for the next day or day $d+1$. The bids to calculate the prices are placed on day d and the prices are calculated sometime on the day d , before the next day or day $d+1$. The prices are calculated for all of the 24 hours of the day $d+1$. The forecasting is done for a specific hour the next day, and it requires attributes that are available on day d , so that the prices can be predicted successfully. This problem is presented as a time-series problem where the next item in the series is predicted for a future time. Each item in the series is an hour of the day and the price is measured in Euros per Mega Watt Hour (€/MWh).

This research explored many different Machine Learning models to predict the prices, but the best one seemed to be a Neural Network with 2 layers with 11 neurons each. Deep Learning was also explored, but it didn't produce the same results as Neural Networks with simpler structures. In the results section, there are more details on the models explored and future research directions.

Before we explain how the paper is structured, there is important information to state about this problem that will help the reader to understand the problem into more detail. First of all, the Croatian day-ahead electricity market operated by CROPEX was coupled with the Slovenian day-ahead electricity market operated by BSP South Pool on June 19th 2018. Another information worth to note is that Croatia is trying to increase production of electric energy from renewable sources [8].

The paper is structured as such. First, the introduction presents the problem and important information for future sections. Then, the methodology explains the methodology used in short, because Neural Networks have been explored in thousands of papers before and it would be redundant to go into excessive detail. The third part has interesting visualizations of the data used. The fourth part goes into the data used and the model structure. The fifth part discusses the results of the models tested and touches over what future work can be conducted on this topic.

II. METHODOLOGY

The methodology used for forecasting the prices has been used many times by countless of research papers. This is why I will only touch briefly on the methodology used.

The methodology used for forecasting the prices is the Feed Forward Neural Network, specifically the Multi-layer feed-forward Perceptron (MLP).

The neural networks that are used in this paper are represented as a directed acyclic graph, which has several layers - an input layer, 2 hidden layers and an output layer, which makes the structure a not deep Neural Network. The connection between the input x and the output y are presented with the following equation:

$$y = \sum_{j=0}^h \left\{ w_j x f \left(\sum_{i=0}^d w_{ij} x_i \right) \right\} \quad (1)$$

Where w_j and w_{ij} represent the weight and biases that connect the layers.

Supervised Learning is used and the optimization algorithm used is 'Adam' as presented in [9]. The loss function used for the optimizer was the Mean Absolute Error (MAE) for a regression problem.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2)$$

Where n is the number of observations, y_i is the actual value, and $\hat{y}_i$ is the predicted value.

For evaluating the model's performance, multiple evaluation metrics were used. One of them is the MAE metric represented with (2). The other evaluation metrics are Mean-Squared Error (MSE), Root Mean-Squared Error (RMSE), and Adjusted R2 Score. MSE is represented with this formula:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3)$$

The RMSE is represented with this formula, which is just a square root of the MSE:

$$RMSE = \sqrt{MSE} \quad (4)$$

Finally, the formula for the Adjusted R2 Score is given with this formula:

$$Adjusted\ R2 = 1 - \frac{(1-R^2)(N-1)}{N-p-1} \quad (5)$$

where N is the number of samples or observations, R^2 is the R squared score for the observations, and p is the number of predictors, which in our case is 1.

III. VISUALIZATIONS OF THE DATA

In this section, some interesting data visualizations are presented. The purpose of this section is to describe the data and to explain some facts about the data that will be of use in the later sections and to better understand the problem at hand.

In Fig. 1 below, we can see that the prices range from 300 to around -250 €/MWh, however, this is because there are 2 peaks or outliers in the data that need to be removed so they don't affect the models. This is the reason they are called Old prices. The outliers were replaced by the value above itself, so that the time-series isn't affected too much.



Fig. 1. Old graph for all of the prices in the dataset from 11.02.2016 to 24.08.2021. The x axis presents the time and the y axis presents the prices of electricity measured in €/MWh.

In Fig. 2 the new prices are shown with the 2 outliers removed around the start of 2017 and the start of 2021, positive and negative values respectively. The new graph shows that the prices range from 250 to around -70 €/MWh. We can also see a lot of spikes of the prices and some zero prices and negative prices as well.

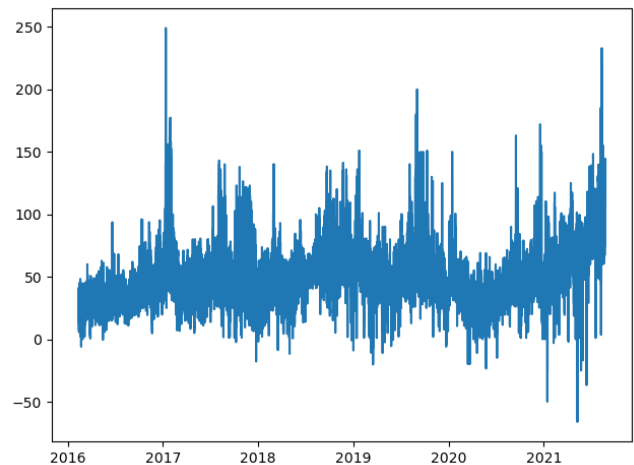


Fig. 2. New prices from 11.02.2016 to 24.08.2021 without the 2 outliers from the previous graph. The x axis represents the time and the y axis represents the price of electricity in €/MWh.

Fig. 3 shows the mean price for every day of the week from Monday to Sunday, represented as numbers from 1 to 7 respectively. We can see a drop in the price as the weekend

approaches. The price seems to be the lowest on Sunday and the highest on Wednesday, right in the middle of the week.

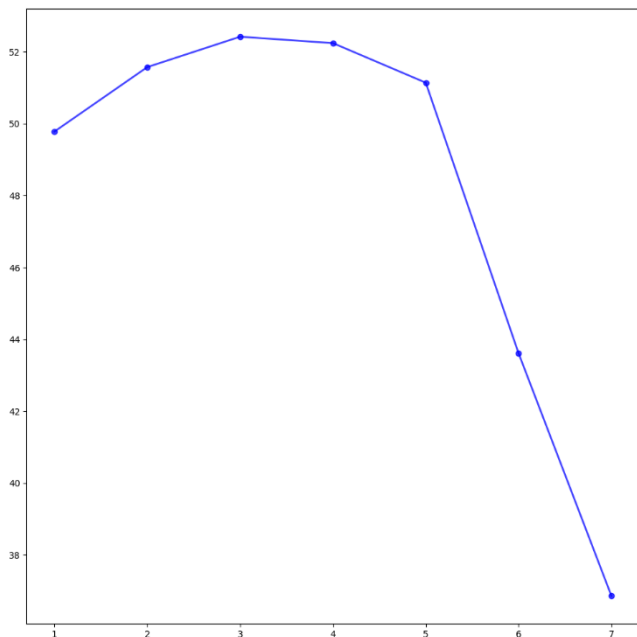


Fig. 3. Mean of the prices for every day of the week from Monday to Sunday, represented as numbers from 1 to 7 respectively. The x axis presents the day of the week from 1 to 7 and the y axis presents the mean price in €/MWh.

In Fig. 4 the mean price for every month is presented. The months are numbered from 1 to 12 presenting the months from January to December respectively. We can see in Fig. 4 that the prices tend to be higher in the middle of winter and in the middle of summer, when people use electricity to heat or cool their homes. More and more people use electricity for heating or cooling, which makes the prices higher due to the increase in demand.

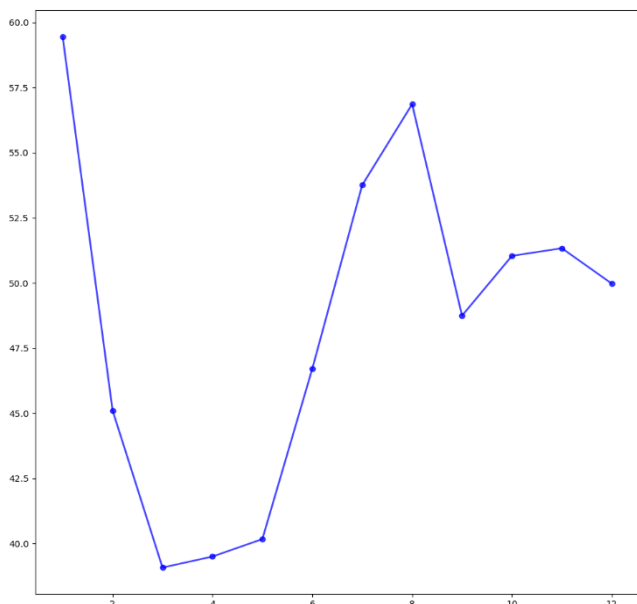


Fig. 4. Mean of the prices for every month of the year from 1 to 12 for January to December respectively. The x axis presents the months from 1 to 12 and the y axis presents the mean of the prices calculated for every month in the dataset.

The last graph, Fig. 5, shows the volume of electricity generated with respect to time. The volume of electricity

produced is presented with Megawatt hour (MWh). We can see a sharp increase in the volume produced from around the middle of 2018. This is due to the aforementioned coupling of Croatia's day-ahead electricity market and Slovenia's day-ahead electricity markets on June 19th 2018. We can see that the volume produced (Fig. 5) and the prices of electric energy don't have much similarities and it shows in the results later that the volume of electricity produced isn't a very good predictor or attribute in the Neural Network model.

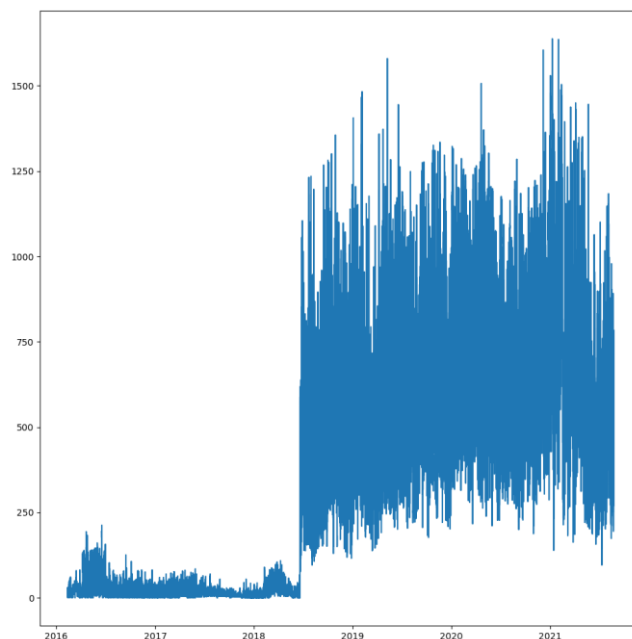


Fig. 5. Volume of electric energy produced from 11.02.2016 to 24.08.2021. The x axis presents the time from 2016 to 2021 and the y axis presents the volume of electricity produced in MWh.

IV. INPUT DATA FOR THE NEURAL NETWORK

The input data was collected from the official CROPEX website where they post the prices and volumes of electricity generated for every hour from 11.02.2016 to 24.08.2021 [7]. The latter is decided to be August 24th, because the research was conducted in August 2021 and the data was collected up until that point. There is more data to collect for the year 2021 and 2022 to see how the prices were affected from the COVID19 pandemic, if at all.

Additionally, weather data was also collected from the website Weather2Umbrella [10]. The data collected was the wind strength, weather description, and temperature for the city of Zagreb. The problem with this data was its quality, because it had a lot of missing values and incorrect values. This data throttled the potential of the model and it is suspected that better quality weather data would lead to better performances, especially because Croatia is switching to more renewable energy sources.

The last data source was the website Entso-E Transparency program, which had data on the Hourly Load Forecast for the country of Croatia [11]. The data was collected for the same time period from 11.02.2021 to 24.08.2021.

From all of these data sources, these were the attributes or features that were created to forecast the electricity prices:

- Price - The target variable y that is measured in €/MWh.

- Price of day 1 – 7 before – The prices for the same hour, but 1 to 7 days before. There are 7 features for the price for every day from 1 to 7.
- Volume of day 1 – 7 before – The same as above, but for the volumes of electricity produced.
- Average price yesterday – The average price from the day before or day d.
- Average volume yesterday – Same as above, but only for the volumes of electricity produced on day d.
- Hour – The hour in the day from 0 to 23.
- Day – The day in the month from 1 to 31.
- Day of the week – The day of the week from 1 to 7 for every day of the week respectively.
- Month - The month from 1 to 12 for every month of the year.
- Year – The year from 2016 to 2021.
- Hourly Load Forecast – The hourly forecast for the load for every hour on day d+1. This data column is generated and is measured in Megawatts (MW).
- Wind – Wind gust or the wind strength for that hour.
- Weather description – The weather description such as “Sunny” or “Cloudy” etc.
- Temperature – The forecasted temperature on day d+1 measured in degrees Celsius (°C).
- Maximum and minimum temperature – The forecasted maximum and minimum temperature of day d+1 measured in °C.
- Holiday – True or False if the date is a national holiday in Croatia.

Some of the collected features were recommended by Jesus Lago et al in [2]. Of these input variables, only 14 were used and found to be useful for the model’s predictions. These features were the Hour, Day, Day of the week, Month, Year, Holiday, Price of day 1-2 day before, Price of day 6-7 before, Hourly Load Forecast, Maximum and Minimum Temperature, and Weather description.

Not all of the data was used for training and not all of the available data for training was used. The data for the year 2021 was not used, because it caused overfitting in many of the experimented models. The data from the year 2021 is most likely different from the other years, because of the rise of energy prices that started around May 2021 [14]. With the sudden rise of energy prices in Europe and the war that started in 2022, it is expected that the prices in the years 2021 and 2022 are different from previous years. Another model to accommodate the new changes in the prices can be created. It would make the prices of the electricity more predictable in irregular periods. This model was shown to not accommodate the changes that happened in the year 2021 as it was trained on years without the irregularities mentioned above.

The data was split 80:20 and exactly 1 year was used for the testing data. The testing year was the year 2020. When

using the year 2021, the model trained on part of 2020 where it suffered from overfitting and achieved worse results seen from the error rates (MSE, RMSE, and MAE). The training data was the data from 11.02.2016 to 31.12.2019, while the testing data ranged from 01.01.2020 to 31.12.2020. At least one year was used for testing purposes and MAE was used as an evaluation metric as suggested by Jesus Lago et al in [2].

The model structure was a Neural Network with 14 input neurons in the input layer, 2 hidden layers with 11 neurons each, and 1 neuron in the output layer. The model structures that were experimented with were pretty simple and other models such as Linear Regression achieved good results as seen from the Adjusted R2 score of 0.79. The problem with the Linear Regression model was the fact that it performed poorly as seen by the error metrics (MSE, RMSE, and MAE). Very Deep Neural Network structures were not explored due to computational limits, and only some shallow Deep Neural Networks were explored with about 10 layers with 200 neurons each.

V. RESULTS AND DISCUSSION

The results of the model were evaluated with multiple evaluation metrics to see the performance of the models clearly. For this model, MSE, RMSE, MAE and Adjusted R2 Score were used as evaluation metrics on the testing dataset.

The results for the Neural Network with 2 hidden layers with 11 neurons each, were:

- MSE 78.41822
- RMSE 8.855406
- AR2 0.721466
- MAE 5.83894.

Unfortunately, MAPE couldn’t be used as an evaluation metric, because there were multiple zero values or close to zero values, which make the MAPE value very high since it is dividing with zero values or close to zero values. Using MAPE, we could compare this model with the model proposed by Aleksandra et al in [12] where the model used data from only the year 2016.

The main factor for determining the success of the model was the amount of data used for training. If less data is used for training, then that significantly lowers the performances of the model and this is very impactful, because not much data has been collected in Croatia around electric energy. We suspect that as Croatia collects more data around electric energy production that could benefit researchers trying to build Machine Learning models to benefit Smart Cities in Croatia.

The weather data didn’t have the best quality which throttled the model. We suspect that better quality data around weather conditions in Croatia, specifically for the weather conditions around power plants that collect energy from renewable sources, would be of great use. As Croatia is switching to more renewable energy sources, we suspect that it could be beneficial for the country to collect as much data as possible as soon as possible to make prices from renewable energy sources more predictable in the long run. Data is a fundamental part of Smart Cities and collecting more data can benefit researchers that are trying to create Machine Learning models.

As previously mentioned, very Deep Neural Networks weren't explored out of computational limitations. These models were very good at predicting the spikes in the price of electric energy, but unfortunately had a higher error rate. When experimenting with different model structures, it seems that simpler structures are favored and they achieve better performances than more complex Neural Network structures. Future work can be done using very Deep architectures for forecasting the prices for the day-ahead electricity market in Croatia.

Another model that wasn't explored was the LSTM proposed in [13] that is made to predict time-series. Future work can be done using this type of model to forecast the prices on the day-ahead electricity market in Croatia.

In Fig. 6, a comparison of the predicted and actual values is shown. As we can see from Fig. 6, the predicted value fits very well under the actual values.

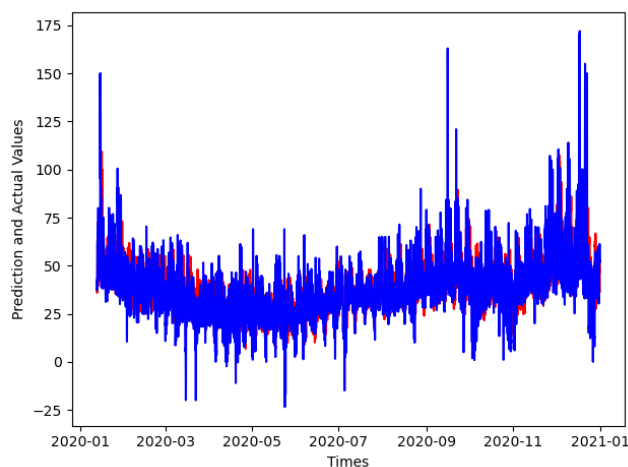


Fig. 6. Comparison between the predicted values from the model and the actual values. The x axis is the time period from the test values and the y axis is the price of electric energy in €/MWh. The blue line is the actual value and the red line represents the predicted values.

Another look at the results is shown on Fig. 7, Fig. 8, and Fig. 9. The graphs show the mean of the actual electricity prices compared with the model's predicted price means for every hour, weekday, and month respectively.

We can see that the model fits all of the lines very well. The only noticeable exception can be found on Fig. 7 for the per hour comparison. We can see that the model makes the biggest mistake from around 7 to about 17, when the people are usually away at work.

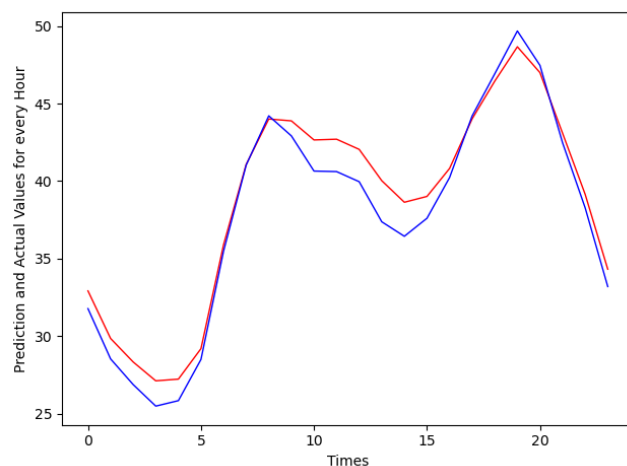


Fig. 7. Comparison between the actual electricity prices (blue line) and model predicted prices (red line) for every hour. The x axis presents the hours from 0 to 23 and the y axis presents the mean of the prices for all of the hours.

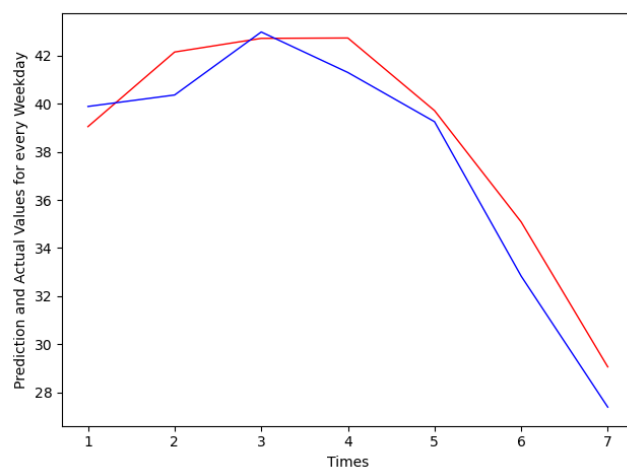


Fig. 8. Comparison between the actual electricity prices (blue line) and model predicted prices (red line) for every weekday. The x axis presents the weekdays from 0 to 7 for Monday to Sunday and the y axis presents the mean of the prices for all of the weekdays.

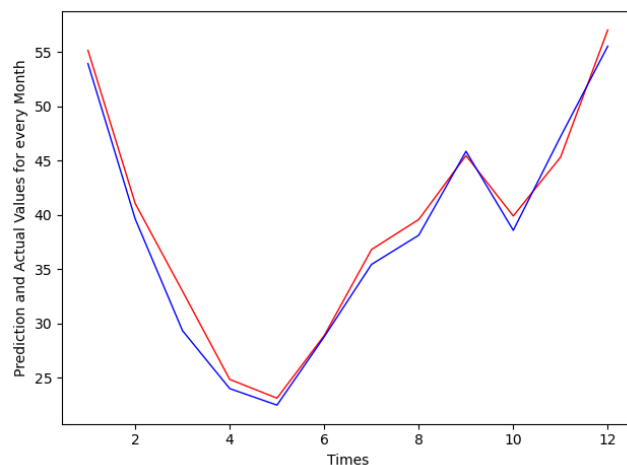


Fig. 9. Comparison between the actual electricity prices (blue line) and model predicted prices (red line) for every month. The x axis presents the months from 1 to 12 for January to December and the y axis presents the mean of the prices for all of the months.

The last thing to note is the fact that the volume of produced electric energy gave the model worse performances and it wasn't of any use at all. This is seen in Fig. 5 where the volume of electric energy produced doesn't seem to have any correlation visually with the prices of electric energy.

VII. CONCLUSION

In this paper we discussed an approach to forecast the electricity prices on the day-ahead electricity market in Croatia. The data used was from the period of 11.02.2016 to 24.08.2021 and the attributes for the model were features that can help in describing the demand and production of electricity, which in turn, determines the price of electricity. Predicting the electricity prices becomes an important task to predict less predictable sources such as renewable energy sources, which is why the model was built. The results are favorable, however, they cannot be compared to previous models on this topic, because the metrics to evaluate the models differ and the prices have zeros, making it impossible to calculate an accurate MAPE as in the work by Aleksandra et al [12].

ACKNOWLEDGMENT

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius" University.

REFERENCES

- [1] R. Weron, "Electricity price forecasting: A review of the state-of-the-art with a look into the future," *International Journal of Forecasting*, vol. 30-4, pp. 1030-1081, 2014.
- [2] Jesus Lago, Grzegorz Marcjasz, Bart De Schutter, Rafał Weron, "Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark", *Applied Energy*, Volume 293, 2021, 116983, ISSN 0306-2619, <https://doi.org/10.1016/j.apenergy.2021.116983>.
- [3] Brancucci Martinez-Anido C, Brinkman G, Hodge B-M. The impact of wind power on electricity prices. *Renew Energy* 2016;94:474–87. <http://dx.doi.org/10.1016/j.renene.2016.03.053>.
- [4] Gianfreda A, Parisio L, Pelagatti M. The impact of RES in the Italian day-ahead and balancing markets. *Energy J* 2016;37:161–84. <http://dx.doi.org/10.5547/01956574.37.si2.agia>.
- [5] Grossi L, Nan F. Robust forecasting of electricity prices: Simulations, models and the impact of renewable sources. *Technol Forecast Soc Change*, 2019, 141:305-18. <http://dx.doi.org/10.1016/j.techfore.2019.01.006>.
- [6] Maciejowska K. Assessing the impact of renewable energy sources on the electricity price level and variability – A quantile regression approach. *Energy Econ* 2020;85:104532. <http://dx.doi.org/10.1016/j.eneco.2019.104532>.
- [7] Croatian power exchange (CROPEX) <https://www.cropeex.hr/hr/trgovanja/dan-unaprijed-trziste/rezultati-dan-unaprijed-trzista.html> (accessed August 24, 2021)
- [8] Gelo, T.; Šimurina, N.; Šimurina, J. The Economic Impact of Investment in Renewables in Croatia by 2030. *Energies* 2021, 14, 8215. <https://doi.org/10.3390/en14248215>.
- [9] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).
- [10] Weather data, Weather2Umbrella, <https://www.weather2umbrella.com/pr%C3%A9visions-m%C3%A9t%C3%A9orologiques-zagreb-croatia-fr/m%C3%A9teogramme> (last accessed 24.08.2021)
- [11] Entso-E Transparency program, Hourly Electricity Load forecast, <https://transparency.entsoe.eu/> (accessed 30.10.2021)
- [12] Dedinec, Aleksandra and Aleksandar Dedinec. "Electricity price forecasting of the South East European power exchanges." (2017).
- [13] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. *Neural computation*. 9. 1735-80. 10.1162/neco.1997.9.8.1735.
- [14] Eurostat website, Energy prices on the rise in the euro area in 2021, <https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20220210-2> (last accessed)

Framework for Using BigchainDB in Software Application

Avni Rustemi

*Faculty of Electrical Engineering and
Information Technology (FEIT)
Ss. Cyril and Methodius University in
Skopje
Skopje, North Macedonia*

Vladimir Atanasovski

*Faculty of Electrical Engineering and
Information Technology (FEIT)
Ss. Cyril and Methodius University in
Skopje
Skopje, North Macedonia*

Aleksandar Risteski

*Faculty of Electrical Engineering and
Information Technology (FEIT)
Ss. Cyril and Methodius University in
Skopje
Skopje, North Macedonia*

Abstract—Recently, cryptocurrencies have become very important, as people are investing a lot of money in them, and people slowly are entering in virtual world where physical money is getting lost and all trade is being done with cryptocurrencies which we neither see nor touch, respectively they don't exist in the form of money as we are accustomed to. Cryptocurrencies are numerous and their values change day by day depending on the interest of people investing in them. In this paper we will present some of the most used cryptocurrencies and their purchase, the energy that these cryptocurrencies consume during the execution of transactions, taking as an example Ethereum. It will be given our models of how the process of cryptocurrency transactions is executed, taking also as an example Ethereum network. We will talk about the future of cryptocurrencies, architecture and data security etc. Our main focus will be on BigchainDB and the possibility of implementing the same by developers, and in particular we will provide a framework on how BigchainDB can be implemented by developers, to create their own software applications and to see in the future the possibility of how the same can be implemented in practice. Our main goal through this paper is to break the taboos of databases that we are used to working with when creating software applications, and introduce as a challenge for us the use of a new database which is less used by developers, but which is more secure, ie. BigchainDB. And the framework that we will present will be as a reference point for researchers on how BigchainDB can be used in their applications.

Keywords—*ethereum network, mining, purchase, energy consumption, architecture, data security, BigchainDB, framework.*

I. INTRODUCTION

Thanks to technological developments that have caused industrial revolution, on one side, people's lives are being facilitated in many areas, but on the other side, people are doing their best to invest as much as possible in the advancement of these technologies, always based on the fact that every technological change that happens is like consequence of previous changes. Blockchain technology is one of the main topics of most people today and based on the application of the same and the interest of the people in it, it is thought to be among the biggest technological revolutions after the internet, causing major changes in the global economy. Although it is a profitable technology and people are investing a lot of money in it, in many aspects still remains an unsolved mystery as to who invented and who administers such a complex system with delicate operating structure. It was very difficult to gain people's trust in this technology, as it was thought not to be a fraudulent system aimed only at collecting money from people and ultimately failing like many fraudulent systems they had faced and are facing even nowadays. What was even more interesting and more difficult

to gain people's trust was the fact that people are used to do business with money, either through physical contact or through banking transactions, but in this technology, transactions are performed only through cryptocurrencies that you need to buy in advance and the same change their value depending on the investment of people in them. Blockchain technology is a decentralized technology, which if we compare it with the way it works, to have it simpler, it resembles Google Doc, where all people can access the document at the same time, but cannot copy or transfer it to others. This creates a decentralized distribution chain that gives everyone access to the document at the same time. And this is where the essence of blockchain functionality lies, where everyone can see the blocks, access them but never modify or get the same. This creates a centralized system of operation of this technology. Every time someone performs a transaction, they go online and computer algorithms determine the authenticity of the transaction. Once the transaction is verified, this new transaction is linked to the previous transaction thus forming a chain of transactions. Blockchain technology rather encourages the redistribution and decentralization of power across wide communities of peers interacting with one another. And therefore, eliminates fraudulence or any chances of corruption entertained by mediators, consequently from time to time proving itself to be a philanthropist technology [1].

II. CRYPTOCURRENCIES AND THEIR PURCHASE

Bitcoin, Ethereum, Tether, BNB, and several other cryptocurrencies seem to experience a large-scale growth in price, market capitalization and widespread public acceptance. The cryptocurrencies tend to offer a lot of features and functions that are altering the way we do and implement things. Cryptocurrencies are undoubtedly extremely intervening the global economy. Even though most investments are made in relatively established cryptocurrencies, including Bitcoin (BTC) and Ethereum (ETH), there are thousands of other smaller cryptocurrencies. These currencies are prime targets for manipulation by scammers, as evidenced by the proliferation of pump and dump schemes. Pump and dump schemes are those in which a security price inflates due to deliberately deceptive activities. Those fraudulent schemes originated in the early days of the stock market and are now growing rapidly in the cryptocurrency market [2]. Market capitalization of the crypto asset is calculated by multiplying the existing reference price of the crypto asset by the current circulating supply. The price for each individual market pair is calculated by taking the unconverted price reported directly from the exchange and converting it to USD using CoinMarketCap's existing reference prices [3].

TABLE I. PURCHASE OF CRYPTOCURRENCIES

Coin	Price(\$)	Market cap(\$)	Main Markets
BTC	39,177	742 billion	Binance, Coinbase Pro, OKEx, Kraken, Huobi Global, Bitfinex
ETH	2,589	310 billion	Binance, Coinbase Pro, OKEx, Kraken, Huobi Global
USDT	1.00	80 billion	Binance, OKEx, HitBTC, Huobi Global
BNB	378.08	62 billion	FTX
USDC	1.00	52 billion	Poloniex, Binance, OKEx and Bitfinex
XRP	0.791	37 billion	Binance, Bitumb, KuCoin, Huobi Global, Bybit, Gate.io, FTX, Bitstamp, Bitfinex, Kraken
LUNA	89.85	33 billion	Huobi, Bitfinex and Upbit
ADA	0.8	26 billion	Binance, Bittrex, eToro, and HitBTC
SOL	81.27	26 billion	Binance, FTX, Bilaxy and Huobi Global
AVAX	72.20	19.22 billion	Binance, Bitfinex, Gate.io and Kucoin

The table above is based on the ranking of the most used cryptocurrencies, but always keeping in mind that they can vary depending on the sale and purchase of cryptocurrencies, where for the top ten cryptocurrencies are ranked the details of price, market cap and the place where the same can be bought.¹ When it comes to buying and selling cryptocurrencies, respectively the realization of cryptocurrency transactions, what is most important is the trust that people have or should create towards the resources that carry out these transactions. Lankton and McKnight [4] proposed that trust in technology "reflects beliefs that a specific technology has the attributes necessary to perform as expected in a given situation in which negative consequences are possible". Decentralized blockchain systems, such as Bitcoin or Ethereum, do not rely on a trusted third party in order to validate the transactions made by users or to maintain the security of the system [5][6]. The first thing you need to do to perform cryptocurrency transactions is to define the platforms for conducting transactions. A very important feature that should be noted is that not all countries have the same platforms offered and portfolios with which transactions can be realized through these decentralized systems. The process of creating an account is similar to many systems, however it differs in the way the data is stored and secured. Some systems require more data to be more secure in the future by the user, while others have a lower level of security. Once you have an account created, then you can start performing cryptocurrency transactions. Transactions can be performed only after you have filled your account with money, which as usual is done through your credit card. In these systems trading starts with buying cryptocurrencies like Ethereum, Bitcoin depending on the money you have available as each of them has its own price. The processing time of transactions depends mainly on the quantity you buy, the cryptocurrencies you buy, the platform you use to conduct trade, etc. Once you have completed some transactions, and you already own the cryptocurrencies, you can place them in the wallet or portfolio where you can more easily control any changes to your cryptocurrencies. Within a cryptocurrency, there are more assets that can be used to conduct transactions. You can even create so-called tokens, with which you can

make transactions, but they should definitely be verified and secure so that you can exchange cryptocurrencies with other people.

III. ENERGY CONSUMPTION AND THE FUTURE

Costs of processing cryptocurrency transactions are currently very high and volatile. Resolving energy spending concerns without sacrificing security and decentralization is an important challenging technique and has been the focus of studies and developments for years. Why do cryptocurrencies spend so much energy? For example, Ethereum, as a blockchain, requires that transactions and blocks to have to be validated by miners. In addition, every computational step carried out requires also some extra-charges to be paid by users. This kind of fuel is called gas, which is a unit of measuring the computational work of running transactions or SCs in the Ethereum network. In particular, gas is expressed through gwei, a subunit of ether [7]. According to a recent report by Eurelectric, the union of the electricity industry, more than 1000 projects are currently using Ethereum [8]. Energy provides the grid. Ethereum transactions are validated by miners. They merge transactions into numbered blocks and add them to the Ethereum blockchain. The blocks are then transmitted to 'node operators' who verify them separately whether they are valid or not. The possibility of a mine adding new blocks only works if there is a cost associated with mining. These conditions are met by imposing proof-of-work (PoW). To be eligible to submit a block of transactions, a miner must solve a computer puzzle faster than anyone else. Solving this puzzle creates competition between mines and costs in the form of energy expenditure. Ethereum has used proof-of-work since its inception. Migration outside of proof-of-work has always been an essential goal of Ethereum. But, technically it has been challenging because the other successful alternatives all sought to compromise the core principles and decentralization of Ethereum [9].

IV. ETHEREUM BLOCKCHAIN ARCHITECTURE AND DATA SECURITY

The data is obviously stored in certain databases, be it local databases, cloud databases, etc. It is important the data not to be misused by a third party. An Ethereum network is composed of a set of nodes running an Ethereum client (Fig. 1). Each of these nodes have a copy of the blockchain, which contains a list of all operations performed on the network. This enables nodes to prevent fraudulent activities, such as counterfeiting and duplicating cryptocurrencies, as well as containing an auditable record of all transactions performed on the network. Due to the decentralized nature of the network, the Ethereum framework preserves user pseudonymity (a weaker form of anonymity), as each user's identity is given by a public key. This enables users on the platform to perform functions such as transferring money, buying and selling, and much more [10]. The generalized structure of an Ethereum network is provided in figure below.

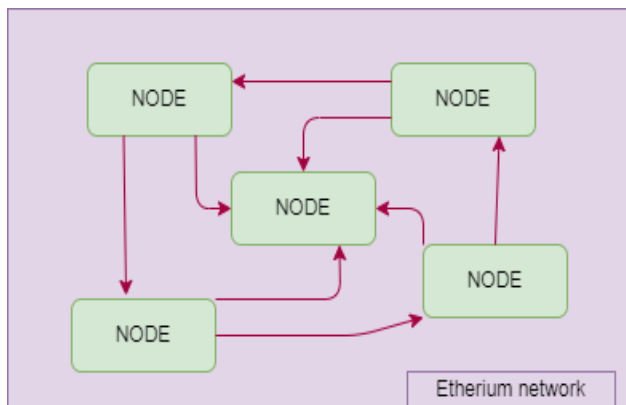


Fig. 1. Ethereum network

What exactly happens between the user and the Ethereum network is shown in figure 3 to [10], where it is clearly explained that between the user and the Ethereum network we have the browser through which we access this network. It is clearly clarified that between the user and the browser there are several interfaces that serve to provide a consumer access to their Ether wallets, enabling them to use Ethereum to pay for goods online. Ethereum provides an interface with the Ethereum network for developers in the web3.js API, which is actually a collection of libraries which allow you to interact with a local or remote Ethereum node, using an HTTP or IPC connection [11]. This allows web applications to interpret events sent by the Ethereum network and submit transactions to the network. Another more elaborate version which shows the process of performing the transaction in Ethereum is shown in Figure 2. Where among other things it is clarified that the first job is for the user to create an account, and only after the account is verified you can access the execution of transactions in the Ethereum network. Access is done through certain browsers or interfaces that can be installed within them such as MetaMask or Web3.js for developers to connect with Ethereum network. Undoubtedly, in order to carry out transactions, ETH must be purchased first, where the price and the places where it can be bought are also given in Table 1. And transactions can only be done if you have enough ETH in your portfolio for the transaction you want to make. Before we continue to elaborate on whether the blockchain network can be adapted by programmers to the use of any other software application, we will first clarify some details about BigchainDB.

V. BIGCHAIN DATABASE

With the rapid development of science and technology, especially Internet technology and computing storage capacity in the last decade, data have become the most important means of production. With the rapid development of cloud computing, the database technology has become the most basic data technology facility. Especially, the Distributed Databases have been widely used to store and process large-scale data, which are applied in different scenarios than relational databases [12]. The BigchainDB is an extensible blockchain database that is the first decentralized database system to fill in the blanks of decentralized databases. It is based on a distributed database with the features of blockchain, so it has both the advantages of distributed database and blockchain. The BigchainDB is based on MongoDB. Therefore, it has these advantages: high throughput, massive data storage capabilities, NoSQL query language support, efficient query and rights management. The

BigchainDB network consists of multiple nodes. Each node is a virtual concept consisting of three parts: the MongoDB database, the BigchainDB server, and the Tendermint communication node. Figure below shows the architecture.

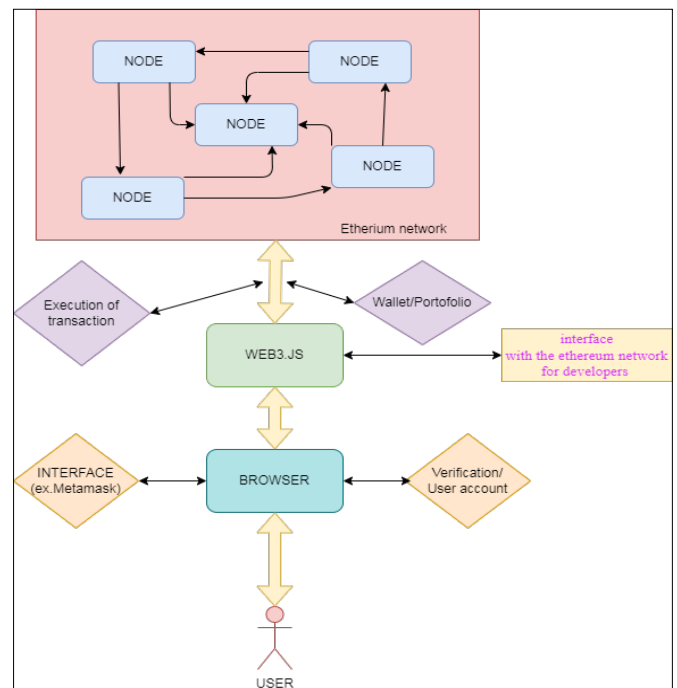


Fig. 2. Process of transaction execution in Ethereum network

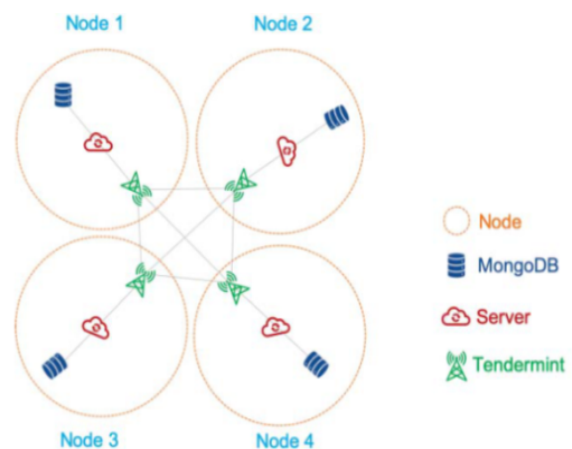


Fig. 3. A four-node BlockchainDB network

Source: <http://docs.bigchaindb.com/en/latest/query.html>

Rather than trying to enhance blockchain technology, BigchainDB starts with a big data distributed database and then adds blockchain characteristics - decentralized control, immutability and the transfer of digital assets. BigchainDB combines the key benefits of distributed DBs and traditional blockchains. Blockchain technology gives control back to you. On the internet, siloed, walled gardens lock up your personal data. This means it's next to impossible to port your data if you want to leave or to gain visibility into what is stored where. Blockchain technology enables sovereign personal data, a new way to manage your identity, medical records, and credentials [13]. In BigchainDB, one will store their confidential data off-chain and can also store their confidential data on-chain in an encrypted format. Through extensible

capability, the legitimately binding contracts, certificates could also be kept into the blockchain database. The authentication system permits the features starting from non-public enterprise level blockchain databases to open, government blockchain databases. If we are interested to deploy any kind of BigchainDB then we need to deploy it in the government version only. BigchainDB is taken into account in the top level distributed computing platforms like Eris/Tender mint [14], [15].

VI. PROPOSED FRAMEWORK FOR USING BIGCHAINDB IN SOFTWARE APPLICATION

The problem of data security and big data is more and more obvious, therefore more and more attempts are being made to create new data management opportunities and to provide more secure data transfer. The benefits of the application of blockchain are applied in three areas namely for decentralized data storage, performing blockchain enabled data analysis and finally in maintaining blockchain enabled data security. Some of the blockchain projects for big data applications are: Story, Omnilytics, Rubix, Provenance, File Coin and Datum [16]. BigchainDB, a scalable blockchain for distributed big data was proposed by integrating the traditional blockchain with distributed databases to attain scalability through faster querying mechanisms. Blockchain network allows a user to join the chain based on the consensus. The process of communication with the set of nodes in BigchainDB are often done by employing a BigchainDB HTTP API [17] just like the BigchainDB Python Driver. There are many BigchainDB drivers provided in BigchainDB docs [18]. BigchainDB offers developers, startups and enterprises unique solutions to develop their concepts, platforms, and applications successfully. One of the main benefits of the BigchainDB is its ability to deploy a Blockchain-like decentralized system using production-ready technologies. It can be used to create new private Blockchains to take advantage of Blockchain capabilities at scale, or as a service for developers who desire the benefits of Blockchains without the stress of setting up private networks.

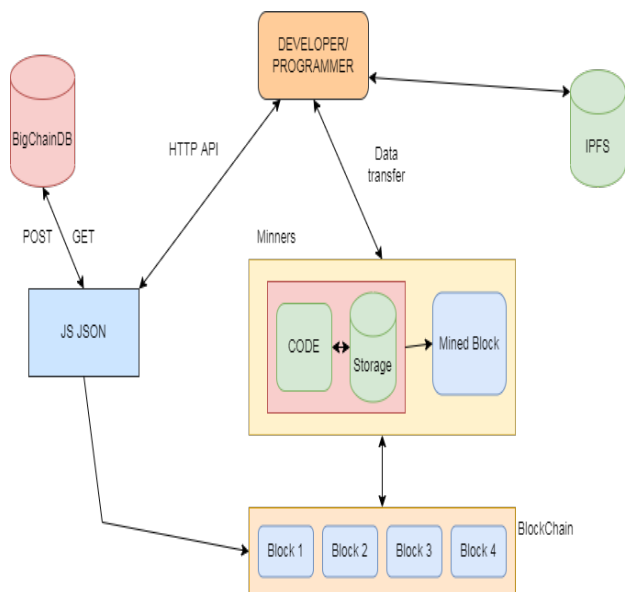


Fig. 4. Framework for using BigchainDB

Without going into the details of how BigchainDB works, we will present a framework on how it can be used the same by blockchain developers. It should be noted that it is a

database with a more delicate structure, and requires good knowledge of the structure of the blockchain operation, data structures, understanding of cryptography and etc, so that it can be used rationally. Developers transfer data to the blockchain network through the mining process. Process of adding transaction records to public ledger is termed as mining where miners need to solve a mathematical problem based over cryptography hash algorithm. Each cryptocurrency has its own unique hashing algorithm. For instance, bitcoin uses the SHA-256 algorithm, Monero uses Crypto Night, and Ethereum uses Ethash [19]. Once a miner finds the correct input, they have achieved Proof of Work (PoW). IPFS try to keep item data stored securely in a decentralized fashion. IPFS is used to help with data storage. Not all data are stored in the blockchain, but only the hash value gained while storing data in IPFS are stored into BigchainDB. Through back-end developers, the transactions are transferred to the BigchainDB, using HTML, CSS, JavaScript, and heavy use of web3js [20]. GET and POST are the two transactional methods for retrieving and validating the transactions within the BigchainDB [21]. BigchainDB transactions are within the type of JSON strings. Using BigchainDB is not at all easy. This is because you should also have knowledge of the UBUNTU (Linux) operating system, as it is the most suitable for its use. BigchainDB Server doesn't work on Windows or macOS (unless you use a VM or containers). The idea of the framework is this: developers create applications in a programming language suitable for BigchainDB, whether Java or Solidity. They build applications based on blockchain architecture and protocol. Their applications run on the blockchain too, but in advance using data selection techniques. Data transfer to the BigchainDB is done using web3.js, but always based on the way the blockchain operates. IPFS in this case it is used only as an auxiliary method for faster data processing by blockchain. Reasons why BigchainDB is being attempted to be used is the high security that blockchain technology has shown and is showing even today. Since blockchain is supposed to be extremely secure and unalterable, many individuals have dubbed this technology as "unhackable". However, recent incidents have unfortunately shown that hackers can access blockchains in certain situations. However, for now it remains one of the most secure databases of all existing databases.

VII. CONCLUSION

BigchainDB and distributed databases play a very important role and are seen as the future of data management, as the amount of data and their processing increases. It is very difficult to penetrate into the depths of blockchain operation because still remains a mystery, however what is possible is to use it and modify it in our applications without going into the details. Through this paper, we tried to bring details of blockchain technology, to provide explanations for the most used cryptocurrencies and how their market values change, energy consumed by cryptocurrencies, concluding with a framework on how BigchainDB can be utilized using the network blockchain system to create our applications. We have tried to give detailed explanations regarding the process of communication between the developer and BigchainDB. Although it is still a term which is being developed, which requires greater programming skills and greater knowledge of the structure of blockchain, it remains our next challenge to get into the details of data storage and processing in the BigchainDB by describing in detail the process of searching and executing queries, as well as its connection to software

applications. And finally we will try to create our own software application using BigchainDB and describe the challenges we will face.

REFERENCES

- [1] [1] Samyak Jain, Umang Rastogi, Nikita Bansal, Gagandeep Kaur, "Blockchain based Cryptocurrency for IOT", 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN), 13 May 2019, Noida, India, DOI: 10.1109/SPIN.2019.8711727.
- [2] [2] Mehrnoosh Mirtaheri, Sami Abu-El-Haija, Fred Morstatter, Greg Ver Steeg and Aram Galstyan, "Identifying and Analyzing Cryptocurrency Manipulations in Social Media", IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, VOL. 8, NO. 3, JUNE 2021, DOI: 10.1109/TCSS.2021.3059286.
- [3] [3] CoinmarketCap, online (<https://support.coinmarketcap.com/hc/en-us>) Accessed: March 2022.
- [4] [4] N. K. Lankton and D. H. McKnight, "Do people trust facebook as a technology or as a 'person'? Distinguishing technology trust from interpersonal trust,". Conference: Learning from the past & charting the future of the discipline. 14th Americas Conference on Information Systems, AMCIS 2008, Toronto, Ontario, Canada, August 14-17, 2008.
- [5] [5] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," Available online (www.bitcoin.org). p. 9, 2008, Accessed: March 2022.
- [6] [6] Ethereum Foundation, "Ethereum," 2022. [Online]. Available: <https://Ethereum.org/>. [Accessed: 30-Jan-2022].
- [7] F. Buccafurri, G. Lax, L. Musarella, A. Russo, An Ethereum-based solution for energy trading in smart grids, Digital Communications and Networks (2022), doi: <https://doi.org/10.1016/j.dcan.2021.12.004>.
- [8] [8] Max N. Luke, Stephen J. Lee, Zdenek Pekarek, Anna Dimitrova, "Blockchain in Electricity: a Critical Review of Progress to Date", May 2018.
- [9] [9] Ethereum energy consumption, available online (<https://Ethereum.org/en/energy-consumption/>), Accessed: Jan, 2022.
- [10] [10] R. Vishnu Prasad, Ram Dantu, Aditya Paul, Paula Mears, "A Decentralized Marketplace Application on The Ethereum Blockchain", 2018 IEEE 4th International Conference on Collaboration and Internet Computing, October 2018, DOI:10.1109/CIC.2018.00023.
- [11] [11] Ethereum JavaScript API, available online (<https://web3js.readthedocs.io/en/v1.2.6/>), Accessed: Jan, 2022.
- [12] [12] Yajun Wang, Chao-Hsien Hsieh, "Research and Analysis on the Distributed Database of Blockchain and Non-Blockchain", 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics, April 2020, DOI:10.1109/ICCCBDA49378.2020.9095589.
- [13] [13] BigChainDB Features and Uses cases, available online (<https://www.bigchaindb.com/features/>), Accessed: February 2022.
- [14] [14] Eris industries, available online (<http://erisindustries.com/>), Accessed: February 2022.
- [15] [15] Tendermint. Available online (<http://www.tendermint.com/>), Accessed: February 2022.
- [16] [16] Deepa N, Quoc-Viet Pham, Dinh C. Nguyen, Sweta Bhattacharya, B. Prabadevi, Thippa Reddy Gadekallu, Praveen Kumar Reddy Maddikunta, Fang Fang, Pubudu N. Pathirana, "A Survey on Blockchain for Big Data: Approaches, Opportunities, and Future Directions", Elsevier, Volume 131, June 2022, Pages 209-226.
- [17] [17] BigchainDB HTTP API. Available: (<https://docs.bigchaindb.com/projects/server/en/latest/http-clientserver-api.html>), Accessed: February 2022.
- [18] [18] BigchainDB Drivers & Tools. Available: (<https://docs.bigchaindb.com/projects/server/en/latest/driversclients/index.html>), Accessed: February 2022.
- [19] [19] Bhadoria, R.S., Arora, Y., Gautam, K. (2020). Blockchain Hands on for Developing Genesis Block. In: Kim, S., Deka, G. (eds) Advanced Applications of Blockchain Technology. Studies in Big Data, vol 60. Springer, Singapore. https://doi.org/10.1007/978-981-13-8775-3_13.
- [20] [20] web3.js - Ethereum JavaScript API, online (<https://web3js.readthedocs.io/en/v1.7.1/>), Accessed: February 2022.
- [21] [21] R. Vamshidhar Reddy, V. Jaya prakash Reddy, E. Madhusudhana Reddy, "Automatic Face Expressions and Gesture Detection System Using Blockchain Security", 2020 International Conference on Intelligent Engineering and Management (ICIEM), 17-19 June 2020.

Missing Data in Longitudinal Image Retrieval for Alzheimer's Disease

Katarina Trojchanec Dineva
Faculty of Computer Science and Engineering
Skopje, Republic of North Macedonia
katarina.trojchanec@finki.ukim.mk

Ivica Dimitrovski
Faculty of Computer Science and Engineering
Skopje, Republic of North Macedonia
ivica.dimitrovski@finki.ukim.mk

Ivan Kitanovski
Faculty of Computer Science and Engineering
Skopje, Republic of North Macedonia
ivan.kitanovski@finki.ukim.mk

Suzana Loshkovska
Faculty of Computer Science and Engineering
Skopje, Republic of North Macedonia
suzana.loshkovska@finki.ukim.mk

*for the Alzheimer's Disease Neuroimaging Initiative

Abstract— The paper is focused on the missing scans in the context of longitudinal image retrieval for Alzheimer's Disease. Namely, we explore the influence of missing data on the retrieval results when the subjects are represented by the longitudinal changes calculated on the basis of the within-subject template generated using the available time points. To evaluate the effect of the missing scans, we defined two (most characteristic and most common) scenarios, in which missing scans at a specific time point are considered, and one scenario that is based on complete data used as a baseline to compare against. Additionally, we increased the number of patients with missing scans from 10% to 50% and evaluated its impact on the retrieval results.

The evaluation showed that from the examined types of feature vectors, concatenated longitudinal changes of the volumes of the cortical and sub-cortical structures are superior and robust. In the case when the dimensionality of the descriptor is an important criterion, we recommend the usage of the percent change or symmetrized percent change of the volumetric measures. Additionally, the influence of the missing scans on the retrieval results is lower when incomplete data occurs in the early time points, rather than in later ones. Moreover, very little or no performance reduction was detected by increasing the number of subjects with missing scans. In general, the evaluation showed very small or no performance degradation in the retrieval process in the scenarios with missing scans, in comparison to the scenario with fully complete data.

Keywords— *Missing Data, Longitudinal Images, Longitudinal Image Retrieval, Alzheimer's Disease, Magnetic Resonance Imaging.*

I. INTRODUCTION

Alzheimer's Disease (AD) is a progressive neurodegenerative disease and the most common cause of dementia [1]. Considering this, early diagnoses of the disease, monitoring the patient's condition or the disease progression, finding powerful diagnostic or prognosis biomarkers, identifying the patients who are most probable to ultimately

develop AD, as well as reaction to the therapy are active research fields.

Advances in genetics and medicine, as well as the rapid evolution of technology and neuroimaging techniques increase the amount of generated data of the medical cases for AD. Among all the data generated within the medical cases for AD, Magnetic Resonance Images (MRI) offer extremely good opportunities and are found to have a key role in early diagnosis, prognosis and monitoring of AD [2]. Considering that the information extracted from the brain images provide precise and consistent markers for diagnosis and monitoring the development of the disease [3-4], our research is towards medical case retrieval by using medical images as input queries.

The image retrieval process involves representation of the query image with a descriptor and then comparing it with the descriptors of all images in the medical data- base. The result is a list of all images sorted in the database by similarity, so that the most similar one is at the top [6]. In that sense, for a given medical case of a patient for whom an MRI was acquired, the image is given as a query to the system. The retrieval system provides a sorted list of all images of other patients according to their similarity to the query. The main challenge is to get the result that is semantically relevant [7-8].

Taking into consideration the progressive nature of AD, longitudinal data are extremely important. Longitudinal images (images acquired at multiple consecutive time points) are meant to reflect the disease progression. However, a key problem that arises considering longitudinal images is missing data, i.e. lack of scan/s for one or more time points. The reason for this might be the patient's inability to undergo scanning at the predefined period, disability to continue the examination, poor quality of the scan, etc. [9]. In addition, the occurrence of missing data where the time difference between time points is not uniform is very common [10].

One possible solution is to exclude from the research all patients with incomplete data. The main limitation of this solution is that it significantly reduces the number of patients included in the examination, excluding possibly important knowledge that might be extracted from them. One way to overcome this limitation is to use data imputation [5,9, 11-12]. This would complement those features that correspond to the missing time point, in order to maintain the same dimension of the vector for all scans and to artificially

* Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

supplement the missing information. In [13] multivariate normal imputation (MVNI) method and fully conditional specification (FCS) method are compared and discussed. This study is the first that uses real data as part of an ongoing trial to make the validation and comparison. Detailed review of the hierarchy of missing data mechanisms and their relationship to likelihood-based methods and a series of simulation studies with designs common in longitudinal neuroimaging studies is provided in [14]. Although the challenging problem of missing data has been researched for over forty years [14], it still an active research field and a big challenge.

The aim of the paper is to investigate the influence of missing scans on the retrieval performance in the context of longitudinal image retrieval for AD. Several key questions arise here:

1. What kind of features are more powerful and robust when missing scans are present?
2. How does retrieval results change with increasing the number of patients with missing data?
3. Does the time point at which the scan is missing affect the retrieval results?
4. How does a missing scan at a different time point affect the retrieval results in comparison to fully complete dataset?

A research that answers these questions in the context of image retrieval for AD is still not performed. Thus, our research is aimed to provide insight into the problem of missing data in this context as well as answers to the aforementioned questions.

Moreover, taking into consideration the diversity of the datasets and/or the sample selection for the evaluation process, we cannot provide an objective comparison with other research involving addressing the problem of missing data.

The paper is organized as follows. We present the methods used in this study in Section 2. The experimental results and discussion are provided in Section 3. The concluding remarks and future work are summarized in Section 4.

II. MATERIALS AND METHODS

A. Participants and Inclusion Criteria

The performed research is based on the scans and data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies, and non-profit organizations as a \$60 million, 5-year public-private partnership. The aim of the initiative is to enable research on whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, such as cerebrospinal fluid (CSF) markers, APOE status and full-genome genotyping via blood sample, as well as clinical and neuro-psychological assessments can be used together to indicate and estimate the progression of mild cognitive impairment (MCI) and Alzheimer's Disease (AD). Finding appropriate markers indicating early AD progression is aimed towards

development of new treatments, improving the process for monitoring treatments effectiveness, and reducing the time and cost of the clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California - San Francisco. Many coinvestigators from a broad range of academic institutions and private corporations have given their support and contribution to make ADNI a valuable product. Subjects have been recruited from over 50 sites across the U.S. and Canada, initially supposed to be 800 subjects in total. However, ADNI has been followed by ADNI-GO and ADNI-2, with over 1500 participants, namely in fact adults in the age group of 55 to 90 years. The protocols for ADNI-1, ADNI-2, and ADNI-GO specify the three groups (Cognitively normal individuals, adults with early or late MCI, and people with early AD) available in the dataset with different follow up duration of each group. For up-to-date information, see <http://www.adni-info.org>.

For this research, we used the standardized list from ADNI-1, containing images acquired at multiple time points. From this list, we selected the subjects that have available scans at baseline (TP1), and the 6-month (TP2), 12-month (TP3), and 24-month (TP4) follow-ups and belong to AD or normal control (NL) group. In this way, we selected a total of 267 subjects, 168 in AD group, and 99 in NL group. Patients' demographics information can be found in [15]. The timing of scans per time point by clinical group is provided in Table I, also available in [15], but included in this paper as well for clarity.

We chose exactly this selection criteria because of the following reasons: (1) more time points (ex. 36-month follow-up) is not available for the AD group of patients and, additionally, the total number of patients for whom all the scans are available is reduced by more than 12%; (2) a smaller number of time points would not give enough space for research and an opportunity to have a good insight into the problem; (3) with the selected time points, we have an opportunity to analyze the problem in the case of equally and unequally spaced available time points (depending of which time point is missing).

TABLE I. TIMING OF SCANS PER TIME POINT BY CLINICAL GROUP

Time Point	Time from baseline (years)	
	AD	NL
TP1 (Baseline)	0	0
TP2 (Month 6)	0.57 (0.05)	0.58 (0.06)
TP3 (Year 1)	1.07 (0.05)	1.08 (0.07)
TP4 (Year 2)	2.09 (0.14)	2.1 (0.11)

Abbreviations: TP = Time Point.

The time from baseline is given in mean (standard deviation).

B. Longitudinal image representation

This study is based on longitudinal data for patients that undergo examination for Alzheimer's Disease. This kind of data are characterized by that the outcome variables are measured repeatedly on the same cohort of individuals at multiple time points. This enables identifying the influence of the changes in the derived measurements over time to the examined clinical, biological or experimental factor, possibly reflecting the diseases progress and/or reaction to the treatment. Longitudinal studies provide direct assessment of

within-subject changes across different time points, free of any between-subject variability [10, 20]. Several distinctive characteristics should be emphasized for longitudinal data [10]:

1. Longitudinal measurements are ordered in time. They reflect the temporal trajectory of an underlying non-stationary continuous process.
2. Typically, serial measurements obtained for a single subject are positively correlated caused by the smooth trajectory of the underlying biological process.
3. Between-subject variance is not usually constant over the duration of the study.
4. Missing data and non-uniform timing are extremely common, particularly for longitudinal studies of larger duration.

This study takes into account the aforementioned characteristics and focuses specifically on the fourth. Namely, when longitudinal images are available for the patients, the information contained in all of them should be properly extracted and utilized. This is a crucial step to be able to get the semantically more relevant retrieval result. A significant challenge that arises in this case is how to deal with the missing scan/s for some of the patients.

In fact, to deal with the problem of missing scans for Alzheimer's disease, we used the feature extraction method that we explored in our previous research [15] and evaluate the impact of the missing time point when using this approach. Namely, for each patient who has a missing scan at a specific time point, we processed the data longitudinally and generated a within-subject template only from the available time points for a particular patient and not for all possible time points. For example, if for a particular patient only the scan at baseline (TP1), the 12-month (TP3), and the 24-month (TP4) follow-ups are available, then we generate the template on the basis of only these time points, instead of on the basis of all the possible four time points (scans at baseline (TP1), the 6-month (TP2), 12-month (TP3), and 24-month (TP4) follow-ups)). The longitudinal scheme for generating the template is designed to be unbiased with respect to any time point and there was no initialization with information from a specific time point [16].

We performed this processing with the FreeSurfer's fully automated longitudinal pipeline [16]. Fig. 1 shows a schematic illustration of the main steps in the longitudinal pipeline. The full list of methods applied to the examined dataset is included in [15] and proposed and explained in more details in [16].

After this kind of processing, we generated the feature vector by using longitudinal changes of the volume of the cortical and sub-cortical regions. The following reliable estimates of the longitudinal changes on the bases of the available time points were calculated and used as features [15-17]:

- VolumesRC - rate of change (RC) of the volumes (calculated as a difference per time unit). The statistic is derived from the volumetric temporal information per subject for each cortical (34 measures for each hemisphere) and sub-cortical region (55 measures), 123 measures in total;

- VolumesPC1fit - percent change of the volumes (calculated with respect to the value obtained from the linear fit at baseline, i.e. percent thinning/volume loss per year) The statistic is derived from the volumetric temporal information per subject for each cortical and sub-cortical region, 123 measures in total;
- VolumesSPC – symmetrized percent change (the rate with respect to the temporal average). The statistic is derived from the volumetric temporal information per subject for each cortical and sub-cortical region, 123 measurements in total.

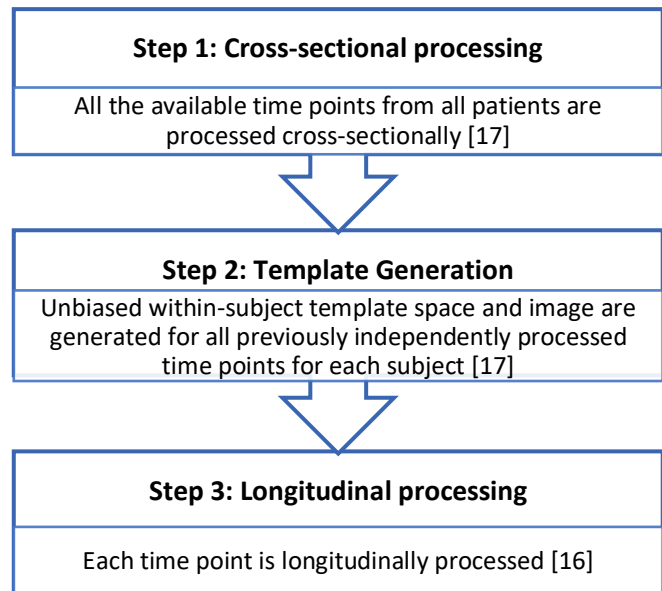


Fig. 1. Schematic illustration of the main steps in the FreeSurfer longitudinal pipeline

These features reflect the speed and degree of degeneration caused by the disease, in fact the disease severity and progression. This way, for a patient with missing data, a representation extracted only from the available scans will be generated and will be of the same type and with the same dimension as those for the patients with complete data. In fact, the dimension of all descriptors for all patients will be the same, regardless the number of the available time points. For example, if a scan taken at the screening 6 months after the first visit to the hospital is missing, then the scans at TP1, TP3, and TP4 are longitudinally processed and a template on the basis of these scans is generated and subsequently the longitudinal changes of the volume of the cortical and sub-cortical structures as well as of the cortical thickness are calculated. For a patient with complete data for instance, all the time points are longitudinally processed, and the same statistics of the same regions are generated. However, the dimension of the feature vector is the same as it is for the patient with three available time points.

Although it is also possible to calculate longitudinal changes of the cortical thickness measures, the reason for using the volumetric changes is that in our previous research [15] they have proven to be superior over the longitudinal changes of the cortical thickness in the context of image retrieval for AD.

We chose this strategy because it has two main advantages: (1) it can operate only with the available data independently for each subject, whether or not they have a missing scan at some point in time (2) it provides the same

descriptor dimensionality for all patients, regardless of the number of the available time points.

Additionally, we applied the Correlation-based Feature Selection (CFS) method [18] with the aim to reduce the feature vector dimensionality and to select the most relevant features, because it provided superior results on the experimental basis for the longitudinal image retrieval for Alzheimer's disease that we reported in [19]. To provide relevant and unbiased results, we performed the feature selection method independently of the query subject information. In fact, we obtained a specific feature subset for each query subject using the information of all other subjects similarly like in [15].

C. Evaluation of the influence of the missing data on the retrieval results

To be able to evaluate the influence of the missing data on the retrieval results, we defined two most characteristic and most common scenarios with missing time points and one scenario for reference:

- M1 – from the available four time points, scans at TP2 were excluded from $x\%$ randomly selected patients, where $x \in \{10, 20, 30, 40, 50\}$, keeping at least 50% of the patients with fully complete data.
- M2 – from the available four time points, scans at TP4 were excluded from $x\%$ randomly selected patients, where $x \in \{10, 20, 30, 40, 50\}$, keeping at least 50% of the patients with fully complete data.
- M3 – fully complete dataset that we used as a baseline to compare against.

It should be noted that both M1 and M2 scenarios contain the first time point. In the clinical environment, this is most often the case because it is a screening made during the first visit to a medical institution and in most cases, it is available with good quality. We did not evaluate the scenario in which the third time point is missing, because during the longitudinal processing of the other available scans (when the third TP is excluded), there was an error that required manual interventions in the processing in some of the cases. We wanted to keep the processing fully automated, so as to be able to reduce the subjectivity of the human factor and to perform fair comparison. Another characteristic of M1 and M2 is that the time points available are at the same time distance (12 months for M1 and six months for M2). Additionally, only the M1 scenario contains the time point that is the most temporally distant from the first visit to a doctor, that is the fourth TP. This is important to be noted because according to the previous experiments it has an impact on retrieval results when all time points are available [15]

In this examination, we did not consider templates and accordingly, scenarios, with only two time points, because in order to generate the template, a voxel-wise median was used. But, when only two time points are available, median is equal to mean, hence resulting in blurry, instead of crispy edges, which is the case when more time points are available to build the template [16].

To evaluate the proposed strategy, we used the standard evaluation metric for quantitative measurement of the retrieval performance, Mean Average Precision (MAP). It was calculated as the mean of the average precision scores for

each query, evaluation metric for the general retrieval performance. It is meant to favor retrieval systems that return more relevant subjects at the top of the list. The retrieved subject is assumed to be relevant if the patient has the same diagnosis as the query one. The higher the relevant subjects are in the retrieved list, the higher the value of the precision is.

Considering the small number of patients included in this study, we used leave-one-out strategy. This means that each patient's representation was used as a query against all other representations stored in the database.

Considering that a random selection of patients with missing data was performed, and to get representative results, we did 10 repetitions of the random selection and we repeated the experiment for each case accordingly. Then we calculated the mean of the MAP for each of the 10 repetitions. Subsequently we compared those results with the retrieval results obtained by the scenario based on the complete dataset (M3).

III. EXPERIMENTAL RESULTS AND DISCUSSION

This section summarizes the results from the evaluation of the presented strategy for dealing with missing data in the context of longitudinal image retrieval for AD.

Table II shows the results associated with the M1 scenario for a different percentage of patients with a missing scan at TP4. According to the results, the rate of change of the volume of the cortical and sub-cortical regions provides the same value as in the case of fully complete data when the number of patients with missing data is not more than 30% of the total number of patients in the examined dataset. When 30-50% of the subjects have missing scan at TP4, the value of MAP is lower by 1%. The value of VolumesPCfit and the VolumesSPC remains constant as the number of patients with missing data increases up to 50%, which makes them stable features, suitable in the case of missing data. It should be noted that VolumesPCfit led to a 1% lower value of MAP in comparison with the case of fully complete data. The combination of all calculated longitudinal changes led to the best result in terms of MAP. In this case, the value of MAP remained the same, even though the number of patients with missing data was increased up to 50%. Moreover, it is the same as the corresponding value in the case of fully complete data.

TABLE II. EVALUATION OF THE SCENARIO M1 BASED ON THE VALUE OF MAP – MISSING SCAN AT TP2 IN $x\%$ RANDOMLY SELECTED PATIENTS, WHERE $x \in \{10, 20, 30, 40, 50\}$

	M1 (10%)	M1 (20%)	M1 (30%)	M1 (40%)	M1 (50%)	M3
VolumesRC	0.76	0.76	0.76	0.75	0.75	0.76
VolumesPCfit	0.78	0.78	0.78	0.78	0.78	0.79
VolumesSPC	0.78	0.78	0.78	0.78	0.78	0.78
All volume rates - concatenated	0.79	0.79	0.79	0.79	0.79	0.79

Regarding the second scenario M2 (Table III), the obtained results indicate reduction of MAP value by 1-4% when the number of patients with missing data increases to 50%, when the patients are represented by the rate of change of the volumetric measures. Even when 10% of the patients

are with missing data, the value of MAP is reduced. By increasing this percentage to 20, there was again a reduction in MAP by 1%, but it remained the same when the number of patients with missing data was increased to 30%. Then again, the value of MAP decreased by 1% in case of 40% of patients with missing data. This trend was also the case when the number of patients with missing data was increased to 50%. When PCfit and SPC of the volumes of cortical and sub-cortical structures were used to generate the descriptors, the value of MAP started at 0.77 and decreased by 0.01 as the number of patients with missing data increased to 20%. It remained the same in the case of 30% of patients with missing data. Then it decreased firstly by 1% and then, by 2% as the number of patients with missing data increased to 40% and 50% respectfully. When concatenation of all the longitudinal changes was used to generate the feature vector, it led to the best results regarding the M2 scenario but showed the same trend of decreasing the values of MAP.

TABLE III. EVALUATION OF THE SCENARIO M2 BASED ON THE VALUE OF MAP – MISSING SCAN AT TP4 IN X% RANDOMLY SELECTED PATIENTS, WHERE $X \in \{10, 20, 30, 40, 50\}$

	M2 (10%)	M2 (20%)	M2 (30%)	M2 (40%)	M2 (50%)	M3
VolumesRC	0.75	0.74	0.74	0.73	0.72	0.76
VolumesPCfit	0.77	0.76	0.76	0.75	0.73	0.79
VolumesSPC	0.77	0.76	0.76	0.75	0.73	0.78
All volume rates - concatenated	0.78	0.77	0.77	0.76	0.74	0.79

According to the performed evaluation, we can note that the presence of a time point that is more distant from the first visit to a medical institution, makes the features more relevant, more stable and more powerful to deal with missing scan. In fact, the ability of the algorithm to handle missing data in the context of image retrieval is closely related to the longitudinal processing of the scans, i.e. the stability and relevance of the results of the longitudinal processing using the available points. Namely, according to the previous research [15], it was concluded that the features extracted from the scans at time point 1, 3 and 4 are more promising and lead to greater accuracy compared to those extracted on the basis of the scans at time point 1, 2 and 3. This was directly reflected in the results obtained from the evaluation of the scenarios with missing data.

The main findings of our research on the bases of the evaluation of the defined scenarios are:

1. From the examined features vector types, VolumesPCfit or VolumesSPC, or concatenated feature vector of all longitudinal change, VolumesPCfit and VolumesSPC proved to be more efficient, stable and powerful to deal with missing scans when separate longitudinal changes are used.
2. Very small or no performance reduction as the number of cases with missing data increases. This means that the strategy for representing the subjects with longitudinal changes of the volumetric measures provides reliable and robust way of image representation, and the whole setup that we used for image retrieval provides promising way for facing the problem of missing data.
3. When the scan is missing in the earlier time points, then the influence to the retrieval results is smaller, rather than in the case of missing data at the later time points. This is highly expected because the changes in the brain as the time passes carry more significant information about the progression of the disease.
4. There is no or small degradation in the retrieval performance in cases with missing scans in comparison to the scenario with fully complete data.

According to the obtained results, in the case of missing data we suggest representing the patients' scans by feature vector comprised of concatenated longitudinal changes of the volumes of cortical and sub-cortical structures. In case when the feature vector should be kept as short as possible, then VolumesSPC or VolumesPCfit of the volumetric measurements are advisable as they turned out to be superior and more robust.

IV. CONCLUSION

Missing data is a serious problem when longitudinal data are considered. In this paper, we evaluated the influence of the missing data in the context of image retrieval for Alzheimer's disease. Our results showed that the retrieval performance is closely related to time points at which the scans are missing and to the reliability of the longitudinal data extracted from the available time points. The features extracted from the scans at the later time points, proved to be more efficient, stable and powerful to deal with missing scans. From the examined types of feature vectors, concatenated longitudinal changes of the volumes of cortical and sub-cortical structures provided superior results. In the case when the dimensionality of the descriptor is important factor, VolumesSPC or VolumesPCfit of the volumetric measurements are advisable to be used as more powerful and robust. In all cases, we noticed very small or no degradation in the retrieval performance when we used these features as the number of cases with missing data increased.

In the future, we plan to extend the evaluation of this strategy on a bigger cohort, to include more available time points and also to consider scenarios with more than one missing scan per patient. Moreover, we are going to examine deep neural network architectures developed in a longitudinal manner to provide a solution for dealing with missing data.

ACKNOWLEDGEMENT

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). This study was partially funded by the Faculty of Computer Science and Engineering, Skopje through the SatView project.

REFERENCES

- [1] Alzheimer's Association, 2019. 2019 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 15(3), pp.321-387. [Last access: 01.03.2022]
- [2] Ye, J., Wu, T., Li, J. and Chen, K., 2011. Machine learning approaches for the neuroimaging study of Alzheimer's disease. *Computer*, 44(4), pp.99-101. [Last access: 01.03.2022]
- [3] Bron, E.E., Klein, S., Reinke, A., Papma, J.M., Maier-Hein, L., Alexander, D.C. and Oxtoby, N.P., 2022. Ten years of image analysis and machine learning competitions in dementia. *NeuroImage*, 253, p.119083. [Last access: 15.04.2022]
- [4] Hedges, E.P., Dimitrov, M., Zahid, U., Vega, B.B., Si, S., Dickson, H., McGuire, P., Williams, S., Barker, G.J. and Kempton, M.J., 2022.

- Reliability of structural MRI measurements: The effects of scan session, head tilt, inter-scan interval, acquisition sequence, FreeSurfer version and processing stream. *NeuroImage*, 246, p.118751. [Last access: 15.04.2022]
- [5] Peng, L., Lin, L., Lin, Y., Chen, Y.W., Mo, Z., Vlasova, R.M., Kim, S.H., Evans, A.C., Dager, S.R., Estes, A.M. and McKinstry, R.C., 2021. Longitudinal Prediction of Infant MR Images With Multi-Contrast Perceptual Adversarial Learning. *Frontiers in neuroscience*, p.1114. [Last access: 15.04.2022]
- [6] Piras, L. and Giacinto, G., 2017. Information fusion in content based image retrieval: A comprehensive overview. *Information Fusion*, 37, pp.50-60. [Last access: 14.03.2022]
- [7] Faria, A. V., Oishi, K., Yoshida, S., Hillis, A., Miller, M. I. and Mori, S., 2015. Content-based image retrieval for brain MRI: An image-searching engine and population-based analysis to utilize past clinical data for future diagnosis. *NeuroImage: Clinical*, 7, pp.367-376. [Last access: 14.03.2022]
- [8] Qayyum, A., Anwar, S. M., Awais, M. and Majid, M., 2017. Medical image retrieval using deep convolutional neural network. *Neurocomputing*. [Last access: 14.03.2022]
- [9] Thung, K.H., Wee, C.Y., Yap, P.T. and Shen, D., 2016. Identification of progressive mild cognitive impairment patients using incomplete longitudinal MRI scans. *Brain Structure and Function*, 221(8), pp.3979-3995. [Last access: 17.03.2022]
- [10] Bernal-Rusiel, J.L., Greve, D.N., Reuter, M., Fischl, B., Sabuncu, M.R. and Alzheimer's Disease Neuroimaging Initiative, 2013. Statistical analysis of longitudinal neuroimage data with linear mixed effects models. *Neuroimage*, 66, pp.249-260. [Last access: 17.03.2022]
- [11] Lei, B., Yang, M., Yang, P., Zhou, F., Hou, W., Zou, W., Li, X., Wang, T., Xiao, X. and Wang, S., 2020. Deep and joint learning of longitudinal data for Alzheimer's disease prediction. *Pattern Recognition*, 102, p.107247. [Last access: 24.03.2022]
- [12] Liu, X., Chen, K., Wu, T., Weidman, D., Lure, F. and Li, J., 2018. Use of multimodality imaging and artificial intelligence for diagnosis and prognosis of early stages of Alzheimer's disease. *Translational Research*, 194, pp.56-67. [Last access: 24.03.2022]
- [13] Rosato, R., Pagano, E., Testa, S., Zola, P. and di Cuonzo, D., 2021. Missing data in longitudinal studies: Comparison of multiple imputation methods in a real clinical setting. *Journal of Evaluation in Clinical Practice*, 27(1), pp.34-41. [Last access: 24.03.2022]
- [14] Matta, T. H., Flournoy, J. C. and Byrne, M. L., 2018. Making an unknown unknown a known unknown: Missing data in longitudinal neuroimaging studies. *Developmental cognitive neuroscience*, 33, pp.83-98. [Last access: 24.03.2022]
- [15] Trojchanec, K., Kitanovski, I., Dimitrovski, I. and Loshkovska, S., 2017. Longitudinal brain MRI retrieval for Alzheimer's disease using different temporal information. *IEEE Access*, 6, pp.9703-9712. [Last access: 24.03.2022]
- [16] Reuter, M., Schmansky, N.J., Rosas, H.D. and Fischl, B., 2012. Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage*, 61(4), pp.1402-1418. [Last access: 24.03.2022]
- [17] Reuter, M., Rosas, H. D., Fischl, B., 2010. Highly Accurate Inverse Consistent Registration: A Robust Approach. *Neuroimage* 53 (4), pp.1181-1196. [Last access: 25.03.2022]
- [18] Hall, M., and Holmes, G., 2003. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering*. 15(6), 1437-1447. [Last access: 01.04.2022]
- [19] Trojcanec, K., Kitanovski, I., Dimitrovski, I. and Loshkovska, S., 2015, October. Medical image retrieval for Alzheimer's disease using data from multiple time points. In *International Conference on ICT Innovations* (pp. 215-224). Springer, Cham. [Last access: 01.04.2022]
- [20] Fitzmaurice, G.M., Laird, N.M. and Ware, J.H., 2012. *Applied longitudinal analysis*. John Wiley & Sons. [Last access: 01.04.2022]

Creating a distributed, privacy-aware e-learning system based on item response theory

Bojan Kostadinov
Software Engineering
Cloud Solutions LLC
Skopje, Macedonia
bojankostadinov@gmail.com

Irena Stojmenovska
School of Computer Science and
Information Technology
University American College Skopje
Skopje, Macedonia
irena.stojmenovska@uacs.edu.mk

Abstract—The quality of education is an important objective for governments and organizations around the world, as education has the potential to open new opportunities for people, improve their lives, and make them a productive and useful member of society. One of the ways to improve education is through the use of technology, and by making sure children have access to knowledge and information. In this paper, we will describe a distributed e-learning system based on item response theory, which can intelligently offer students appropriate questions and tasks to solve, and adapt the difficulty of the presented test items based on the students' previous answers and activity. Similarly, after each test item, the system presents the user with a solution to the problem, and the theory behind it. Our data and survey scores show positive results, and outline that students find the system engaging and want to continue to use it in the future.

Keywords—distributed systems, e-learning, item response theory.

I. INTRODUCTION

Governments and organizations throughout the world have recognized education as an important tool to reduce poverty, teach values, create a more productive and confident workforce, and to promote equality by empowering people from various underprivileged groups. Through education, people learn and acquire the necessary facts, principles, standards, norms and skills in order to be a successful, confident, useful and productive member of society. An educated workforce is crucial for economic growth.

The use of technology and e-learning is one of the main methods that allows us to bridge the gap between people that come from different backgrounds and socioeconomic status – i.e., to allow people to have similar access to learning opportunities. The ability of e-learning systems and technology tools to scale, to assist groups and individuals in learning at their own pace, to cover material according to an individual schedule, and to digest up-to-date content, are just some of the advantages of electronic learning. Therefore, it is very important to analyze their capabilities and usage in both formal and informal education. One of the current research areas in this field is related to the use of intelligent and adaptive e-learning systems, which have the ability to automatically adjust to students' capabilities [1].

In this paper, we will describe a new distributed, privacy-aware e-learning system based on item response theory, including data related to it's usage for improving students' scores in mathematics.

First, in Section II, we will look at present research related to distributed systems, privacy, item response theory and existing e-learning systems. Next, in Section III, we will present the architecture of the newly created e-learning system, as well as it's features. In Section IV, we will outline

data gathered from using the system in the course of one week, as well as results from a survey conducted among students who used the system during that period. In Section V, we will identify our ideas for future work - and finally, in Section VI, we will write a conclusion about our findings.

II. LITERATURE REVIEW

A. E-learning

The role that e-learning software plays in supporting and improving educational systems across the world continues to increase every year. This includes their usage in both primary schools, high schools, universities and informal settings like online courses and certifications. The global COVID-19 pandemic has expanded this growth even further, with some organizations and institutions even providing entire curriculums online. In some countries, like Macedonia, an entire school year was conducted online.

E-learning has several advantages compared to traditional learning, although in most cases they run in parallel – i.e., e-learning systems are commonly used to augment traditional learning. Blended learning is often considered to be a better and more effective alternative to both pure online classes and pure face-to-face classes [2].

E-learning systems offer students and teachers the ability to communicate and exchange messages asynchronously, reduces costs, and allows students to review and learn information at their own pace. Additionally, e-learning is scalable, (often) offers personalization features for the students, and analytics (data) for the teachers and administrators. This can lead to better learning outcomes, and better retention – as different students have different learning styles.

Learning management systems might be designed to handle various types of courses, or they might be targeting a specific area, market or expertise. For example, Moodle is often considered the most popular general-purpose Learning Management System [3], with some data showing that it has as much as 64% of the European LMS market share. It is an open-source platform, which is easy to setup and maintain. It offers a personalized dashboard for users, file management, calendar, notifications, multimedia, assessments, custom integrations (through plugins), and a lot more. From a technological point of view, it requires PHP and a working web server like Apache, as well as a relational database (MySQL, MariaDB or PostgreSQL).

An example of a LMS that is used in a specific area (computer programming) is MENDO [4], which allows students to read lectures which contain interactive elements, solve tasks with automatic grading, participate in competitions and forum discussions, etc.

Some systems, called intelligent e-learning or intelligent tutoring systems, are able to provide immediate and customized feedback to students, and can be utilized even when the physical instructor/teacher is absent [5]. Some examples of intelligent tutoring systems are those that can offer automatic and individual assistance, hints or adaptable task selection. In this paper, we are going to describe how our system automatically selects tasks for each student and enables them to learn and improve in the specific areas that are most important to them.

B. Distributed systems

Most e-learning systems are installed on a single web server, either with a relational database management system running on the same machine or in the same datacenter. For example, using Amazon Web Services, Moodle can be installed on a single EC2 instance, and then connect to a database running on RDS (in the same region). This kind of setup is very simple to configure, manage and has very little network overhead.

However, distributed systems – i.e., systems whose components might span across different machines and networks - have certain advantages, which include efficiency, scalability, and reliability, amongst others. In a distributed system, the autonomous nodes have local memory, and communicate and coordinate amongst each other through messages. Commonly, each node might have just a limited view of the system, or a specific function that it executes. In the latest version of the MENDO e-learning system for example, there are nodes called graders, whose job is simply to safely and securely grade a students' submission on a series of test cases (inputs and expected outputs). Because these nodes execute a single function, they might be automatically scaled up and down, as traffic increases or decreases.

A different usage of distributed systems is to share and synchronize content and courses among users, institutions or organizations [6], which leads to better content, network robustness, scalability, and improved privacy protections.

C. Privacy

Data privacy is important to both institutions and users. Recent regulations, such as, for example, GDPR, focus on empowering users in managing and protecting their data by giving them an option to intervene/control how and when their data needs to be removed, transparency (how data is being used), and minimization (i.e. systems should not ask for more personal information than is necessary for the purpose of the service) [7].

For e-learning, privacy norms also require that personal student information is only used for educational purposes, and that a student profile is only used in that context. Most software organizations also adhere to common security practices and follow guidelines and reports like OWASP Top 10 – a list of the most critical security findings and recommendations. Another common approach to tackling and limiting problems is to require and store only the information that is crucial for a system to provide the service for which it was designed – for example, if a student name or address is not required for an e-learning system to function correctly, then a simple username or student code can be used instead.

D. Item response theory

Item response theory is a paradigm of testing that focuses on items – as opposed to tests. IRT is concerned with the development of test items, where each item is not assumed to be equally difficult to solve. It is often considered as a superior alternative to classical test theory – for example, it is used in important tests like Graduate Management Admission Test (GMAT), Graduate Record Examination (GRE), Scholastic Aptitude Test (SAT), and others [8]. Often, for simplicity, we assume items to correspond to multiple-choice questions – but IRT can also work with scales and other complex scores.

For common tests used in schools, Item Response Theory leads to more accurate results with a smaller number of questions. For example, in Classical Test Theory, we can consider that the observed total score is based on the person true score (T) and a simple error (E), which is everything that is unrelated to the value we are measuring. On the other hand, we can consider IRT as a theory that allows us to create adaptive tests, and then to use the student's responses to improve the pool of test items. The standard mathematical model is based on the cumulative form of the logistic function – and based on the number of parameters, we can talk about a one parameter (1PL), two parameter (2PL) or three parameter (3PL) logistic model.

$$p_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta - b_i)}}$$

Here, for the three parameter model, we outline that the probability of a correct response to an item i , depends on the item difficulty b_i (also called location), discrimination (a_i), and chance or pseudo-guessing (c_i). In the 2PL model, we assume that the data has no guessing – so it's appropriate for items where guessing is unlikely (the student has to enter an "answer", or a multiple-choice question which has many options to select from). Likewise, the 1PL model assumes that all items are similar in terms of discrimination.

Using this general idea, we can start with an estimated student ability (center, based on previous test results, or ask the user about their opinion), calculate which test item from our pool of items would be best and give that item as the next question/task for the student to solve. After the student completes the first test item, we update their estimated ability based on all previous responses, calculate the next best test item and repeat the procedure. This process continues until 1) the student completes a fixed number of questions, or 2) alternatively, the test might be limited in time.

In other words, when a student starts taking a test, an item of average difficulty can be presented to them. If they answer it correctly, they will get a more difficult item – or, if they answer incorrectly, they will get an easier item. This is the basic idea of Computerized Adaptive Testing (CAT), and many such systems rely on IRT.

One of the main goals of Item Response Theory is to update or improve the measurement accuracy and reliability of tests, either dynamically or using previously collected knowledge about student responses to certain test items. Comparably, with Classical Test Theory, reliability is much harder to estimate, because that would require knowledge about the student's true scores (minus errors).

In Item Response Theory, several assumptions are made in order to simplify the logic: 1) that a single trait, attribute or

ability decides the item responses and that it can be measured on a scale, 2) the local independence of items, and 3) that the response can be modeled by a mathematical item response function.

Some disadvantages of Item Response Theory are that it's relatively harder to compute and explain by hand (in comparison to scoring with classical test theory, which doesn't require any complex formulas), so therefore IRT tests are almost exclusively conducted with the help of computers. Additionally, IRT tests require a much larger database of test items, and a much larger number of students in order to obtain stable and valid results (the model and values improve over time).

Because of their ability to adapt to students, IRT is useful for developing and improving intelligent e-learning (tutoring) systems. Often, in e-learning systems, instructions are combined with assessments in order to increase retention in students, and to continuously evaluate and improve student models. For example, test items might be tagged/linked to certain instructions and lectures, or there might be a separate agent in the system which provides learning recommendations based on the IRT test scores (as shown in [9], where artificial neural networks are used for that purpose). The goal is for the e-learning system to guide the student to the correct instructions, and to adapt to their knowledge and skills – similarly to a human instructor.

III. ARCHITECTURE AND FEATURES

We have developed a distributed e-learning system based on item response theory, which allows students to learn different subjects and signup with either their name, email or – for privacy reasons - a username of their choosing (the choice can be made in consultation with their teachers). The system is composed of one main server component, developed with Node.js (using Express as the web framework) and utilizing a PostgreSQL relational database management system, where the main data is located (such as, for example test items), as well as gathered (anonymous) information from students utilizing the system. As a content management system (allowing both creation and easy editing) we decided to use WordPress. This content is then loaded in the main system by detecting changes through a sitemap.

For privacy reasons, and in order to make sure the system can be used even when schools have connectivity issues, each organization or school can run a separate web agent on their premises, which students connect to. These separate agents download the latest test items at configurable regular intervals (for example, daily), and can (optionally) send anonymized data to the main server component in order to improve the test items. The agents are also developed with Node.js (Express), but they utilize an SQLite database management system so that installation is easy. Both the main server and the web agents can be run inside Docker containers, and that is the preferred way of managing and running them.

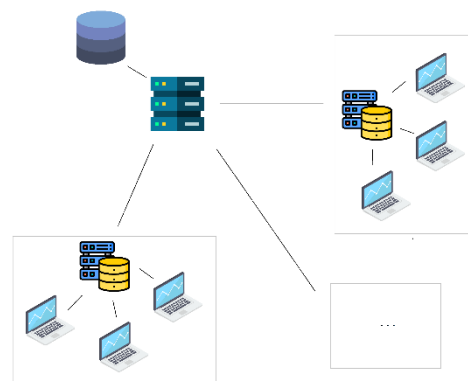


Fig. 1. High-level view of the system architecture

The client side was developed with Bootstrap and jQuery. We have made this decision based on our previous experience running other e-learning systems in Macedonia, and our gathered data showing that some operating systems and browsers used in schools are extremely outdated, so special care must be taken in order to make sure that students can use these systems both from home and from school.

Local storage is used to store a token that can be used to authorize and authenticate users in the system, and no tracking cookies or outside scripts are ever loaded when a student opens the web application. Both the main server system and the web agents use extensive caching, and we have tested and verified that the system can sustain more than a thousand requests per second – which can correspond to tens of thousands of active users at the same time for one agent (this number can be even larger, but it's dependent on the test items themselves, and the average time that users take to solve them – in addition to the server hardware). In our tests with real students, for which data is presented in the next section, we ran the main server system on Hetzner Cloud in one virtual private server, and the resource usage was minimal. This is one of the advantages of the event-driven and asynchronous nature of Node.js, making it suitable for real-time web applications which can sustain a very large number of active users at the same time.

The system uses AJAX requests in order to simplify the user interface and guarantee the best experience for the students. In the first version of the system, we only support one language (Macedonian), but both WordPress (used as a content management system) and the web server can easily be extended to support more languages – the main workload is actually to translate the entire database of test items.

Based on the architecture described above, and using item response theory, the system shows students test items (most of which are multiple-choice questions). After a student answers a question, (regardless if their answer was correct or not) the system shows them the correct answer and brief explanation of the idea, theory and solution (We have tried to keep this content brief, under 3 paragraphs, and to include images whenever possible to make learning more fun and engaging). Afterwards, based on the student response to the test item (correct or not), the system will offer them an easier or harder question. Because answers are saved in the agent database, the student can take breaks and use the system on different days if necessary – this is clearly one of the use cases for the system – i.e., for it to be used as an intelligent tutoring system. Before offering the student another question – the

system might decide to go with a question that might be slightly easier or harder than the optimal question to ask at that point, in order to guarantee that two students don't get the same list of questions if they make exactly the same answers (for example, imagine two students sitting next to each other, getting the same first question and making the same answer – to prevent cheating or gaming that system, in this case, our system might give a different question – which is based on slight randomization in the algorithm).

Note: It should be pointed out that in our tests and gathered data, the agents weren't running in schools, but on different Docker containers in one Virtual Private Server – with several subdomains pointing to a proxy running on the server, which then routed traffic to the correct agent/container. This was done in order to simulate a real working environment, but without asking schools and teachers to actually setup web servers and networks in their schools, as they might not be able to or have the necessary time or skills. Therefore, the system can be run either in schools, or on other servers as required.

IV. DATA AND SURVEYS

A group of 236 students participated in this research, 107 of them were part of the control group, and 129 tested the newly created e-learning system. Originally, 310 students signed up for participation, split evenly and randomly into the two groups, but some of them didn't complete all the necessary steps and surveys (outlined below) to be taken into consideration. Considering this was a weeklong experiment and didn't influence the students' grades (they were simply offered a certificate for participating in the experiment), it was expected that some of them wouldn't complete everything that was planned. The difference between the students from the control group completing the experiment (69%) versus the students using the system completing the experiment (83%) might indicate that the students found the e-learning system as offering real value for them and which will help them learn faster, but it might also be explained by temporary interest in something that is new to them.

Both the students from the control group and the students using the system were offered a web application with tests and study materials related to mathematics (due to the difficulty associated with creating a large set of test items and content related to their solutions, we decided to initially focus on one subject). The architecture and features of the newly created system were described in the previous section. For the control group, apart from test questions, the students had both pdf versions of learning materials and web pages containing the same lectures (as embedded pdf viewers) that they might reference while learning (to simulate standard learning from home). They were able to easily jump to any material or chapter after working on any question/task, or simply continue with the next one. Note that this is radically different than what the newly created system does – which automatically guides the students by choosing the tasks most appropriate for them, and then explaining the solution and brief theory behind it (all of which is previously created).

After the week-long experiment, students were asked if they agree with the premise that the web application that they used would be useful as an addition to traditional school learning. The results clearly indicate that students that used the newly created system showed much higher enthusiasm

and the scores in that group were much better compared to the control group.

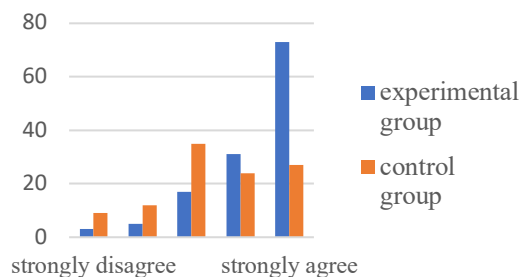


Fig. 2. Students' opinion about the willingness to continue using the web applications

Similar results were obtained when students were asked if they would like to use the system with other subjects besides mathematics, with the total scores being 4.132 for the group using the new system (if we consider 1 to indicate strong disagreement, and 5 strong agreement), compared to 3.318 for the control group using a simple web application with a classical testing functionality and static content.

Google Forms was used to conduct the survey, and students weren't asked for any personal information. Additionally, at the start of the form there was a message indicating that the survey is conducted anonymously.

Next, by analyzing the final data, we have noticed that students that used the newly created system spent 31% more time studying, which is a very significant difference. It should be noted that this estimate is based on the timestamps of submitted answers to test items, because we intentionally didn't measure other data and didn't track other student activities like the amount of time they stay on a certain web page, mouse movements, scrolling or clicks.

On the final day, we sent a test with 10 questions to both groups, and asked them to complete it in 30 minutes by themselves. These questions were actually taken from a real mathematics test, so that we don't pick questions ourselves and (possibly) introduce bias in the results. The experiment group showed better results than the control group, with the averages being 47 points and 41 points respectively (i.e. the experiment group had around 15% better scores). The maximum number of points was 100. Note that the student scores are positively affected by the fact that all tasks were converted to multiple-choice questions, and negatively affected by the fact that these tests didn't really influence the students or their grades in any way, so it is expected that they prepare much more for real tests later on.

Finally, with regards to the system itself, no significant downtime or issues were discovered during the testing period. UptimeRobot was used for monitoring, showing 100% uptime and expected/low response times. Similarly, the tracked cpu, memory and disk utilization was low – as expected. We also verified the communication between the web agents and the main server, and deliberately simulated communication errors (both by changing the api key, and blocking traffic), and all of them showed no errors or downtime. As soon as we reverted the api key and resumed traffic, the next scheduled data synchronization completed successfully.

V. FUTURE WORK

In the future, we plan to test this system with other subjects as well – we are already in the process of testing students in physics (where we also have the ability to add interactivity, and make a system that is much more interactive compared to reading a book or pdf). Similarly, we have plans for preparing materials for other subjects and classes in the near future. Although the system is designed in such a way that it can be utilized easily no matter the content - preparing a series of questions and solutions, as well as making the initial assessments of difficulty and discrimination, is extremely time consuming.

Additionally, the system currently only supports one language, and we plan to add support for more languages in the near future. Plans for future research involve adding more interactivity, notifications and viewing how they impact students, an ability for them to share their progress and possibly earn badges, and more.

VI. CONCLUSION

E-learning systems enable teaching with the help of electronic technology and resources, and are commonly used in combination with traditional methods of learning. There are several well-known learning management systems, some of which are open-source and freely available. Some systems, called intelligent tutoring systems, aim to provide automated and adaptable feedback to users, without the help or intervention from a human teacher. Privacy is also a key factor when talking about any web-based systems, but this is particularly important with regards to children and their education.

In this paper, we presented a distributed e-learning system based on item response theory. The system uses WordPress for preparing and managing content, and then functions through a main server component, and several agents which can be run either in the cloud or directly in schools. Clients connect to the agents, which can function even when they don't have communication with the main server component, as they store data locally in a SQLite database, and only communicate with the main server to fetch the latest test items, and to submit anonymized data so that the system can improve. Our tests and surveys indicate a very positive reception among students, and willingness to use such a system in addition to their traditional learning methods.

REFERENCES

- [1] Jando, Emanuel, Achmad Nizar Hidayanto, Harjanto Prabowo, and Harco Leslie Hendric Spits Warnars. "Personalized E-Learning model: A systematic literature review." In 2017 International Conference on Information Management and Technology (ICIMTech), pp. 238-243. IEEE, 2017.
- [2] US Department of Education OoP, Evaluation, Policy Development. Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies. Washington: Office of Planning, Evaluation and Policy Development, 2009.
- [3] Cavus, Nadire, and Teyang Zabadi. "A comparison of open source learning management systems." *Procedia-Social and Behavioral Sciences* 143, pp. 521-526, 2014.
- [4] Kostadinov, Bojan, Emil Stankov, Marija Mihova, and Marjan Gusev. "State Competitions in Informatics and the Supporting Online Learning and Contest Management System with Collaboration and Personalization Features MENDO", 2013.
- [5] Bhattacharya, Suman, and Sayan Nath. "Intelligent e-learning systems: An educational paradigm shift.", 2016.
- [6] Van Steen, Maarten, and Andrew S. Tanenbaum. "A brief introduction to distributed systems." *Computing* 98, no. 10, pp 967-1009, 2016.
- [7] Zeide, Elana, and Helen Nissenbaum. "Learner privacy in MOOCs and virtual education." *Theory and Research in Education* 16, no. 3, pp. 280-307, 2018.
- [8] Chen, C. M., Liu, C. Y., & Chang, M. H. Personalized curriculum sequencing utilizing modified item response theory for web-based instruction. *Expert Systems with Applications*, 30, pp. 378–396, 2016.
- [9] Baylari, Ahmad, and Gh A. Montazer. "Design a personalized e-learning system based on item response theory and artificial neural network approach." *Expert Systems with Applications* 36, no. 4, pp. 8013-8021, 2009.

System for reproducible code environments - Dohrnii

Dimitar Mileski

Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering
Skopje, North Macedonia
dimitar.mileski@ieee.org

Marjan Gusev

Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering
Skopje, North Macedonia
marjan.gushev@finki.ukim.mk

Abstract—The reproducibility of code and the whole environment in which that code needs to be executed takes up much of the time of software engineers, computer scientists, and anyone who writes code. There are several reasons why we can not get the same results if the original data is available to us. This may be because the environment is not configured the same way as the source code author, we do not have enough knowledge of the technology used, poorly written documentation, lack of documentation, different versions of libraries, tools, development frameworks, mistakes made by one who wants to reproduce the code or there are errors in the code or errors that occurred after a certain period of time. In this paper, we introduce a new system for reproducible coding environments called Dohrnii, as a cloud-based solution. The main purpose is to save the time lost when we want to reproduce the results of a project for the first time. Every Dohrnii environment contains a video, description, instance (virtual machine), resources, and evaluation. This means that with this system it will be possible to reproduce environments in which the code needs to be executed, for that there will be a video where it will be shown how to reproduce the code, instance (virtual machine), description, additional resources, and part for an evaluation to determine if what is in the video corresponds to what is in the instance.

Index Terms—code reproducibility, code environment, web platform, cloud computing

I. INTRODUCTION

The reproducibility of the code is the process by which we will get the same results as the author of the code if we have access to the original data [1].

It takes time to configure and set up the environment when we try to reproduce code, to get the same results using the original data. In many cases, we spend a lot of time so that in the end we can not reproduce the results even though we have the original data. This is a big problem because software engineers, computer scientists, and anyone who writes code spends a lot of time trying to reproduce the results. Part of the job of software engineers and computer scientists is to try out different projects. Many hours and even days are spent trying to reproduce some code, which in the end may not be reproducible. There are several reasons why we can not get the same results as the author of the paper, or the project.

A. Code reproducibility problems

There are a number of errors that can lead to not reproducing the results. This is because the environment in which the code is executed is not the same as the environment in which the author of the code executed it and obtained the results. Common problems are identified by:

- not being familiar with technology,
- not having the opportunity to take the time to learn new technology,

- trying to see the end results, insufficient expertise in the field.

The problem is usually found in a new programming language, a new library, or a new software development framework. Extensive and complex documentation, poorly written documentation, and often incomplete or no documentation at all can be a big problem. Different versions of libraries, programming languages, and software development frameworks are one of the most common reproducibility issues. Software version incompatibility is often a problem. In order to solve this problem, the author of the source code needs to write text documentation write requirements, Docker File [2], or similar configuration files where the versions will be documented. Often package management tools do automatic version updates which can cause conflicts. The one who wants to reproduce the code also makes mistakes. In many cases, there are errors in the code from the start, and we spend time trying to reproduce the results from the code that is not working. Library support ends after a certain period, there are errors in the library or others problems with the libraries used. Some systems are not reproducible because they depend on a service that may have problems and cause errors. In all these cases, the computer scientist, the computer engineer, and all those who write the code will spend time without getting the same results as the author of the project.

B. Methods to reproduce the results

When we try to reproduce code we first see if the author of the code has documentation that we can follow. If it is not there or it is not complete, problems with reproducing the code are possible. In such cases, we can ask for help from the official documentation of the programming language, the library, and the development framework. More and more video platforms are used on which there are videos showing software development. Forums and blogs are also used so that we can solve a certain problem or replace part of the code. If the problem is in the environment in which the code is executed, which is initially set before the code is executed over the data set, configuration files, installations, tools, programs, integrated development environments, development frameworks, and libraries can solve the problem, and finally reproduce the code. If the code has documentation that lists the libraries used and their versions and all the details about the environment in which the code needs to be executed, we can reproduce the results.

For some projects, library version file, such as requirements.txt, package.json, pom.xml, and others is sufficient. Docker containers [2] make the environments in a container easily reproducible. Google Colab [3], and Jupyter Notebook

[4], are also platforms for reproducibility of code that is stored and executed in the cloud. The problem arises because the code and data are often not sufficient to obtain the same results as those obtained by the author. The environment in which the solution is developed is different from the environment in which we try to reproduce the results.

C. How the Dohrnii system solves the problem

The system for reproducible code environments (Dohrnii) tries to solve the problem with a unique combination of the previous methods. The system is based on Dohrnii environments. Each environment is composed of a video recorded by the author of the code using the screen recording tool on the Dohrnii platform. The video explains how to reproduce the results.

Short description of the Dohrnii environment, a description of the code, the data set, and the results to be obtained at the end. Each environment has link to the code repository. There is a list of additional resources. This section lists links to additional resources that are not in the code repository. It could be a link to a data set, a link to documentation, or some other useful link. All previous parts of the Dohrnii environment will help us to be able to reproduce the code in a virtual machine created by the Dohrnii system itself, and that virtual machine is in the cloud. SSH can be used to connect to the instance (virtual machine) [5] SSH Web client of choice or one proposed in Dohrnii system. If the instance has a graphical user interface, then we can connect to with Remote Desktop [6], Web or Desktop version of Remote Desktop [6]. Finally, in each Dohrnii environment, there is a section for evaluation. In the evaluation section, two answers are possible, the environment is reproducible or the environment is not reproducible, i.e. it means that what is in the video corresponds to the current state of the instance (virtual machine). This combination of screen recording, description, instance (virtual machine), code repository, additional resources, and evaluation section should provide a Web-based, cloud-based solution for reproducible code environments.

The paper is further organized as follows. An overview of similar systems for reproducible coding environments is given in Section II. The system is described in Section III with details on architecture, implementation technologies, and functional description. Section IV addresses the conclusions, including a discussion on advantages and disadvantages, and reveals a perspective on future work.

II. RELATED WORK

This section will look at existing solutions that try to solve the problem of reproducibility. The Jupyter Project [4] is a project and community that aims to develop open-source software, open standards, and interactive computing services for multiple programming languages. The Jupyter project is made up of several subprojects such as Jupyter Notebooks [4], Jupyter Hub [7], Jupyter Lab [8]. Jupyter Notebook is an open-source web application that allows you to create and share documents with code, equations, visualizations, and text. Usage includes data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and more. Jupyter Notebooks if it is hosted can be a good platform for reproducibility. JupyterHub is a platform for hosting Jupyter Notebooks. It is used by students, scientists, and software developers.

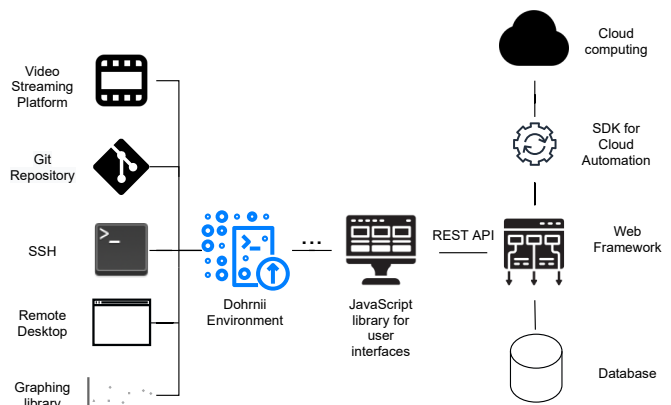


Fig. 1. System Architecture

Another project based on Jupyter is Binder [9]. Binder is an open-source web service that allows users to create interactive, reproducible cloud environments. Some of the technologies included in Binder are JupyterHub, Kubernetes [10], Repo2Docker [11]. Binder works with pre-existing community projects to create interactive versions of code repositories that exist on sites like GitHub [12].

Another cloud system based on Jupyter projects that provides computer resources such as GPU and TPU is Google Colab [13]. Collaboratory, or Colab for short, is a Google Research product that allows developers to write and execute Python code through their browser. These are Jupyter notebooks that do not require installation and have a free version that gives free access to Google's computer resources, such as GPUs and TPUs. Another such project is JupyterBook [14]. Jupyter Book is a collection of Jupyter notebooks.

There are also interactive environments for executing code in the field of Web development. Some of the client-side code development environments are:

- CodeSandbox is an online web development editor. CodeSandbox [15] is a tool for quickly building prototypes, experimenting, and sharing code with other developers. Used to create static Web pages, Web applications as a whole, or components of any Web browser device.
- JSFiddle [16] is an online service, integrated development environment, and community for testing and displaying user-generated code snippets. HTML, CSS, and JavaScript code are written and shared via JSFiddle. JSFiddle also enables the simulation of AJAX calls.
- CodePen [17] is a system for testing and displaying code snippets created by developers. It functions as a code editor and open-source learning environment, where developers can create snippets of code called "pens" and test them.

III. DESCRIPTION OF THE DOHRNII SYSTEM

A. System architecture

Figure 1 shows the system architecture with the following components:

- **Dohrnii environment** - Dohrnii environment is an abstraction of all the technologies needed to have a reproducible project.
- **Video platform** - Screen recording and instructions on how to reproduce the environment are uploaded on a

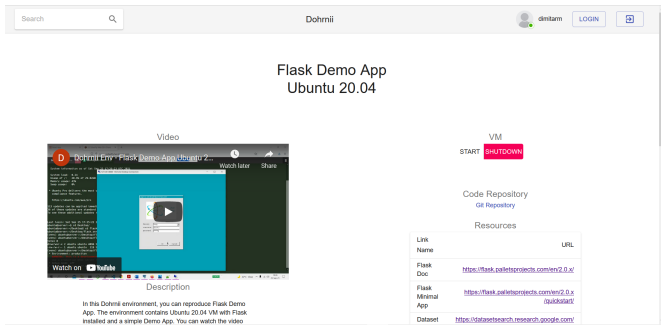


Fig. 2. Dohrnii environment - video, description, VM controls, code repository, resources

video platform. These are external video platforms that are not part of Dohrnii.

- **Git repository** - Code is versioned with Git [18] and is hosted by external Git repository hosting services.
- **SSH** - SSH connections to instances (virtual machines) are made with Web-based SSH clients, which are free to use and open source.
- **Remote Desktop** - In order to be able to see the graphical user interface of the instances (virtual machines), we need to use the Remote Desktop client. SSH and Remote Desktop clients may not be Web-based, but Web-based RD clients are recommended to have a fully Web-based system.
- **Visualization Library** - Graph for the evaluation section of the Dohrnii environment is plotted with the help of visualization library.
- **Javascript library** - The client application is built with Javascript library, which uses the REST API [19] to perform basic CRUD operations, create, read, update and delete. This library includes other packages such as a screen capture package using the Screen Capture API [20], an icon package, and a Bootstrap [21] CSS framework.
- **Web framework** - The web framework should be compatible with the database. The choice of the framework also depends on the programming language in which the framework is written and the programming language of the library for cloud automation, starting and stopping instances (virtual machines) in the cloud.
- **Database** - The database stores user information so that we can implement authentication and authorization. Data for Dohrnii environments such as title, description, evaluation section data, link to Git repository, and links to additional resources are persisted in the database.
- **Cloud computing** - Cloud Instances (virtual machines) are created, started, stopped, and cloned for the users to reproduce the results.
- **Cloud computing automation** - The Automation Part uses a Software Development Kit (SDK) for automating cloud operations.

B. Implementation technologies

Technologies used for each of the components of the system architecture are given in this section. As shown in Figure 2, the Dohrnii environment contains title, video, description, instance (virtual machine) controls, code repository and resources.

- **Video platform** - We use YouTube for the video platform in the Dohrnii [22]. YouTube is a Google-owned video-sharing platform for the Internet and social media. The advantage of an existing platform is that users can monetize videos through existing platforms. The user who creates the Dohrnii environment should upload the screen recording with the code reproducibility instructions on the YouTube platform and insert the link to the YouTube video when creating the Dohrnii environment.
- **Git repository** - A link to GitHub is placed in each Dohrnii environment. One Dohrnii environment can have multiple code repositories. Other services such as GitLab [23] and BitBucket [24] can also be used.
- **Visualization Library** - To display the evaluation section shown in Figure 3 of the Dohrnii environment we will use the Plotly [25] library for data visualization. Plotly is an open-source library for data visualization. Plotly libraries are available for multiple programming languages including Python, Julia, R, ggplot2, MATLAB, and FSharp. We will use the plotly.js React component from Plotly to visualize the Dohrnii environment evaluation section.
- **Screen capture** - Figure 4 shows the screen capture interface for creating the Dohrnii environment. The Dohrnii screen capture uses a React js component called the React-media-recorder [26]. React-media-recorder is a React component that can be used to record audio, video, and screen. React-media-recorder uses the MediaRecorder API [27]. The MediaRecorder interface of the MediaStream recording API provides easy media recording functionality. Created using the MediaRecorder () constructor. An option that is important to us is that the recorded video is available for download.
- **Javascript library** - The React js [28] library is used to create the client application. React (also known as React.js or ReactJS) is a free, open-source JavaScript library for building user interfaces or components. React can be used as a basis for developing single-page applications or mobile applications. The most important React components we have used are React-media-recorder component and Material-UI. Material-UI [29] is a simple and customizable component library for building React applications.
- **Web framework** - We have used Django for the Web framework [30]. Django is a free web framework written in Python that follows the model-template-views (MTV)

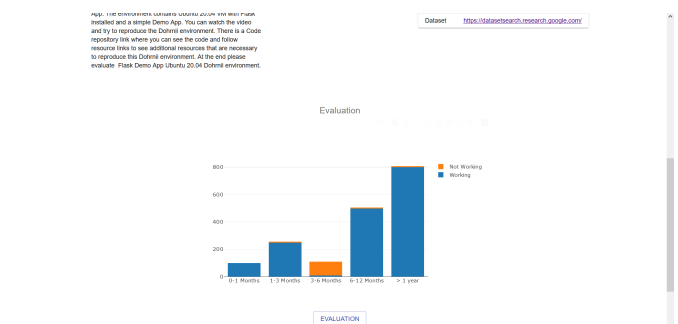


Fig. 3. Dohrnii environment - evaluation

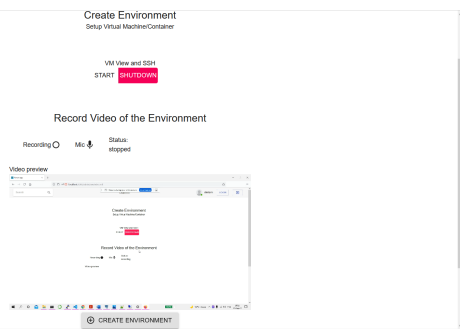


Fig. 4. Create Dohrnii environment view

architecture. The framework emphasizes the usability and "connectivity" of the components, less code, fast development, and the principle of non-repetition. Python is used for settings, files, and data models. Django also provides an optional administrative interface for creating, reading, updating, and deleting dynamically generated data.

- **Database** - PostgreSQL is used as the database in the Dohrnii system [31]. Django has built-in support for PostgreSQL which facilitates the process of developing Django applications with PostgreSQL. PostgreSQL, also known as Postgres, is a free and open-source relational database management (RDBMS) that emphasizes scalability and is SQL compliant. PostgreSQL features transactions with properties of atomicity, consistency, isolation, endurance (ACID), automatic updates, materialized views, external keys, and stored procedures. It is designed to handle a wide range of workloads, from single machines to data warehouses or Web services with multiple concurrent users.
- **Cloud computing** - The instances that are created, started, and terminated on the Dohrnii system are AWS EC2 [32] instances. Amazon Elastic Compute Cloud (EC2) is part of Amazon.com's cloud computing platform, Amazon Web Services (AWS), which allows users to rent virtual machines to run their own computer applications. EC2 encourages the scalable deployment of applications by providing a Web service through which the user can upload an Amazon Machine Image (AMI) to configure a virtual machine, which Amazon calls an "instance" that contains the desired software. The user can create, start and terminate instances in the cloud.
- **Cloud computing automation** - The AWS SDK for

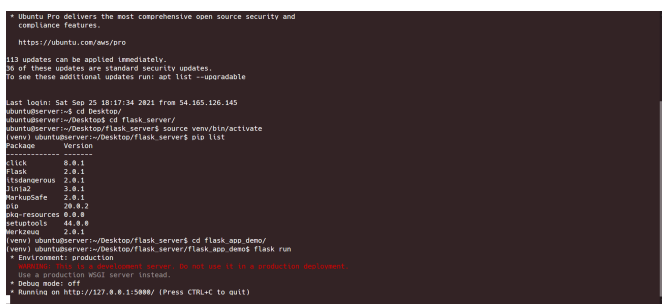


Fig. 5. SSH to Dohrnii environment

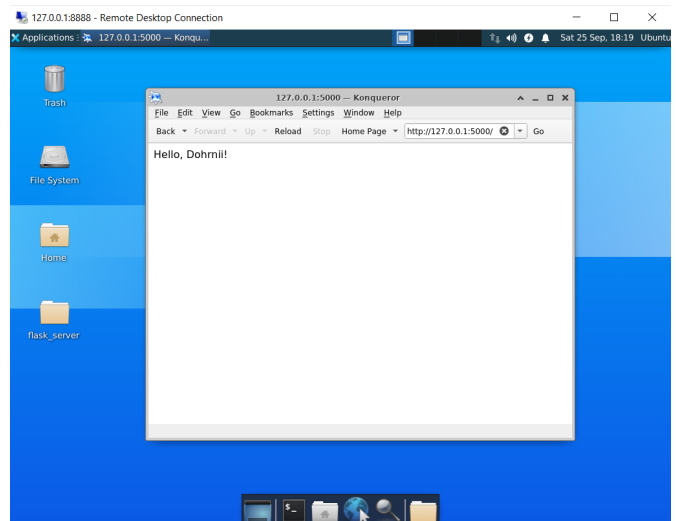


Fig. 6. Remote desktop connection to Dohrnii environment

Python (Boto3) [33] is used to create, configure, and manage AWS services such as Amazon Elastic Compute Cloud (Amazon EC2) and Amazon Simple Storage Service (Amazon S3). The SDK provides an object-oriented API as well as low-level access to AWS services. Boto3 is a software development kit for Amazon Web Services (AWS) using the Python programming language, which allows Python developers to write software that uses services such as Amazon S3 and Amazon EC2. Listed below are just some of the features of the AWS SDK for Python (Boto3). Manage Amazon EC2 instances, get basic information about your Amazon EC2 instances, start and stop detailed instance monitoring of Amazon EC2, start and stop an instance of Amazon EC2, restart an instance of Amazon EC2, working with Amazon EC2 key pairs, get information about your key pairs, generate a pair of Amazon EC2 instance access keys, delete an existing pair of keys, working with security groups in Amazon EC2, get information about your security groups, create an Amazon EC2 instance access security group, delete an existing security group.

- **SSH** - Figure 5 shows Sshwifty, SSH and Telnet [34] client made for the Web. Can be deployed to our server to provide SSH and Telnet access interface for any compatible (standard) Web browser. Sshwifty is an open-source Web SSH client that allows you to connect to SSH on a remote machine without downloading additional software.
- **Remote Desktop** - Remote Desktop Protocol (RDP) is a commercial protocol developed by Microsoft that provides the user with a graphical interface for connecting to another computer over a network connection. The user in Figure 6 uses RDP client software for this purpose, while the instance (virtual machine) must be an RDP server. RDP client exist for most versions of Microsoft Windows (including Windows Mobile), Linux, Unix, macOS, iOS, Android, and other operating systems. RDP servers are built into Windows operating systems. An RDP server for Unix and OS X also exists. By default, the server listens on TCP port 3389 and UDP port 3389. To use RDP with Dohrnii environments it

is recommended to use Web RDP clients to have a completely web-based experience.

C. Functional description

The Dohrnii system is designed to be used by a variety of users who write code from developers, software developers, computer scientists, electrical engineers, scientists in other fields who write code.

Examples of use cases for the Dohrnii code reproducibility system include:

- Users can browse the system and reproduce the code in the environments. Developers who work in a company can share the environments so that they can reproduce a project. These could be projects in Web Development, Artificial Intelligence, or other areas.
- Professors, assistants, and laboratory assistants can insert link to the Dohrnii environment at the end of the presentation or exercise so that students can reproduce the results.
- Scientists can use Dohrnii environment as an open-source environment and enable others to reproduce the results. The link from the Dohrnii environment should be placed on the platforms for hosting scientific papers, or on the servers for the preprinted versions of the papers.
- Developers who ask questions on one of the software development question platforms, where the answer with the most votes to the question can be a link to the Dohrnii problem-solving environment or there is an environment that has a similar solution for the problem.
- Users can also access the Dohrnii system from existing code repositories with a link to the Dohrnii environment.
- Content creators on one of the video platforms that create videos in the field of computer science and computer engineering to put a link in the video description to the Dohrnii environment.

Use of the Dohrnii system requires basic knowledge and understanding of SSH, Remote Desktop Protocol, SSH Tunneling, and Port Forwarding.

IV. CONCLUSION

A new system for reproducible coding environments Dohrnii was introduced. The main purpose of the system is to save the time lost when we want to reproduce the results of a project.

The system was developed with sophisticated cloud-based architecture supported by modern technologies.

The advantages of the Dohrnii system are that it is a complete Web-based solution that uses cloud computing. All the tools needed to successfully reproduce an environment are available in the Web version. Such are screen capture, SSH, and Remote Desktop, clients.

The biggest disadvantage of the Dohrnii system is that it requires a large amount of resources. Because to reproduce the environment we use instances (virtual machines) that need working memory, storage, and processing resources. Disadvantages of the system include the disadvantages of virtual machines as a technology. The cost of creating and launching multi-user instances can be very high, depending on the cloud. The system is intended to be able to support the concurrent creation, start-up, and termination of instances for more users at the same time and the demand for the resources can be high, thus the cost of operating the system.

The Dohrnii system now uses external SSH Web clients and external Web Remote Desktop clients. This is a potential security issue for the system as instance connection information will be shared with systems other than Dohrnii.

Possible improvements to the system include additional development including cloud automation with SDK Boto3. There would be a limit to the number of instances that can be run by a single user, we can also limit the time that each instance has from the time it is started in order to save resources. One of the priorities would be to have an SSH client built into the system. Additionally, SSH clients can be deployed supporting the free and open-source. Remote Desktop client built into the system to avoid using external clients. Consider possible ways to improve visualization in the Dohrnii environment evaluation section. Development of a built-in program for editing screen recording. That program would be Web-based and would have only basic features like trimming the screen recording and inserting text over the video itself if something needed to be further explained.

REFERENCES

- [1] V. Stodden, "The scientific method in practice: Reproducibility in the computational sciences," 2010.
- [2] C. Anderson, "Docker [software engineering]," *Ieee Software*, vol. 32, no. 3, pp. 102–c3, 2015.
- [3] M. J. Nelson and A. K. Hoover, "Notes on using google colaboratory in ai education," in *Proceedings of the 2020 ACM conference on innovation and Technology in Computer Science Education*, 2020, pp. 533–534.
- [4] F. Pérez and B. E. Granger, "Ipython: a system for interactive scientific computing," *Computing in science & engineering*, vol. 9, no. 3, pp. 21–29, 2007.
- [5] T. Ylonen, "Ssh—secure login connections over the internet," in *Proceedings of the 6th USENIX Security Symposium*, vol. 37, 1996.
- [6] R. Rahim, J. Simarmata, A. Purba, M. A. Prayogi, A. Sapta, O. Krianto, and S. Suharman, "Internet based remote desktop using indy and socket component," *Int. J. Eng. Technol.*, vol. 7, no. 2.9, pp. 44–47, 2018.
- [7] L. Fernández, R. Andersson, H. Hagenrud, T. Korhonen, E. Laface, B. Zupanc *et al.*, "Jupyterhub at the ess. an interactive python computing environment for scientists and engineers," in *This conference*, 2016.
- [8] B. Granger and J. Grout, "Jupyterlab: Building blocks for interactive computing," *Slides of presentation made at SciPy*, vol. 2016, 2016.
- [9] B. Ragan-Kelley and C. Willing, "Binder 2.0-reproducible, interactive, sharable environments for science at scale," in *Proceedings of the 17th Python in Science Conference (F. Akici, D. Lippa, D. Niederhut, and M. Pacer, eds.)*, 2018, pp. 113–120.
- [10] D. Bernstein, "Containers and cloud: From lxc to docker to kubernetes," *IEEE Cloud Computing*, vol. 1, no. 3, pp. 81–84, 2014.
- [11] J. Forde, T. Head, C. Holdgraf, Y. Panda, G. Nalvarete, B. Ragan-Kelley, and E. Sundell, "Reproducible research environments with repo2docker," 2018.
- [12] L. Dabbish, C. Stuart, J. Tsay, and J. Herbsleb, "Social coding in github: transparency and collaboration in an open software repository," in *Proceedings of the ACM 2012 conference on computer supported cooperative work*, 2012, pp. 1277–1286.
- [13] T. Carneiro, R. V. M. Da Nóbrega, T. Nepomuceno, G.-B. Bian, V. H. C. De Albuquerque, and P. P. Reboucas Filho, "Performance analysis of google colaboratory as a tool for accelerating deep learning applications," *IEEE Access*, vol. 6, pp. 61 677–61 685, 2018.
- [14] D. Mileski, M. Jovanovik, and D. Trajanov, "Semantic web and data science integration using computational books," 2021.
- [15] "Codesandbox, together." [Online]. Available: <https://codesandbox.io/>
- [16] JSFiddle. [Online]. Available: <https://jsfiddle.net/>
- [17] "Codepen," Feb 2022. [Online]. Available: <https://en.wikipedia.org/wiki/CodePen>
- [18] S. Chacon and B. Straub, *Pro git*. Springer Nature, 2014.
- [19] M. Masse, *REST API design rulebook: designing consistent RESTful web service interfaces*. " O'Reilly Media, Inc.", 2011.
- [20] "Screen capture api - web apis: Mdn." [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/API/Screen_Capture_API
- [21] J. Spurlock, *Bootstrap: responsive web development*. " O'Reilly Media, Inc.", 2013.
- [22] J. Burgess, "Youtube," in *Oxford Bibliographies Online*, L. H. Meyer, Ed. United Kingdom: Oxford University Press, 2011, pp. 1–1. [Online]. Available: <https://eprints.qut.edu.au/46719/>

- [23] J. M. Hethey, *GitLab Repository Management*. Packt Publishing, 2013.
- [24] Atlassian, “The git solution for professional teams.” [Online]. Available: <https://bitbucket.org/product>
- [25] J. M. Perkel *et al.*, “Data visualization tools drive interactivity and reproducibility in online publishing,” *Nature*, vol. 554, no. 7690, pp. 133–134, 2018.
- [26] “React-media-recorder.” [Online]. Available: <https://www.npmjs.com/package/react-media-recorder>
- [27] “Mediarecorder - web apis: Mdn.” [Online]. Available: <https://developer.mozilla.org/en-US/docs/Web/API/MediaRecorder>
- [28] E. Elrom, “React router and material-ui,” in *React and Libraries*. Springer, 2021, pp. 79–113.
- [29] “The react component library you always wanted.” [Online]. Available: <https://mui.com/>
- [30] A. Holovaty and J. Kaplan-Moss, *The definitive guide to Django: Web development done right*. Apress, 2009.
- [31] B. Momjian, *PostgreSQL: introduction and concepts*. Addison-Wesley New York, 2001, vol. 192.
- [32] S. Ostermann, A. Iosup, N. Yigitbasi, R. Prodan, T. Fahringer, and D. Epema, “A performance analysis of ec2 cloud computing services for scientific computing,” in *International Conference on Cloud Computing*. Springer, 2009, pp. 115–131.
- [33] “Boto3 documentation.” [Online]. Available: <https://boto3.amazonaws.com/v1/documentation/api/latest/index.html>
- [34] J. Postel, “Telnet protocol specification,” in *RFC 854, ISI*. Citeseer, 1983.

logs2graphs: Data-driven graph representation and visualization of log data

Stefan Andonov

*Faculty of Computer Science
and Engineering*

*Ss. Cyril and Methodius University
Skopje, North Macedonia*

stefan.andonov@finki.ukim.mk

Viktor Jovev

*Faculty of Computer Science
and Engineering*

*Ss. Cyril and Methodius University
Skopje, North Macedonia*

viktor.jovev@students.finki.ukim.mk

Aleksandar Kitanovski

*Faculty of Computer Science
and Engineering*

*Ss. Cyril and Methodius University
Skopje, North Macedonia*

aleksandar.kitanovski@students.finki.ukim.mk

Aleksandar Krsteski

*Faculty of Computer Science
and Engineering*

*Ss. Cyril and Methodius University
Skopje, North Macedonia*

aleksandar.krsteski@students.finki.ukim.mk

Gjorgji Madjarov

*Faculty of Computer Science
and Engineering*

*Ss. Cyril and Methodius University
Skopje, North Macedonia*

gjorgji.madjarov@finki.ukim.mk

Abstract—In recent years, AIOps has helped a lot with the exploration of different types of resources, in the processes of optimization and automation of complex IT operations. One of the main resources that AIOps is exploring is system logs. There are many techniques based on machine learning in AIOps that help in logs anomaly detection, logs prediction, and root cause analysis guided by logs, but a majority of them are considering log messages either individually or as log sequences, without exploring the relationships between different types of logs. We believe that those relationships can be expressed via using graph representations of log messages and those representations can be utilized in almost any AIOps operation. Therefore in this paper, we present logs2graphs, an open-source system for the creation and visualization of such graph representations of log messages, which is compatible with several publicly available log sources and expandable to other log sources.

Index Terms—AIOps, logs, graphs, visualization, software engineering, design patterns

I. INTRODUCTION

A. AIOps

AIOps is a field that became popular after 2017 when the term AIOps was coined by Gartner in [8]. As AIOps is a modern concept, there is no widely adopted definition yet. [11] For the purpose of this paper we will use the definition by Nedelkovski et al. in [10]: "Artificial Intelligence for IT Operations (AIOps) combines big data and machine learning to replace a broad range of IT Operations tasks including availability, performance, and monitoring of services. By exploiting log, tracing, metrics, and network data, AIOps enable detection of faults and issues of services". According to the definition, one of the 4 important resources that are explored by AIOps are logs, so therefore this paper focuses on a graph representation of log data which can later be used in various operations related to AIOps. There are many papers on the topic of AIOps which are focused on logs and how they can be used to detect future anomalies or determine the root cause of incidents, but mainly all of them are considering logs just as a sequence of logs. We believe that log messages can be represented as graph structures in which the relationship between log message types will be highlighted. This concept

was also recently introduced in [12]. As our goal is a graph representation of logs, the next subsection will give a short introduction to logs and graphs.

B. Logs

The background of computer systems generates logs in order to record notable events and executions indicating what has happened. Therefore, these logs are precious sources to check and investigate system status and detect anomalies in them. Logs usually are in a structured format that is defined on a system-level i.e all the logs that are generated by one system will be in the same format. Log formats are varying depending on the system, but in general they all mandatory contain a timestamp (the moment when the log was produced on the system), content (textual explanation of what happened in the system), type (information about the severity of the log per ex. error, information, etc.)

C. Graphs introduction

A graph is a mathematical structure used to model pairwise relations between objects. There are different types of graphs, the ones used in this project are directed graphs. Directed graphs are formally defined as an ordered pair $G = (V, E)$, where V is a set of nodes (or vertices), and E is a set of links (or edges) which are ordered pairs of nodes (v_1, v_2) meaning there is a link from node v_1 to v_2 . For the purposes of our paper, we build graphs from the logs as described in [12] - each alert type is a node, and there is a link between two nodes (r_1, r_2) if an alert of type r_1 appears before an alert of type r_2 . A detailed explanation of the graph representation of log sequences is provided in Section 4.

In the following, Section 2 provides an overview of the relevant work for log collection, log parsing, and their application in AIOps. Section 3 offers detailed descriptions of some publicly available log sources which will be available for exploration in our system. Section 4 presents technical details on the system design, architecture, and implementation

of all subsystems part of the logs2graphs tool. Section 5 demonstrates the user interface of our system as well as the visualization of the graph representation of different types of logs, which is the primary goal of our paper. Section 6 states our plans for utilization of this application in future research related to building models for graph representation learning. Section 7 concludes the paper and its results.

II. RELEVANT WORK

Logs are an important tool for monitoring and troubleshooting computer system behavior [5, 4]. As a result, there has been substantial work on automated log analysis. Techniques have been proposed for highlighting anomalous messages [3, 2] or patterns of messages [4]. Generally, these techniques are evaluated on proprietary log sources described in fairly general terms. Many recent papers describe the logs in question with a handful of excerpted lines plus a few aggregate statistics [5]. This is unfortunate because it makes it hard to reason about the relationship between log structure and analysis quality. Many research results are based on measurements taken at particular sites with highly customized software environments; there is often a justifiable reluctance to reveal operational details. Additionally, the general workflow is adapted when working with logs as shown in [6]. The phases relevant from that workflow for this work are:

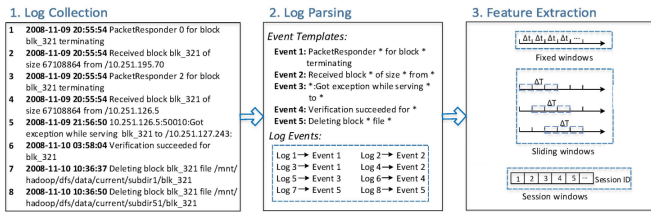


Fig. 1. Workflow (Source: [6])

- Log Collection (This phase is generally completed because in our paper we are using publicly available datasets explained in Section 3)
- Log Parsing (There are several techniques for parsing of the logs and extraction of generic event templates from log collection. In our work we are using DRAIN which will be explained shortly in Section 4)
- Feature Extraction (Based on the extracted event templates, log data is being windowed with the usage of tumbling windows, sliding windows, and session windows, if applicable. Features are being extracted from the logs in the created windows and afterward they are used for supervised and unsupervised tasks in Machine Learning.)

In this work, we are replacing the phase of feature extraction with the phase of the creation of graph representation for logs, as feature extraction is a stage that belongs to a more general terminology related to traditional machine learning processes.

III. LOG MESSAGE SOURCES

A. BGL

BGL represents a log message source that contains logs collected from the Blue Gene/L Supercomputer made by IBM. The supercomputer has 131072 processors and 32768GB of memory. This source has around 4.8 million

alert and non-alert messages identified by category tags, collected between 03.06.2005 and 04.01.2006 (nearly 6 months). Details about the Blue Gene/L System, methods of log collection, and alert identification can be found in [1].

B. HDFS

The logs were already published online from [5]. There are 24 million lines of logs. Most of the events in logs indicate run-time performance problems that have been confirmed by Hadoop developers. All errors, warnings, and information are collected between 09.11.2008 and 11.11.2008 (about 3 days).

The log message source is generated in a private cloud environment using benchmark workloads. The logs are sliced into traces according to block ids.

C. OpenStack Nova

OpenStack is an open-source cloud operating system that can control large pools of compute, storage, and networking resources throughout a data center, all managed and provisioned through APIs. Because it is a complex system, OpenStack is divided into 3 subsystems: Nova (for instances management), Cinder (for volumes management), and Neutron (for virtual networks management). [9] introduced the concept of fault-injection in order to generate anomalies (bugs) in the logs. The fault-injections are done based on the bugs that were most reported on StackOverflow and are done with surrounding buggy lines of codes with if statements.

A total of 911 tests (with IDs from 1-911) with fault injections were conducted on testbeds. In each test the system first starts with the bug disabled, then the bug is enabled and finally, the bug is disabled again. In this paper, for the purpose of simplicity, we only use the logs obtained from the NOVA subsystem of OpenStack.

IV. SYSTEM IMPLEMENTATION

Logs2graphs is a platform that enables multiple different representations of log message sequences from time and session windows. These representations are created upon user request where the user specifies the type of the graph, the log messages' source, and the configuration of the windows. The logs collected in the specified type and length of the window are represented as graphs and are served in a stream environment for further automated processing and extensive analysis. Also, the created logs representations are presented on a rich interactive graphical user interface, so the user can manually inspect, monitor, and analyze the log messages relationships. The current implementation supports the representation of the log message sequences into different graph structures that emphasize the relationships that exist between the individual log events in a given time or session window.

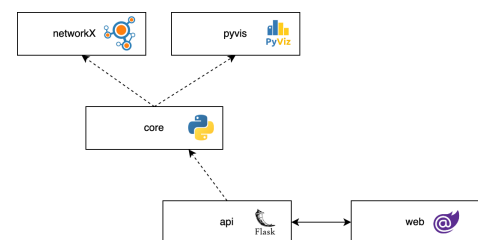


Fig. 2. Diagram of the main system components

A. Software architecture

The software architecture of logs2graphs as shown in Figure 2 is composed of three main components:

- **core** - The core of the system is a Python library (published as logs2graphs on pypi ¹) that is used for the purpose of log message sources initialization (loading and parsing of the log messages), creation of windows (time and session windows) and graphs from the aggregated logs in the windows. The creation of the windows and graphs is entirely dynamic i.e it is performed on each user request. The core library uses the libraries networkX and pyviz for the representation of graphs and their visualization accordingly.
- **API** - The API is built in the Flask (Python) framework and it has several endpoints (creating and listing experiments, listing created windows per experiments, generating data for graph visualizations). The API uses the logs2graphs core library for initialization of the log message sources and dynamical creation of the log representations.
- **web** - The User Interface for the application is created in the .NET Blazor framework and it provides several forms that are used to display the data (including graphs visualizations) obtained from the API or send data to the API when creating experiments.

B. Core library

1) Design of the core library

The core library logs2graphs is a Python package that provides the possibility to download, parse and process all of the 3 log message sources defined in Section 3. This is achieved by the usage of the **Template Method software design pattern** as shown in Figure 3. We have defined

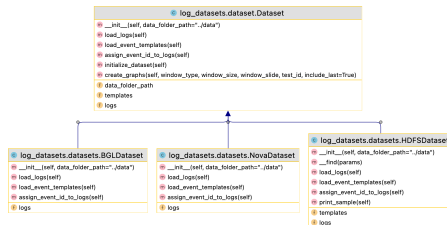


Fig. 3. UML class diagram of the classes used for the representation of the different log message sources in the core library

an abstract class `Dataset` for the representation of a log message source with the possibility to extend the library by deriving classes for other sources. The class `Dataset` has a method `initialize_dataset` which is actually the template method participant in the design pattern. This method includes three steps for initialization of the log message sources:

- `load_logs` - used to download the logs locally
- `load_event_templates` - used to perform the log parsing and extraction of event templates with the DRAIN algorithm.
- `assign_event_id_to_logs` - used to assign event template ID to each of the logs, via matching them with the extracted event templates. This is a very important step in the system as it provides a mapping of millions of

logs into several important categories (templates) which will be used for creation of a more understandable graph representation.

The three steps are abstract methods in the base class `Dataset` and each of them is overridden in the derived classes in a corresponding way.

For the purposes of event extraction from logs, we used Drain [7] which is an online log parsing method that parses logs in a streaming manner. Drain uses a fixed depth parse tree which encodes specially designed rules in its tree nodes, while the leaves contain log groups. When a new raw log message arrives Drain preprocesses it using simple regular expressions based on domain knowledge. Then the tree is searched by using the specially designed rules encoded in the tree nodes for a suitable log group for the given log message, if a suitable log group is not found then a new log group is created.

The graph creation is done in the method `create_graphs` where based on the `window_type` argument, we are creating a corresponding object from class `GraphsCreator`.

We have utilized the Template Method design pattern again in the definition of the classes `GraphsCreator` and `WindowsCreator` as shown in Figure 4 and Figure 5 accordingly.

In the class `GraphsCreator` the template method is the

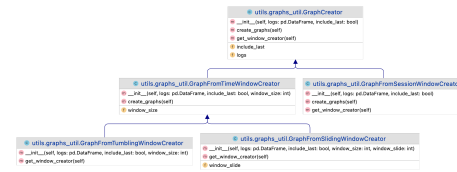


Fig. 4. UML class diagram of the classes used for the creation of graphs

method `create_graphs` while the step that varies in the different graph creators is the `WindowsCreator` object that is used for creation of windows. Therefore, the method `get_window_creator` is overridden differently in all types of graph creator. In the class `WindowsCreator` the template method is the method `create_windows` while the steps that vary in the different windows creators are the start and the frequency (applicable only for the time windows). The creation of the time windows (tumbling and

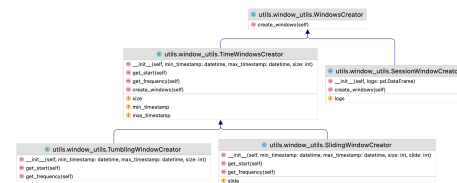


Fig. 5. UML class diagram of the classes used for creation of windows

sliding) depends on the minimal timestamp detected in the logs, as well as on the maximal timestamp.

2) Graph creation

We will explain the process of graph creation for a window of logs through an example. For example, we have obtained a window of 10 logs. We map that sequence of logs to a sequence of the event IDs of the template that was matched

¹<https://pypi.org/project/logs2graphs/0.0.6/>

Edge	Including last event ID		Excluding last event ID	
	Weight	Normalized weight	Weight	Normalized weight
A-B	3	0.33	3	0.375
B-A	1	0.11	1	0.125
B-B	2	0.22	2	0.25
B-C	2	0.22	1	0.125
C-A	1	0.11	1	0.125

TABLE I
WEIGHTS OF THE GRAPH FOR THE EXAMPLE SEQUENCE

with each log in the sequence. Let the sequence of event IDs be [A, B, A, B, B, B, C, A, B, C], which means that the first log corresponds to the event template whose id is 'A', the second log corresponds to the event template whose id is 'B' and so forth.

We create graph nodes for each of the unique event IDs present in the sequence. Regarding our example, the graph will have 3 nodes - one for event ID 'A', another for event ID 'B', and the third one for event ID 'C'.

Furthermore, we add directed edges between the nodes. The criteria we use for edge extraction from the log window are the following: if a log corresponding to event ID 'B' came after a log corresponding to event ID 'A' and a directed edge from node 'A' to node 'B' does not exist yet, we add a directed edge from node 'A' to node 'B' with weight 1. Otherwise, if a log corresponding to event ID 'B' came after a log corresponding to event ID 'A' and a directed edge from node 'A' to node 'B' already exists, then we only add 1 to the current weight of the edge. The final step we perform is weight normalization. In this step, we compute the sum of the weights of all edges present in the graph and then we normalize (divide) the weight of each edge with this sum.

The logs2graphs core library allows the users to create graphs from windows with or without the last event ID. As shown in Table I, if we exclude the last event ID we have a smaller weight in the edge B-C. This also causes the normalized weights of the graph to be different.

V. RESULTS

In this section, we will present the user interface of the system with some interesting graphs visualization created and displayed by our system. On Figure 6 a form to submit a new experiment is demonstrated. The user can choose the log messages source (HDFS/BGL/NOVA), the windowing type (tumbling/sliding/session), and other relevant parameters. In this case, the user is submitting an experiment for the creation of graph representation for the BGL log message source which should create sliding windows with a size of 2 hours and a slide of 1 hour. Also, the user wants to exclude the last event ID from the windows when generating the graph representations.

After the request is submitted, the user can open the list of

Fig. 6. Form to submit a request for new graph representation

submitted requests for graph representations. When the graph representation creation is completed, the user can open the list of windows created for the requested graph representation as shown in Figure 7. The user on this page can download the full generated graphs data for this graph representation in JSON format, or can select a window and visualize it. In Figure 8, the visualization of the graph created from a selected window is shown.

Time window (start,end)	SHOW VISUALIZATION
2005-07-08 22:00:00_2005-07-09 00:00:00	SHOW VISUALIZATION
2005-11-04 13:00:00_2005-11-04 18:00:00	SHOW VISUALIZATION
2009-07-20 04:00:00_2009-07-20 06:00:00	SHOW VISUALIZATION
2009-10-31 02:00:00_2009-10-31 04:00:00	SHOW VISUALIZATION
2009-12-14 17:00:00_2009-12-14 19:00:00	SHOW VISUALIZATION
2009-06-17 18:00:00_2009-06-17 20:00:00	SHOW VISUALIZATION
2009-12-28 03:00:00_2009-12-28 05:00:00	SHOW VISUALIZATION

Fig. 7. List of windows generated for the graph representation created in Figure 6

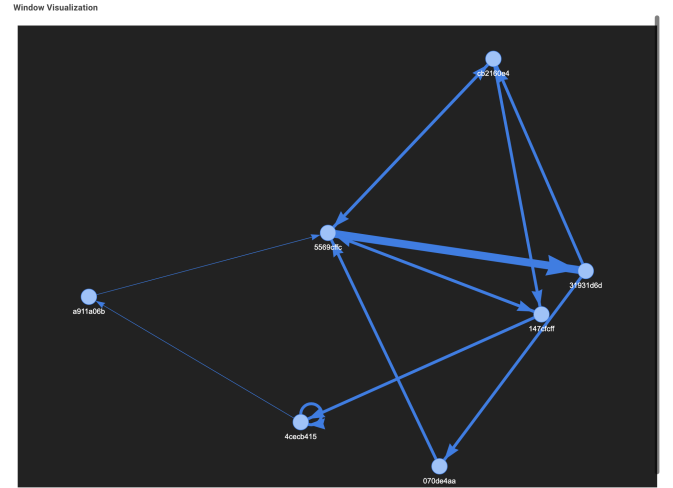


Fig. 8. Graph of the BGL logs in the interval [2005-06-16 18:00:00, 2005-06-16 20:00:00) from the graph representation requests in Figure 6

VI. FUTURE WORK

Furthermore, this graph representation of log data can be used for training and testing graph neural networks (GNNs) in order to detect anomalies in logs, predict future logs or perform root cause analysis. Therefore, the future work includes different architectures like GCN (Graph Convolution Networks), GGSNN (Gated Graph Sequence Neural Network), GAT (Graph Attention Networks).

VII. CONCLUSION

The amount of log data created by supercomputers, distributed systems, cloud systems, etc. is massive and cannot possibly be processed by humans, but because the uninterrupted work of these computer systems is crucial for many businesses, hospitals, governments, etc., ways of automated processing and anomaly detection/prediction must be developed. In this paper, we gave an introduction to AIOps as a field, logs and graphs, and an overview of relevant work in those fields. We presented a way to connect the log messages into a graph representation and provided a system with several

components that can be used to create and visualize the graph representations. Finally, we have defined the need for this system in the actions that we plan on taking in our future research.

ACKNOWLEDGEMENTS

This research was partially funded by the Faculty of computer science and engineering, Ss. Cyril and Methodius University in Skopje.

REFERENCES

- [1] Adam Oliner and Jon Stearley. “What supercomputers say: A study of five system logs”. In: *37th annual IEEE/IFIP international conference on dependable systems and networks (DSN’07)*. IEEE. 2007, pp. 575–584.
- [2] Sivan Sabato et al. “Analyzing system logs: A new view of what’s important”. In: *2nd USENIX workshop on Tackling Computer Systems Problems with Machine Learning Techniques*. 2007.
- [3] Michal Aharon et al. “One graph is worth a thousand logs: Uncovering hidden structures in massive system event logs”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2009, pp. 227–243.
- [4] Weihang Jiang et al. “Understanding customer problem troubleshooting from storage system logs”. In: *Proceedings of the 7th conference on File and storage technologies*. 2009, pp. 43–56.
- [5] Wei Xu et al. “Detecting large-scale system problems by mining console logs”. In: *Proceedings of the ACM SIGOPS 22nd symposium on Operating systems principles*. 2009, pp. 117–132.
- [6] Shilin He et al. “Experience report: System log analysis for anomaly detection”. In: *2016 IEEE 27th international symposium on software reliability engineering (ISSRE)*. IEEE. 2016, pp. 207–218.
- [7] Pinjia He et al. “Drain: An online log parsing approach with fixed depth tree”. In: *2017 IEEE international conference on web services (ICWS)*. IEEE. 2017, pp. 33–40.
- [8] Andrew Lerner. *AIOps Platforms—Gartner*. <https://blogs.gartner.com/andrew-lerner/2017/08/09/aiops-platforms/>. [Online; accessed 15-April-2022]. 2017.
- [9] Domenico Cotroneo et al. “How bad can a bug get? an empirical analysis of software failures in the openstack cloud computing platform”. In: *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 2019, pp. 200–211.
- [10] Sasho Nedelkoski, Jorge Cardoso, and Odej Kao. “Anomaly Detection and Classification using Distributed Tracing and Deep Learning”. In: *2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*. 2019, pp. 241–250. DOI: 10.1109/CCGRID.2019.00038.
- [11] Paolo Notaro, Jorge Cardoso, and Michael Gerndt. “A Survey of AIOps Methods for Failure Management”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 12.6 (2021), pp. 1–45.
- [12] Yi Wan et al. “GLAD-PAW: Graph-Based Log Anomaly Detection by Position Aware Weighted Graph Attention Network”. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2021, pp. 66–77.

Artificial Intelligence in Robotics

Martin Stefanov

Faculty of Computer Science
and Engineering, Skopje, Macedonia
martin.stefanov@students.finki.ukim.mk

Nikola Stojanovski

Faculty of Computer Science
and Engineering, Skopje, Macedonia
nikola.stojanovski.1@students.finki.ukim.mk

Vlatko Spasev

Faculty of Computer Science
and Engineering, Skopje, Macedonia
vlatko.spasev@finki.ukim.mk

Abstract—Artificial intelligence plays a key role in robotics. The artificial intelligence in a robot makes its actions autonomous and improves the efficiency of the robot. This advancement in robotics has had considerable influence in the modern world. There are primitive robots which have no artificial intelligence implemented in them, but they follow a predetermined set of instructions which may not be the optimal set of instructions. Robotics are involved in almost every field of work and the majority of them use artificial intelligence or machine learning. In this paper we will discuss the various fields robotics with artificial intelligence are being used and how these robots operate. We will also discuss the influence the artificial intelligence in robotics has had in society and the ethical questions and dilemmas that have been raised by this technology.

Index Terms—Artificial intelligence, Neural network, Robot;

I. INTRODUCTION

Artificial intelligence in robotics has long been a subject of science fiction novels. These science fiction novels display robots as man-like machines who display intelligence like humans would or mimic some cognitive function of the human brain. The most famous science fiction writer who wrote about robots was Isaac Asimov, who established in his book the Three Laws of Robotics, which are 1. A robot may not injure a human being or, through inaction, allow a human being to come to harm; 2. A robot must obey the orders given to it by human beings except where such orders would conflict with the First Law; 3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law [1]. Robotics, however, is not exactly like this although there are robots who are being developed to act like humans. Before we see what artificial intelligence is, we will discuss what robotics is. Robotics is the study of robots and it is a branch of technology that deals with designing, construction, and applications of robots. Robots are machines, which are used to do repetitive or tasks that may be dangerous for humans. Most of the robots nowadays are doing their job themselves without any specific directions while other robots must have a person telling them what to do [2]. Before we look at the implementations of artificial intelligence in robotics, we will discuss artificial intelligence as a concept. Artificial intelligence is intelligence proved by machines, which is able to perform various tasks for the needs of human intelligence. Some activities of computers with artificial intelligence are learning, speech recognition, problem-solving and planning [2]. Artificial intelligence describes the work processes of machines that would require intelligence if performed by humans. The term ‘artificial intelligence’ thus means ‘investigating intelligent problem-solving behavior and creating intelligent computer systems’. There are two kinds of artificial intelligence: 1. Weak artificial intelligence: The computer is merely an instrument for investigating cognitive

processes – the computer simulates intelligence; 2. Strong artificial intelligence: The processes in the computer are intellectual, self-learning processes. Computers can ‘understand’ by means of the right software/programming and are able to optimize their own behavior on the basis of their former behavior and their experience. This includes automatic networking with other machines, which leads to a dramatic scaling effect [3]. Artificial intelligence is a very powerful technology which with a large amount of data and computing power it can do various actions that humans do in a fraction of the time. The aforementioned actions that artificial intelligence can perform fit perfectly to what a robot should do such as speech recognition, face recognition, problem solving etc. From [2], we can define five functions which the robot ponders and makes a decision about which steps to take about an activity. They are planning, acting, perceiving, monitoring, and goal reasoning. In planning, the agent (robot) deliberates with the help of the combination of prediction and searches which are used for synthesizing a trajectory in an abstract space with the help of predictive models of feasible action and environment. The function of acting is based on implementing an online close-loop feedback function. In order to refine and control the achievement of planned actions, this function processes streams of sensory stimulus to actuators commands. In perceiving, the agent can extract the features of the environment for identifying the events and the states, and other situations which are relevant to the tasks. It also combines bottom-up sensing from sensors to meaningful data, with top down focus mechanisms, planning for gathering information and sensing actions. The monitoring function combines and observes the differences between predictions and observations. It also performs trigger recovery actions and diagnosis. With goal reasoning, the agent combines and observes the differences between predictions and observations. It also performs trigger recovery actions and diagnosis. From aforementioned discussions we can see that artificial intelligence and robotics work really well together and we will see this collaboration in working fields where robotics is being used.

II. RELATED WORK

Robotics, information technology, communication technology, and artificial intelligence have all become more interwoven since the dawn of the artificial intelligence era. After a long evolution in the electrical and digital eras, robots are entering into a new era of intelligence. Three components of the entire process are reflected: To begin with, traditional industrial technologies such as the controller, servo motor, and reducer have evolved into artificial intelligence technologies such as computer vision, natural language processing, and deep learning; second, robots have become more deeply

integrated into human society, attracting attention from both industrial and commercial, family, and individual users. Finally, what were formerly mutually autonomous human-robot connections are being replaced by a bond of collaboration and interaction [2]. John McCarthy is the man who came up with the concept of AI in 1955, assuming that every component of learning and other domains of intelligence can be characterized so precisely that they can be mimicked by a machine. However the phrases 'artificial intelligence' and 'intelligent human behavior', are not well defined [3]. People often imagine a robot rolling across the floor responding automatically when they think of artificial intelligence robotics. And, undoubtedly, the humanoid robot is thrilling and popular. However, there are numerous types, forms, and uses of artificial intelligence in robots that are transforming our business landscape [4]. So what is an artificial intelligence robot? Artificial intelligence robots combine AI and robotics by incorporating AI software into robot systems. The most advanced robots are those controlled by AI that can learn from their environment and experience and then build on their capabilities based on that knowledge. A survey has been made on 12,000 experts and members of the interested public to share their opinions on the likely future of the Internet by Aaron Smith and Janna Anderson conducted between November 25, 2013, and January 13, 2014 [5]. By 2025, the clear majority of respondents expect robotics and artificial intelligence will pervade many aspects of daily life, with significant ramifications for areas such as health care [6], transportation and logistics [7] [8], customer service [9], and home maintenance [10]. They are significantly divided on how improvements in AI and robotics will affect the economy and employment situation over the next decade, despite being fairly consistent in their projections for the future of technology itself. The point of discussion in [11] is how the application of artificial intelligence onto the robotics of manufacturing could lead to a new type of manufacturing called intelligent manufacturing. In a published article Ethics of Artificial Intelligence and Robotics is debated about issues that arise with certain uses of the technologies which would not arise with others. Main debates in this article are about Privacy Surveillance, Manipulation of Behavior, Opacity of AI Systems, Bias in Decision Systems, Human-Robot Interaction, Automation and Employment and Artificial Moral Agents [12]. In [13] we see the results of 3D-printed habitats for Martian or Lunar space missions using artificial intelligence and robotics. In [14] it is shown the specifics of the Pepper robot. In [15] we see the way the Shadow Dexterous Hand works. In [16] it is discussed about the potential and the challenges being faced with the emergence of artificial intelligence-assisted surgery. In [17] it is being shown a robotic hand solving a Rubik's cube.

III. IMPLEMENTATIONS OF AI IN ROBOTICS

A. Fourth industrial revolution

The fourth industrial revolution is a new chapter in human development which is a product of technological advances. It is going to represent the fundamental change in our daily work routine, the way we live and interact with others and how robots and the progress of AI will substitute a lot of labor work. The digital, physical, and biological worlds are all altering as a result of the fourth industrial revolution. It is offering us new possibilities and promising a brighter future.

Artificial intelligence, machine learning, augmented reality, and virtual reality are just a few of the technologies that make up the fourth industrial revolution. Machine learning is already present in our daily lives, unlocking our phones with a peek or a touch, recommending music we enjoy, and teaching automobiles to drive themselves. The proliferation of data is at the heart of it all. Data is expanding at a breakneck pace. Manufacturing is no longer the same as it was some decades ago. We are seeing greater automation with production lines consisting of various machines mixed in amongst fewer workers, rather than factories overflowing with busy workers. In today's production sector, industrial robots have grown extremely ubiquitous. There has been a decline in the workforce for manual labor in recent years, particularly for skilled positions. Robotic automation bridges the labor and skills gap, avoiding costly delays and downtime caused by labor shortages. Robotic automation is becoming increasingly accessible to more businesses as the cost of robots has decreased and robotic technology has advanced. Today's robot prices are much more reasonable, and there is also the option of purchasing a secondhand robot.

B. Transport and Logistics

Part of the reason for the automation and modernization of production processes in industrial plants is to improve transportation operations inside the industry itself. One of the most significant advancements in the automation of transport operations in manufacturing, assembly lines, and storage is the introduction of autonomous robots. These are autonomous service robots that move using an automatic control system, which is usually powered by a battery or an electric motor. The robot employs sensors on its infrastructure to receive information on the vehicle's location and velocities, which the control system then uses to transmit the right commands to the vehicle, allowing it to follow the proper paths. The application of autonomous service robots for logistics involves transport, handling, packaging, sorting and delivery of products. These robots are typically found in industrial plants, where they are used to transport working materials, such as boxes, pallets, and tools, from one machine to another, as well as between shipping areas and storage areas. Apart from their employment in the automation of transport operations in manufacturing, service robots are also utilized in offices, hospitals, post offices, airports, and other public institutions where the transportation and delivery of various items is required. The United Nations Economic Commission for Europe (UNECE) and IFR began working on terminology unification in 1995, and the outcome was the new ISO-Standard 8373 definition, which went into force in 2012. This classification divided the service robots in two groups: service robots for personal/domestic use and service robots for professional purposes. Given that the service robots for transportation and logistics are classified in the group for professional use, we will state their usage in real world problems. This group of robots are included in: military service robots, field robots (agriculture, cattle breeding, forestry, etc.), logistics service robots, medicine service robots, professional cleaning service robots, inspection and maintenance service robots, rescue and monitoring service robots, underwater systems, mobile platforms for general use, public relations service robots etc [18]. An article posted from the International Federation of Robotics (IFR) on Oct 28, 2020 states

that the sales value of professional service robots increased by 32% to USD 11.2 billion worldwide (2018-2019). The COVID-19 pandemic will further boost the market. High demand for robotics disinfection solutions, robotic logistics solutions in factories and warehouses or robots for home delivery are examples of this trend.

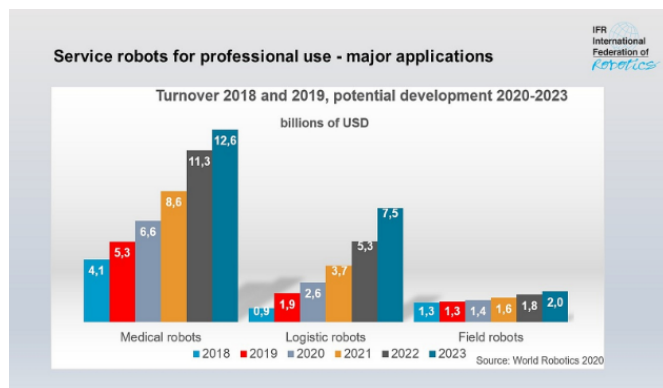


Figure 1. Annual supply of service robots for professional use

1) *Pick-it-Easy Robot*: The Pick-it-Easy Robot is an automated single-item picking robot that can also be used to induct items from pre-picked batch containers into overhead or sorting systems. The Pick-it-Easy Robot's various grippers, AI-based item recognition – the so-called Covariant Brain – and grasp point specification allow it to handle a wide range of articles for a variety of applications in various industries. The Pick-it-Easy Robot knows just where to grasp objects with shiny surfaces and gently deposits them where they belong. It is designed using a system that is always learning and improving as it works.

2) *RUNPICK*: RUNPICK stands for Robotic Universal Picker and was created specifically for the food retail industry. The completely automated system is capable of handling a wide range of products. Large handling units of fast-, medium-, and slow-moving fresh and shelf-stable items are picked and palletized by the robot. RUNPICK builds mixed pallets and roll containers, making it ideal for offering businesses with a variety of structures.

3) *Automatic palletizing and depalletizing*: Because of the heavy lifting involved, loading and unloading pallets is a physically demanding duty for workers. As part of creative automation solutions, robots can substantially lighten the work here. They depalletize totally mechanically and prepare the articles for storage in the goods-in area. They assemble stable pallets for shipping in goods-out. Robots can also fully automate the stacking and de-stacking of cartons and standard handling units, as well as the loading of dollies. This complete automation relieves staff of their responsibilities while also increasing productivity.

4) *Open Shuttle*: Open Shuttle is an autonomous mobile robot which connects different areas within facilities quickly and efficiently, bringing materials and components where they are needed. They act completely autonomously, do not require any assistance, and are completely safe for people thanks to their innovative technology. Furthermore, they are adaptable and simple to use.

5) *Self-Driving cars*: Autonomous driving is the most exciting, and at the same time, most controversial of im-

plementing Artificial Intelligence in transportation. Today we don't have fully automated driving systems, rather we have driver-assistance systems that support the driving which reduces the workload on the driver. Although no carmaker currently sells a fully driverless vehicle, others are pursuing the technology. One such project is Waymo, a Google-owned company, which is testing autonomous driving vehicles in Phoenix.

6) *Commercial motor vehicles*: Another interesting potential about the AI are self driving trucks, especially truck platooning. Truck platooning is the process of syncing multiple trucks in a convoy by using connective technologies and automated driving systems. It allows the trucks to travel at the same speed as dictated by the leader of the convoy. When the truck in front slows down, it sends an immediate signal to the platoon to automatically brake. The autonomous system's reaction time would theoretically be five times faster than that of a human driver, allowing for substantially shorter following distances.

C. Construction and manufacturing industry

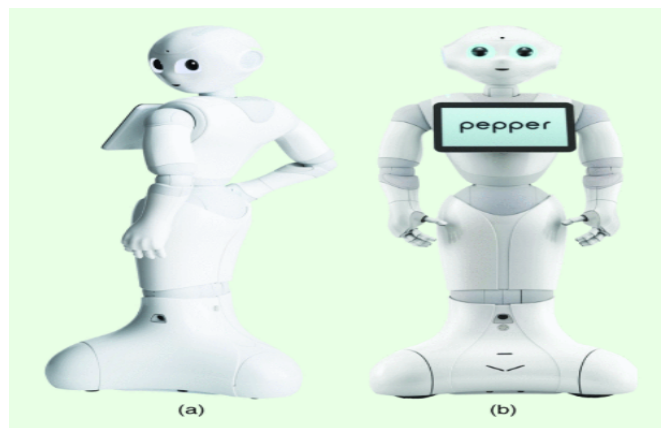
In construction and in manufacturing, robots are being used for doing precision work or jobs that are dangerous for a human being to perform. In Fig. 1 we can see the amount of money being invested in the United States of America in field robots. It is evident that the investment being made in field robots is not at the same level as logistic and medical robots but the investment in field robots for construction and manufacturing is rising. The projection for 2022 is that the amount of money being invested in field robots will be 2 billion dollars which is more than 200 million dollars from 2021. One example of a robot using artificial intelligence being used in construction is the Site Monitoring Robot by Scaled Robotics [4]. The purpose of this robot is to monitor the progress of a construction site. This robot scans the site and makes a detailed survey of the very same site and then analyzes the data it has collected which it then sends it to a cloud platform to compare to the building model in order to see the progress. Users can also examine the issues on a special color-coded 3D model. This process helps you prevent costly errors and even shows you Health or Safety problems, such as holes in edge protection. This robot is a very powerful tool for a construction project as it lowers the time being spent for inspection by a human and it ensures a more reliable quality assurance of the project. A big role in construction and manufacturing will play 3D printing. This technique in construction and manufacturing works very well with robotics and artificial intelligence. This construction type is already being used for habitats that will be used in future space missions. In [13], the paper discusses the results of the 2015 3D-Printed Habitat (3DPH) competition, which was a competition to develop housing solutions for extended-duration missions on planetary surfaces using advanced additive construction technology. The projects entered in the competition used robotics with artificial intelligence to construct habitats autonomously. In 2015, Foster + Partners competed in Phase 1 of the NASA Centennial 3D-Printed Habitat Challenge, securing second-place for their proposal that used an array of semi-autonomous robots with artificial intelligence to 3D-print habitats for eventual human occupation [13]. One of the winners of the competition, AI Space Factory, saw that this technology could be used in the very expensive con-

struction industry on our planet. AI SpaceFactory recognized that by addressing off-grid power, autonomous construction, and in situ resource utilization for Mars, it had developed technologies that can be directly applied to the construction industry on Earth. The other winner of the competition, The Pennsylvania State University, had designed a different solution to the challenge. Penn State's research approach was to model and coordinate the complex interrelationship between multiple and associated variables, including those related to material properties, printing system, and design. BIM technology and AI techniques were used to encode this model and concurrently design the operation logistics, the printing system and setup, the toolpath and printing process, as well as the design of the habitat. This approach permits identifying optimized solutions for both Earth and Mars. The challenge demonstrated processes and equipment for large-scale vertical autonomous construction using robots, diversity and innovation in viable designs for realistic planetary habitats, and of new software and artificial intelligence control algorithms for depositing material in a non-two-dimensional layer. In the manufacturing industry, the automation process is being improved by artificial intelligence. The automation process is a combination of robotics and equipment performing an activity with a predefined set of instructions and a small number of fail safe scenarios. However, robotic machines implemented with artificial intelligence don't just follow the rules but intelligently understand and recognize patterns within the data, learn from past experiences and improve future performance. Therefore, artificial intelligence enables the same set of robots and equipment to identify solutions to complex issues within a given solution space. Intelligent manufacturing is a new manufacturing model and the technical means by which new information and communication technology, intelligent science and technology, large manufacturing technology (including design, production, management, testing, and integration), system engineering technology, and related product technology are integrated with the whole system and lifecycle of product development. Artificial intelligence is applied in the field of intelligent manufacturing through the intelligent manufacturing system. An intelligent manufacturing system is characterized by autonomous intelligent sensing, interconnection, collaboration, learning, analysis, cognition, decision-making, control, and execution of human, machine (robot), material, environment, and information in the whole system and life cycle [11].

D. Assistance

The invention of robotics and its improvement over the years have significantly influenced our society. It has improved the effectiveness and efficiency of various fields which assist in our lives or provide some sort of service in our lives. With the emergence and astronomical rise of artificial intelligence, robotics are now even more capable to assist in our lives as they now possess the ability to recognise and solve a problem autonomously. In the previous sections we have discussed the assistance robotics have made in our lives in fields such as transportation, logistics, manufacturing etc, while in this section we will discuss the assistance in more niche areas of our everyday lives.

Figure 2. The appearance of the Pepper robot



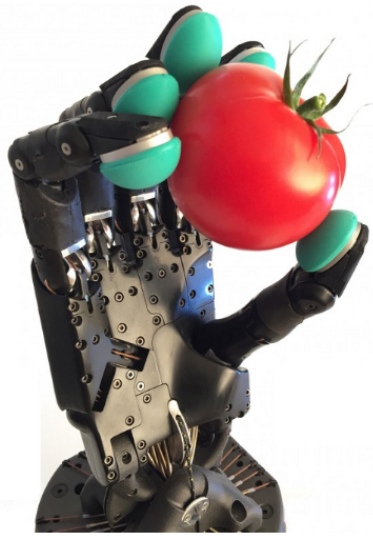
One robot that provides a service in day-to-day activities of humans is the robot Pepper. Pepper is an industrially produced humanoid robot launched in June 2014 that was first created for Buisness-to-Buisness needs and later adapted for Business-to-Consumers purposes. The Pepper robot is capable of exhibiting body language, perceiving and interacting with its surroundings, and moving around. It can also analyze people's expressions and voice tones, using the latest advances and proprietary algorithms in voice and emotion recognition to spark interactions. The robot is equipped with features and high-level interfaces for multimodal communication with the humans around it [14].

In Fig. 2 we can see the appearance of the robot. Pepper is four feet tall, has a tablet (that displays information) in the center of its chest, can gesticulate and talk in several languages. The robot has no sharp edges, there are soft parts of the cover, and the center of mass is at the base to keep the robot from falling over. The Pepper robot has 20 degrees of freedom (DoF) for motion in the whole body (17 joints) and omnidirectional navigation (three wheels). The DoFs include two in the head, two in each shoulder (left and right), two in each elbow (left and right), one in each wrist (left and right), one in each hand (five-fingered left and right hands), two in the hips, one in the knee, and three in the base. The actuators of the robot are DC motors and a magnetic rotary encoder-based sensor. The robot has a system that ensures its autonomy. The whole system was designed to balance the software and hardware loads and achieve a battery life of up to 12 h. In addition, a specifically designed docking station for autonomous charging was developed. The Pepper robot is equipped at both the hardware and software application programming interface (API) levels to provide good functionality for perceiving humans. The multimodal perception components are primarily intended to discern people's presence and avoid collisions with the environment during body movement. The NAOqi's (The operating system of the robot) People Perception module provides a list of in-built APIs to help in the development of high-level reasoning and behavioral capabilities [14].

Figure 3. Shadow Dexterous Hand

The artificial intelligence implemented in the robot enables the robot to make a dialog-based interaction with humans and recognise an emotion of the human it converses with.

Artificial intelligence in robotics could help in the pharmaceutical and agricultural sector and could potentially improve the lives of the physically disabled people. One such example



is the Shadow Dexterous Hand. This robotic hand can grasp objects like human would and it is being used for delicate work in pharmaceutical environments which are harmful to humans and it is being used for soft picking fruits in agriculture because of its precision and delicacy. In [15], it is being proposed an adaptive, compliant grasping strategy using only tactile feedback. Using artificial intelligence, the Shadow Dexterous Hand can grasp objects of varying shapes, sizes and weights without having a priori knowledge of the objects. In Fig. 3 we can see the Shadow Dexterous Hand.

Table 1. Experiments conducted with the hand

Objects	Type	Pass	Fail	Success
Cup	Cylinder	4	1	80
Glass	Cylinder	4	1	80
Football	Sphere	3	2	60
Softball	Sphere	4	1	80
Sugar	Cuboid	2	3	40
Electronic box	Cuboid	4	1	80

In Table 1. we can see results of the experiments conducted in [15]. The success field is expressed in percentage. For every object the Shadow Dexterous Hand had 5 attempts to grasp the given objects. With the cylinders like the cup and the glass it only failed once which means it has an 80% success. With the spheres it had an average of 70% success because the hand failed twice to grasp the football as opposed to softball, which it failed once. The worst success the Shadow Dexterous Hand had with the sugar, which it failed three times, because of its unique surface. With the cuboid shapes it had an average success percentage of 60%.

In the proposed work from [15] they developed a simple closed-loop formulation to grasp and manipulate an object with just tactile feedback using Shadow Dexterous Hand with BioTac Tactile sensors. They achieve a human-like form factor without compromising on the accuracy. This shows promise in the advancement of the robotic hand along with the artificial intelligence.

The Shadow Dexterous Hand can also solve complex puzzles such as the Rubik's cube. In [17] it is demonstrated that models trained only in simulation can be used to solve a manipulation problem of unprecedented complexity on a real robot. This is made possible by two key components: a novel algorithm, which we call automatic domain randomization

(ADR) and a robot platform built for machine learning. In Fig. 4 it is visualized the movements the Shadow Dexterous Hand would make to solve the Rubik's cube. The most important field in which robotics with artificial intelligence plays a part is medicine. Surgeries require extreme steadiness, patience and the surgeon being impervious to the pressure of life and death situations.

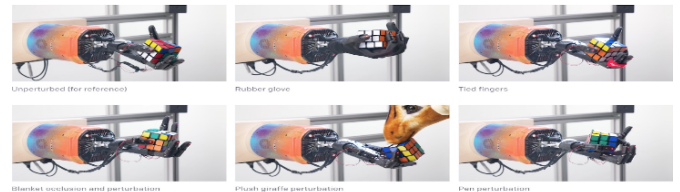


Figure 4. A five-fingered humanoid hand solves Rubik's cube

Robots fit these requirements and surgeries in advanced countries have been robotically assisted. In recent years, robot assisted surgeries have become the preferred approach in developed countries such as the United States and the United Kingdom; have attracted considerable attention in many other parts of the world. In order for the robot with artificial intelligence to assist in surgeries it requires access to comprehensive data in a sensor-enhanced operating room and enrichment of those data with surgical knowledge by annotation to make them usable for machine learning methods that provide AI assistance to the surgeon. Artificial intelligence may also change practice in robot-assisted surgery towards cognitive surgical robotics. Today, clinically used surgical robots are mere telemanipulators without any autonomous activity. In research, robotic systems have been developed for situation-aware automatic needle insertion [16].

IV. ETHICS

As robotics technology improves, ethical considerations grow in importance: If this is even possible, should robots be trained to obey a code of ethics? Is it dangerous to build emotional relationships with robots? What impact will robotics have on society and ethics? The ethics of AI and robotics are frequently centered on various "concerns," which is a common reaction to new technologies. Some technologies, such as nuclear power, automobiles, and plastics, have sparked ethical and political debate as well as considerable regulatory initiatives to limit their trajectory, usually after some damage has been done. New technologies, in addition to "ethical issues," challenge present norms and conceptual frameworks, which is of particular interest to philosophy. As the robots have been developed more and more, philosophy was also applied to robotics. One of the first publications addressing and setting the foundation of robot ethics was "Runaround", a short science fiction story written by Isaac Asimov which featured Three Laws of Robotics. In a series of scenarios, Asimov demonstrated how, despite their hierarchical organization, conflicts between these three rules will make it difficult to apply them. In this section we will discuss the ethical use of AI and robotics that can be more or less autonomous. We will look into the benefits or issues that we can receive from developing this kind of machines. The first general discussion that we will mention will be about privacy and surveillance in information technology. By 2025,

it is expected that 75 million smart linked gadgets will be in use in our homes, offices, and elsewhere, each having the ability to sense, analyze, and make decisions on their own. It is self-evident that security, privacy, and surveillance standards must be prioritized throughout the creation and implementation of all new technology. In today's society, our data is collected at remarkable rates and in unexpected ways. A lot of data appears to be collected in exploitative methods, abusing human flaws and addictive tendencies, in some of the largest and most popular firms such as Amazon, Google, Apple, Facebook, and Microsoft. At this exact moment there are no worldwide laws and regulations regarding privacy and surveillance in the world of Artificial Intelligence. But the EU, US and other countries have already begun with this imperative process.

Opacity of Artificial Intelligence (AI) Systems, also known as transparency of AI systems, is the idea that what happens inside AI algorithms, and how they arrive at their conclusions, is unknown to people who use the AI, but often even to the designers. This lack of transparency exists for a variety of reasons: human cognitive abilities simply cannot comprehend massive, complex algorithmic models and datasets; the inherent nature of self-learning algorithms, whose decision logic evolves without human input; a lack of appropriate visualization tools for mammoth-sized code and datasets; poorly structured code and data that effectively renders them impossible to read, and so on. This is important because if the code of the algorithm is not understandable by the user and the designer, an opacity even experts cannot penetrate, then bias like this simply cannot be controlled. In the EU, some of these issues have been taken into account with the (Regulation (EU) 2016/679), which foresees that consumers, when faced with a decision based on data processing, will have a legal "right to explanation"—how far this goes and to what extent it can be enforced is disputed. AI decision support systems and "predictive analytics" operate on data and create an "output" in the form of a decision. This output could be anything from "this restaurant matches your preferences" to "the patient in this X-ray has completed bone growth," "credit card application declined," "donor organ will be given to another patient," "bail is denied," or "target identified and engaged." In business, healthcare, and other industries, data analysis is frequently utilized in "predictive analytics" to predict future changes. Unfortunately, bias has been discovered in several AI decision-making systems. Amazon's automated AI system for recruitment, which was discontinued in 2017, is a significant example. Because Amazon had a history of discriminating against women in the recruiting process, when their ten-year hiring data was used to construct their AI algorithms, this bias was passed on to the AI system. Because human-robot contact is becoming more prevalent and will continue to become more widespread, we believe it is critical to address Human-Robot interaction (HRI) ethics. Robot ethics looks at how robots can help or hurt humans, how they can alter individual autonomy, and how they might affect societal justice. Humans are quick to ascribe mental traits to objects and empathize with them, particularly when the items' exterior appearance resembles that of live beings. This can be used to trick humans (or animals) into giving robots or AI systems more intellectual or even emotional significance than they deserve. Two examples are Care Robots and Sex Robots. Current systems include robots that assist human

caregivers, robots that enable patients to complete some tasks on their own, and robots that are provided to patients as companions and comfort. It's not apparent whether there's a problem here, given that the conversation largely centers on the fear of robots dehumanizing care, yet the actual and anticipated robots in care are assistive robots for traditional technical work automation. Also it has been argued by several tech optimists that humans will likely be interested in sex and companionship with robots and be comfortable with the idea. Humans have long had deep emotional attachments to objects, so perhaps companionship or even love with a predictable android is attractive, especially to people who struggle with actual humans, and already prefer dogs, cats, birds, a computer or a tamagotchi.

V. CONCLUSION

In conclusion, AI, like all significant revolutionary events in human history, will undoubtedly have far-reaching consequences. However, with appropriate planning and following the proper ethics, these issues can be overcome. The benefits of AI in improving human lives will certainly outweigh the concerns in the short to medium term. Robots will help us with many things where we won't be needed. Every wave of automation and computerization has enhanced production without lowering employment in the past, and there is no reason to believe that this time will be any different. The new wave, in particular, is likely to boost our personal or professional productivity (e.g., self-driving cars), but not necessarily to eliminate jobs (e.g. chauffeur). While robots may eliminate certain manual tasks, the impact should be similar to that of past waves of automation in factories and other settings. Someone will, on the other hand, have to develop and build the new tools, which will almost certainly result in a new wave of innovations and jobs. Any big concept that has a broad impact on people's lives, if not handled appropriately, can be extremely dangerous.

REFERENCES

- [1] R. Clarke, "Asimov's laws of robotics: Implications for information technology," 1994.
- [2] K. Anitha and M. Sabri, "Artificial intelligence and robotics," *SSRN Electronic Journal*, 6:1323–1327, 2018.
- [3] U. B. A. M. G. N. G. J. S. Gerlind Wisskirchen, Blandine Thibault Biacabe and B. von Brauchitsch, "Artificial intelligence and robotics and their impact on the workplace," *IBA Global Employment Institute*, 11(5):49–67, 2017.
- [4] [Online]. Available: <https://www.intellspot.com/artificial-intelligence-rob>
- [5] A. Smith and J. Anderson, "Ai, robotics, and the future of jobs," *Pew Research Center*.
- [6] [Online]. Available: <https://www.howtorobot.com/expert-insight/healthcare-robotshead1>
- [7] [Online]. Available: <https://www.automate.org/blogs/robots-in-logistics-and-transportation>
- [8] [Online]. Available: <https://www.dhl.com/global-en/home/insights-and-innovation/thought-leadership/trend-reports/robotics-in-logistics/robotics-in-logistics-why-the-time-is-now.html>
- [9] [Online]. Available: <https://www.automate.org/a3-content/service-robots-customer-service>
- [10] [Online]. Available: <https://porch.com/advice/robots-home-maintenance>
- [11] W.-t. Y. X.-b. L. Bo-hu Li, Bao-cun Hou and C. wei Yang, "Applications of artificial intelligence in intelligent manufacturing: a review," *Frontiers of Information Technology Electronic Engineering*, 18(1):86–96, 2017.
- [12] V. C. Muller, "Ethics of artificial intelligence and robotics."
- [13] M. F. S. N. A. A.-P. B. M. B. Monserrate Roman, Melodie Yashar and J. Ballard, "3d-printing lunar and martian habitats and the potential applications for additive construction," *2020 International Conference on Environmental Systems*, 2020.

- [14] A. K. Pandey and R. Gelin, "A mass-produced sociable humanoid robot: Pepper: The first machine of its kind," *IEEE Robotics Automation Magazine*, 25(3):40–48, 2018.
- [15] K. B. K. C. F. Kanishka Ganguly, Behzad Sadrfaridpour and Y. Aloimonos, "Grasping in the dark: Compliant grasping using shadow dexterous hand and biotac tactile sensor," *arXiv e-prints*, pages arXiv:2011, 2020.
- [16] B. P. M. -S. J. W. Sebastian Bodenstedt, Martin Wagner and S. Speidel, "Artificial intelligence-assisted surgery: potential and challenges," *Visceral Medicine*, 36(6):450–455, 2020.
- [17] M. C. M. L.-B. M. A. P. A. P. M. P. G. P. R. R. Ilge Akkaya, Marcin Andrychowicz, "Solving rubik's cube with a robot hand," *arXiv preprint arXiv:1910.07113*, 2019.
- [18] T. Zielinska, "Professional and personal service robots," *International Journal of Robotics Applications and Technologies (IJRAT)*.

SHORT PAPERS

Wearable Patch for Mass Casualty Screening with Graphene Sensors

Teodora Vićentić

Center for Microelectronic
Technologies

Institute of Chemistry, Technology and
Metallurgy

Belgrade, Serbia

teodora@nanosys.ihtm.bg.ac.rs

Milena Rašljić Rafajilović

Center for Microelectronic
Technologies

Institute of Chemistry, Technology and
Metallurgy

Belgrade, Serbia

ORCID 0000-0001-8170-746X

Stefan Ilić

Center for Microelectronic
Technologies

Institute of Chemistry, Technology and
Metallurgy

Belgrade, Serbia

ORCID 0000-0002-1721-9039

Bojana Koteska

Faculty of Computer Science and
Engineering, "Ss. Cyril and Methodius"
University

Skopje, R.N. Macedonia

bojana.koteska@finki.ukim.mk

Ana Madevska Bogdanova

Faculty of Computer Science and
Engineering, "Ss. Cyril and Methodius"
University

Skopje, R.N. Macedonia

ana.madevska.bogdanova@finki.ukim.
mk

Igor Pašti

University of Belgrade
Faculty of Physical Chemistry
Belgrade, Serbia

ORCID 0000-0002-1000-9784

Fedor Lehocki

STU

Faculty of Informatics and Information
Technologies

Bratislava, Slovakia

fedor.lehocki@stuba.sk

Marko Spasenović

Center for Microelectronic
Technologies

Institute of Chemistry, Technology and
Metallurgy

Belgrade, Serbia

spasenovic@nanosys.ihtm.bg.ac.rs

Abstract—Wearable sensors are reaching maturity, at the same time as technologies for communicating physiological data and those for analyzing massive amounts of data. The combination of the three technologies invites for applications in mass screening of personal health through smart algorithm deployment on data from wearable patches. We propose and present an architecture for a wearable patch to be used in mass casualty emergency situations, or for hospital bedside monitoring. The proposed patch will contain multiple sensors of physiological parameters. We propose to create respiration and heartbeat sensors made of laser induced graphene. We show that graphene on flexible substrates can be utilized in conjunction with the Python heart rate analysis toolkit - HeartPy to reliably acquire physiological data from human subjects.

Keywords—wearables, graphene, Python, sensors, electronics

I. INTRODUCTION

Wearable sensors are a growing field of research, with applications in telehealth [1], fitness tracking [2], and mass casualty incident management [3]. Various forms of wearable sensors of physiological parameters have been developed, such as patches [4]–[7], bands [8]–[10] or watches [11]. Although wearable sensors have been made to monitor numerous different physiological parameters, such as blood pressure, heart rate, ECG, sweat composition, and respiration rate and volume, monitoring several different parameters with the same device would be of interest for severely ill patients and for use in emergencies. In situations where patients suffer from life-threatening conditions, it is of vital importance to continuously monitor respiration, heartbeat, blood pressure and ECG. We propose to create a device that contains multiple sensors integrated in a wearable patch, connected to power and communication modules, paired with advanced data analysis based on artificial intelligence, as depicted in Figure 1.

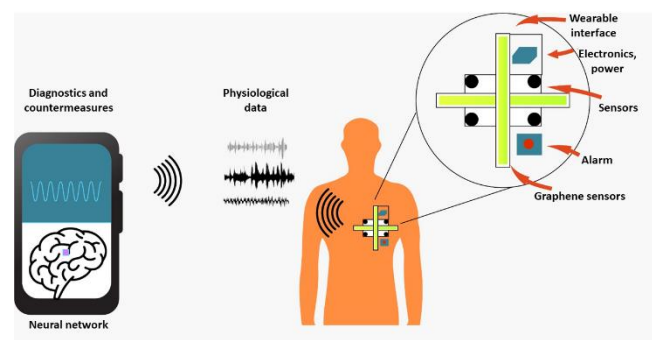


Figure 1: A sketch of the proposed telehealth monitoring system.

As steps towards this goal, we developed sensors for heart rate and respiration monitoring based on the novel material graphene [12].

The paper is organized as follows. In Section II we discuss the background of graphene, with particular focus on laser-induced graphene sensors. We also discuss commercially available wearable respiration and heartbeat sensors, including their advantages and limitations. We propose a wearable patch that contains graphene sensors for monitoring several physiological parameters at once. In Section III we describe the fabrication of graphene sensors, their use for physiological parameter monitoring, and signal analysis using the HeartPy module, followed by conclusions in Section IV.

II. BACKGROUND AND STATE OF THE ART

The discovery of graphene [13] has caused a research flurry, due to its unique and favorable electronic, optical, chemical and mechanical properties. The material is abundant, thin, flexible, electrically conductive, with a possibility to tailor its chemical reactivity by geometry and functionalization. Various embodiments of graphene sensors have been made for respiration monitoring [14]–[18]. Most recently, laser-induced graphene (LIG) has emerged as a

platform for sensors that can be made in custom shapes and dimensions with a facile fabrication process [19]. The use of LIG for respiration monitoring has been demonstrated [20]–[22]. We propose to make use of LIG respiration sensors as a viable technology that is competitive to state-of-the-art respiration monitoring technologies, both on the market and in development.

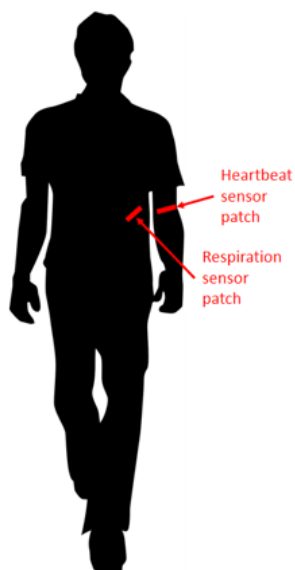


Figure 2: Sketch of the graphene sensor placement for respiration and heartbeat monitoring.

There are few scientific and commercially available solutions to monitor human respiration. One notable technology that has potential for use in wearable respiration sensors is piezoresistive thin films. The technology has been utilized in conjunction with a Bluetooth integrated system, as a standalone wearable respiration monitoring system [23]. The system can simultaneously measure both respiration rate (if only one sensor is attached to the body) and volume (if two sensors are attached to the body) with high fidelity. The standalone nature of the system, coupled with low invasiveness, is prospective for monitoring respiratory disease in real time and over extended periods. However, the piezoresistive sensors have a complex structure, requiring special laboratory conditions for fabrication. There are also a number of commercially available piezoresistive thin film products that can be utilized as components of such a wearable device.

Another technology that has been demonstrated in academic research is based on the triboelectric effect [24]. The solution makes use of a multilayer structure that has triboelectric properties. The structure is attached to a specially designed belt which is worn around the abdomen. As the wearer inhales and exhales, the belt stretches, pulling on the triboelectric structure. Electrodes connected to the structure are used to detect compression and strain due to respiration. The data is transferred wirelessly by connection with a Wi-Fi module. Advantages of this solution include the ease of fabrication and low cost, as the structure makes use of widely available foil materials such as copper, nylon, and PTFE. Also, the solution has been shown to have little interference from wearer activity, such as walking and standing up. Disadvantages include the belt format, which is restrictive and may not be applicable in emergency situations.

Aside from these academic research results, several solutions exist on the market or in commercial development. One of the most complete systems that is directly commercially available is respiBAN [25] — a professionally designed wearable respiration sensor with real-time acquisition and high respiration and motion data performance. The respiBAN has data-logging capabilities, allowing out-of-the-lab signal acquisitions without needing a permanent Bluetooth connection with a computer. It allows data acquisitions with up to 16-bit resolution at up to 3000 Hz sampling rate and continuous data streaming via Bluetooth for ten hours. The system houses three accelerometers and an inductive sensor, and the provided desktop software offers the option to export the raw signals for further analysis, as well as real-time visualization. Additional input/output ports are available for connecting other sensors, with the same company offering ECG, PPG, body temperature, and other sensors of physiological parameters that are compatible with the system. As with the solution based on the triboelectric effect, one disadvantage is that the belt is necessary for respiration measurement. A second disadvantage is the price, coming at ~2,000 EUR per piece.

RESPA [26] is a respiration sensor designed for fitness. The device has the shape of a miniature clip-on box that can be attached to the collar of a sweatshirt. The box can contain, apart from the respiration monitor, a microphone, motion sensor, or elevation sensor. All the sensors connect wirelessly to an app installed on the wearer's smart portable device. The app is geared towards fitness monitoring, tracking parameters such as ventilatory threshold. At the moment of writing it is not clear whether this solution is fully commercially available. There are several disadvantages to this solution, including the fact that there is no local storage option (the wearer's smartphone must be nearby during exercise, which most amateur and professional athletes will not agree to), and that the app does not provide access to the raw data files for custom analysis.

The wearable respiratory monitoring system from Resmetrix Medical [27] is geared towards detecting respiration abnormalities associated with conditions such as asthma, Chronic Obstructive Pulmonary Disease (COPD), or COVID-19. This product is designed to provide early warnings of acute episodes, alerting patients and their doctors. Also, real-time monitoring and connecting to a smartphone is available. Resmetrix makes use of a proprietary AI-powered algorithm that analyses patient condition and progress. The device comes in the form of a patch that is attached to the chest. As many other solutions, this one is not available for commercial purchase at the moment of writing, and it also does not seem to offer user access to the raw data for custom analysis.

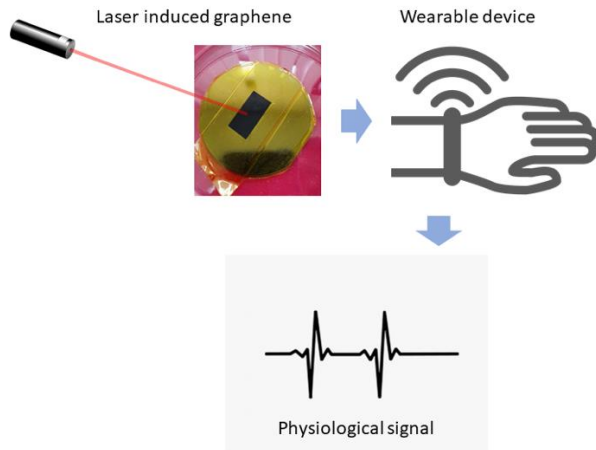


Figure 3: The process of making laser induced graphene, producing a wearable sensor, and obtaining physiological parameter data.

An athletic wearables company, Sweetzpot, makes a device (FLOW) to measure oxygen intake from breath [28]. Nevertheless, the device is not commercially available, and the performance or usage methodology are not clearly indicated for this device.

There are also solutions that are specialized for certain applications, such as the iBrevé [29], which tracks breathing, activity and stress levels. The device is geared towards “achieving calmness” and is paired to an app that trains the user for relaxed breathing. Such specialized devices are not compatible with use in other case scenarios because they most commonly do not let the user directly access the data, and are designed with limited use in mind.

Our proposal is to make LIG-based sensors that will independently measure heartbeat and respiration. This can be performed by attaching LIG sensors to two different parts of the body, as in Figure 2. A sensor attached to the forearm, across the median cubital vein, can detect heartbeat as the vein pulses. The LIG acts as a flexure sensor, employing the piezoresistive property of LIG on a flexible substrate. LIG has been used before to measure heartbeat and even ECG [22]. The same physical principle can be utilized to detect respiration, when the sensor is placed on the person’s ribcage. As the person breathes, the strain across the sensor changes, which causes changes in the measured resistance which are proportional to chest volume changes. The LIG sensor can be made to be stand-alone, attached to a portable battery supply and a communication module, similar to what has been shown with other, more complex sensors [23].

In a second iteration, the two different graphene devices can be integrated in the same patch that could be used to monitor both respiration and heartbeat simultaneously. It is likely that advanced signal processing would be necessary in that case, in order to filter out the heartbeat signal which will most likely be weak when measured on the chest.

III. EXPERIMENTS AND DISCUSSION

We produce laser-induced graphene in a similar manner as described before [19]. LIG was produced by scanning a CO₂ laser beam across the surface of polyimide tape. The laser used was a DBK FL-350 with maximum power of 60 W, with power set to 20%, scanning speed 900 mm/s, and resolution of 600 DPI. The devices were formed by laser-writing LIG in the shape of rectangles with dimensions 10x20 mm. Electrical

contacts were made to the LIG by attaching wires with silver paste at the ends of the device, or by attaching conductive copper tape. Material analysis, including Raman spectroscopy, scanning electron microscopy, electron diffraction spectroscopy, and Fourier transform infrared spectroscopy confirms the presence of graphene, with a level of oxide depending on the laser illumination conditions. Graphene devices were attaching the adhesive side of the polyimide tape to an adequate place on a subject’s body. For heartbeat measurements, the sensor was attached to the cubital vein on the forearm. For respiration monitoring, the sensor was attached to the ribcage. Wires that are connected to the sensor were connected to a Keithley 2450 SMU, which was interfaced to a desktop computer. Measurements were performed in the constant current mode with current set to 1 mA, and the voltage was measured over a period of several minutes. As a reference, the heart rate was measured with a free app installed on a smartphone. The reference respiration rate was obtained by manual counting. The process flow of the experiment is shown in Figure 3.

In the case of respiration monitoring, there was immediate and good agreement between the respiration rate measured with our sensor and that acquired by manual counting. In the case of heartbeat, the data was processed with the HeartPy module for Python [30]. The HeartPy process function allows different pre-processing options to clean up signals, including finite impulse response (FIR) filtering and outlier detection. The filtering was necessary when the signals contained noise due to subject motion or other external parameters. This function requires the input signal and frequency as input parameters and returns 13 different signal features, including bpm which represents the heart rate. It also returns the feature breathing rate in Hz which can be used for comparison with the sensor respiratory rate. Preliminary results indicate that both respiration and heartbeat monitoring are performed well by our LIG sensor, which is lightweight, unobtrusive, flexible, and inexpensive. An example of the value of physiological parameters measured in this way, returned by HeartPy, is shown in Table 1. The algorithm measures a heart rate of 81.50 bpm, which is in agreement with that measured during the experiment with the reference app.

ibi: 736.2	sdnn: 153.3	sdsd: 109.9	rmssd: 214.0
pnn20: 0.917	pnn50: 0.833	hr_mad: 122.45	sd1: 149.4
sd2: 154.0	s: 72295	sd1/sd2: 0.970	br: 0.20

Table 1: Parameters returned by HeartPy. Ibi: inter-beat interval, sdnn: standard deviation of intervals sdsd: standard deviation of successive differences between adjacent R-R intervals, rmssd: RMS of successive differences between adjacent R-R intervals, pnn50/pnn20: proportion of differences greater than 50ms / 20ms, hr_mad: median absolute deviation, Poincare analysis (sd1, sd2, s, sd1/sd2), br: breathing rate in Hz.

IV. CONCLUSION

The proposed graphene sensors are a good solution for wearable monitoring of physiological parameters such as heartbeat and respiration. Combined with the Python toolkit HeartPy, the signals can be processed in a satisfactory level as the existing commercial Heart rate and Respiratory rate sensors. Even more, several sensors integrated into a single

wearable smart-patch, together with power and communication modules, and coupled with smart algorithms are a very promising solution for rapid and continuous mass casualty monitoring [31]. A drawback of using LIG sensors is the large accompanying noise in the signal, which needs to be removed.

V. ACKNOWLEDGMENT

This paper was written thanks to the support of the "Smart Patch for Life Support Systems" - NATO project G5825 SP4LIFE, and support of the National project IBS4LIFE of Faculty of Computer Science and Engineering, at Ss. Cyril and Methodius University in Skopje. We acknowledge funding from the Ministry of Science, Education, and Technological Development of the Republic of Serbia, through grant no. 451-03-68/2022-14/200026.

VI. REFERENCES

- [1] X. Zhu and A. Cahan, "Wearable Technologies and Telehealth in Care Management for Chronic Illness," *Healthcare Information Management Systems: Cases, Strategies, and Solutions: Fourth Edition*, pp. 375–398, Sep. 2016, doi: 10.1007/978-3-319-20765-0_22.
- [2] A. Farrokhi, R. Farahbakhsh, J. Rezazadeh, and R. Minerva, "Application of Internet of Things and artificial intelligence for smart fitness: A survey," *Computer Networks*, vol. 189, p. 107859, Apr. 2021, doi: 10.1016/J.COMNET.2021.107859.
- [3] F. Misichroni et al., "A Novel, Reliable and Real-Time Solution for Triage and Unique Identification of Victims of Mass Casualty Incidents," *Engineering Proceedings 2021*, Vol. 6, Page 72, vol. 6, no. 1, p. 72, May 2021, doi: 10.3390/I3S2021DRESDEN-10180.
- [4] S. Anastasova et al., "A wearable multisensing patch for continuous sweat monitoring," *Biosensors and Bioelectronics*, vol. 93, pp. 139–145, Jul. 2017, doi: 10.1016/J.BIOS.2016.09.038.
- [5] M. Etemadi, O. T. Inan, J. A. Heller, S. Hersek, L. Klein, and S. Roy, "A Wearable Patch to Enable Long-Term Monitoring of Environmental, Activity and Hemodynamics Variables," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 2, pp. 280–288, Apr. 2016, doi: 10.1109/TBCAS.2015.2405480.
- [6] R. G. Haahr et al., "An electronic patch for wearable health monitoring by reflectance pulse oximetry," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 6, no. 1, pp. 45–53, Feb. 2012, doi: 10.1109/TBCAS.2011.2164247.
- [7] T. Elfaramawy, C. L. Fall, S. Arab, M. Morissette, F. Lellouche, and B. Gosselin, "A Wireless Respiratory Monitoring System Using a Wearable Patch Sensor Network," *IEEE Sensors Journal*, vol. 19, no. 2, pp. 650–657, Jan. 2019, doi: 10.1109/JSEN.2018.2877617.
- [8] Q. Zhang, D. Zhou, and X. Zeng, "Highly wearable cuff-less blood pressure and heart rate monitoring with single-arm electrocardiogram and photoplethysmogram signals," *BioMedical Engineering Online*, vol. 16, no. 1, pp. 1–20, Feb. 2017, doi: 10.1186/S12938-017-0317-Z/FIGURES/7.
- [9] V. P. Rachim and W. Y. Chung, "Multimodal Wrist Biosensor for Wearable Cuff-less Blood Pressure Monitoring System," *Scientific Reports 2019 9:1*, vol. 9, no. 1, pp. 1–9, May 2019, doi: 10.1038/s41598-019-44348-3.
- [10] Y. Zheng, C. C. Y. Poon, B. P. Yan, and J. Y. W. Lau, "Pulse Arrival Time Based Cuff-Less and 24-H Wearable Blood Pressure Monitoring and its Diagnostic Value in Hypertension," *Journal of Medical Systems 2016 40:9*, vol. 40, no. 9, pp. 1–11, Jul. 2016, doi: 10.1007/S10916-016-0558-6.
- [11] J. H. Moon, M. K. Kang, C. E. Choi, J. Min, H. Y. Lee, and S. Lim, "Validation of a wearable cuff-less wristwatch-type blood pressure monitoring device," *Scientific Reports 2020 10:1*, vol. 10, no. 1, pp. 1–9, Nov. 2020, doi: 10.1038/s41598-020-75892-y.
- [12] A. K. Geim and K. S. Novoselov, "The rise of graphene," *Nature Materials 2007 6:3*, vol. 6, no. 3, pp. 183–191, Mar. 2007, doi: 10.1038/nmat1849.
- [13] K. S. Novoselov et al., "Electric field in atomically thin carbon films," *Science (1979)*, vol. 306, no. 5696, pp. 666–669, Oct. 2004, doi: 10.1126/SCIENCE.1102896/SUPPL_FILE/NOVOSELOV.SOM.PDF.
- [14] Y. Pang et al., "Wearable humidity sensor based on porous graphene network for respiration monitoring," *Biosensors and Bioelectronics*, vol. 116, pp. 123–129, Sep. 2018, doi: 10.1016/J.BIOS.2018.05.038.
- [15] S. Andrić et al., "Ultrafast humidity sensor based on liquid phase exfoliated graphene," *Nanotechnology*, vol. 32, no. 2, p. 025505, 2021, doi: 10.1088/1361-6528/abb973.
- [16] Y. Su et al., "Surface acoustic wave humidity sensor based on three-dimensional architecture graphene/PVA/SiO₂ and its application for respiration monitoring," *Sensors and Actuators B: Chemical*, vol. 308, p. 127693, Apr. 2020, doi: 10.1016/J.SNB.2020.127693.
- [17] Y. Wang, L. Zhang, Z. Zhang, P. Sun, and H. Chen, "High-Sensitivity Wearable and Flexible Humidity Sensor Based on Graphene Oxide/Non-Woven Fabric for Respiration Monitoring," *Langmuir*, vol. 36, no. 32, pp. 9443–9448, Aug. 2020, doi: 10.1021/ACS.LANGMUIR.0C01315/SUPPL_FILE/LA0C01315_SI_001.PDF.
- [18] V. Selamneni and P. Sahatiya, "Bolometric Effect Enhanced Ultrafast Graphene Based Do-It-Yourself Wearable Respiration Sensor for Personal Healthcare Monitoring," *IEEE Sensors Journal*, vol. 20, no. 7, pp. 3452–3459, Apr. 2020, doi: 10.1109/JSEN.2019.2961156.
- [19] R. Ye, D. K. James, and J. M. Tour, "Laser-Induced Graphene," *Accounts of Chemical Research*, vol. 51, no. 7, pp. 1609–1620, Jul. 2018, doi: 10.1021/ACS.ACCOUNTS.8B00084/ASSET/IMAGES/ACS.ACCOUNTS.8B00084.SOCIAL.JPEG_V03.
- [20] Y. Yang et al., "A laser-engraved wearable sensor for sensitive detection of uric acid and tyrosine in sweat," *Nature Biotechnology 2019 38:2*, vol. 38, no. 2, pp. 217–224, Nov. 2019, doi: 10.1038/s41587-019-0321-x.
- [21] Y. Huang et al., "Health Monitoring and Automatic Notification Device Based on Laser-Induced Graphene," *IEEE Transactions on Electron Devices*, vol. 67, no. 10, pp. 4488–4492, Oct. 2020, doi: 10.1109/TED.2020.3015458.
- [22] B. Sun et al., "Gas-Permeable, Multifunctional On-Skin Electronics Based on Laser-Induced Porous Graphene and Sugar-Templated Elastomer Sponges," *Advanced Materials*, vol. 30, no. 50, p. 1804327, Dec. 2018, doi: 10.1002/ADMA.201804327.
- [23] M. Chu et al., "Respiration rate and volume measurements using wearable strain sensors," *NPJ Digit Med*, vol. 2, no. 1, Dec. 2019, doi: 10.1038/S41746-019-0083-3.
- [24] H. Zhang et al., "Waist-wearable wireless respiration sensor based on triboelectric effect," *Nano Energy*, vol. 59, pp. 75–83, May 2019, doi: 10.1016/J.NANOEN.2019.01.063.
- [25] "Plux Biosignals," <https://www.pluxbiosignals.com/collections/shop> (accessed Apr. 08, 2022).
- [26] "RESPA: World's 1st Breathing Sensor for Fitness — Zansors," <https://www.zansors.com/respa> (accessed Apr. 08, 2022).
- [27] "Wearable monitor could prevent rehospitalization for COPD - ISRAEL21c," <https://www.israel21c.org/wearable-monitor-could-prevent-rehospitalization-for-copd/> (accessed Apr. 08, 2022).
- [28] "FLOWTM - The First Consumer Wearable To Master The Art Of Breathing," <https://www.pnnewsire.com/news-releases/flow---the-first-consumer-wearable-to-master-the-art-of-breathing-300605830.html> (accessed Apr. 08, 2022).
- [29] "iBreve – Wearable helping you to relax, tracks stress & monitor your breathing - Innovation World Cup Series," <https://www.innovationworldcup.com/ibreve-wearable-helping-you-to-relax-tracks-stress-monitor-your-breathing/> (accessed Apr. 08, 2022).
- [30] P. van Gent, H. Farah, N. van Nes, and B. van Arem, "HeartPy: A novel heart rate algorithm for the analysis of noisy signals," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 66, pp. 368–378, Oct. 2019, doi: 10.1016/J.TRF.2019.09.015.
- [31] F. Lehoccki et al., "SmartPatch for Victims Management in Emergency Telemedicine," 2021 13th International Conference on Measurement, 2021, pp. 146–149, doi: 10.23919/Measurement52780.2021.9446791.

Activation functions' impact on regularization of deep neural networks application in atomic simulations

Ljubinka Sandjakoska
Faculty of Computer Science and Engineering
UIST St Paul the Apostle
Ohrid, Republic of North Macedonia
ljubinka.gjergjeska@uist.edu.mk

Ana Madevska Bogdanova
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, Republic of North Macedonia
ana.madevska.bogdanova@finki.ukim.mk

Abstract—When it comes to atomic simulations, the regularization of the deep neural networks is key to its successful application. The generalization capability of deep network depends on some factors. This paper aims to show that activation function is one of the most important factors that influence to decreasing the generalization error. For that purpose, several experiments were performed. Moreover, new approach for choosing the activation function is proposed. The purpose of the activation mechanism is not to find a universal, new activation function, although this is not excluded, but the most appropriate for the given task and for the given data set. The obtained results show that using the proposed activation approach, a decreasing of the mean absolute error compared to the benchmark set is achieved.

Keywords—activation, regularization, deep neural networks, atomic simulations

I. INTRODUCTION

Nowadays, developing advanced machine learning based approaches in many domains is more than a trend. Deep learning, as sophisticated concept of machine learning, allows solving the most specific issues related to the data. Those issues in the cheminformatics are striking and the automatic construction of complex features is crucial [1]. Due to the ability of deep learning to provide a hierarchical representation of a compound, where higher levels present more complex concepts [2], its application in atomic simulations give advantages that could not be achieved otherwise. That is noticeable in tasks such as: prediction of absorption, distribution, metabolism, excretion and toxicity (ADMET) properties [3], prediction of molecular activity, toxicity [4], quantitative structure activity relationship (QSAR) predictive modeling [5], prediction of the drug-target interaction [6], virtual screening [7] etc.

When it comes to atomic simulations, the regularization of the deep neural networks is key to its successful application. Regularized deep neural network is capable to predict on previously unseen data with a small error. Usually regularization is viewed as modification to the learning algorithm to reduce its generalization error but not its training error [8]. In [9] more general definition is given. The authors in [9] include various properties of the loss function, the loss optimization algorithm, or other techniques and define the regularization as any supplementary technique that aims at making the model generalize better. In this research, we also follow that broader definition.

The regularization methods can be classified into several groups: *methods that affect data* (generic data-based methods [10] - [14], [15], [16] and domain-specific data-based methods [17], [18], [19]), *methods that affect the network architectures*

[20], [21], [22], [23], *error terms* [24], [25], *regularization terms* [23], [27], and *optimization procedures*. The generalization capability of deep network depends on some factors. This paper aims to show that activation function is one of the most important factors that influence to decreasing the generalization gap defined as difference between the models' performance on training data and on new data.

The paper is organized as follows: in the second section the deep neural network architecture and the mechanism for the activation are presented. The third section includes the details of experimental design: description of the data sets, evaluation measures and the experimental setup. Results of the experiments followed by discussion are provided in the fourth section. At the end of this paper the concluding remarks are given.

II. METHODS

A. Deep Tensor Neural Network (DTNN) architecture

The deep architecture which deploy the proposed mechanism for choosing activation function is based on the architecture in [28]. The reference DTNN use *tanh* (*hyperbolic tangent*) activation function, which has S-shape and take real values as inputs and outputs in range -1 to 1. For consistency, we use data division: 80/10/10, training/test/validation. The prediction of the atomic properties using DTNN (Fig.1) is realized by: computing features, generating, embedding, feedforwarding, gathering and backpropagation. Since the nature of the atomic data, there is a need for special procedures to make the data suitable for manipulation. Those procedures do not include standard preprocessing such as denosing; elimination of redundancy; reduction; identifying and dealing with missing data; standardization; solving the problem with multicollinearity of data etc., but *featurization*, and *splitting*. Featurization is procedure for effective encoding of the molecules in the string with fix length or vectors.

The Deep Tensor Neural Network (Fig.1) does not include explicit binding information, the atomic features are updated using all other atoms based on their physical distances. In Table 1 details for the parameters of DTNN are given.

B. Activation mechanism

The proposed mechanism chose the activation function $f_A(x)$, which can have three general forms. The occurrence of each form is equally probable. The forms are:

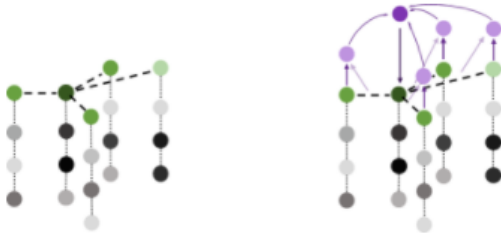


Figure 1. DTNN atomic model [29]

$$\begin{aligned} f_A(x) &= U_i(x), \\ f_A(x) &= U_i(U_{i+j}(x)) \text{ and} \\ f_A(x) &= B(U_i(x), U_{i+j}(x)) \end{aligned}$$

where $U_i(x)$ is unary function, while $B(U_i(x), U_{i+j}(x))$ is binary function where $i \in (1,15)$ and $j \in (1,14)$.

TABLE I DTNN ARCHITECTURE DETAILS

DTNN_model
n_tasks (int)
n_embedding (int, optional)
n_hidden (int, optional)
n_distance (int, optional)
distance_min (float, optional)
distance_max (float, optional)
mode (str)
compute_features_on_batch(X_b)
default_generator
dataset (Dataset)
epochs (int)
mode (str)
deterministic (bool)
pad_batches (bool)
DTNNEmbedding
n_embedding: int, optional
periodic_table_length: int, optional
init: str, optional
DTNNStep
n_embedding: int, optional
n_distance: int, optional
n_hidden: int, optional
init: str, optional
activation: str, optional
DTNNGather
n_embedding: int, optional
n_outputs: int, optional
layer_sizes: list of int, optional(default=[1000])
init: str, optional
activation: str, optional

The values of i and j are determined by the number of unary functions that are used. The set of unary functions used in this research contains the following functions 0, 1, x , $-x$, $|x|$, x^{-1} , x^2 , $x^{\frac{1}{2}}$, e^x , $e^x - 1$, $\log(e^x + 1)$, $x \cdot \sigma(x)$ (known as Swish), where $\sigma(x) = (1 + e^{(-x)})^{-1}$ is sigmoid function,

$x \cdot \tanh(\ln(1 + e^x))$ (known as Mish) and $\frac{x}{\frac{x}{\alpha} + e^{-\frac{x}{\beta}}}$, $\alpha, \beta > 0$ (as Soft-Root-Sign). The set of binary function include the following functions: $U_1 + U_2$, $U_1 - U_2$, $U_1 \cdot U_2$, $\frac{U_1}{U_2}$, $U_1^{U_2}$, $\max(U_1, U_2)$, $\min(U_1, U_2)$.

First of all, the search space is formed by the specified unary and binary functions as the sum of the permutations for all three general forms (Fig.1) in which the activation function occurs. The activation mechanism automatically selects one of the possible activation functions from the search space, formed

- For the first general form of the activation function, defined with $f_A(x) = U_i(x)$, we have to choose one from 15 unary functions. Or permutations $P(15,1) = \frac{15!}{(15-1)!} = 15$;
- For the second general form, defined with $f_A(x) = U_i(U_{i+j}(x))$, where $i \neq j$, we chose 2 from 15 unary functions, or $P(15,2) = \frac{15!}{(15-2)!} = \frac{15 \cdot 14 \cdot 13!}{13!} = 15 \cdot 14$ functions;
- For the third general form, $f_A(x) = B(U_i(x), U_{i+j}(x))$, we use 15 unary and 7 binary functions and obtain $7 \cdot \frac{1}{15} P(15,15) = 7 \cdot \frac{1}{15} \frac{15!}{(15-15)!} = 7 \cdot \frac{1}{15} \frac{15 \cdot 14!}{0!} = 7 \cdot 14!$ functions.

The total number of candidate functions is $15 + 15 \cdot 14 + 7 \cdot 14!$ i.e. 610 248 038 625. Due to computer resources, the number of used unary and binary functions is limited. The large number of possible activation functions further increases the complexity and causes difficulties in the activation mechanism. It should be emphasized that the purpose of the activation mechanism is not to find a universal, new activation function, although this is not excluded, but the most appropriate for the given task and the given data set. The activation mechanism solves a classic optimization problem, with iterative search - attempts to find the best of the candidates solutions according to the selected criterion.

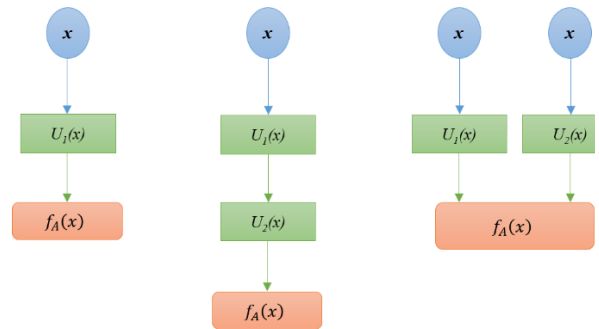


Figure 2. Combining functions

a) First general form b) Second general form c) Third general form

III. EXPERIMENTAL DESIGN

We use DeepChem¹ package, which is result of project that aims to create high quality, open source tools for drug

¹ <https://deepchem.io/>

discovery, materials science, quantum chemistry, and biology. In the implementation are used Scikit-learn, TensorFlow [30] and Keras [31]. For experiment tracking wandb.ai tool is deployed. In order to show the influence of choosing the activation function, well established activation functions are applied in different experiment runs.

A. Datasets

For the purpose of the research presented here we use four datasets from MoleculeNet² aggregator [29]: QM7, QM7b, QM8 and QM9. The datasets belong to quantum mechanics category for solving regression task.

1) QM7 and QM7b data set

The data sets contain 3D Cartesian coordinates and electronic properties of 7165 molecules. The main goal is to predict the following electronic properties:

- Atomization energy - PBE0
- Excitation energy of maximal optimal absorption - ZINDO
- Highest absorption - ZINDO
- HOMO - ZINDO
- LUMO - ZINDO
- 1st excitation energy - ZINDO
- Ionization potential - ZINDO
- Electron Affinity - ZINDO
- HOMO - KS
- LUMO - KS
- HOMO - GW
- LUMO - GW
- Polarizability - PBE0
- Polarizability - SCS

2) QM8

This data set is consisted of low-lying singlet-singlet vertical electronic spectra of over 20 000 synthetically feasible small organic molecules. QM8 data are used for prediction of electronic spectra in a high-throughput manner across chemical space:

- E1 - CC2
- E2 - CC2
- f1 - CC2
- f2 - CC2
- E1 - PBE0
- E2 - PBE0
- f1 - PBE0
- f2 - PBE0
- E1 - CAM
- E2 - CAM
- f1 - CAM
- f2 - CAM

3) QM9

QM9 data include geometric properties and energetic, electronic and thermodynamic properties in order to predict atomic properties using geometric properties (atomic coordinates) which are integrated into features

- mu
- alpha
- HOMO
- LUMO
- gap
- R2
- ZPVE
- U0
- U
- H
- G
- Cv

B. Evaluation measures

The proposed approach is evaluated using Mean Absolute Error (MAE), as a golden standard for regression task. The MAE is calculated using the formula $MAE = \frac{\sum_{i=1}^n |\hat{y}_i - y_i|}{n}$, where $\hat{y}_i$ denotes the predicted value for a single data using the model, while the actual value from the training set is denoted by y_i . Thus, the term *error* depict the difference between the predicted value for a variable and its actual (observed, measured) value.

C. Experimental setup

The optimization of the hyper parameters is done by Gaussian Process Optimization in maximum 20 iteration. The training of the models is limited and is not longer than 10 hours.

IV. RESULTS AND DISCUSSION

In order to depict the effectiveness of proposed mechanism for activation of the neurons and the impact on regularization in general, several experiments were obtained. The first experiment includes prediction of the atomic properties using 3 architectures: ANI-1 [32] - Extensible neural network, DTNN [28], MPNN [33] - Message Passing Neural Network. All models are graph-based, as a more convenient for modeling molecules, since the fact that the molecules and their features are often represented by graphs. The results are given in Table II and Fig.2.

TABLE II PERFORMANCE OF THE REFERENCE VS PROPOSED APPROACH

Dataset	Best performance	MAE kcal/mol [29]	MAE kcal/mol [proposed approach]
QM7	ANI-1	2.8600	0.80960
QM7b	DTNN	1.7700	0.03656
QM8	MPNN	0.0143	0.00129
QM9	DTNN	2.3500	2.14323

² <https://moleculenet.org/>

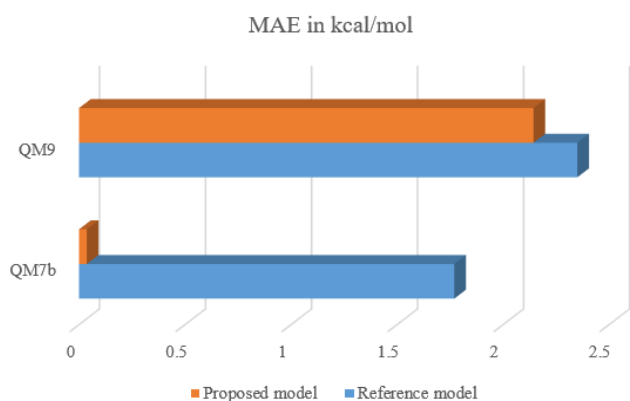


Figure 2. Deep Tensor Neural Network performance – MAE in kcal/mol: Reference vs. Proposed model

If we compare the performance of ANI-1 architecture and the DTTN architecture with applied proposed activation mechanism on QM7 dataset we can see improvement of the performance for more than 3.5 times. For the QM7b dataset the improvement is most noticeable (Fig. 2). We have decreasing of MAE value for more than 40 times. When it comes to MPNN, the obtained results with the applied activation approach also outperform the benchmark. For QM9 the difference is small, but in terms of atomic simulations small differences can have a big impact. Decreasing the MAE value, as indicator for regularized neural network, show us that activation function has a regularization effect on the network.

V. CONCLUSIONS

In this paper a different approach to regularization is presented. The main idea is to show that – the selection process of the activation function has a big influence to the need for regularization of the network. The new approach, named as an activation mechanism, don not try to find a universal, new activation function, although this is not excluded, but the most appropriate one for the given task and for the given data set. The obtained results show that the activation function can be viewed as an implicit regularizer.

REFERENCES

- [1] Bengio, Y., Courville, A. and Vincent, P., 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8), pp.1798-1828.
- [2] Bengio, Y., 2013, July. Deep learning of representations: Looking forward. In *International Conference on*
- [3] Hughes, T.B., Miller, G.P. and Swamidass, S.J., 2015. Modeling epoxidation of drug-like molecules with a deep machine learning network. *ACS central science*, 1(4), pp.168-180.
- [4] Unterthiner, T., Mayr, A., Klambauer, G. and Hochreiter, S., 2015. Toxicity prediction using deep learning. *arXiv preprint arXiv:1503.01445*.
- [5] Ma, J., Sheridan, R.P., Liaw, A., Dahl, G.E. and Svetnik, V., 2015. Deep neural nets as a method for quantitative structure–activity relationships. *Journal of chemical information and modeling*, 55(2), pp.263-274.
- [6] Unterthiner, T., Mayr, A., Klambauer, G., Steijaert, M., Wegner, J.K., Ceulemans, H. and Hochreiter, S., 2014. Deep learning for drug target prediction. *Work. Represent. Learn. Methods complex outputs*.
- [7] Hamanaka, M., Taneishi, K., Iwata, H., Ye, J., Pei, J., Hou, J. and Okuno, Y., 2017. CGBVS-DNN: Prediction of Compound-protein Interactions Based on Deep Learning. *Molecular informatics*, 36(1-2), p.1600045.
- [8] I. Goodfellow, Y. Bengio and Aaron Courville (2016). Deep learning. (MIT Press.)
- [9] J. Kukačka, V. Golkov, and D. Cremers. “Regularization for deep learning: A taxonomy”. [Online]. Available: ArXiv:1710.10686
- [10] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, (2012). “Improving neural networks by preventing co-adaptation of feature detectors”. ArXiv: 1207.0580.
- [11] X. Bouthillier, K. Konda, P. Vincent, and R. Memisevic, (2015). “Dropout as data augmentation.” ArXiv: 1506.08700.
- [12] P. Morerio, J.Cavazza, R. Volpi, R. Vidal, and V. Murino, (2017).Curriculum dropout. ArXiv:1703.06229.
- [13] S. Maeda, (2014). A Bayesian encourages dropout. ArXiv:1412.7003.
- [14] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, (2016a). “Densely connected convolutional networks.” ArXiv: 1608.06993
- [15] T. DeVries, and G. W. Taylor, (2017). “Dataset augmentation in feature space”. In Proceedings of the International Conference on Machine Learning (ICML), Workshop Track.
- [16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, (2016). “Rethinking the inception architecture for computer vision.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp: 2818–2826.
- [17] A. Fawzi, S. Horst, D. Turaga, and P. Frossard, (2016). “Adaptive data augmentation for image classification.” In Proceedings of the IEEE International Conference on Image Processing (ICIP), pp:3688–3692
- [18] J. Salamon, and J. P. Bello, (2017). “Deep convolutional neural networks and data augmentation for environmental sound classification.” *IEEE Signal Processing Letters*, 24(3) pp: 279–283.
- [19] D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber, (2010). “Deep big simple neural nets excel on handwritten digit recognition.” *Neural Computation*, 22(12) pp: 1–14.
- [20] V. Dumoulin, and F. Visin (2016). “A guide to convolution arithmetic for deep learning.” ArXiv: 1603.07285.
- [21] F. Yu, and V. Koltun (2015). Multi-scale context aggregation by dilated convolutions. ArXiv: 1511.07122.
- [22] J. Long, E. Shelhamer, and T. Darrell, (2015). “Fully convolutional networks for semantic segmentation.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp: 3431–3440
- [23] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, (2011c). Contractive autoencoders: Explicit invariance during feature extraction. In Proceedings of the International Conference on Machine Learning (ICML), pp: 833–840
- [24] S. Ruder, (2017). An overview of multi-task learning in deep neural networks. ArXiv: 1706.05098
- [25] F. Milletari, N. Navab, and S. A. Ahmadi, S. A. (2016). “V-net: Fully convolutional neural networks for volumetric medical image segmentation.” In Proceedings of the International Conference on 3D Vision (3DV), pp: 565–571. IEEE
- [26] M. Sajjadi, M. Javanmardi, and T. Tasdizen, (2016). “Regularization with stochastic transformations and perturbations for deep semi-supervised learning.” In *Advances in Neural Information Processing Systems* (NIPS), pp: 1163–1171
- [27] B. Neyshabur, “Implicit Regularization in Deep Learning,” Ph.D. dissertation, TOYOTA Technological Institute at Chicago, 2017.
- [28] Schütt, K.T., Arbabzadah, F., Chmiela, S., Müller, K.R. and Tkatchenko, A., 2017. Quantum-chemical insights from deep tensor neural networks. *Nature communications*, 8(1), pp.1-8.
- [29] Wu, Z., Ramsundar, B., Feinberg, E.N., Gomes, J., Geniesse, C., Pappu, A.S., Leswing, K. and Pande, V., 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical science*, 9(2), pp.513-530.
- [30] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M. and Ghemawat, S., 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.
- [31] <https://keras.io/>
- [32] Smith, J.S., Isayev, O. and Roitberg, A.E., 2017. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical science*, 8(4), pp.3192-3203.
- [33] Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O. and Dahl, G.E., 2017. Neural message passing for quantum chemistry. *arXiv preprint arXiv:1704.012*

STUDENT PAPERS

Virtualization technology and possibilities of its use in education

Bojan Karkalashev

*Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
bojan.karkalashev@students.finki.ukim.mk*

Marjan Gushev

*Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
marjan.gushev@finki.ukim.mk*

Abstract—In this research, we review the possibilities and benefits of using virtualization technologies in a secondary education environment. All tests were made on-site at the "Orce Nikolov" High School in Skopje, using real equipment and in real classroom scenarios. Our main goal with this research was to provide creative and better ways for the school to: manage their IT equipment, cut costs, provide better learning environments for students, but also for teachers.

Index Terms—Education, Human-Computer Interaction, Virtualization

I. INTRODUCTION

Over the past years and especially lately with the current COVID-19 world situation, most educational institutions have experienced many problems related to moving to the cloud, but also with the system administration of IT equipment and laboratories. Many of these problems have found elegant solutions in the deployment of virtualization suites. Providing schools with solutions for the problems mentioned above was the main idea behind this research.

Hardware virtualization refers to the creation of a virtual machine that acts like a real computer with an operating system. Software executed on these virtual machines is separated from the underlying hardware resources. For example, a computer that is running Microsoft Windows may host a virtual machine that looks like a computer with the Ubuntu Linux operating system.

Desktop virtualization means that each desktop lives in a virtual machine, which is partitioned off from other virtual machines in the server. The operating system and applications are not shared resources, but they still physically live on a remote server. These virtualized resources can be accessed from any device that is able to connect to the server.

A thin client is a computer that runs from resources stored on a central server instead of a localized hard drive. Thin clients work by connecting remotely to a server-based computing environment where most applications, sensitive data, and memory, are stored.

II. RELATED WORK

M. Klement [1] explored ways of integrating virtualization in education, the costs and benefits from it. A. Gaspar, S. Langevin, W. Armitage, R. Sekar and T. Daniels [2] explored the role virtualization has in education while mostly focusing on virtualization of the IT classroom and laboratory from two perspectives: the perspective of the system administrator and the perspective of the instructor.

Chandra and Malaya [3] were focused on how cloud computing and virtualization can help with the increasing pressure educational institutions have to deliver more for less

and the need to find ways to offer rich, affordable services and tools. High energy costs and IT equipment costs are also one of the biggest problems facing IT in education, the Erskine and Füstös [4] in their case-study, researched solutions such as using thin-clients instead of full desktop computers, delivering educational desktop environments to students home or dorm room computers and so on. Anderson, Joines and Daniels [5] study might be one of the most interesting articles currently because of the problems students and universities are facing not being able to provide education in computer labs on campus with the current COVID-19 situation.

Other studies such as the ones from Karlov [6], Barionuevo, Gil, Giribaldi, Suarez and Taffernaberry [7], Xu, Yang and Lei [8], Lundsford [9] and Sultan [10] have also inspired and improved the research in this paper.

III. USE-CASE SOLUTION ARCHITECTURE

This paper focuses on the use of virtualization technologies in during a IT course. Usually, educators use a dedicated computer laboratory with computers isolated from the campus networks for IT courses; however, in case a dedicated computer laboratory was not available due to construction delays. As a result, this educator decided to test the use of virtualization technologies in this course.

The computer laboratory used for this course consisted of student workstations and an educator's workstation connected to the campus network. Students used the computer laboratory heavily throughout the day. Most classes made full use of the computer laboratory with all available computers taken. As a result, it was not possible to dedicate some computers to this course. Fortunately, the educator's workstation had sufficient resources to run VMWare Workstation 12. The educator's workstation consisted of an Intel 6-core Xeon Processor (3,20 GHz) with 32 GB RAM, 512GB Solid State Drive, 23,8" flat-panel display, and an Internet connection via a local area network. The student's workstations are meant only to be used as Remote Desktop Connection clients. Each computer consisted of an Intel dual-core processor (2.90 GHz) with 2 GB RAM, an 120 GB Solid State Drive, 21,5" flat-panel display, and a restricted Internet connection via a local area network. Using Remote Desktop Connection, each student can connect in a personalized hosted virtualization environment depending on the lesson, employing hardware virtualization on the educator's workstation to enable desktop virtualization.

This IS educator selected VMWare Workstation as the virtualization platform for several reasons. First, VMWare provides a range of virtualization products, including server

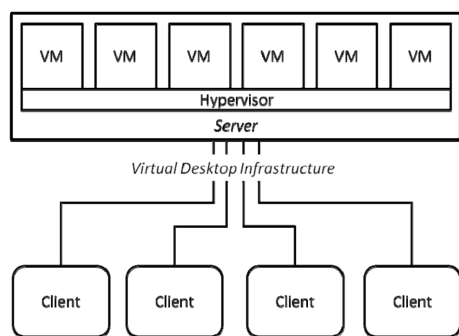


Fig. 1. Computer Lab (Classroom) Virtualization Architecture

and desktop virtualization. Second, VMWare Workstation supports a range of hardware, including universal serial bus (USB) 2.0 storage devices natively, while other virtualization products require the user to setup a shared folder with the host operating system to access these devices. Third, VMWare Workstation supports a wide range of guest operating systems, including most versions of Microsoft Windows, Linux distributions, and several other operating systems. Fourth, VMWare has an active user community providing an assortment of virtual appliances, which are pre-configured virtual machines for various applications. Finally, VMWare operates an academic program (<http://www.vmware.com/partners/academic/>) that makes select VMWare products available to the academic community free of charge for instructional purposes.

Most virtual machines used in the course were pre-build ones, specially designed for the course. These virtual machines are designed to be a closed environment where students don't have many permissions. Though sometimes students work in a virtual machine of their own creation, they have complete control over the virtual machine. This is desirable in settings where the student is acting as a system administrator; however, this can also make it difficult for the educator to monitor the students' activities or assist students in solving some problems. In other settings where the student does not require administrator control over the virtual machine, the IS educator may choose to distribute a virtual machine with a pre-defined student account with the appropriate level of access. In this course all students were authenticated using their Active Directory authentications or a predefined password.

IV. DISCUSSION

For the duration of this research the our experience with involved parties was: The school's administration benefited and expressed satisfaction with the lower costs for energy and IT equipment, they were satisfied that they can use their current equipment and network infrastructure for virtualization. The educator benefited and expressed satisfaction with the course and the use of virtual machines. In addition to the ease of management of the student lab and monitoring of the student activities, he experienced better results from students. The class preparation was much easier because the educator needed to prepare only several virtual machines for the course, instead of a dedicated computer for each student. The also students expressed satisfaction with the course and the use of virtual machines. In addition to developing technical skills, the students demonstrated an increased

level of confidence in working with advanced features of Windows and other applications over the course. One student commented that the student felt using the virtual machine helped the student become more comfortable and confident with existing and new skills, because of the bigger freedom.

The main benefits from using virtualization in education are:

- Using a centralized system for Operating System & application updates;
- Migration can be done regularly by a IT technician at the data center collectively for a group of virtual desktops;
- Desktop Virtualization solutions integrate with corporate directories like Active Directory and User Access Policies can be applied to all users based on their role, location, type of device, etc.;
- Enhanced security as the administration is centralized – Antivirus/ Firewall policies can be applied and monitored from the data center. For example the applications that can be accessed and run by the students can be restricted and even tracked;
- Faster, reliable and easier backup/ recovery of all the user data is possible due to centralized hosting and administration;
- Desktops, Laptops, Thin-Clients, Tablets or even Smart Phones could be used by users (where ever they are) to access the centralized hosted applications and desktop environments;
- For applications that require heavy CPU or GPU processing, the performance with Desktop Virtualization might be better, as the applications are now being processed by the more powerful server processors, than the desktop processors;
- Desktop Virtualization software generally has settings to allow administrators to allocate or limit server resources that can be used per user, so that an individual user doesn't over-consume resources.

While most of the experiences by everyone involved were generally positive, the use of virtual machines in an instructional setting introduces some challenges and limitations over the use of dedicated computers.

The main disadvantages from using virtualization in education are:

- The educational institution needs to buy the Desktop Virtualization Software licenses, Servers, Centralized Storage infrastructure, Upgrade Network infrastructure to support more bandwidth, etc. in addition to buying computers or thin-clients for each user;
- IT administrators and teachers need to learn how to manage the Virtualization software which can be difficult for some and professional training can be expensive for the educational institution;
- There is no reduction in the number of end-user client machines that are needed in the network;
- The licenses for Operating Systems, applications, etc., still needs to be bought for each user (mostly) and there is no reduction of costs there (except in some application that benefit from discounts for Educational purposes);
- The thin-clients are sometimes as expensive or even more expensive than individual computers;
- The network infrastructure needs to be able to handle all the extra bandwidth that Desktop Virtualization is going

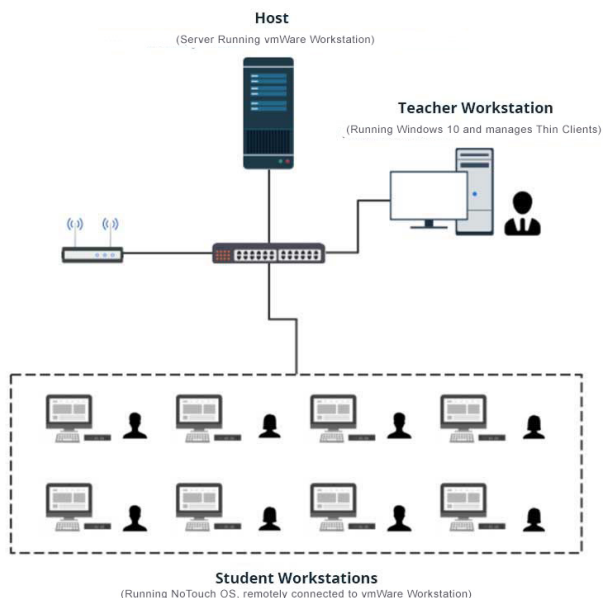


Fig. 2. Complete schema of the setup used on the test site, with a server, workstation PC for the teacher and thin clients as student workstations all of them connected by a Local Area Network (FTP Cat6A)

to introduce. Otherwise, it has to be upgraded (which can be costly);

- The WAN links need to have sufficient bandwidth to handle all the remote users, as well;
- If the bandwidth on the remote end is not sufficient or if there is congestion in Local Area Network, the display quality may not be as good as processing and viewing applications directly from a desktop;
- Its difficult to handle graphics and high-definition video with Desktop Virtualization.

V. CONCLUSION

Information Technologies are the future of education and Virtualization and moving to the cloud is the future for Information Technologies. Both students and teachers can benefit from the introduction of Virtualization and Cloud technologies to the classroom, but this need to be done right with carefully planned needs and resources, proper budgets for the equipment and licenses and professional training for everyone involved. The school's administration needs to understand the benefits and disadvantages of Virtualization and then decide if they are ready to go "cloud". While our experience was mostly positive and new technology was easily adopted and appreciated by most, the use of virtual machines in an educational setting still left a small number of people sceptical.

In our final meetings with the school's administration and IT personnel we discussed doing a "Phase 2" of this research where they could try and move their current applications that they use in their finance, archive, library and educational departments to the cloud. They also provided a lot of interest in applying desktop virtualization to their Android based tablets for students and teachers that are currently not utilized enough by students because of the operation system restrictions.

- [1] M. Klement, "Models of integration of virtualization in education: Virtualization technology and possibilities of its use in education," *Computers & Education*, vol. 105, pp. 31–43, 2017.
- [2] A. Gaspar, S. Langevin, W. Armitage, R. Sekar, and T. Daniels, "The role of virtualization in computing education," *ACM SIGCSE bulletin*, vol. 40, no. 1, pp. 131–132, 2008.
- [3] D. G. Chandra and D. B. Malaya, "Role of cloud computing in education," in *2012 international conference on computing, electronics and electrical technologies (ICCEET)*. IEEE, 2012, pp. 832–836.
- [4] M. A. Erskine and J. T. Füstös, "Survey of desktop virtualization in higher education: An energy-and cost-savings perspective," *aisel.aisnet.org*, no. 8, 2013.
- [5] B. R. Anderson, A. K. Joines, and T. E. Daniels, "Xen worlds: leveraging virtualization in distance education," in *Proceedings of the 14th annual ACM SIGCSE conference on Innovation and technology in computer science education*, 2009, pp. 293–297.
- [6] A. Karlov, "Virtualization in education: Information security lab in your hands," *Physics of Particles and Nuclei Letters*, vol. 13, no. 5, pp. 640–643, 2016.
- [7] M. Barrionuevo, C. Gil, M. Giribaldi, C. Suarez, and C. Taffernaberry, "Virtualization in education: Portable network laboratory," in *Argentine Congress of Computer Science*. Springer, 2017, pp. 90–98.
- [8] Z. Xu, L. Yang, and J. Lei, "Conception and design of desktop virtualization cloud platform for primary education: based on the citrix technology," in *2015 International Conference of Educational Innovation through Technology (EITT)*. IEEE, 2015, pp. 226–230.
- [9] D. L. Lunsford, "Virtualization technologies in information systems education," *Journal of Information Systems Education*, vol. 20, no. 3, p. 339, 2009.
- [10] N. Sultan, "Cloud computing for education: A new dawn?" *International Journal of Information Management*, vol. 30, no. 2, pp. 109–116, 2010.

Open source BI tools for SMEs: Comparative analysis of Pentaho CE and Knowage CE stacks

Igor Jakimovski

Faculty of Computer Science and Engineering

Skopje, Republic of North Macedonia

jakimovski.igor.1@students.finki.ukim.mk

Abstract—It is well known that Business Intelligence (BI) plays an important role in helping companies in dialing with competition and gain better success in business as many published successful world companies stories show. But how many companies in the domestic market have implemented it? BI is a discipline that transforms data into information that can be used to accelerate the organization's business success. In other words, BI uses data with a bigger historical diapason for comparison with the present ones, for responsible companies people, be able to draw conclusions on what actions to take in the near future or which steps to choose for making better business results or keeping business on the wanted level. This paper makes a proposal while evaluating of two different concrete implemented BI stacks, that could be implemented for small and medium-sized organizations. Evaluation is based on the experience while implementation, testing, and of course manuals of the developers as of the community support.

Keywords—BI, DW, ETL, PDI, TOS, OLAP

I. INTRODUCTION

The importance of information is well known. But how do to collect really important information that can help do better business? Finally, how can do use that information at all? According to the parts of [1] and [2], BI covers all processes beginning with collecting data from the sources to the final visual presentation of the information, that can be used for decision support or others.

Although in more developed business environments around the world, BI has been implemented for years, and the positive experiences can be read, our domestic business environment does not emphasize the BI at the level it deserves to be, nor much has been researched for technology comparisons.

The purpose of this paper is as a manual that is based on personal experience to present with and know the domestic small and medium-sized companies and their Managements through IT sectors with the BI, and trigger them to try to implement it. For this purpose, two free-of-charge and enough serious BI stacks (their Community Edition (CE) versions can be used in production very well), were used. Let me repeat, the evaluation was made on Pentaho CE and Knowage CE (with Talent Open Studio for ETL) with all processes from ETL to visualization. As in [3] and following the processes, four different Data Warehouses were created, two using the PDI tool with loading data in MySQL and PostgreSQL databases and another two using the TOS tool also loading the data in MySQL and PostgreSQL databases. A comparative analysis is made for each performed BI process which highlights the more specific features or shortcomings of the respective tools if any. As a final result, the paper proposes the best combinations of tools, from several aspects, that could help the future BI implementers in selection.

The paper begins with a brief theoretical description of the topic, after which goes through all the BI processes with theoretical and practical explanation, to move to a comparative analysis of technologies, after which it ends with a concrete conclusion.

II. IMPLEMENTATION

Before starting the implementation, first have to learn which kind of data the users like to have available in the Data Warehouse DW, or the users' requirements as in [2][5]. As a trading company [3],

had financial and commercial data available, but generally would be very good the DW contains manufacturing data too. Next, to discover all tables and find what data from the existing tables would fit in the DW tables, a very precise analysis of the transaction operational database should be made. If there is a need for data from other sources, should be found and analyzed too as well. First, the solution to the dimension tables should be found. For example, every transaction operational database of an organization contains some kinds of tables with the data for articles, clients, and so on. Every one of those tables fits in some dimension would be needed should be analyzed, and minimum possible fields to be chosen. A similar should be done for the fact tables later too. In this case [3], dimension tables for articles, clients (including suppliers in the same dimension), accounting codes, company organization units, document type, and calendar as a date dimension. Once the tables and the fields had been chosen, the ETL process can start organizing.

A. Architecture

As in [3], the architecture of implemented BI system contains two layers, the bottom one that houses the DW connected to the data sources through the ETL process. The upper layer houses the user applications on the upper part and OLAP servers on the lower one, connected with the DW databases. DW is practically a database, multidimensional or here relational, optimized for reading, aggregating, and searching big amount of data that is updated regularly in certain terms, usually when the sources from which it is charged are free and which is designed to store a long-range of historical data. As in [1] the tables of the DW database are connected in specific schematic structures, from which the most known are the star, snowflakes, and constellation schemas. As the name says, the star schema looks like a star, with a fact table in the middle, and dimension tables on the angles, connected directly to the fact table forming composite primary keys using their foreign. The snowflake schema looks like the star one but allows connections between dimensions or has a higher normalization level, and the constellations schema allows more than one fact table and different connections. The point of a DW database is faster return results of the queries, so the database should be as less normalized. As for the tables, the fact ones contain the measures and the dimension tables contain attributes that describe the measures.

B. ETL process

As described in [1], ETL – extract, transform, loading means that we have to choose the data from the sources, extract them on the stage area where will do cleansing and transformation before loading the ready data in the DW database. The ETL process is very important and statistically can spend 70-80% of the time needed to develop a BI. After the tables and the fields of the tables that would be used for the dimension tables are found, analyzing table by table to search for dirty data (refers to data that contains erroneous information) could start. The data can be dirty because for different reasons, writing data in wrong fields, wrong names, dates formats and etc, but all those have to be corrected. How this is done in practice? As in [3], PDI and TOS too, contain different steps for different actions. The job in TOS or transformation in PDI for ETL data in dimension tables (except calendar one) starts in the same way, with reading (extract) data from the original tables from the basic data source. As a row by row is read, the row flows to the first of the cleansing steps, then the second, third, then some other

purposed steps, and so on until finished all programmed for cleansing and transforming the data. The last step is loading in the DW database, in the given dimension table. As in [4], specific for the dimension tables is that they should contain information on the changes of the data that could occur through time and the way to handle them. That is the part managed by the slowly changing dimension (SCDI) named principle that offers seven ways how we can handle the changes, starting from SCDI 0 as ignoring the changing, SCDI 1 as overwriting the changing with deleting the history, to keeping history details in new created database's tables. The most used is SCDI 2 which keeps history with the versions and dates of change added in newly created columns of the existing tables. So for the final loading of the data in the dimension table, the SCDI loading step with set the type for all columns from rows would be used.

According to [5] and [6], both PDI and TOS handle all steps from reading to loading very well, in their own way. They have different steps for different actions according to their working principles. The creation of the end tables, in PDI is going by pressing a button while setting the transformation while TOS does it automatically. For reading the sources, both tools have steps for it, for working with strings, PDI has separate steps while TOS mainly uses the tMAP step for different actions including the mentioned one. Both tools allow using regular expressions but with their own syntax, also java code expressions. The steps in the dimension in both tools are starting with reading the original tables from the source databases listed by the years we set, sorted from the oldest ascending, followed by one or more cleansing steps as strings operations depending on a cleansing procedure, ending with SCDI loading steps. The exception of the normal dimensions is the calendar one that has to be created from the scratch setting the start date as the attribute for both tools, starting with generate rows step, then adding date sequence and calculate dates steps followed by select values and calculate additional fields steps, ending with the loading of the record as the last step in PDI version. The TOS version starts with the tGenerateRows step followed by the tMap step (that is doing all other PDI steps job), ending with the step that loads the data in the table. That means this transformation or job is not dependent on source data. In that case, the loading step should not be SCDI as not depends on the changes, and the normal loading step works very well in TOS too. Once the dimension tables loading has finished, the same with the facts tables should be done too. The principle is the same with the extra steps for joining the dimensions with the facts, calculating final measures, and the last step which is normal loading in the database table. The source tables should be searched between the ones containing input and output transactions as for receipt and sale of goods for the commercial part, same for input and output of accounting document for the financial part. In this case, the commercial part transformations in PDI or jobs in TOS begin with the reading of two source tables which continue in two separate streams, then a few filters on both streams followed by joining of the two streams, then again filtering after which coming joining with the dimensions, then calculating measurements step and at the end loading in the table step. As to the financial fact tables transformations or jobs, the same starts with a reading source table followed by four filters according to dates and accountant codes, then dimension joining followed by calculating measurements step and at last the step for loading data in the table.

According the [3], the only problem that could be found while executing was a bad speed performance for the SCDI loading step using TOS. That step loads so slow that is not worth using. For comparison, PDI executes all dimensions transformations connected serial for seven to eight minutes while TOS executes only the smallest one for a similar time. According to the explanation, the recommendation is to do the complete ETL process with the PDI tool, or if any want to do a mixture can do with the TOS tool only the ETL process for the facts.

C. Data Warehouse

As the transformations and the jobs have been created and executed in the order dimensions then facts, four loaded databases for four DW have been created. The last step of the DW part is the scheduling of updating new data. Depending on the usual data amount, the ETL process of only fiscal year data can be scheduled

to run in some therm out of the working hours once a week or oftener, as in [7]. With all this previous, the DW part has finished and the next BI parts can be worked on.

D. OLAP

According to [1], DWs that are low-level in BI architecture are complemented by a set of top-level technologies that provide specialized functionality for different BI scenarios through user applications and tools that lie on top of them. OLAP is just one of such technologies which is the background of many BI applications. The OLAP server is a core component of OLAP technology that sits between the DW and front-end client applications and provides functionality for aggregating, calculating, and analyzing historical data stored in the DW. It optimizes the questionnaires it downloads from client applications, redirects them to the DW database, and sends the data received from the DW to the same client applications. As a result, the end-users receive the required information presented in an easily recognizable form in seconds.

One of the most popular OLAP servers is Mondrian which supports Relational OLAP. OLAP offers a wide range of operations.

Drill down operation refers to dis-aggregation, viewing data at the level of magnified detail. The roll-up operation performs aggregation of the data cube by climbing the hierarchical tree, Slice and Dice operations are performed so that a large segment of data is cut into smaller pieces and this process is repeated until the right level of detail for analysis is reached. Pivot or rotation is a visualization operation that rotates the data axes to provide an alternative presentation of the data. It may contain replacing rows and columns or moving one of the row dimensions to the column dimensions. As in [1], The most common OLAP architectures encountered are ROLAP - Relational OLAP, applications based on relational databases, MOLAP - Multidimensional OLAP, applications based on multidimensional databases, and HOLAP - Hybrid OLAP, applications use both relational and multidimensional techniques.

For practical implementation, the first step is creating an OLAP schema [8] that represents an XML file containing a description of the OLAP environment with the dimensions and facts the cube contains. The file can be written in some text editor or through an application built for it as SchemaWorkbench SW is [8], developed to build Mondrian schemas. In the implementation, the SW application has been used to create the schemas and a text editor to create virtual cubes in the schemas as the application does not do it. The SW application allows checking the schemas using MDX statements writing and executing through the executor the application contains. Once the OLAP schema has been created, should be loaded into the Mondrian servers. In the Pentaho CE environment that operation is going through the Pentaho console choosing "add a new source" and importing the existing schema in the repository, while in the Knowage environment the schema should be uploaded through the "Mondrian schemas catalog" sub-menu from the administration's menu of the server. Now, that OLAP schemas have been uploaded, the transformations of the data into the information using visualization tools can be started

E. Visualization

As in [2], the upper end of the BI architecture is the part where business users use client tools for business intelligence. The responsibility of these tools is to perform clear transmission and visualization of data from data sources through DW to end-users. This level is sometimes called the presentation of knowledge because it is responsible not only for presenting the data but also for presenting it in a format that is easy and understandable to consume. Typical BI implementation does not use a single type of presentation software but includes more specific tools for different purposes. For example, if the CEO of a company prefers a high-level business overview for which he sees data in high visual formats, such as dashboards, a well-trained financial analyst may prefer a powerful spreadsheet or spreadsheet-like format instead of simple graphical representations. This is an example of why most BI software implementations provide a mixed set of tools tailored not only to specific functionalities but also to user types. Some of the more used and standard BI tools are spreadsheets, querying and

reporting software, OLAP analyzers, digital boards and scorecards, business performance management, ETL tools, and others.

According to [9] and [10], using both technologies, the next visualization documents have been created:

Reports - two tabular reports, the first one shows the top buyers for a chosen year, and the second gives an analytic view of the buyers also for a chosen year, and two reports with charts, the first one shows the perceptual sales for all cities for a chosen year presented with a pie chart, and the second sales per cities presented with bar chart also for a chosen year. Should be mentioned that the created DW contains 20 years of history and the reports are parameterized with the year parameter that has to be chosen when the report is run. With that, the reports have some kind of interactivity.

Dashboards - two types static and interactive. The static dashboards are for use in management (showing present and historic financial data for comparison) and for use in commerce (showing buy/sale data). Both contain different elements like charts, tables, and maps. The interactive dashboards are for use in commerce and contain elements like charts, tables, and Pivot tables.

OLAP analytic documents - for use in the commerce department and show Pivot tables and charts (in Pentaho CE).

The procedure of creating is very long for this paper so will not be explained, but the conclusions of the creation and use will be summarized in comparative analyses of the stacks.

III. COMPARATIVE ANALYSES OF THE BI STACKS

A. Functionalities of the Business Intelligence stacks

As in [11], Hitachi Ventura as the owner of Pentaho, as time goes by mentions the CE version less and less. Unlike the Pentaho EE stack, which automatically integrates all modern BI functionalities, with the installation of the Pentaho CE stack you get only: Pentaho BA Platform - Server as a base software part that houses content created on the server itself and includes functions for managing security, executing reports, displaying whiteboards, bursting reports, scripted business rules, OLAP analysis, and scheduling. Additional functionalities are upgraded through commercial plugins from the Pentaho market. As standalone software tools that can be installed and used for free are: Pentaho Data Integration PDI with Pentaho for Big Data as a plugin, Pentaho Report Designer PRD, Pentaho Data Mining Weka, Pentaho Metadata Editor PME, Pentaho Aggregate Designer PAD, Pentaho Schema Workbench PSW and Pentaho Design Studio PDS, as from Pentaho market plugins which are numerous, the relevant ones for the case are: CTools with Community Dashboard Editor CDE, Community Dashboard Framework CDF, and Community Graphic Generator CGG. Ivy CDE connectors complete as CTools plugins, Pivot4J Analytic, Datafor Visualizer, and BTable.

With all summarised [3], Pentaho CE has the following functionalities: ETL - with PDI, Reports - creating reports with PRD, publishing on Pentaho server and report viewer from the same, Dashboards - Creating with CDE and Datafor Visualizer and display on Pentaho Server, OLAP Analytics - built-in Mondrian Server with Pivot4J Analytic, PAD, PSW, Datafor Visualizer, Metadata - Creating Business Models with PME and Predictive Analytics - with Weka.

Summary: It does not have all the functionalities that Knowage CE has but what it has is enough to use in production. The main components are updated regularly with new versions, the auxiliary ones are less frequent or are stalled, but all our work is based on the basic ones.

According to [10], Knowage, on the other hand, has different principles for limiting CE from EE versions. Namely, it has all the necessary modules but there is a limitation of functionalities in them except ETL which we work with Talend and has the ability to physically add as an external module in the stack. Otherwise, the architecture of the Knowage stack is presented through modules in which all the functions of the stack are arranged and which should be grouped to obtain the desired functionalities, which are called sub-products and with which specific actions are performed. During the installation, can be chosen which modules to install, but ER (Enterprise Reporting) and SI (Smart Intelligence) are considered as initial and can be used alone or in a combined way, sharing all the common features. The SI module can be enriched

with four different plugins such as SD (Smart Data), LI (Location Intelligence), PM (Performance Management), and CA (Custom Analytics) which provide additional features. As in [3], functionalities of the Knowage CE stack here according to the modules are: Knowage CORE - ETL with Talend, the module that contains the functions of menu management, document search, profiling and security, drilling everywhere (for all documents), and other very important for server administration. Some of the essentials are also missing such as map navigation, tracking and monitoring, repository management, line data connection, and others. As second is Knowage ER - ER reporting tool which is part of the module of the same name, and is part of the stack with the stand-alone tool Knowage Report Designer KRD which is used for creating reports, publishing on the Knowage server, and so on. As the third one is the Knowage SI - SI product which is part of the module of the same name and which covers: 1. Dashboards - the part for creating interactive Cockpits, 2. OLAP analytics, 3. Metadata - creating business models, 4. As a plugin of this module should mention LI which serves to link business data with spatial or geographical information, data used in analytical documents. Not available functions: advanced graphical editing with pixel options, Creating a data set from federated, template designer for OLAP documents, wizard for creating OLAP analysis template, etc. The fourth one is Knowage PM - a plugin that enables performance management work. Unavailable features: Scorecards and Alarms. As the last one is Knowage CA - a plugin that enables Predictive Analytics with what-if scenarios. Unavailable features: Pre-built analytics features, Pre-built R / Python advanced visualizations.

Summary: It has more functionality than Pentaho CE but lacks features. What it offers is still enough to function in production. Updating versions is done regularly.

B. Scalability of views

Pentaho CE – according to the tools [9], PRD and CTools (CDE Dashboards) is known for the so-called pixel editing of their components. Pentaho OLAP analyzers offer template/background editing - stylish according to the tools themselves, that is enough. The first two offer a large amount for the selection of components such as spreadsheets, charts, other elements, and the third one small number of charts. Regarding data sources, PRD and CDE Dashboards offer the use of a lot of ones, based on SQL and MDX statements, using GUI too. Pentaho OLAP analyzers use MDX statements or drag and drop of cubes and dimensions using GUI. PRD make export in all standard export documents, CDE Dashboards only on the basis of individual components in standard forms, and Pentaho OLAP analyzers in xls,xlsx, and pdf. Automate filtering offer only CDE Dashboards for the charts in the case of series and labels. Automatic drilling is not available for all. Knowage CE – as for the tools [10], KRD offers template editing - pixel according to certain rules, The Cockpits according to the capabilities of the template designer himself, and OLAP Analyzers, manually making a template in XML format and editing options as such (if any) without giving an example in the tutorial. The first and second offer a large number of components such as spreadsheets, charts, cross-tabulations, and others even with quite big restrictions for the second one. The third one offers only Pivot table. For the data sources, KRD and Cockpits offer sufficient, with SQL statements as well as sub-statements directly or with SQL Query Designer, Business Model-Based Knowage Metadata (QbE) Statements, Dynamic Statement Styling. The OLAP Analyzers accepts only MDX statements manually entered in the template. For an export preview, the first one offers export in all standard documents, the second allows export of the whole cockpit in pdf and separate components in Excel spreadsheets, and the third one only classic print can also be in pdf. Automating filtering offers the Cockpit only for the charts with series and labels same as Automatic drilling - only for some charts with series and labels.

C. Smart visualization

As in [3], Pentaho CE has two-way interactivity with all tools except the standalone OLAP analyzer which has a one-way (has two-way included in the CDE Dashboard) with settings using script encoding.

As in [3], Knowage CE has two-way for all its components with setups using GUI.

D. Integration with Existing Transaction Processing Systems

As in [13], Pentaho CE offers direct embedding / merging with existing Web applications - the various components of the BI Platform can be combined to be added to an external application with more or less optional functionality, which is up to 100% of the platform's capabilities. The platform can also be expanded with its own components and services.

As in [10], Knowage CE offers Direct embedding/ integration with existing Web applications - a Web services called the BI Service and the BI API, which are both ways of communicating with the Knowage delivery layer, allow stack components to communicate with external applications or integrate with custom business applications with or without a graphical user interface.

E. Possibility to use mobile devices

As in [9], Pentaho CE - While the EE version has a module for the mobile version, the CDE component offers the size of window panels for viewing on mobile devices depending on the size of the devices, which means the size of the elements is adjustable.

As in [10], Knowage CE - With the help of BI Mobile also one of the ways to communicate with the delivery layer and thanks to the interaction between the Knowage server and the client interface remotely, reports, cockpits, or other useful documents can be accessed and displayed on mobile devices.

F. Security, managed ability to distribute documents by users

According to [9], Pentaho CE allows the management of user authentication and authorization types by the role as well as permissions to access folders and documents in the repository. Authentication types can be standard on Pentaho and LDAP which are managed directly from the Pentaho User Console (PUC), CAS for which settings must be added directly to the security files in the Pentaho settings folder as well as a username and password from another application that is sent with a URL when called. A user with administrator privileges can manage existing and create new roles and users directly from the PUC as well as set up access to documents and folders directly from the repository browser.

As in [10], Knowage CE also allows for the management of types of authentication, authorization of users by roles and permissions to access folders and documents in the repository as well as the use of analytical drivers and value lists. Types of authentication can be standard on Knowage, via CAS, LDAP, Google Sign-In, Azure Sign-In, and with the help of Javascript code authentication through a user application, but the authorization must be pre-set by the administrator and used only for login.

G. Ease of use of the tools

As in [3], Pentaho Reporting Design PRD - for developers, medium-difficult to create, SQL can be entered manually or you can work with a designer but all others manually and each individual action must be programmed. Pentaho CDE Dashboard - for developers, medium-difficult to create, all data sources are entered manually and each individual action needs to be programmed. Pentaho OLAP Analyzer - for developers with Pivot4J and Datafor quite easy, set upping via GUI with drag and drop facts and dimensions from the existing ones, except for interactivity that has to be programmed. Users - easy for all tools. Knowage Report Designer KRD - for developers, medium-difficult to create, SQL can be entered manually or with a designer but all else manually. Base interactivity is set via the GUI, and any more advanced action needs to be programmed. Knowage Cockpits / Dashboards - for developers, quite easy, all data sources are selected from previously entered data sets that are created manually or with GUI through a business model. Interactivity through GUI other than advanced things to be programmed. Pentaho OLAP analyzer - for developers, medium easy with inserting a hand-crafted template as the designer in CE does not work, but even if it works it is not easy to use. The rest with all the interactivity is

realized through GUI. Users - easy for all tools. Infrastructure, both stacks work on the most of existing operating systems and servers.

H. Customer support

Pentaho CE - has pretty good user support from developers, users who write by experience but also a lot of third parties who have described many use cases. Knowage CE - has limited user support mainly from partially responsive developers.

IV. CONCLUSION

No doubt is that is useful for organizations to implement BI. If there is any decision or interest, as in [3], here are a few words as conclusions to be considered before.

Regarding databases, PostgreSQL is giving better performances in the speed and stability used in all BI processes in comparison with MySQL. On the other side, MySQL has a well know syntax and manipulation tools, but recommend PostgreSQL.

Regarding the ETL tools, as already mentioned, the TOS problem is a slow SCDI step but TOS has problems also with the MySQL database in some situations too. So the recommendation is to use PDI.

For the BI stacks, if one prefers more OLAP processing better chose Pentaho CE as offers more OLAP-like options. All his tools use MDX language even the ones that use OLAP with drag and drop in GUI environment. Of course, Pentaho CE offers SQL language too between all other options and works very well and stable.

On the other hand, if you prefer easier development choose Knowage CE. Creating Dashboards and setting the interaction is easier in Knowage CE than in Pentaho CE where all actions should be coded. Knowage CE also works very well and is stable.

For the user's support, Pentaho CE offers better.

For maturity, Knowage is a younger project than Pentaho, so should expect better things in the future.

REFERENCES

- [1] F.Almeida, "Concepts and Fundaments of Data Warehousing and OLAP," INESC TEC and University of Porto, 2017
- [2] Cindi Hovson, "Successful BUSINESS INTELLIGENCE Secrets to Making BI a Killer App," The McGraw-Hill Companies, 2008
- [3] I. Jakimovski, "Sistemi za delovna inteligencija/razuznavanje za mali i sredni pretprijatija – analiza na arhitektura i finkcionalnite karakteristiki," unpublished
- [4] R. Kimball and M. Ross, "Kimball dimensional modeling techniques" in The Data Warehouse Toolkit, The Kimball Group, 2013
- [5] R. Holowczak, "Building ETL Transformations in Pentaho Data Integration (Kettle)." holowczak.com
<http://holowczak.com/building-etl-transformations-in-pentaho-data-integration-kettle/> (accessed March 12, 2022)
- [6] <https://help.talend.com/> (accessed March 12, 2022)
- [7] V. Poe, P. Klauer and S. Brobst, "Building a Data Warehouse for Decision Support," 2nd Edition, Prentice Hall, 1998
- [8] <https://mondrian.pentaho.com/> (accessed March 11, 2022)
- [9] <https://help.hitachivantara.com/Documentation/Pentaho/8.2> (accessed March 11, 2022)
- [10] <https://knowage-suite.readthedocs.io/en/7.2> (accessed March 11, 2022)
- [11] <https://www.predictiveanalyticstoday.com/pentaho-community-edition/> (accessed March 12, 2022)
- [12] <https://www.knowage-suite.com/site/licensing/knowage-editions/> (accessed March 13, 2022)
- [13] Integrating Pentaho Software and Content, Pentaho Corporation, USA, 2011

Trends in Autonomous Vehicles

Vesela Trajkoska

"Ss. Cyril and Methodius" University
Faculty of Computer Science and Engineering
Skopje, Macedonia
vesela.trajkoska@students.finki.ukim.mk

Abstract—This essay explains some of the current technological trends in autonomous vehicles. Most of the research is regarding the artificial intelligence of autonomous vehicle systems, but also considers strategies for proper tuning of sensors, algorithms for communication between vehicles and their energy efficiency. Some important research questions in the field of cybersecurity in autonomous vehicles are also listed. Finally, there is a comparison between three currently developed vehicle models. The general conclusion is that there is still no real autonomous vehicle, i.e. we are between the second and third level of automation (partial and conditional). It is also necessary to regulate the use of autonomous vehicles and to define procedures for when a system error occurs.

I. INTRODUCTION

People have always strived to achieve balance between increased productivity and improved efficiency. With the incorporation of machines and computers in every sphere and the automation of lots of difficult and repetitive tasks, life today is unimaginable without the technological devices that make our daily lives easier. However, there are still many tasks that require complex thought processes and intelligence, and despite all the efforts, machines and artificial intelligence still cannot perform them.

This includes driving, due to the fact that it happens in an unpredictable, stochastic and partially-observable environment. Still, many companies are actively trying to develop systems that will be able to drive completely autonomously without the help of humans. In this essay I will analyze the current technological trends in autonomous vehicles and currently developed vehicle models, as well as important open questions for research in this field.

II. DEFINING AUTOMATION

Six technical levels of driving automation [1] [2] have been defined in terms of the roles of the human driver and the vehicle's system while performing the driving actions. The levels start at 0 with no automation, then driver assistance, partial, conditional, high-level, and finally, full automation at level 6.

Following this classification, current autonomous driving systems are located between the second and third levels – partial automation (the system controls the vehicle's movement but the driver detects and reacts to events) and conditional automation (the system performs all necessary driving actions, but the driver must pay attention and be ready to take over in case of error or system failure). The fourth and fifth levels, which include actual autonomous behavior, are still unattainable [3].

III. TECHNOLOGICAL TRENDS

A. Research on artificial intelligence systems

Most of the research in the last ten years related to autonomous vehicles is based on artificial intelligence, es-

pecially after the rise of deep learning [3]. Fields where artificial intelligence can mainly contribute to improvement include control, perception, prediction, and modeling driving behavior.

1) *Control*: One of the main problems of autonomous vehicles from the aspect of control is controlling the speed of the vehicle in parallel with the rest of the dynamic driving tasks. To enable this, PID controllers are commonly used, but various methods are constantly being explored to optimize them, as well as reduce their errors [4]. Professors from the Department of Artificial Intelligence at the Technical University of Madrid have devised a function for global error that allows the use of various heuristic functions to regulate and optimize the control parameters of PID controllers. Using genetic algorithms, memetic algorithms and network adaptive direct search, they have shown increased controller accuracy compared to classical optimization based on other methods.

2) *Perception*: Artificial neural networks play a leading role in improving the perception of autonomous vehicles. In particular, the application of deep learning methods is explored in object detection, semantic scene segmentation, examples segmentation and lane line segmentation [5]. The application of the Q-learning algorithm, which can be used to solve stochastic problems without a model, is also explored for driving in roundabouts [6], as well as the introduction of a so-called blind intersection planner in the system controller [7].

3) *Prediction*: Neural networks are also used to improve agent predictions in autonomous driving. One of the most important problems in this area is the prediction of the trajectory of other traffic participants. Different machine learning methods are used to solve this problem. Professors at Hanyang University in South Korea have developed a method based on the Random-forest algorithm, which they have tested in different scenarios. The results showed improved predictions with increased robustness [8]. Scientists at Shanghai University of Science and Technology have developed a model for the sequential dynamics of motion estimation, which has better performance than currently used models for monocular visual odometry, or the use of image information obtained with a single lens [9].

4) *Driving behavior modeling*: Modeling the behavior of other agents in the environment during autonomous driving is possible only after their successful detection and furthermore, accurate prediction of their movement. The most challenging task appears to be modeling the behavior of the vehicle in front of the agent, especially when they suddenly shift into the lane, brake, or change speed. Many researchers are experimenting with different methods to improve the accuracy of such predictions [10] [11].

B. Sensors

The movement of autonomous vehicles is highly dependent on their ability to orient themselves and navigate around the environment. Simultaneous localization and navigation algorithms show satisfactory results [12]. Various algorithms that use deep learning for vision (rare spatial convolutional neural network) and sensor integration algorithms (sensor-weighted integrated field) are also explored. These algorithms mainly use four sensors: vision, light detection, range of vision, and a GPS sensor [13]. To improve the precision of obstacle avoidance while maintaining a smooth and uninterrupted ride, sensor-based feedback strategies are being explored [14].

Additionally, proper sensor tuning is crucial to ensure their functionality and then to optimize it. In autonomous vehicles, the placement of the cameras has a direct impact on the vehicle's ability to properly orient and navigate the environment. New interactive camera calibration methods have been developed, adaptable to different models, which reduce the complexity and improve the accuracy of the process [15].

C. Communication between vehicles

One of the benefits of autonomous vehicle traffic is their ability to communicate quickly and therefore operate more efficiently. But secure communication between vehicles is still evolving because their systems are at a certain level of risk of cyberattacks. Attempts have been made to develop different models for risk analysis in the early stages of system design [16].

Scientists from the Eindhoven University of Technology have presented a framework for cooperative assessment of the condition of autonomous vehicles. The assessor is analyzed through simulations of three different scenarios (straight line driving, constant curve rotation, and turns in the form of the number 8), which show benefits of vehicle-to-vehicle communication, such as accurate measurements of the components of the states of both vehicles and a reduction in measurement error of over 50%. Additionally, this framework uses a combination of different sensors together with information received from the other vehicle, leading to greater overall robustness of the system in the events of sensor failures or data loss [17].

D. Energy efficiency

As the amount of information processed by the systems of autonomous vehicles increases, so does the energy consumption, i.e., the amount of electricity required for their calculations. Currently used SRAM architectures offer great accuracy, but also consume a lot of energy, which is why methods are being explored that would reduce energy consumption without reducing accuracy. Field effect transistors on carbon nanotubes or fin field effect transistors are potential solutions for high-consumption artificial intelligence devices, but there is also new research on SRAM devices that combine these transistors and offer a low-energy solution [18]. Designs for electronic control units of embedded electronic systems that increase the security of the agent's entire energy management system are explored as well [19].

IV. OPEN RESEARCH QUESTIONS

A. Human factors in cybersecurity of autonomous vehicles

When using autonomous vehicles, there is risk of a cyberattack on the software that controls the vehicle. In such

cases, appropriate action should be taken to neutralize the attack or minimize its consequences, which requires human intervention. But compared to machines, humans are far more prone to errors, so to increase efficiency in dealing with software threats and attacks, it is necessary to minimize human errors.

It is crucial to identify potential cyberattack scenarios and the people who are most vulnerable to them, and to create educational materials specifically designed for them [20]. In my opinion, narrowing down the most vulnerable groups and targeting them during any informative processes could result in more efficient appropriation of funds and resources for education on this topic.

Furthermore, it is important to lay the groundwork for what cybersecurity is and what risks there are to autonomous vehicle software. An important research question is how to explain these terms to the average citizen [21], or even exactly how exposed is the average citizen to such threats.

B. Defining regulations

It is important to define the level of attention required by passengers in autonomous vehicles. Research shows that different drivers have vastly different reactions and reaction times to unexpected events while driving [22] [23] [24]. Researchers need to determine the maximum permissible level of distraction of the driver to enact proper legislation [20]. This is especially important for vehicles of levels 2 through 4, since they are not fully automated and require human assistance or takeover in certain cases, which differ through each of the levels.

Lastly, it is important to note that the findings of some of these studies would be truly necessary and useful only when there is daily and widespread use of autonomous vehicles [20]. But even so, I believe that researchers should be encouraged to explore these areas continuously. It is far better to have the answers to these questions early on, so that they can fundamentally affect the development and design of such autonomous systems.

V. CURRENTLY DEVELOPED SYSTEMS

A. Autopilot

Autopilot is a collection of driver assistance software features currently developed by Tesla, intended to increase road safety [25]. They currently offer three system packages:

- 1) Autopilot – steering assistance and car-speed matching assistance with the speed of surrounding vehicles
- 2) Enhanced Autopilot – upgraded to include additional assistance with lane changing and parking, as well as a “Summon” feature which can move the vehicle out of a parking spot or tight space
- 3) Full Self-Driving Capability – includes all aforementioned features and additionally all necessary hardware for fully autonomous driving.

As of 2019, every Tesla car is equipped with the standard Autopilot package, while the models containing necessary hardware can be upgraded to the additional two packages. The systems of Model 3 and Model Y use 8 external cameras, 12 ultrasonic sensors and an on-board computer which uses neural network processing. The systems of the newer Model S and Model X are additionally equipped with a radar [26].

It is important to note that all system packages require active vehicle control and supervision by the human driver,

especially keeping their hands on the steering wheel. The current capabilities of this system are at level 2 driving automation.

B. Toyota Teammate

Toyota Teammate [27] is a suite of driver assistance technologies currently developed by Toyota:

- Advanced Park – performs hands-free parking operations monitored by the human driver
- Advanced Drive – driving assistance including steering, acceleration, breaking, changing lanes and more, with a hands-free option under specific conditions, requiring constant supervision by the human driver [28].

This system is also at Level 2 driving automation, intended to assist human drivers.

C. Drive Pilot

Drive Pilot is an intelligent driving assistance system actively developed and tested by Mercedes-Benz, which has the potential to reach Level 3 driving automation on highways [29]. The user (human driver) needs to be ready to resume driving if the system anticipates that it cannot continue reliably driving. The procedure is precisely defined – the system firstly requests from the user to resume driving, while continuing to operate the vehicle until the user takes over. If the human fails to respond, the system will bring the vehicle to a safe stop and turn on hazard warning lamps.

This system can only be enabled within its Operational Design Domain which includes the required conditions in which it knows how to drive – fully access-controlled highways with medium to dense traffic in fair weather. Additionally, available driving routes are defined within a precise and high-definition map. The system also continuously monitors the user's state and if it detects prohibited actions, such as sleeping, it issues warnings and furthermore, safely stops the vehicle if the user persistently fails to respond [29].

The Mercedes-Benz S-Class models equipped with this system use a variety of sensors, including radar, LiDAR, camera, road moisture, ultrasonic sensors, and microphones, which provide a 360-degree field of view. The input from all sensors is combined to improve the accuracy of the environment. The redundancy of some of the perception sensors ensures proper functioning even in the event of certain sensor failure [29].

Additionally, the system contains thirteen individual functions meant to assist the driver when Drive Pilot is turned off, including speed limit, lane change and keeping, braking, blind spot vision, parking, emergency stop, and more [30].

VI. CONCLUSION

There is no doubt that autonomous vehicles offer many benefits to humans. Statistical estimates show that the vast majority of road accidents occur due to human error. Even if we put car accidents aside, humans are still fundamentally irrational beings who do not have the computing power of a machine and cannot always calculate and take the optimal step. Due to this, there are often traffic jams, delays, and congestion that last for hours and drastically reduce the efficiency of traffic.

Autonomous vehicles, as systems that are completely rational and guided by the rules programmed into them, are a potential solution to many of the traffic problems we

face. They would follow all traffic rules without unnecessary speeding or unplanned lane switching, they are not prone to drunkenness and other intoxicating substances, and they have no egoistic tendencies that would influence their decisions.

However, the technological innovations in this field are still insufficient to achieve the fourth or fifth level of vehicle autonomy. Although experts are optimistic that a solution will be found in the next few decades, this requires a lot of resources and research, but also legislation that will properly regulate their use. We must not allow fully autonomous vehicles to be integrated into our daily lives without being completely sure that they are secure and reliable.

REFERENCES

- [1] Bartneck C., Lütge C., Wagner A., Welsh S. (2021) Autonomous Vehicles. In: An Introduction to Ethics in Robotics and AI. SpringerBriefs in Ethics. Springer, Cham. [CrossRef](#)
- [2] On-Road Automated Driving (ORAD) Committee. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. 2021. [CrossRef](#)
- [3] Talavera, E.; Díaz-Álvarez, A.; Naranjo, J.E.; Olaverri-Monreal, C. Autonomous Vehicles Technological Trends. Electronics 2021, 10, 1207. [CrossRef](#)
- [4] Naranjo, J.E.; Serradilla, F.; Nashashibi, F. Speed Control Optimization for Autonomous Vehicles with Metaheuristics. Electronics 2020, 9, 551. [CrossRef](#)
- [5] Guo, Z.; Huang, Y.; Hu, X.; Wei, H.; Zhao, B. A Survey on Deep Learning Based Approaches for Scene Understanding in Autonomous Driving. Electronics 2021, 10, 471. [CrossRef](#)
- [6] García Cuenca, L.; Puertas, E.; Fernández Andrés, J.; Aliane, N. Autonomous Driving in Roundabout Maneuvers Using Reinforcement Learning with Q-Learning. Electronics 2019, 8, 1536. [CrossRef](#)
- [7] Narksri, P.; Takeuchi, E.; Ninomiya, Y.; Takeda, K. Deadlock-Free Planner for Occluded Intersections Using Estimated Visibility of Hidden Vehicles. Electronics 2021, 10, 411. [CrossRef](#)
- [8] Choi, D.; Yim, J.; Baek, M.; Lee, S. Machine Learning-Based Vehicle Trajectory Prediction Using V2V Communications and On-Board Sensors. Electronics 2021, 10, 420. [CrossRef](#)
- [9] Zhao, B.; Huang, Y.; Wei, H.; Hu, X. Ego-Motion Estimation Using Recurrent Convolutional Neural Networks through Optical Flow Learning. Electronics 2021, 10, 222. [CrossRef](#)
- [10] Cui, G.; Wang, S.; Wang, Y.; Liu, Z.; Yuan, Y.; Wang, Q. Preceding Vehicle Detection Using Faster R-CNN Based on Speed Classification Random Anchor and Q-Square Penalty Coefficient. Electronics 2019, 8, 1024. [CrossRef](#)
- [11] Tsai, W.-C.; Lai, J.-S.; Chen, K.-C.; M.Shivanna, V.; Guo, J.-I. A Lightweight Motion Object Behavior Prediction System Harnessing Deep Learning Technology for Embedded ADAS Applications. Electronics 2021, 10, 692. [CrossRef](#)
- [12] Ren, R.; Fu, H.; Wu, M. Large-Scale Outdoor SLAM Based on 2D Lidar. Electronics 2019, 8, 613. [CrossRef](#)
- [13] Oh, M.; Cha, B.; Bae, I.; Choi, G.; Lim, Y. An Urban Autodriving Algorithm Based on a Sensor-Weighted Integration Field with Deep Learning. Electronics 2020, 9, 158. [CrossRef](#)
- [14] Masood, K.; Molfino, R.; Zoppi, M. Simulated Sensor Based Strategies for Obstacle Avoidance Using Velocity Profiling for Autonomous Vehicle FURBOT. Electronics 2020, 9, 883. [CrossRef](#)
- [15] Lei, W.; Xu, M.; Hou, F.; Jiang, W.; Wang, C.; Zhao, Y.; Xu, T.; Li, Y.; Zhao, Y.; Li, W. Calibration Venus: An Interactive Camera Calibration Method Based on Search Algorithm and Pose Decomposition. Electronics 2020, 9, 2170. [CrossRef](#)
- [16] Luo, F.; Hou, S.; Zhang, X.; Yang, Z.; Pan, W. Security Risk Analysis Approach for Safety-Critical Systems of Connected Vehicles. Electronics 2020, 9, 1242. [CrossRef](#)
- [17] Schinkel, W.; van der Sande, T.; Nijmeijer, H. State Estimation for Cooperative Lateral Vehicle Following Using Vehicle-to-Vehicle Communication. Electronics 2021, 10, 651. [CrossRef](#)
- [18] Kim, Y.; Patel, S.; Kim, H.; Yadav, N.; Choi, K.K. Ultra-Low Power and High-Throughput SRAM Design to Enhance AI Computing Ability in Autonomous Vehicles. Electronics 2021, 10, 256. [CrossRef](#)
- [19] Biba, D.R.; Ancuti, M.C.; Ianovici, A.; Sorandaru, C.; Musuroi, S. Power Supply Platform and Functional Safety Concept Proposals for a Powertrain Transmission Electronic Control Unit. Electronics 2020, 9, 1580. [CrossRef](#)
- [20] Linkov V, Zámecník P, Havlíčková D and Pai C(2018) Human Factors in the Cybersecurity of Autonomous Cars: Trends in Current Research. Front. Psychol. 10:995. [CrossRef](#)

- [21] Brase, G.L., Vasserman, E.Y. and Hsu, W. (2017). Do different mental models influence cybersecurity behavior? Evaluations via statistical reasoning performance. *Frontiers in Psychology* 270 8:1929. [CrossRef](#)
- [22] Dogan, E., Rahal, M.-Ch., Deborne, R., Delhomme, P., Kemeny, A. and Perrin, J. (2017). Transition of control in a partially automated vehicle: Effects of anticipation and non-driving-related task involvement. *Transportation Research Part F* 46, 205-215. [CrossRef](#)
- [23] Dixit, V. V., Chand, S. and Nair, D. J. (2016). Autonomous Vehicles: Disengagements, Accidents and Reaction Times. *PLoS ONE* 11(12): e0168054. [CrossRef](#)
- [24] Gold, Ch., Dambrock, D., Lorenz, L. and Bengler, K. (2013). "Take over" How long does it take to get the driver back into the loop? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 57, 1938-1942. [CrossRef](#)
- [25] Tesla, Inc. 2022, Autopilot and Full Self-Driving Capability, accessed 25 March 2022. [CrossRef](#)
- [26] Tesla, Inc. 2022, Autopilot, accessed 25 March 2022. [CrossRef](#)
- [27] Toyota Newsroom 2021, Teammate Advanced Drive Backgrounder, accessed 25 March 2022. [CrossRef](#)
- [28] Toyota Motor Corporation 2022, Your New Partner for the Road, accessed 25 March 2022. [CrossRef](#)
- [29] Mercedes-Benz 2019, Introducing DRIVE PILOT: An Automated Driving System for the Highway, accessed 26 March 2022. [CrossRef](#)
- [30] Mercedes-Benz Group 2022, From the intelligent to the autonomous car. Intelligent Drive – today and tomorrow, accessed 26 March 2022. [CrossRef](#)

Anonymous Blockchain based model for e-Voting

Marija Taneska

Faculty of computer science and engineering
Ss. Cyril and Methodius, University Skopje, North Macedonia
 marija.taneska@students.finki.ukim.mk

Fat Halimi

Faculty of computer science and engineering
Ss. Cyril and Methodius, University Skopje, North Macedonia
 fat.halimi@students.finki.ukim.mk

Abstract—In this paper, we analyze the provided security, anonymity, and privacy provided by Zcash and Monero. We focus on the description for the e-voting framework based on Zcash and Monero using their anonymity advantages. Analyzing advantages and disadvantages of Zcash and Monero, we make comparison between them, and concluded that Monero is more secure than Zcash, and we recommend using Monero for serious e-Votings. Our key is to give a sparkle to Blockchain developers to create a legitimate framework or system that would be based on Zcash and Monero Blockchains to realize e-voting.

Keywords—e-voting, Blockchain, Zcash, Monero

I. INTRODUCTION

Voting is an action we perform in several different situations: some examples are song contests, reality shows, associations/councils, company decisions, or politics. Wherever allowed, digital voting systems have been introduced as tools to make it easier for voters to express their choice and to reduce the huge costs of voting in person. However, most digital voting solutions in use nowadays affected by several problems. Digital ledger technologies, especially Blockchains, renew interest in e-voting.

In Section II (e-Voting criteria), we describe the criteria for e-Voting, then in Section III (Blockchains overview), we describe Blockchains and conclude what type of Blockchain satisfies those criteria. In the Section IV (Blockchains overview), we describe how Zcash and Monero work and analyze the provided security, anonymity, and privacy. In Section V (e-Voting implementation ideas) is the description of how will be working e-Voting based on Zcash and Monero, in Section VI (e-Voting phases) we describe phases in both implementations. In Section VII (Comparison between Zcash and Monero), we compare them based on their provided security and conclude what is more secure for e-voting. In Conclusion we describe the key features and advantages of using Blockchain for e-Voting.

II. E-VOTING CRITERIA

e-Voting must meet the following criteria: Immutability, Equality, Eligibility, Anonymity, No forgery, Verifiability, Scalability, and Stability.

Immutability means integrity too because this step means that no one can have access to votes in terms of deletion or modification. Using the Blockchain we can satisfy these criteria because once a voting transaction is made it is permanent, unchangeable, undeletable.

Equality assures that each vote once is counted and the choice added to a database, that vote should no longer be

eligible to go through counting again. With a framework, we would make some changes and add “flags” to the votes once they go through counting.

Eligibility assures that the vote comes from the voter and no one else. We provide eligibility by using zk-SNARKs in Zcash based implementation, in Monero based implementation by using Ring Signatures.

Scalability means that the system should be secure enough, professional enough, and have free data space for millions of voters since this system can be used for political elections.

Stability means that the system should be designed to have a long life, stability depends on scalability.

Anonymity ensures that the Voting system user knows that the voter voted, but not what the voter voted for.

No forgery means that only the legal voters can vote and every voter can vote only once.

III. BLOCKCHAIN

Blockchain is a peer-to-peer ledger system that allows peers to transact directly with each other eliminating the need for a central authority. Blockchain is a system for recording information about a transaction in a new decentralized way that makes it difficult or impossible to alter. These transactions are stored on sheets or blocks in a digital ledger that is shared among the participants of the network. Consensus on the transactions brings the peer-to-peer network into an agreement.[1] There are two types of Blockchain - public Blockchain and private Blockchain.

A. Public Blockchain

A Public Blockchain is open and full transparency for all and allows anyone to participate in the network, read or write. We can not use the public Blockchain for e-Voting implementation, since it does not satisfy Anonymity criteria. Also, the public Blockchain is decentralized and does not have a single entity that controls the network. Moreover, Blockchain does not have any regulations that the nodes have to follow.

B. Private Blockchain

A Private Blockchain is a permissioned managed Blockchain and participants must consent to join the network. Only entities participating in the transaction know about the transaction performed, others will not be able to access it. Private Blockchain focuses on privacy concerns and provides high efficiency, faster transactions, and better scalability. We can conclude from the previously said that

private Blockchain satisfies all criteria for e-Voting described in the previous section.

IV. BLOCKCHAINS OVERVIEW

The e-Voting process must be anonymous and private, because of that, we chose Zcash and Monero, which are the most popular private cryptocurrencies, for the base of e-Voting. Below we describe how Zcash and Monero work and offer anonymity and privacy.

A. Zcash

Zcash is a privacy-protecting, digital currency built on strong science. Zcash was first released on October 28, 2016, and it was originally based on Bitcoin's codebase [2]. The Zerocash protocol is being developed into a full-fledged digital currency, Zcash. Zerocash extends the protocol and software underlying Bitcoin by adding new, privacy-preserving payments. Zcash has its unique advantage. Zcash's anonymity relies on a shielded pool, where partial transaction information such as input/output addresses and the transaction value is no more directly available from the Blockchain. Zcash is a cryptocurrency that uses advanced applied cryptography to provide enhanced privacy via shielded addresses. Zcash is the first practical application of zk-SNARKs, a specific type of zero-knowledge proof, a novel form of zero-knowledge cryptography.

zk-SNARK stands for Zero-Knowledge Succinct Non-Interactive Argument of Knowledge, and refers to a proof construction where one can prove possession of certain information, e.g. a secret key, without revealing that information, and without any interaction between the prover and verifier [3]. Zero-knowledge proof allows the prover to prove the verifier that a statement is true, without revealing any information beyond the validity of the statement itself. Succinct means that the zero-knowledge proof can be validated within a few milliseconds, even in the case of statements related to large-scale programs. Non-interactive refers to a zero-knowledge protocol where the prover and verifier have little to no interaction. This means they can only exchange one proof. Zk-SNARKs records only the proof of the transaction on the Blockchain node, safeguarding the identity of the sender, receiver, and other details associated with the transaction.

Zcash addresses are either private (z-addresses) or transparent (t-addresses). Z addresses are private addresses and only the user can see the balance in the wallet.

Zcash has 5 types of transactions [2], we will describe only the most secure type – shielded transactions that will be used in e-voting. A shielded (private) transaction is a transaction from z-address to z-address. In these transactions, the addresses, transaction amount, and the memo field are all encrypted and not publicly visible. The fact that the transaction has happened is recorded on the ledger, but the sending and receiving addresses and the amount sent are not revealed to the public.

B. Monero

Monero is the leading cryptocurrency, launched in April 2014 and focused on private and censorship-resistant transactions. Monero transactions are confidential and untraceable. Monero is the only cryptocurrency where every user is anonymous by default. The sender, receiver, and amount of every single transaction are hidden through the use of three important technologies: Stealth Addresses, Ring Signatures, and RingCT. [4]

Stealth addresses are an important part of Monero's inherent privacy. They allow and require the sender to create random one-time addresses for every transaction on behalf of the recipient. The recipient can publish just one address, yet have all of his/her incoming payments go to unique addresses on the Blockchain, where they cannot be linked back to either the recipient's published address or any other transactions' addresses. By using stealth addresses, only the sender and receiver can determine where a payment was sent. [5]

Ring signatures are anonymous digital signatures from a member of the group, but they do not reveal which member signs the transaction. To generate a ring signature, Monero uses a combination of keys from the sender's account and links them to the public keys on the Blockchain. This makes it unique, but also private. It hides the identity of the sender, as it is computationally impossible to determine which of the group members' keys was used to produce the complex signature. [6]

RingCT, short for Ring Confidential Transactions, is how transaction amounts are hidden in Monero. RingCT introduces an improved version of ring signatures called "A Multi-layered Linkable Spontaneous Anonymous Group signature" [6], which allows for hidden amounts, origins and destinations of transactions with reasonable efficiency and verifiable, trustless coin generation. [7]

V. E-VOTING IMPLEMENTATION IDEAS

The implementation of e-voting, in the best case, would involve two systems that makeup one Voting system. Those would be the system at the end of the user - voter and the system at the end of the state voting organization - The voting system.

We have the Voting system (through which is implemented voting) and two parties: The voting system user (for example State Election Commission) and the Voters for implementing e-voting. In contrast to the traditional Zcash and Monero coins transaction where the user mine coins and then gets them, in e-voting, all transactions are to the Voting system user.

For an implementation based on Zcash, we use the zkSnark protocol and Shielded (private) transactions. We use this type of transaction as the most secure, because the addresses, transaction amount, and the memo field are all encrypted and not publicly visible. As we said all transactions are to the Voting system user and all voters will know that address.

For an implementation based on Monero for the transaction (exchange of the vote from the state commission to the voter and then back to the state commission, we would use Ring CT because thus the way we hide the content of the vote. The content or the vote should not be public because no one should know which person has voted, that's why we would use ring CT and thus way modify the algorithm to show only what is voted but not by who.

VI. E- VOTING PHASES

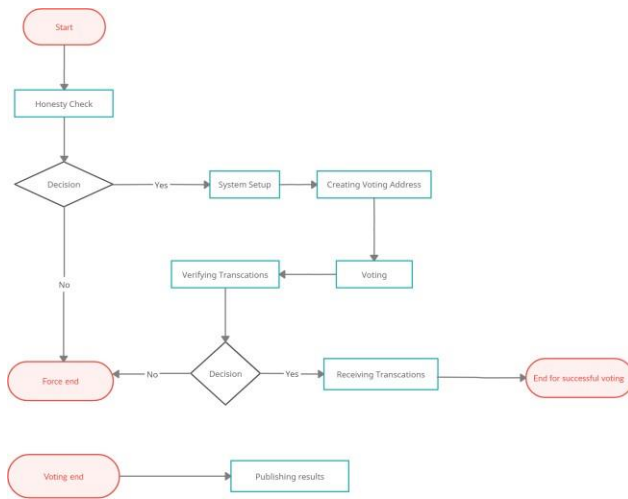


Fig. 1. E-Voting procedure

We have seven phases Honesty Check, System Setup, Creating Voting Address, Voting, Verifying Transactions, Receiving Transactions, and Publishing results. In Fig.1 we show the e-Voting procedure.

1. Honesty Check

We need to ensure that each voter will vote only once, because of that we store IDs of voters that voted and we start the e-voting procedure with the Honesty Check phase. In this phase we check Is the ID of the voter is in the database? If it is, that means that the voter has already voted and has no right to vote again, so the procedure ends here. Otherwise, we move on to the second phase.

2. System Setup

For an implementation based on Zcash, the System Setup phase is identical to the phase with the same name in Zcash coins transactions [8]. During this sub-phase, the KeyGen function samples the proving key(encrypted proving polynomial) and the verification key (encrypted target polynomial) and stores it as the output. This is done using the RSA protocol [9] by the Trusted Third Party. Next, we store these keys in the ppar string that is stored on the Blockchain, so can be easily accessed by both the prover and the verifier.

The traditional Monero implementation does not request the trusted setup, but a trusted setup is necessary for e-voting

3. Creating Voting Address

In Zcash based implementation Creating a Voting, Address is identical to the Creating Payment Address phase in Zcash coins transactions. Using the ECIES encryption [10] we create a public and a secret key for a single user. Following these keys are used in conjunction with a pseudo-random function to generate a public key address or public/private key address pair.

In Monero based implementation Creating Voting Address for the voters is identical to the traditional coin transaction we create random one-time addresses for every voter's transaction (if a voting transaction was not successful, the voter still have the right to vote and his voting address will be different from the previous address). For the Voting system user, we need a static unique address on which it will receive all transactions.

4. Voting

The Voting phase as the name suggests is voting from a voter's private address(only known by the voter), in both implementations Blockchain voting token is sent to the voter.

The input in this phase is ppar string if we use Zcash based implementation, for Monero based implementation we do not need the input. This phase in both implementations provides an output tuple that stores the vote and Voting system user. The voting phase is not executed by a Voting system user.

5. Verifying Transaction

In Zcash based implementation this phase has two sub-phases, first, the sub-phase is executed by a prover sub-phase, and second by a verifier. The first sub-phase is executed by a prover and takes as input the ppar string. During this sub-phase, the Prover (voter) executes the Prove function. The zk-SNARK proof has a ppar string as input for the second sub-phase. The second sub-phase is executed by a verifier and takes as input the ppar string and zk-SNARK proof which is the result from the previous sub-phase. This sub-phase is similar to Verifying Transaction phase in Zcash coins transaction with pour ledger. During this sub-phase, the Verifier executes the Verify function using the zk-SNARK proof, the public input as stored inside the transaction ledger, and the verification key (part of the ppar string). The Verify function returns a boolean value depending on whether the verifier is convinced that the prover's assertion is genuine. If the verifier is convinced of the zk-SNARK proof (and that the digital signature is valid), the Verifying Transactions phase returns TRUE. If the result is false the procedure ends here.

In Monero based implementation we create the signature for the transaction after the voting, using a random sampling algorithm, the wallet selects a set of global indexes from the histogram provided by the daemon, then the daemon will need to provide the corresponding outputs. The index of the real output is also included in the request for two reasons: as a test whether the daemon sends the correct outputs and to obfuscate the final ring signature from a curious daemon. We use the Multilayered Linkable Spontaneous Anonymous Group signature (MLSAG)[6] which is a generalization of using a key vector. The point of using a key-vector is so that a user can prove they simultaneously have knowledge of both the secret key of the input address and the secret. The above is de-facto the Verifying Transaction.

6. Receiving Vote

Following the TRUE result from the previous phase in both implementations, we get to Receiving Vote phase. As we said before the receiver for all voting transactions is a Voting system user. The voter used the Blockchain vote token and the voting system stores voter's username in our database .

7. Publishing results

Publishing results is the last phase and it is executed only once by the Voting system user when voting is officially over. The voting system user summarizes votes, adding flags to the votes once they go through counting. The results and according to the appropriate voting protocol announces the winner(s).

VII. COMPARISON BETWEEN ZCASH AND MONERO

As we discussed in the topic both Monero and Zcash provide great privacy, anonymity, and a possible framework to be made using these Blockchains for e-voting. Both of these Blockchains offer very good privacy but in realization, they are completely different. When we are on Zcash we can say that Zcash is more opt-in but on the other hand on addressing known issues Monero is more practical. As it is already discussed, we know that Monero uses Zero-knowledge proofs with actually it is in a way mathematically provable system to hide details of the transaction, transaction meaning the transfer of a vote but for sure this is not enough to consider protecting privacy. On the other hand, Monero uses older layers but those layers are better studied. Monero's layers tend to obfuscate the source of funds which are ring signatures, hide the amount (the vote in our case) which means we have a confidential transaction, and also the destination which is a stealth address. Even though their layers give extraordinary privacy, in theory, this can never be zero-knowledge, but Monero is slightly on advantage because we don't have proof that Zcash is a bug-free implementation of the theory [11]. Monero is more secure than Zcash, therefore we recommend using Monero for serious e-votings.

CONCLUSION

Blockchain-based e-voting is the best solution and we can conclude that from key features. No external entity can add, remove or modify votes because it is native to Blockchain technologies. With zk-SNARKs and Ring Signatures of voters, we confirm that is not possible and no one can falsify an identity using another voter identity by abusing Blockchain. Because of token fungibility, each vote is equal to the other.

- [1] LinuxFoundationX Course, Blockchain: Understanding Its Uses and Implications, available at <https://learning.edx.org/course/course-v1:LinuxFoundationX+LFS170x+2T2021/home>.
- [2] Official Zcash documentation, available at <https://zcash.readthedocs.io/en/latest/>
- [3] Zcash – "What are zk-SNARKs?", available at <https://z.cash/technology/zksnarks/>
- [4] Monero – "What is Monero (XMR)?", available at <https://www.getmonero.org/get-started/what-is-monero/>
- [5] Moneropedia – "Stealth Address", available at <https://www.getmonero.org/resources/moneropedia/stealthaddress.html>
- [6] Nisansala Perera, A Survey on Group Signatures and Ring Signatures: Traceability vs. Anonymity, January 2022
- [7] Moneropedia – "RingCT", available at <https://www.getmonero.org/resources/moneropedia/ringsignatures.html>
- [8] Aritra Banerjee, Michael Clear, Hitesh Tewari, Demystifying the Role of zk-SNARKs in Zcash, 2020 IEEE Conference on Application, Information and Network Security (AINS), November 2020
- [9] Shireen Nisha, Mohammed Farik, RSA Public Key Cryptography Algorithm – A Review, July 2017
- [10] Víctor Gayoso Martínez, L. Hernández Encinas, Carmen Sánchez Ávila, A Survey of the Elliptic Curve Integrated Encryption Scheme, January 2010
- [11] Casaleggio, D., Nicola, V. D., Marchesi, M., Missineo, S., & Tanelli, R. (n.d.). A digital voting system for the 21st Century. Retrieved March 27, 2022, from https://www.strateghia.it/wp-content/uploads/FINALE-A_digital_voting_system_for_the_21th_century.pdf J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73

Perception of benefits and costs of computer games among high school students in Macedonia

Marina Dodevska

Faculty of Computer Science and Engineering

St. Cyril and Methodius University
Skopje, Republic of North Macedonia
marina.dodevska@students.finki.ukim.mk

Nikola Janev

Faculty of Computer Science and Engineering

St. Cyril and Methodius University
Skopje, Republic of North Macedonia
nikola.janev@students.finki.ukim.mk

Vladimir Trajkovikj

Faculty of Computer Science and Engineering

St. Cyril and Methodius University
Skopje, Republic of North Macedonia
vladimir.trajkovikj@finki.ukim.mk

Abstract— The internet and computers are constantly evolving. This means the people are always staring into the monitors. They can do different things on the internet like learning new stuff, watching videos, playing games, chatting with their friends, etc. But sometimes they forget about the time spent with their computers, doing their stuff, and sometimes their friends or parents will tell them to get off their computers because they can hurt their eyes, or they don't have a social life. Last year we surveyed this topic, but goal was high school students in North Macedonia. So, in this paper, we will explain what are the benefits and what are the costs of playing computer games for high school students.

Keywords—computer games, internet, high school students

INTRODUCTION

People play games for different reasons, some of them play for fun, some of them play to relieve stress, and some people play games to spend time. Playing games can have both, advantages and disadvantages, for a high school student, like having bad grades if they don't know how to manage their time, problems with their parents [8] because they stare too much into the computer screens, not going out to catch a breath, violence, wasting their time, or if they switch to the good side of the book, they will see that if they play games, their IQ will increase, or they can learn new stuff and this depends on the genre, like cooking, building, etc. Today there have tons of platforms to play games, like mobile phones, computers, and laptop computers, consoles, decks. Every year tons of games are coming out from various developers around the world, some of them are free, and some of them cost money.

METHODOLOGY

Last year, we surveyed 188 respondents. The students who answered this survey are from different high schools. Most students are male 75.5% and female 24.5% and they are between 15-18 years old. We shared our survey on several news websites, such as e-sports websites [1] and classic traditional websites. Through this survey, we had several difficulties reaching the students in North Macedonia, because most of the high schools don't have a website or they have inactive social media pages like Facebook page or Instagram page. So we need to contact the headmasters through their personal contact information like Facebook profiles. We conducted this survey for several months, and in the end, we collected the answers. We also had several hypotheses which we wanted to confirm if they are true or fake.

The hypotheses we wanted to confirm:

1. Gaming disrupts the physical and mental condition of people.
2. Gaming improves communication and thinking skills.
3. Gaming socializes people and connects them with people of other nationalities and countries.

To confirm these hypotheses, we had several questions which some of which are with the Likert scale, to see their responses more correctly.

The questions that were in our survey:

1. What is your gender?
2. What year are you in high school?
3. Type of secondary school?
4. Do you play video games?
5. How much time in your day, do you spend playing video games?
6. Have you ever had a physical or mental disorder from playing video games (eg, kyphosis, damaged vision)?
7. How often do you interact with your teammates in the games?
8. Do you play online or offline games?
9. Do you think that by communicating with people from another country you are learning something new (eg new language, their culture)?
10. Which type of genre do you play?
11. Have you ever lied that you can not go out with your friends to play games at home?
12. How often do you hear rude words from your teammates about you or another player?
13. Do you think that playing video games improves the way of thinking and coping in certain situations?
14. Did you make a friendship with another player during your gameplay?
15. Do you watch e-sports or traditional sports?

RESULTS

When we asked them about the year, we got 27.7% of students are in the first year, 31% second, 26.1% in the third, and 15.2% are in the fourth year of their education.

This can be seen in Fig. 2 where we have our pie chart presented.

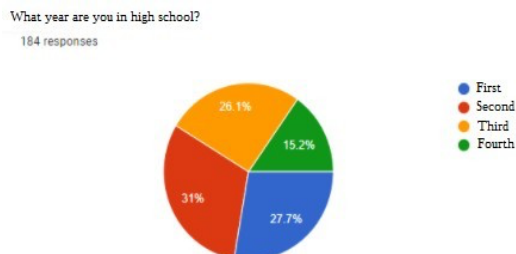


Fig. 2 What year are our respondents in high school

In Fig. 2 we can see that most of our respondents who play video games are in the second year of their education, and second place takes students who are in the first year of their education. This chart maybe can tell us that students who are in the early years of their education are more represented, or we have a small number of respondents that are in the fourth or third year. Our next question was about the type of their secondary education to see from which type of school our respondents are coming. As we can see from Fig. 3 76.1% are in vocational education and 23.9% are from the professional school where students are not focusing on one profession, but they are learning everything.

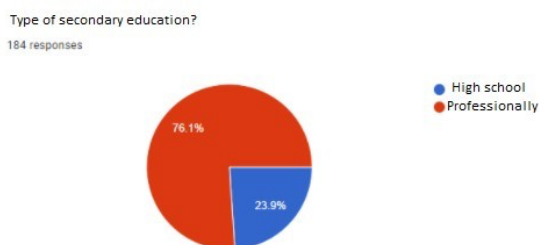


Fig. 3 Type of secondary education

The next question we asked was to see if the students play video games or not. Because we didn't want to have false answers. So from the chart (Fig. 4), we can see that 76.1% of the students are playing games, and 23.9% are not. This indicates that most of the students are playing games in their free time.

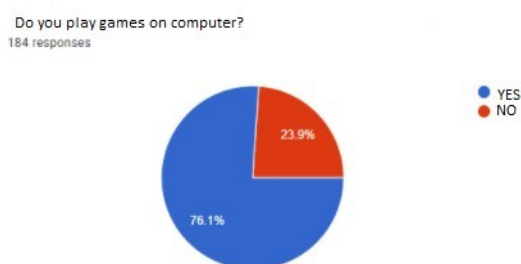


Fig. 4 Do our respondents play video games

Our next question is interesting for time management. Because we wanted to know how much of our respondents play games in their free time, and from Fig. 5 we can see that 44% of our respondents play more than one hour per day second part from the chart is 24.2% of our respondents play less than one hour, which is good and they do something else than playing games. And we have 15.9% who play exactly one hour. These students can say that they have great time management. And 15.9% of the students don't know their average playing time or they don't want to say it.

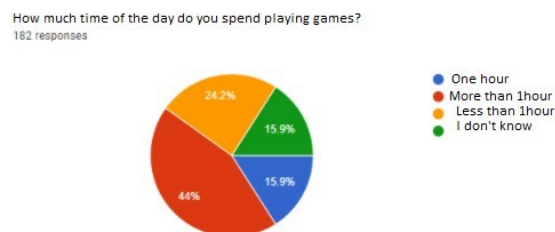


Fig. 5 How much time our respondents spend playing video games

Our next question was a little medical because we asked our respondents did they had disturbed physical or mental health from playing video games for example kyphosis, pain in their back or eyes, etc. In Fig. 6 it seems that 79.9% of our respondents didn't experience any disorders, which is good. But 20.1% of our respondents experienced some disorders.

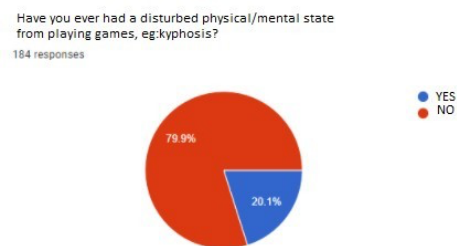


Fig. 6 Did our respondents experienced any disorders?

From the chart, in Fig. 6 we see that 20.1% said yes. But for example, if they got a kyphosis because they play games it's not true. Because they decide how to sit and how far away they are from their computer screens. They can get kyphosis if they write their homework without computers or watch videos or message their friends on Facebook. For our next question, we introduced the Likert scale and we have minus two respondents from our previous two questions. And we wanted to know how often our respondents interact with their teammates if they play online multiplayer games. We can see that 42.9% always communicate with their teammates during the gameplay, and 36.3% said that they sometimes communicate with their teammates. Maybe this means they don't want to talk too much, or no one is talking to them. And 20.9% said that they never communicate in the game, which indicates that maybe they are shy, or don't know the language they hear.

How often you interact with your players during the game?
182 responses

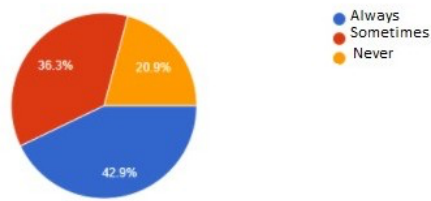


Fig. 7 How often our respondents communicate

In the next question we were interested in the games our respondents played, whether it's online or offline. For offline games they can't communicate with the players who are in the game, but they can communicate with a friend or the person next to them if the game is played with two players.

Do you play online or offline games?
177 responses

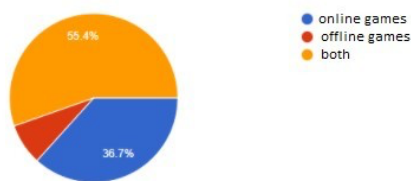


Fig. 8 Do our respondents play online or offline games

From Fig. 8 we can see that students who are in high school play online and offline games i.e. 55.4%, 36.7% play online games, and 7.9% play offline games. This chart indicates that most of the students are playing online games, which means they are active on the internet. This means that they can communicate with their teammate if the game can do that.

Do you think that by communicating with people from another country you are learning something new(eg: new language, culture,...)?
184 responses

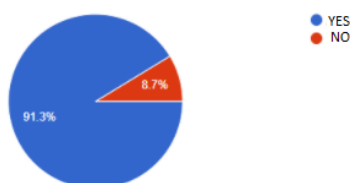


Fig. 9 What do our respondents think about learning new stuff from video games

In the chart from fig. 9 our respondents 91.3% said that they can learn something new while playing videos games. For example a new language or more information about their teammates. If the game is online, the players can be from another country and they will speak different languages which the player can learn more about that.

What games do you play?
180 responses

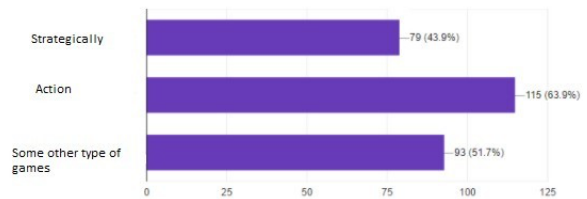


Fig. 10 The game genre

In Fig. 10 and our 10-the question was about the genre that our respondents play. We put an accent on strategy and action games because we believe that our brains are developing faster playing this type of genre. So 43.9% or 79 respondents said that they play strategic games, 115 respondents play action games, and 51.7% said they play something else. Sometimes they can hear rude words when they communicate with the players, friends, or someone else. Of course, this is present in the games, so we asked our respondents how often they hear rude words.

How often do you hear rude words from your teammates about you or another player?
183 responses

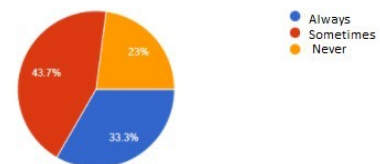


Fig. 11 How often do you hear rude words

Sometimes when people play games, they can make a friendship with someone else if that person is friendly. To confirm this, we introduced a question. As we can see in Fig. 12, 80.9% formed a friendship from playing games, which means, yes, people who play games can form a friendship with somebody.

Have you formed a friendship with another player while playing games?
183 responses

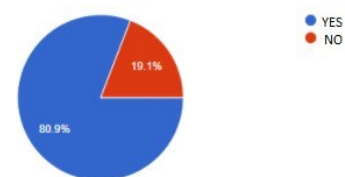


Fig. 12 Chart about forming a friendship in games

And for the final question for our research, we had to ask the students, do they watch e-sports or traditional sports? And for our luck 43.5% watch e-sports which is a new way of entertainment the younger and for the older generation.

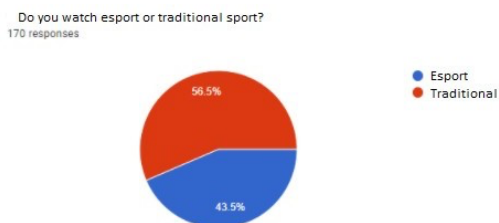


Fig. 13 Do high school students watch e-sports or traditional sports

DISCUSSION

With the first hypothesis we wanted to confirm does playing games can disturb the mental and physical condition of people, so we asked if they ever had a physical or mental disorder from playing video games, and most of them hadn't. To confirm the second hypothesis, we asked two questions: how often they interact with their teammates, and how they think that gaming can improve their way of thinking and coping in certain situations. The third hypothesis is linked with question 9 and question 14.

In one article from Dailymail [2], we can see that a man goes blind in one eye after playing video games on his mobile all night. He stared at the tiny screen of his phone all night. Yes, playing video games can damage the eyes, if the person is addicted to games and can stop. But this is not because of the games, but by looking at the screen for a long period, the eye can be damaged whether it is reading an e-book, watching videos, etc.

The third hypothesis is also true. If students play online games, they connect with different people around the world. This means they meet new people every day and if the other person is friendly, they can make a good conversation, whether it is for the game or their private life. Yes, in online games students often play games in teams, which means they can benefit their social development if its done in moderation [5]. Video games also provide the players with the opportunity to take a leader's position, which requires greater social networking skills and teamwork to keep other players happy. These skills are important the outside of the world for creating and maintaining friendships, especially in school and work environments.

In one study from pewresearch [7], more than half of the teens have made a new friendships online, and a third of them said that they met their new friend or friends while playing video games.

In another study from CNN [3] and our answers from the survey, we can also confirm that the first hypothesis is not true. In our survey and the survey from the study, we can see that many people who are in high school said that playing video games helps them either to deal with stress and difficult emotions or escape it from them. Especially during COVID-10 lockdowns. They also said video games allowed them to maintain friendship bonds in a multifaceted way.

Students must know how to limit themselves from playing video games for a long period. In research by Knepublishing [6] and from our question five on how much students spend their time playing games we can confirm that they know how to limit their time to function well in class.

From an article on the hightechwiz website [8], we can see several advantages and disadvantages which can be taken as an example. The first advantage is concentration. The levels in video games are designed with challenges and require the player to pay attention to the game. With concentration, the players can get analytical skills to solve those levels and their problems. This helps high school students to improve their thinking skills, which can be great for their future. Also, there is stress relief. If they play computer games, their brain release a hormone called dopamine which means they are happy and they can calm down. For the disadvantages, the article shows addiction which we wanted to confirm and other disadvantages for example socialization or health problems.

According to a survey conducted in the paper published by the National Library of Medicine [10], there is concern about the potential negative impact of video games on youth. Among gamers, 4.9% reported problematic gaming, defined as reporting trying to cut back, experiencing an irresistible urge to play, and experiencing a growing tension that could only be relieved by playing. Boys were more likely to report these problems (5.8%) than girls (3.0%). Correlates of problematic gaming included regular cigarette smoking, drug use, depression, and serious fights.

Jonathen Bartholomew, Board-certified Pediatrician in his article [9] highlights the pros and cons of video games for children. According to him although video games get a lot of bad press, they do have some benefits. The trick is to strike the right balance of good content and appropriate limits. After all, video games should enhance a child's life, not take it over. Video games can play a positive role in children's lives, but people need to be there to help set limits. If the parents are worried that their child is becoming too obsessed with games, then it's time to sit down and talk. They should be sensitive and nonjudgmental. If people need help, they can contact their pediatrician. He or she may have great tips and can even recommend a therapist who specializes in tweens or teens.

CONCLUSION

As we can see from our survey students who are in high school, can see that playing video games is not a time-wasting, i.e. by playing video games students can find a new way to communicate with others in a time when we have the Covid-19 pandemic. Also while playing video games, students don't go outside and they don't hang up with people in huge groups. By playing video games, they have a new way to socialize and if they use it with a good purpose, they can learn different skills from the games that simulate a real-life job. And they can boost our intelligence if they play mind games, or they can play games to earn some money, which can be used to pay student bills, etc. High school students must arrange their time and set their school tasks first, then if there was free time left, they can use that time to play games.

REFERENCES

- [1] Igor N. 2021, Survey: Benefits and costs of computer games in high school students: www.esports.mk/anketa-prednosti-i
- [2] Tracy Mailonline, 2019, Man goes blind in one eye after playing mobile games: <https://www.dailymail.co.uk/news/article-7681855>

- [3] /, 2020, *Video games help children improve literacy, communication and mental well-being*, survey finds: <https://edition.cnn.com/2020/09/05/health/video-games-literacy>
- [4] Steven Dupon, 2020, *Be social, make friends and play video games*: <https://www.videogames.org.au/social-development>
- [5] Dennis Dumrique, 2017, *Online gaming: Impact on the Academic Performance and Social Behavior of the Students in Polytechnic University of the Philippines Laboratory High School*: <https://knepublishing.com/index.php/Kne-Social>
- [6] Amanda Lenhard, 2015, *Chapter 3: Video games are key elements in friendships for many boys*: <https://www.pewresearch.org/internet/2015/08/06/chapter-3-video-games-are-key-elements-in>
- [7] /, *Adult perception toward video games and physical activity using Pokemon Go*: <https://www.liebertpub.com/doi/abs/>
- [8] Michael Room, 2021, *5 Advantages and Disadvantages of Computer Games | Drawbacks & Benefits of Computer games*: <https://www.hitechwhizz.com/2021/05/5-advantages-and-disadvantages-drawbacks-benefits-of-computer-games.html>
- [9] Jonathen Bartholomew, 2019, *Video Games: Pros and Cons | How to use video games for good*: <https://www.uvpediatrics.com/topics/video-games-pros-and-cons/>
- [10] Rani Desai, Dana Cavallo, Marc Potenza, 2010, *Video games playing in high school students*: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3678538/>

Survey on VoIP call security perception during COVID-19 pandemic

Marija Taneska

Faculty of computer science and engineering
Ss. Cyril and Methodius, University
Skopje, North Macedonia
marija.taneska@students.finki.ukim.mk

Sasho Gramatikov

Faculty of computer science and engineering
Ss. Cyril and Methodius, University
Skopje, North Macedonia
sasho.gramatikov@finki.ukim.mk

Abstract—The goal of this paper is to identify the most popular users' call security worries and users' perceptions about provided call security by popular VoIP-based applications during the COVID-19 pandemic. To accomplish this goal, we did a survey among different age categories in North Macedonia. In this paper, we describe how applications provide call security and we present and analyze the results of the survey.

Keywords—VoIP, security, COVID-19 pandemic

I. INTRODUCTION

Voice over Internet Protocol (VoIP) is a highly effective technology for enabling voice communications and media sessions over Internet Protocol (IP). VoIP offers significant cost savings over traditional landline service, long-distance phone calls become very inexpensive and that has attracted the most attention. Covid-19 has changed how we live, study, work and interact. Following the essential remote working protocols, there was a transition to work-from-home and the use of VoIP-based applications increased significantly.

In our paper, we describe how popular VoIP-based applications: Viber, Facebook Messenger, Discord, Whatsapp, Skype, Instagram, and Telegram provide call security. Our research focuses on the answers to the survey about call security perception from users of VoIP bases applications during the COVID-19 pandemic. We conclude that users consider that applications provide excellent and satisfactory call security and they are right.

II. CALL SECURITY

A. Call encryption

Analyzed applications ensure call security by providing end-to-end encryption (E2EE) or by using Transport Layer Security (TLS) protocol [1].

TLS provides encryption between users and service providers (applications), while E2EE provides encryption directly between users. E2EE ensures that only call parties can listen (and view), and nobody in between. We can conclude that applications that provide E2EE provide more privacy and security than applications that use only TLS.

B. Applications security

Viber [2], Whatsapp [3], FaceTime [4], and Telegram [5] provide end-to-end encrypted by default. Viber uses Salsa20 cipher [6] and other applications use AES-256 cipher [7].

Facebook Messenger [8], Instagram [9], and Skype [10] provide end-to-end encryption using AES-256 cipher, but not by default. To have end-to-end encrypted you have to use Secret Conversation on Facebook Messenger, End-to-end

encrypted chat on Instagram, and Private Conversation on Skype.

Discord [11] does not provide end-to-end encryption.

C. Comparison

Viber, Whatsapp, FaceTime, and Telegram provide more secure and privacy calls than Facebook Messenger, Instagram, and Skype default conversations. All applications (can) provide more security and privacy than Discord.

Viber, Whatsapp, FaceTime, and Telegram provide excellent call security. Facebook Messenger, Instagram, and Skype by default provide satisfactory call security, but there is an option to have excellent call security. Discord provides only satisfactory call security.

III. METHODOLOGY

In our work, we surveyed 161 people in North Macedonia. We were distributing the survey through social media and analyzed applications, with our friends and family, and gathered the responses for about 1 week. Our survey was answered by 36% male respondents and 64% female respondents. Because of the distribution of the survey, most of the respondents 61.49% belong to the 18-24 age group, and other respondents belong to the other age groups with a representation of around 10%. Our survey was answered by students 63.35%, employees 22.36%, retirees 9.32%, and 4.97% not employed. The respondents were asked a different set of questions about their use of the applications and VoIP call security perception. The results of the survey are presented and discussed in the next section.

IV. RESULTS

We asked our respondents which applications they use most for calls and which applications they started using (more) during the COVID-19 pandemic. As we can see from

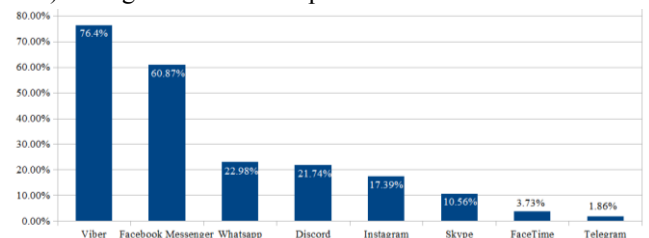


Fig.1 the most used applications for often calling are Viber and Facebook Messenger.

Fig. 1. Use of application for often calling

During the lockdown period in the 2020 year, we almost did not have live communication with family members, friends, and colleagues, though we live in the same city, and

we wanted to talk with them and see them, so we started using analyzed applications for local calls more. Also maybe because of that, our respondents use applications the most for local calls (77.5%), although VoIP at the beginning was the most used for international calls. The responses about which applications respondents started using (more) during the COVID-19 pandemic are shown in Fig.2.

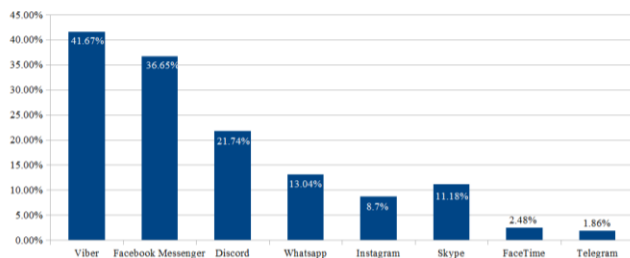


Fig. 2. Increased use of application during COVID-19 pandemic

We can see from Fig.2 that the most used applications Viber and Facebook Messenger have increased in usage for calling during the COVID-19 pandemic, but Skype, Discord, and Telegram have the largest increase. On the report of Viber in 2020 increase was 170% in all group-related activities, including chats, calls, and Communities [12]. As stated by Meta in February 2020 in Italy time in group calling increased by over 1,000% during the last month [13]. According to CNET in March 2020, 70% more people were participating in group video calls using Facebook Messenger, and the amount of time spent on those group video calls has doubled globally [14]. Many people started learning new languages during the COVID-19 pandemic, and the classes of a big part of the courses were conducted on Skype, maybe this is the reason of the large increase in use by our respondents has Skype with 105.87%. As claimed by CNET Skype saw a massive increase in usage as coronavirus spread, Microsoft said 40 million people are using Skype daily in March 2020, up 70% from February 2020 [15]. According to the Statista during the COVID-19 pandemic, there was an increase in playing games [16], and Discord can be described as a gaming social media application, so we can suppose that is the reason for a large increase in usage with 100%. The most important conclusion from Fig.2 is that all applications have an increase in usage for calling, and the respondents' reason for using the application is described below.

VoIP has become a widely used technology because it offers free calls. So we asked our respondents about the reasons Why they use these VoIP-based applications for calling? The most common reason for using applications is offered Free calls (77.9%). Also, the common reason is offered the Video call option (76.5%), accordingly most respondents make video calls, as we can see in Fig.3, and another not common reason is offered the Screen share option (22.1%).

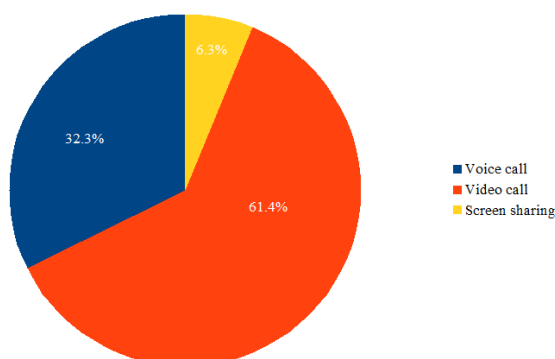
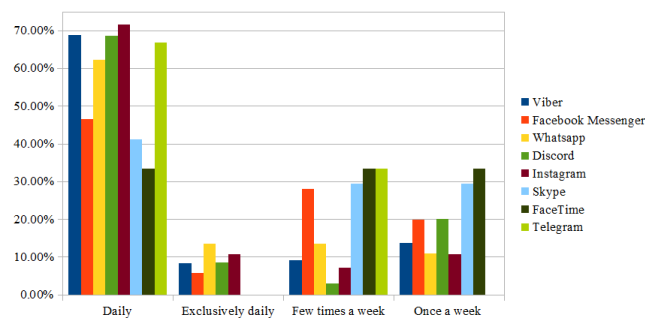


Fig. 3. Respondent makes the most video call, voice call or screen sharing

Almost all respondents (99.38%) use their mobile phone



for VoIP calls, because it is always with them (99.38%) and it is practical for calls (89.44%).

Fig. 4. Frequency of use of application for calls

We asked our respondents how often they use applications for calls during the COVID-19 pandemic and Fig.4 is a detailed view of their answers. We conclude that applications are used mostly daily by our respondents, but also they often use applications in general. Accordingly, respondents are concerned about call security. Also, we asked respondents about their most VoIP call duration during the COVID-19 pandemic. The most common call durations are less than 30 minutes (39.7%) and 30 minutes - 1 hour (36.8%). On the other hand, less represented are calls duration 1 hour - 2 hours (13.2%) and 2 (and more) hours (10.3%).

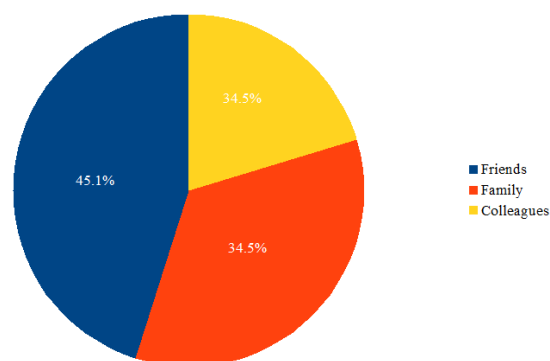


Fig. 5. Respondent use applications the most for call with friends, family or colleagues

Our respondents were asked With whom they communicate the most (during the COVID-19 pandemic) through applications. We can conclude from Fig.5, that our respondents the most communicate with their friends, appropriately, they started to use applications more for a calls with their friends during the COVID-19 pandemic, as we can see from Fig.6. It is important to say that there is an increase in the use of applications to communicate with family and colleagues too.

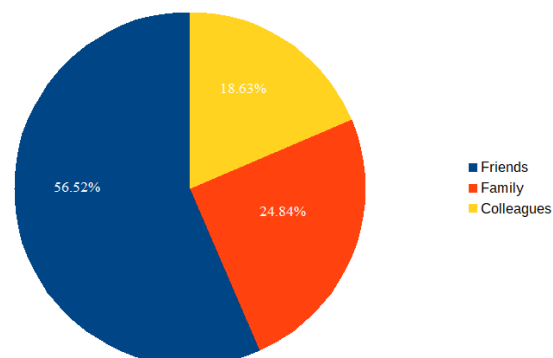


Fig. 6. Respondents use applications the most for call with friends, family or colleagues during COVID-19 pandemic

Heretofore we presented our necessary general analysis of the use of popular VoIP-based applications. Next, we present the analysis of call security worries of respondents and a discussion of the reasons for those concerns, relying on a general analysis.

Our respondents were asked:

- For what type of communication are you most worried about the security of the conversation?
- For what type of call are you most worried about the security of the conversation?
- What level of security do you think the applications provide?

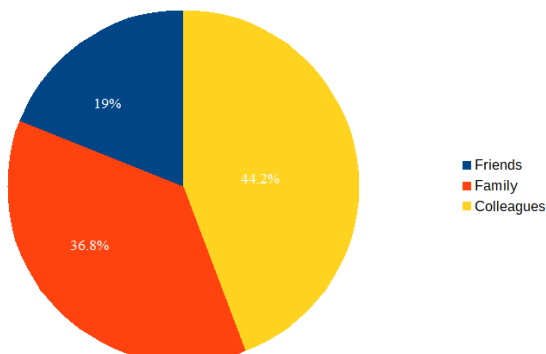


Fig. 7. Respondent is most worried about call security with colleagues, family or friends

We can conclude from Fig.7 that our respondents are the most worried about call security with colleagues, although they communicate the most with their friends, as we concluded before. This may be due to that they may be talking with colleagues about work or they maybe work from home.

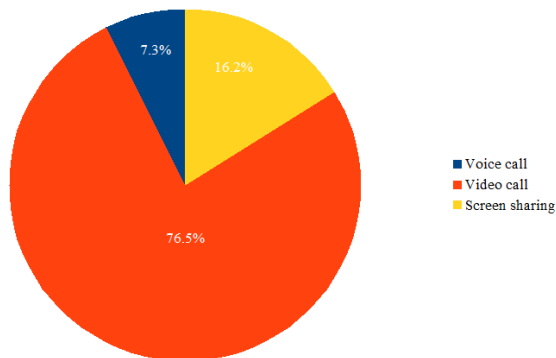


Fig. 8. Respondent is most worried about video call security, voice call security or screen sharing security

According to Fig.8 we conclude that the most common concern is video call security.

For video calls and screen sharing calls, we have a causal-consequential relationship between the use of the call type and concern about that call security type, for voice calls, there is no dependence between the use of voice call and concern about voice call security. All respondents that mostly make video calls and screen sharing calls are the most worried about video call security and screen sharing call security. Out of all respondents that mostly make voice calls, 47.1% are worried the most about video call security, 30.3%

are worried the most about screen share call security, and only 22.6% are worried the most about voice call security.

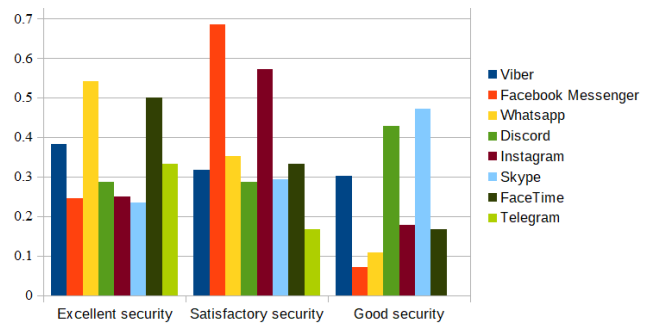


Fig. 9. Respondent's opinion about call security level

From Fig.9 we can conclude that our respondents consider that Excellent security is provided the most from Whatsapp, FaceTime, Viber, and Telegram. Satisfactory security they think is provided by Facebook Messenger and Instagram and Good security is provided the most by Discord and Skype.

According to the conclusions that we made in Section II about provided call security by applications, we can say that our respondents are right about provided security by Viber, Whatsapp, FaceTime, Telegram, Facebook Messenger, and Instagram, but they are not right about provided security by Discord and Skype. Discord provides Satisfactory call security, Skype provides Satisfactory call security by default, but there is an option to have Excellent call security.

CONCLUSION

In our work, we describe how popular VoIP-based applications: Viber, Facebook Messenger, Discord, Whatsapp, Skype, Instagram, and Telegram provide call security. Viber, Whatsapp, FaceTime, and Telegram provide excellent call security. Facebook Messenger, Instagram, and Skype by default provide satisfactory call security, but there is an option to have excellent call security. Discord provides only satisfactory call security.

We did a survey among 161 people in North Macedonia about their use of applications during the COVID-19 pandemic, their call security worries, and their call security perception. We concluded that the respondents started use more applications during the COVID-19 pandemic. The most of respondents use applications daily, accordingly, they are concerned about call security. Respondents use applications the most for calls with friends, but they are the most worried about call security with friends. Most of the respondents use applications for video calls and they are most worried about video call security. Opinion of our respondents about provided call security is correct for Viber, Whatsapp, FaceTime, Telegram, Facebook Messenger, and Instagram, but it is not correct for Discord and Skype.

REFERENCES

- [1] Ashutosh Satapathy, Jenila Livingston, "A Comprehensive Survey on SSL/TLS and their Vulnerabilities", International Journal of Computer Applications, November 2016
- [2] Viber encryption overview, available at <https://www.viber.com/app/uploads/viber-encryption-overview.pdf>
- [3] WhatsApp Security, available at [WhatsApp Security](https://www.whatsapp.com/security)
- [4] FaceTime & Privacy, available at <https://www.apple.com/legal/privacy/data/en/face-time/>

- [5] Telegram End-to-End Encrypted Voice Calls, available at <https://core.telegram.org/api/end-to-end/voice-calls>
- [6] Daniel J. Bernstein, "The Salsa20 family of stream ciphers", December 2007
- [7] Massimo Bertaccini, "Cryptography Algorithms", 2022
- [8] Messenger Secret Conversations Technical White Paper, available at [Messenger Secret Conversations Technical Whitepaper \(fb.com\)](https://www.facebook.com/secretconversations/whitepaper)
- [9] What is end-to-end encryption on Instagram?, available at <https://help.instagram.com/491565145294150>
- [10] Does Skype use encryption?, available at <https://support.skype.com/en/faq/FA31/does-skype-use-encryption>
- [11] Discord Developer Documentation, available at <https://discord.com/developers/docs/intro>
- [12] Viber - A Year of the "New Normal" – How People Stayed Connected, available at <https://www.viber.com/en/blog/2021-03-24/a-year-of-the-new-normal-how-people-stayed-connected/>
- [13] Meta - Keeping Our Services Stable and Reliable During the COVID-19 Outbreak, available at <https://about.fb.com/news/2020/03/keeping-our-apps-stable-during-covid-19/>
- [14] CNET - Facebook sees 70% increase in Messenger group video calls following coronavirus outbreak, available at <https://www.cnet.com/tech/services-and-software/facebook-sees-surge-of-engagement-worldwide-following-coronavirus-outbreak/>
- [15] CNET - Microsoft's Skype sees massive increase in usage as coronavirus spreads, available at <https://www.cnet.com/tech/mobile/microsofts-skype-sees-massive-increase-in-usage-as-coronavirus-spreads/>
- [16] Statista - COVID-19 impact on the gaming industry worldwide - statistics & facts, available at https://www.statista.com/topics/8016/covid-19-impact-on-the-gaming-industry-worldwide/#topicHeader_wrapper

Forecasting the prices of the day-ahead electricity markets using real data from SEEPEX

Pavle Petkovski

Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
pavle.petkovski@students.finki.ukim.mk

Aleksandra Dedinec

Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
aleksandra.kanevche@finki.ukim.mk

Abstract—Forecasting the day-ahead electricity prices can be significant for every business involved in the electricity market. In this paper, we compare different machine learning techniques and algorithms using real data from Serbian Power Exchange, weather data from Serbian capital city Belgrade and generation per production type data for Serbian electricity production. Then on this data, we train different machine learning models: Linear Regression, Decision Trees, Support Vector Machines, Random Forest models, Extreme Gradient Boosting models, Deep Learning models. Metric that we used for comparison between models is the coefficient of determination.

Index Terms—electricity, price forecasting, machine learning, deep learning

I. INTRODUCTION

The importance of forecasting prices comes from the volatile nature of electricity production and consumption. Electricity cannot be stored as easily as gas, it is produced at the exact moment of demand. The purpose of this research is to try to predict electricity prices for day-ahead using historical exchange data and other useful data. The electricity price can be affected by many factors, some of which are historical price, volume, day of a week, season, is it a holiday, hour of the day. With the increasing usage of renewable energy sources, the effect of weather conditions such as wind, rain, sunlight on electricity prices also increases.

With the rise of popularity of machine learning there had been rise of interest to research electricity price forecasting in the last few years. The paper [1] compares common statistical method Lasso Estimated Auto Regressive(LEAR) with deep neural networks on various power exchanges historical data. Hybrid methods are gaining popularity one of them is [2] which combines Deep belief networks, Long short-term memory and Convolutional neural networks to make a powerful forecasting model. These papers, use data for markets that exist for many years and are mostly located in developed countries.

The data used for this research is gathered from the Serbian Power Exchange containing hourly prices and volumes from 2016-02-18 to 2021-08-16, historical weather data for Belgrade in the same period and historical electricity production data for Serbia from Enstso-e transparency platform. The Power Exchange works in a way that during the day based on the traders bids the prices for next day (from 00:00 to 23:00) is set. The goal of this research is to predict prices for all 24 hours. SEEPEX data and weather data were gathered using web scrapping tools more precisely Selenium WebDriver and BeautifulSoup Python library and production data was freely available on official site [3]. The weather data for Belgrade contains temperature, air pressure, wind speed and visibility. Electricity production data contain hourly production from

biomass, fossil brown coal (lignite), hydro pumped storage, hydro run of the river and poundage, hydro water reservoir and other resources. With all these parameters, machine learning algorithms should be able to spot patterns that affect electricity prices assuming our data has insignificant noise.

The paper is structured as follows. First analysis of the data set is presented, than a description of the methodology used is given, followed by the results and corresponding discussion and conclusion.

II. DATA SET ANALYSIS

Using statistical tools and visualisations of the gathered data it can be seen that most of the assumptions made previously in this paper are to some degree true. First, we have analysed the average prices at monthly level (Figure 1). Knowing that January is the coldest month it is expected that electricity prices are highest and in summer electricity is needed for cooling so a spike in July and August proves that the assumption is true. The second assumption that the day of the week affects the electricity prices can be justified from Figure 2. The graph plot shows that weekend prices decrease. Significant difference between prices can be seen if a day is a holiday or not (Figure 3). Last visualisation and maybe most valuable for this research is the average price in different hours of the day. Figure 4 shows the price pattern during 24 hours of the day from which it can be seen that morning prices are lowest and the peak is at 19:00.

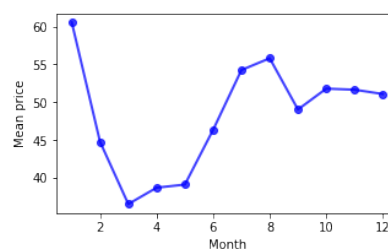


Fig. 1. Visualisation of prices over Months

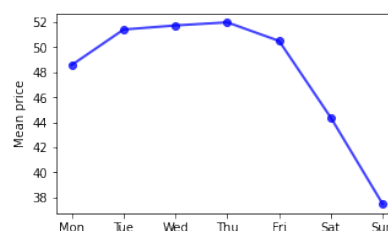


Fig. 2. Visualisation of prices over Days of Week

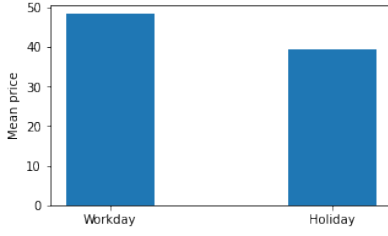


Fig. 3. Electricity prices compared over Workday and Holiday

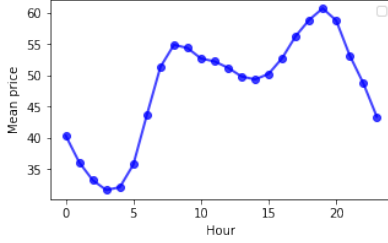


Fig. 4. Visualisation of prices over Hours of the Day

III. METHODOLOGY

Methods used for this research are:

- Linear Regression - widely used statistical method which presents response as a linear function of the inputs using the following equation:

$$y(x) = w^T x + \epsilon = \sum_{j=1}^D w_j x_j + \epsilon \quad (1)$$

where the $w^T x$ represents the inner or scalar product between the input vector x and the model's weight vector w^T , ϵ is the residual error between our linear predictions and the true response and y is electricity price. The goal is to minimise ϵ for given data.

- Decision Trees or Classification and regression trees - defined by recursively partitioning the input space, and defining a local model in each resulting region of input space. This can be represented by a tree, with one leaf per region. It may seem that this approach is not suitable for the regression task but if we get mean value for each region this method can perform decently.
- Support Vector Machines - Support Vector Regression used for solving regression problems means solving

$$\text{minimise}(1/2||w||^2) \text{ subject to } y_i - \langle w, x_i \rangle - b \leq \epsilon \quad (2)$$

where x_i is a training sample with target value y_i , The inner product plus intercept $\langle w, x_i \rangle - b$ is the prediction for that sample and ϵ is a free parameter that serves as a threshold, all predictions have to be within an ϵ range of the true predictions.

- Random Forest - technique used to reduce variance. The idea is to train M different decision trees on different subsets of data chosen randomly with replacement, and then compute the ensemble

$$f(x) = \sum_{m=1}^M \frac{1}{M} f_m(x) \quad (3)$$

where f_m is the m 'th tree.

- Extreme Gradient Boosting (XGBoost) - extremely popular algorithm and was go to algorithm for machine learning competitions until the rise of Deep Learning. It improves the Boosting and Gradient Boosting algorithms. The idea behind boosting is to combine lots of weak learners or classifiers to create strong regressor. At the beginning of training all training samples have same weight but, as training weak learners continues weight of samples that are correctly predicted decreases and weight of incorrect ones increases, in most cases this results lower bias and lower variance. Gradient boosting uses a gradient descent algorithm to minimize error using following equations:

$$w = w + \eta \nabla w$$

$$\nabla w = \frac{\partial L}{\partial w} \quad (4)$$

where L is loss function, w is weight vector and η is learning rate. Improvements that XGBoost brings to Gradient Boost are Tree Pruning, Sparsity Aware Split Finding, as well as computational improvements like parallelization and cache aware optimisation.

- Deep Neural Networks (DNN) - more precisely Feed Forward Neural Networks which moves information only to next layer not to layer behind or to nodes at same layer, this can be represented as directed acyclic graph which has one input layer, one or several hidden layers and one output layer. Between each node from neighboring layers there is weight that affects the output. Using BackPropagation Algorithm DNN model adjusts the weights to fit the training set. The power of DNN comes from the ability to learn non-linear patterns and for that is used in a wide variety of problems including price forecasting. The relationship between input x and output y is presented by this equation:

$$y = \sum_{j=0}^h \left[w_j f \left(\sum_{i=0}^d w_{ij} x_i \right) \right] \quad (5)$$

where w_i and w_{ij} represent the weight and biases that connect the layers.

For loss function Mean absolute error (MAE) was used which can be calculated by the following equation:

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n} \quad (6)$$

where y_i is true value and x_i is predicted value.

During training there was some overfitting present and because of that Lasso Regularisation (L1 Regularization) was used which penalizes large values for weights.

Coefficient of determination (R^2) was used as metric for comparison of machine learning models used in this research. R^2 is a measure of the goodness of fit of a model. R^2 is calculated using the following equation:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - f_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (7)$$

where f_i is predicted value, y_i is true value. An R^2 of 1 indicates that the regression predictions perfectly fit the data.

A. Inputs

After experimenting with lots of different input variables, the following ones were selected:

- Hour of the day
- Day of the week
- Month
- Holiday flag
- Volume
- Average electricity price for previous day
- Electricity price for previous seven days in exact hour
- Daily average temperature
- Daily maximal temperature
- Daily minimal temperature
- Daily average pressure
- Daily average wind speed

IV. RESULTS AND DISCUSSION

The initial data sets were divided into two parts, one for training and one for test. Training data set contained values for the period from 25-02-2016 to 25-12-2019 and test set contained values for the period from 26-12-2019 to 16-08-2021 meaning first 70% of data were used for training and remaining 30% for testing. On this test set all models were tested. The results after training, fine tuning and testing are the following:

- Linear regression model got R^2 of 0.834 on training set and 0.870 on test set. High R^2 value from the test set comes from predicting peak at the end of test period which few of next models did. (Fig. 5)
- Decision tree regressor got R^2 of 0.859 on training set and 0.833 on test set. This model had problems with generalization, it would get high R^2 score on training set but low R^2 score on test. This problem was solved using maximal depth parameter which prevents the tree from overfitting on the training set. (Fig. 6)
- Support vector machine training was time consuming and with that very hard to fine tune the parameters. On training set R^2 was good but, on test set was surprisingly low with maximal R^2 at 0.647. There was no improvement with different kernels and different "C" values hardly affected anything.
- Random forest regressor got R^2 of 0.902 on training set and 0.863 on test set. For fine tuning there was a trade off between number of estimators and max depth and in the end this model did very well. (Fig. 7)
- Extreme gradient boost had the highest R^2 of 0.938 on training set and decent score of 0.867 on test set. Fine tuning was mainly between number of estimators and maximal tree depth as well as learning rate. (Fig. 8)
- Deep Neural Network model got R^2 of 0.871 on training set and on 0.882 test set with 9 hidden layers and 691 nodes. In fine tuning main challenge was to choose the number of layers and number of nodes. With higher number of layers there was increase in R^2 training score, but decrease in R^2 test score. Improvement was made using L1 regularization, which highly improved generalization. (Fig. 9)

It is important to note that great increase was made by adding weather information to data set, almost all model got 0.01 to 0.03 R^2 score boost. (Fig. 10)

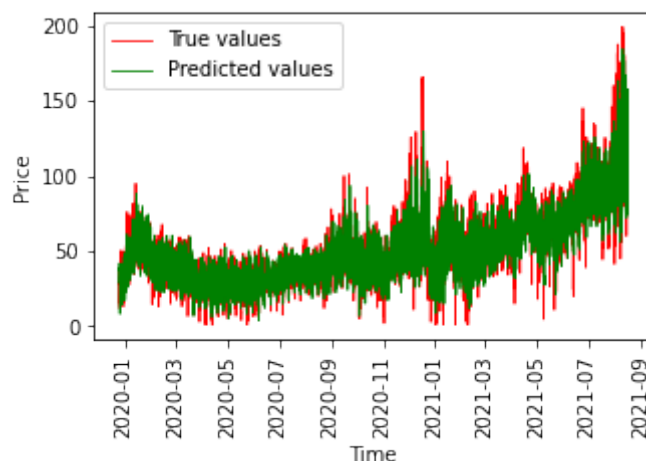


Fig. 5. Comparison between actual and predicted electricity prices using Linear Regression on test data set

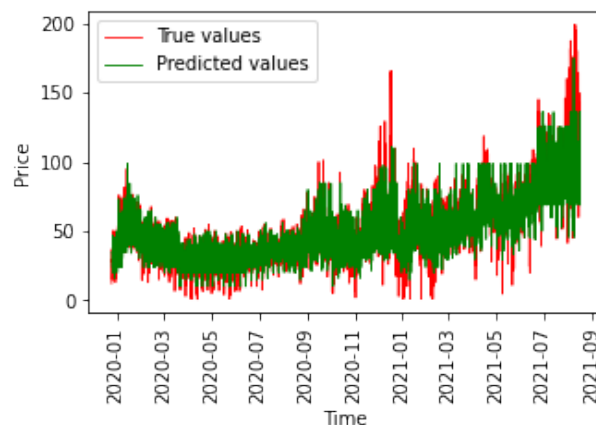


Fig. 6. Comparison between actual and predicted electricity prices using Decision Tree on test data set

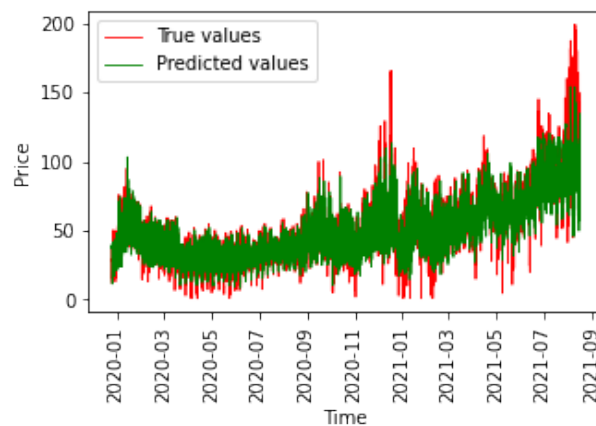


Fig. 7. Comparison between actual and predicted electricity prices using Random Forest on test data set

V. CONCLUSION

Using historical price data and weather data all of the mentioned machine learning models were able to predict the electricity prices. Differences between them still exist and from the obtained results, the most adequate are Deep Neural Networks and Extreme Gradient Boosting. Using the electricity production data by generating unit from ENTSO-E did not improve the results as much and in some situations can act as a noise. The reasons for this are still to be analyzed, including the analyzes of the completeness of the data. It would be very interesting in the future to see these methods used on markets where most of the traded electricity is produced from renewable sources.

ACKNOWLEDGMENT

This work was part of my internship in the Macedonian Academy of Sciences and Arts.

REFERENCES

- [1] Jesus Lago, Grzegorz Marcjasz, Bart De Schutter, Rafał Weron, Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark, *Applied Energy*, Volume 293, 2021, 116983, ISSN 0306-2619, <https://doi.org/10.1016/j.apenergy.2021.116983>.
- [2] R. Zhang, G. Li and Z. Ma, "A Deep Learning Based Hybrid Framework for Day-Ahead Electricity Price Forecasting," in *IEEE Access*, vol. 8, pp. 143423-143436, 2020, doi: 10.1109/ACCESS.2020.3014241.
- [3] <https://transparency.entsoe.eu/>

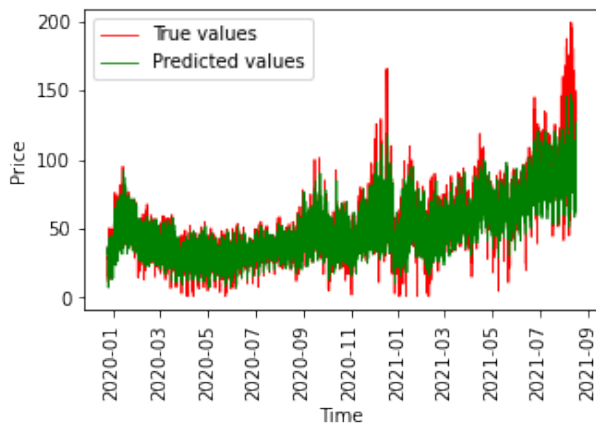


Fig. 8. Comparison between actual and predicted electricity prices using Extreme Gradient Boosting on test data set

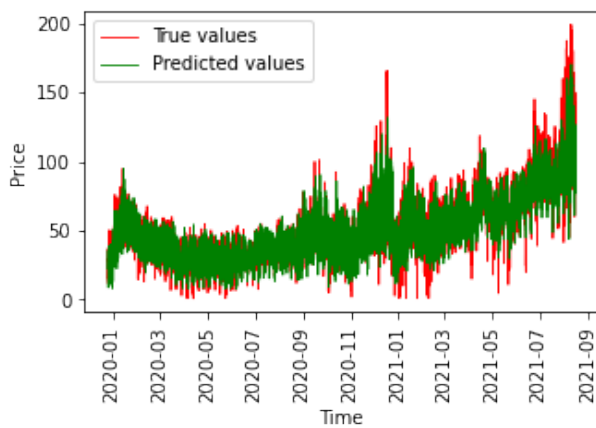


Fig. 9. Comparison between actual and predicted electricity prices using Deep Neural Network on test data set

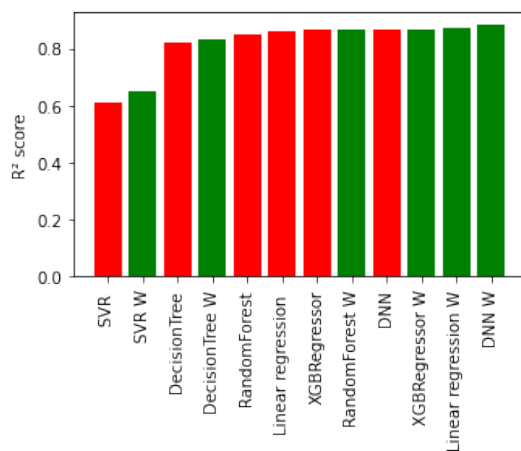


Fig. 10. Comparison between all models made in this research by R^2 score on test set, red bar indicates that model was trained without weather data

Digital Piracy in the Western Balkans with a special reference to North Macedonia

Simon Atanasovski

Faculty of Computer Science and
Engineering-FCSE-UKIM
Resen, North Macedonia

simon.atanasovski@students.finki.ukim.mk

Diellza Sherifi

Faculty of Computer Science and
Engineering-FCSE-UKIM
Resen, North Macedonia

diellza.sherifi@students.finki.ukim.mk

Sasho Gramatikov

Faculty of Computer Science and
Engineering-FCSE-UKIM
Skopje, North Macedonia

sasho.gramatikov@finki.ukim.mk

Abstract— This paper is a brief attempt to understand the current situation with digital piracy on the territory of the Western Balkans with special reference to the situation in North Macedonia.

Digital piracy is a state of owning, using, copying, and distribution of copyrighted material without the permission of its owner. As a phenomenon, piracy has long time been a global problem, which, among other things, causes great material losses to many individuals, and private entire industries. Especially affected are the software industry, music, film, and other industries. For the purposes of this paper, a survey was conducted regarding the use, prevalence, ways, causes, and awareness of the legal consequences of the use of pirated materials in our country. The results of our work should lead to possible directions for further action in order to reduce this problem.

Keywords – Digital piracy, Intellectual property, copyright, ethics.

I. INTRODUCTION

With the fast development of information technologies, their easy availability, and the entering of Internet technology to the most remote areas around the world as one of the global and priority problems in a long period of time the widespread prevalence of piracy is imposed.

Notably, every ("producer"), in this case of computer programs but also a composer and performer of music, film directors, producers, and countless other people and professions who create copyright and intellectual products should receive appropriate material compensation for their work. Proper protection of their labor is a serious challenge for today's information society.

One of the main assumptions for their protection is the creation and adoption of appropriate legislation that should deter potential ("thieves") of people's intellectual labor. Even though in the last years this phenomenon is regulated everywhere within the legal limits, by putting lawbreakers in prison conditions, it is visible that piracy is widespread around the world and especially in our environment in the Western Balkans.

So, even though they are legally regulated there are many other reasons why piracy is generally present in the area. In order to find the reasons for the presence of this phenomenon, we conducted a survey regarding North Macedonia results of which are presented in the upcoming parts. Territorial proximity, the relative similar mentality of the citizens but also the similar capacity of the states and their institutions would probably give an answer to what are the reasons why

piracy is used and above all is widespread in the Western Balkans.

In terms of the research, several papers such as [5] and [6] which treat the topic of piracy in the countries of the Western Balkans, were found and used. However, since no reference data were found on this matter for North Macedonia, we decided to prepare small research in the form of a survey.

II. DIGITAL PIRACY IN THE WESTERN BALKANS

While researching piracy in the Western Balkans, we used analysis by the American International Intellectual Property Association (IPPA) with reference available data for specific countries that are the subject of our paper. It is notable that in the same previous reports some of the countries were on the ("priority list") for observation for a long time [10]. For the same country, there is research by Albanian and Viennese authors (paperwork) which primarily covers digital piracy where is concluded that the problems in the area in that country are noticeable [6].

At the same time, we used data from research of the Business Software Alliance (BSA), according to which the rate of software piracy (especially for unlicensed software) was kept for a long time at 75%, the same as in the Republic of Montenegro, which according to the same institution is extremely high. It should be noted that both Macedonia and Serbia have a relatively high percentage of using unlicensed software [7].

RATES OF UNLICENSED SOFTWARE INSTALLATION				
	2017	2015	2013	2011
CENTRAL AND EASTERN EUROPE				
Albania	74%	73%	75%	75%
Armenia	85%	86%	86%	88%
Azerbaijan	81%	84%	85%	87%
Belarus	82%	85%	86%	87%
Bosnia	61%	63%	65%	66%
Bulgaria	57%	60%	63%	64%
Croatia	50%	51%	52%	53%
Czech Republic	32%	33%	34%	35%
Estonia	41%	42%	47%	48%
FYROM	63%	64%	65%	66%
Georgia	81%	84%	90%	91%
Hungary	36%	38%	39%	41%
Kazakhstan	74%	73%	74%	76%
Latvia	48%	49%	53%	54%
Lithuania	50%	51%	53%	54%
Moldova	83%	86%	90%	90%
Montenegro	74%	76%	78%	79%
Poland	46%	48%	51%	53%
Romania	59%	60%	62%	63%
Russia	62%	64%	62%	63%
Serbia	66%	67%	69%	72%
Slovakia	35%	36%	37%	40%
Slovenia	41%	43%	45%	46%
Ukraine	80%	82%	83%	84%
Rest of CEE	86%	87%	89%	90%
TOTAL CEE	57%	58%	61%	62%

Figure 1 Rates of Unlicensed PC Software Installations (BSA 2018) [7]

At the same time, it is interesting to note that in addition to the widespread use of unlicensed software by citizens and private companies, in these countries there is use of illegal

software by some state and public institutions. As a result in one of the IPPA reports, North Macedonia was warned and was given recommendations to overcome the detected problems [8]. However, according to the latest available report from this institution where North Macedonia is mentioned (2014) and having in mind that certain recommendations have already been implemented in our country, it can be concluded that at least in the legislative part but also in a certain amount of practice North Macedonia has certain progress. The same can be said for Albania and Serbia too. It is interesting that the Republic of Bulgaria, despite the fact being a member of the EU, was on the so-called ("watch list") for 2016, which means that it is in a worse situation with the use of piracy [9].

III. LEGAL PROVISIONS

For the purposes of this paper, only the provisions of the Criminal Codes of several countries in the Western Balkans that regulate this problem were considered and the Republic of Bulgaria was considered as a country for comparison which is a member of the EU. Of course, in all these countries there are ("Laws for protection of copyright and related rights") which, according to the limited space for research, are not taken into detail in this paper.

It can be concluded that in all these countries ("infringement of copyright and other related rights") is generally covered and prohibited. For example in North Macedonia, it primarily means unauthorized publication, display, reproduction, performance, broadcasting, or the manner of encroachment on another's copyright work, performance, or subject of related rights, for which imprisonment from 3 to 5 years is provided. In Serbia, it is sanctioned to put copyright work under one's own name in whole or in part and their publication of copies of another author's work or interpretation and without the author's permission edit or rewrite another work or recorded interpretation, unauthorized publication, recording, duplication or otherwise public announcement copyright, interpretation, phonogram, videogram, show, computer program or database with a fine and imprisonment of between six months and three years [3], [4].

The Albanian Criminal Code provides for these actions based on ("Illegal reproduction of someone else's work"), which means: Full or partial reproduction of a literary, musical, artistic, or scientific work that belongs to another, or if their use is carried out without the consent of the author when his personal and property rights are violated, constitutes a crime and is punishable by a fine or up to two years in prison [1].

As a comparison of these three countries from the Western Balkans, Bulgaria, a country from our closest environment, was taken as representative of the EU. In Bulgaria the above-mentioned regulations are covered by the following provisions:

("A person who records, reproduces, distributes, broadcasts, transmits, or makes any other use subject to copyright or related right without the consent of the owner of such right as required by law, shall be punished by imprisonment of up to five years and a fine of up to EUR 2,553") [2].

It is also sanctioned that anyone who, without the consent of a person, translates carriers of material containing a subject

matter of copyright or related right, which is a large-scale value, or who retains a matrix for the reproduction of such material shall be punished by imprisonment of two to five years and a fine of EUR 1,022 to 2,553. [2].

There is an additional sanction which foresees that if the act from the previous paragraphs is repeated or significant harmful consequences occur, the punishment will be increased from one to six years in prison and a fine of EUR 1,532 to 5,106 and they are sanctioned for damage from larger areas, as well as for lighter cases for which is provided an administrative. [2].

The general conclusion from this brief analysis of the legal provisions that regulate piracy is that (except with a small specificity in Albania) in all these countries the illegal actions regarding piracy are covered to the same degree. They are legally regulated in any case, but the general remark of the public is that there is a lack of serious discovery activity and proper implementation of all these provisions. We will try to give some of the answers in the upcoming part by presenting the results of the conducted survey that refers to North Macedonia. Territorial proximity, the relative similar mentality of the citizens but also the similar capacity of the states and their institutions would probably give an answer to the reasons why digital piracy is widely used and above all in the Western Balkans.

IV. SURVEY ANALYSIS

Given the fact that the last exact data on the use of piracy in North Macedonia for the needs of this paper are nowhere to be found, we conducted a survey to determine the degree of use, the reasons, the familiarity of citizens with the legal consequences, the frequency of use, etc.

The survey was conducted in a range from 22.03.2022 till 27.03.2022 with 346 respondents aged 14 to over 50, covering secondary, higher education, and employees.

The survey was split into several parts and they are as follows:

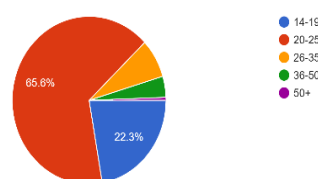


Figure 2. Age Distribution

A. Familiarity and use of illegal content

According to the results from the survey, almost 90% of respondents answered that they have heard of the term piracy. Only 10% answered NO, as shown in figure 3.

From the answer, it can be concluded that a large percentage of respondents know about the term digital piracy, and given that the survey includes people aged 14 years it is expected that some of them only formally have not heard of this term.

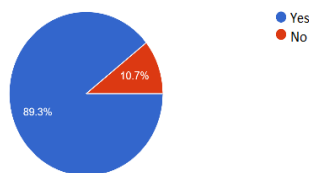


Figure 3. familiarity with the terminology

Figure 4 shows a large percentage of 80.1% of the respondents answered in the affirmative, which refers to the wide range of citizens who admit that they have used illegal content. It should be emphasized that in the framework of surveying by sending the material in groups, in the comments of the respondents, a partial fear and reservation were expressed as to whether their answers could be abused in possible proceedings against them for using illegal content. Accordingly, it is to be expected that the expressed percentage of 80%, although large, is probably not completely accurate, i.e. that the percentage is certainly even higher than shown.

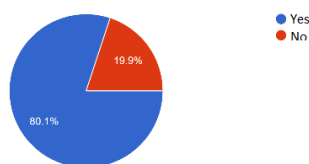


Figure 4. The use of illegal content (Digital piracy)

B. Availability and widespread of digital piracy

- What type of digital piracy have you used?

The results from the question ("What type of digital piracy have you used?") in figure 5, most of the respondents answered (79.2%) that have used movies and TV series, 62.4% answered that it is music, 58.7% answered that they have used pirated video games, 41.3% E-books, etc. which indicates the fact that R.N Macedonia covers all areas of intellectual copyright and related rights violations.

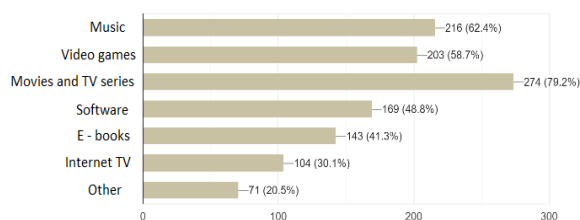


Figure 5. The use of digital piracy in categories

- The next answers obtained from the survey show an almost balanced result in the given possible answers from Everyday (26.3%) to monthly (30.1%) etc. There is a relatively confusing percentage of 21.4% who answered that they do not

use digital piracy which is contradictory with the previous general question for use of illegal content, but it is probably about the age category from the early years of high school who don't know what the term piracy is but later in the survey they saw that they are using.

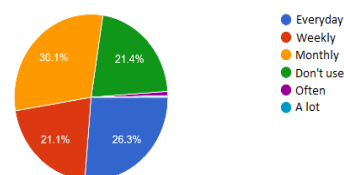


Figure 6. The frequency of use

- The question ("For what reasons did you decide to use pirated content") (figure 7) is showing the expected answer. Namely 76.3% answered that it is free, 41.6% think that the original products are very expensive, 34.4% answered that the material condition does not allow them to buy legal content and 34.1% think that they use it because our country is still not prosecuting for using pirated content.

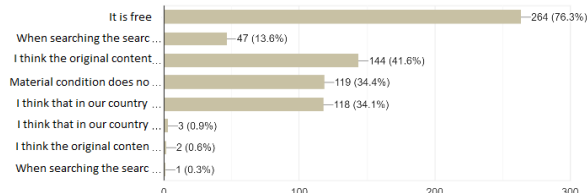


Figure 7 Reasons for using digital piracy

C. legal content

- The set of questions in the legal aspect part ("How often do you use legal content") the answers to the scaling possibility are heterogeneous and it can be concluded that it is often used. The answers to the question ("which legal services do you use") are expected and they move within all the latest modern services, applications, and social networks.

When asked which services are used, the answers are expected. There are almost all current and modern services and applications and of course, about 70% answered positively to the question of whether there is a difference in the quality of use of legal and illegal content.

D. Legal aspect

- 78% are aware that using piracy is an illegal act which indicates that a huge percentage of respondents are quite clear about their choice when using illegal content.

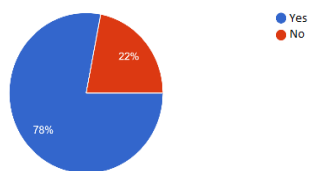


Figure 8. Awareness of illegal act

- We have a defeated percentage of almost 80% of respondents who have not heard in the media about the illegality of using illegal content. This points to the fact that probably there are not enough topics on portals, TV shows, or other magazines that point out the legal procedures and their consequences.

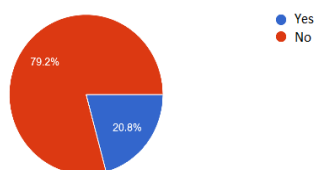


Figure 9. Media coverage

- 80.9% of the respondents have not heard of measures taken by state institutions for use of illegal content which leads to the conclusion that either such measures are almost not taken in our country or if they are taken they are not sufficiently exploited in public.

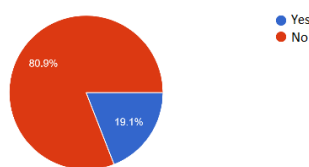


Figure 10. Awareness of taken measures

V. CONCLUSION:

In this paper, we gave an overview of the legal consequences of piracy in the Western Balkans. We also conducted a survey to get a glimpse of the situation in North Macedonia.

In a general conclusion, piracy is widespread in the Western Balkans, for which it is noted and is constantly monitored by relevant international institutions and organizations that have a positive, directing, and advisory role.

Furthermore, it can be concluded that there are relatively good legal preconditions for preventing and punishing it.

In Macedonia, specifically, piracy is also widespread and its breadth and coverage are largely due to the poor economic situation of citizens, the weak media coverage of its impunity, and the weak activity of state institutions committed to preventing and prosecuting this phenomenon.

With the improvement of the last few findings, the use of piracy will likely be reduced to a tolerable level, similar to the advanced Western European countries and the United States.

REFERENCES

- [1] Parliament of Republic of Albania, "Criminal Code of the Republic of Albania", warnathgroup.com, 2015, https://www.warnathgroup.com/wp-content/uploads/2017/11/Albania_CC_1995_am2015_en.pdf
- [2] Parliament of Republic of Bulgaria, "Criminal Code of the Republic of Bulgaria", Legislationonline.org, 2017, https://www.legislationonline.org/download/id/8395/file/Bulgaria_Criminal_Code_1968_am2017_ENG.pdf
- [3] Parliament of Republic of North Macedonia, "Criminal Code of the Republic of North Macedonia", pravdiko.com, 2013, <https://www.pravdiko.mk/wp-content/uploads/2013/11/Krivichen-zakonik-integralen-prechisten-tekst.pdf>
- [4] Parliament of Republic of Serbia, "Criminal Code of the Republic of Serbia", paragraf.rs, 2019, <https://www.paragraf.rs/propisi/krivicni-zakonik-2019.html>
- [5] Miroslav Emanoilov, "Abuse of intellectual property rights on the Internet in Bulgaria", modul.ac.at, 2013, https://www.modul.ac.at/uploads/files/Theses/Master/Thesis-2013-MBA-Miroslav_Emanoilov_t.pdf
- [6] A. Buhajoti, M. Lurie, "Digital Piracy: An Empirical Study On Public Opinion In Albania", jimest.org, 2015, <http://www.jimest.org/wp-content/uploads/JMESTN42350871.pdf>
- [7] Business Software Alliance, "Global Software Survey", 2018, https://gss.bsa.org/wp-content/uploads/2018/05/2018_BSA_GSS_Report_en.pdf
- [8] International Intellectual Property Alliance, "Special 301 Report on Copyright Protection and Enforcement on Macedonia", 2014, https://www.iipa.org/files/uploads/2017/12/2014SPEC301MACEDO_NIA.pdf
- [9] International Intellectual Property Alliance, "Special 301 Report on Copyright Protection and Enforcement on Bulgaria", 2016, https://www.iipa.org/files/uploads/2018/01/rbc20162016SPEC301AN_NEX-2.pdf
- [10] International Intellectual Property Alliance, "Special 301 Report on Copyright Protection and Enforcement", 2014, https://www.iipa.org/files/uploads/2018/01/2014_Special_301.pdf

Privacy Risks and Security Threats in Online Social Networks

Marija Bukalevska

Faculty of AITMIR

University of Information Science and Technology

St. Paul the Apostle

Ohrid, North Macedonia

marija.bukalevska@mir.uist.edu.mk

Aleksandar Karadimce

Faculty of ICS

University of Information Science and Technology

St. Paul the Apostle

Ohrid, North Macedonia

aleksandar.karadimce@uist.edu.mk

Abstract—Online social networks (OSN) have served as a catalyst in increasing social interaction in the 21st century. Because of their nature, social networks attract many people where they can share information such as text and multimedia content using sites such as Facebook, Twitter, Instagram and more. While OSNs provide a simple and fun way to share information about one's self, these networks also give rise to new and serious privacy and security issues that could potentially harm users in serious manners. Phishing attacks, the spread of malware, internet fraud, online stalking are a few examples of the vast landscape of possible security dangers that target OSNs. This paper is intended to reach the general audience and will focus on the privacy concerns and classic security threats in OSNs.

Index Terms—Online Social Networks, Privacy, Security, Risks, Threats, Issues

I. INTRODUCTION

As a result of living in today's digital world and having access to Internet technology, various online social networks have emerged, allowing people to stay connected. While social networks attract a large number of users due to the variety of interesting features and services they provide, their huge number of active users is exactly what makes them more prone and vulnerable to security and privacy risks. Cyber-attackers target websites and social networking platforms that have a large number of users to gain access to user's data and use it for their own personal gain. This data can include: personal information, emails, passwords, credit card credentials and more. There are various ways to illegally obtain information of users, such as phishing attacks, malware, cross-site scripting attacks (XSS), cyber fraud and more. The paper is organized as follows. Section II. lays out the related work. Section III. defines OSNs and their popularity. Section IV. gives a brief discussion of security concerns and privacy issues. Section V. describes classic threats and provides a table of security threats found in OSN platforms using the Internet as a research tool. Section VI. provides the conclusion and future work.

II. RELATED WORK

Fire, Goldschmidt and Elovici [8] provided a comprehensive review by presenting a taxonomy of various security threats that endanger the well-being of OSN users and children, offer an overview of existing solutions that improve OSN user protection, security, and privacy. The authors [7] discuss various security and privacy challenges and threats to OSNs. Furthermore, they present the findings of a survey in which participants were asked how they handled various types of privacy-related inquiries; the results revealed that

the responses were unsatisfactory, as many users did not use the existing privacy settings provided by service providers. Ali, Kamran, Ahmed, Raza, Ilyas [6] explored security and privacy issues of OSNs, presented a taxonomy of OSNs threats by categorizing threats into several categories, and provided privacy recommendations for OSN users. The study [9] examines many security issues and challenges in Social Networking Sites, as well as a classification and explanation of various security risks and threats in Social Networking Sites. Jain, Sahoo, Kaubiyal [10] go through several security and privacy issues, as well as existing defensive solutions that can keep users safe. The popularity, benefits, drawbacks of OSN, as well as key challenges and appropriate security standards for protecting account information, are all examined in this study.

III. DEFINING OSNs AND UNDERSTANDING THEIR POPULARITY

A. Defining Online Social Networks

A growing number of different kinds of services are now accessible via a computer network. These services have created a new type of virtual society known as online social networks. Web-based social networks, computer-assisted social networks, and virtual communities are all terms that have been used to describe them [2]. Boyd and Ellison [3] have defined social network sites as web-based services where individuals can create a public or semi-public profile within a limited system, aggregate a list of other users with whom they share a connection and view that list of connections.

B. The popularity of OSNs

Online social networks being the most widely used, easy to access and low-cost communication and information mediums have reached a remarkable increase in popularity in recent years. People's lives have become increasingly reliant on social networks due to their variety of features. Through social media, people can interact, communicate with their family and friends more easily, meet people with whom they share similar interests and hobbies and as well participate in discussion groups and forums. People can easily access news, information, as well as share, create, and access various types of content such as text, multimedia content, music and more by using social networks. Over the last few years, the popularity of OSNs has greatly increased and they have become a global technology phenomenon, with billions of users accessing them around the world. According to recent statistics there are 4.48 billion social media users around the world in July 2021, equating to almost 57 percent of the total

global population [11]. Table I shows the most popular social networks and their population.

TABLE I
SOCIAL NETWORKS POPULATION

Most Popular Social Networks	Active Users
Facebook	2.853 billion
YouTube	2.291 billion
WhatsApp	2 billion
Instagram	1.386 billion
FB Messenger	1.3 billion
WeChat	1.242 billion
TikTok	732 million
QQ	606 million
Douyin	600 million
Telegram	550 million
Sing Weibo	530 million
Snapchat	514 million
Kuaishou	481 million
Pinterest	478 million
Reddit	430 million
Twitter	397 million
Quora	300 million

* Most popular social networks of 2021 [11].

According to publicly available data up to July, 2021, at least 17 social media networks have 300 million or more monthly active users [11].

IV. PRIVACY AND SECURITY ISSUES ON OSNs FOR USERS

While, indeed, there are many benefits and positive aspects using these networks, there are also certain downsides that can cause severe challenges and problems for both individual users and for the companies that own these networks. In today's society, online social networks have exploded in popularity. People are more inclined to use several networking sites, such as Facebook, Twitter, Instagram, LinkedIn and Reddit [12]. In recent years, privacy and security in social networks has been a major concern. People who use Social Networking Sites expose themselves to a variety of risks, the most common of which being the breach of their privacy [5]. People are sharing an increasing amount of information on social networking sites in different forms, potentially providing access to personal information such as a person's name, gender, location and private images. This content is saved electronically and therefore it can easily be shared with more people than the user intended because many social networks are open to the public, given this, user's information can attract malicious people who want to harm someone [7], [10], [13]. Privacy and security risks can include various data breaches, identity theft, theft of banking information and credit card credentials for financial gain, cyber-bullying, stalking and more [12]. Phishing attacks, malware, cross-site scripting attacks (XSS), internet fraud, are a few examples of the variety of methods used to obtain information from users illegally. The lack of appropriate OSN tools and design, as well as the users' inadequate knowledge and awareness, do little to help the situation [13]. A social network's privacy component is large, with different sub-problems. User privacy, for example, comprises a number of sub-problems such as location data privacy and user personal data privacy [12]. Data in OSNs can include: personal information (derived from the user's profile), connections (such as group of friends or followers), messages (information that can be exchanged between users, but also posted publicly or privately on social

media spaces), multi-media, tags, groups (a collection of users sharing information and resources), behavioral information, login credentials and location data [13]. All of these mentioned types of data can be the subject of privacy and security risks. This information could fall into the wrong hands through browsing, hacker attacks, or simple data sales [13]. There are many ways to illegally obtain information of users, such as phishing attacks, malware, cross-site scripting attacks (XSS), internet fraud and more.

V. SECURITY THREATS FOUND IN ONLINE SOCIAL NETWORKS

According to authors threats can be classified in two main categories: classic threats and modern threats [7], while other authors add a third category, which is targeted threats and/or threats targeting children [8], [10]. The purpose of this paper is to conduct research using the Internet for content analysis to see whether users have ever been exposed and if any of the most popular social networking platforms (sec. III, Table I) have been subject to any classic security threats. Table II provides data for classic threats found on OSNs. The table is based on our findings. The data is obtained by using the Internet as a research tool to see if there have been any reported incidents of classic threats in the twelve OSN platforms given in the table. The data is obtained from references and the table II was created on that basis. [15]–[39].

User's personal information that is published on OSNs can be exploited by classic threats and compromised, but these attacks can be also tailored to target the user's friends.

- **Malware** - Malware is malicious software that is designed to disrupt a computer's operation in order to steal a user's credentials and access their personal information [8]. Because of the key characteristic of online social networks, which is that they connect a large number of users, malware can spread quickly among their users and infect more users [14]. There have been reported incidents of hackers targeting the popular messaging app Telegram with the ToxicEye remote access trojan. The ToxicEye malware can take over file systems, install ransomware, and leak data from victims' computers [38].
- **Phishing Attacks** - Phishing is a type of fraud in which an intruder obtains personal information from a user by impersonating a trustworthy third party using a fake or stolen identity [7]. There are incidents reported on Instagram where hackers have been seen targeting high-profile accounts on Instagram and getting access to accounts using phishing attacks [18].
- **Spammers** - Unwanted messages are known as spam. Spam appears as a wall post or a spam instant message on OSNs. Because users engage more on OSNs than they do on email, spam in OSNs is more hazardous [7]. For example, spam on Facebook and Twitter can involve sending these frauds suggesting that you're in a scandalous video. When you go to watch the video, you're directed to update a program, such as your Flash player. Instead of improving your computer, you're downloading malware that can hijack your Facebook account, collect your personal information, and even harm your computer [23], [24].

- Cross-Site Scripting (XSS) - Cross-site scripting (XSS) is a sort of attack against Web applications, including OSNs, in which the attacker involves injecting malicious script into targeted web-pages and induces the user to run it in order to steal sensitive information [14].
- Internet Fraud - Internet fraud, which is a cybercrime activity, involves using the Internet access to deceive or take advantage of people [8]. For example, fraudsters on Instagram like to impersonate well-known brands in order to trick you into purchasing counterfeit goods. They may also simply want your payment details in order to create bogus charges or empty your bank account [18], [31].

Given Table II we can note that reported incidents of phishing attacks [18]–[22], [34], [35], [37], [39], and Internet fraud [28]–[31], [33], [34], [36], [37], [39], can be found in most OSNs (in nine OSN platforms), followed by malware [15]–[17], [35], [38], [23] (in six OSN platforms) and spam attacks [23]–[27] (in five OSN platforms), whereas no reported incidents of cross-site scripting attacks (XSS) have been found.

VI. CONCLUSION AND FUTURE WORK

OSNs have grown in popularity as a global technology phenomenon, with billions of users throughout the world accessing them. Many individuals are drawn to using them because of the various unique features they provide such as posting and sharing information and multimedia content, staying connected with friends, participating in discussion forums and more. However, while OSNs are entertaining to use, they also raise privacy concerns and security risks that could potentially cause harm to the individuals using them. From Table II we can conclude that most OSNs have been exposed to classic security threats. This is an ongoing research that will explore more content in the future and will be expanded, an extension to the security threat table will be provided where modern threats (such as clickjacking, socialbots, identity clone attacks) will be also added. A more detailed analysis of the privacy awareness of users on these networks and as well as protection recommendations and proposed solutions to better protect OSN users are also planned to be included.

RESEARCH LIMITATIONS

This paper is limited due to the following reasons:

- Research was not made for all most popular social networking sites of 2021 (sec. III, Table I). Included are those sites that have the most international and English-speaking users.
- The data was not acquired from a specific dataset, therefore the exact number of times the threats were encountered in each OSN is not known. Instead, Table II was created and the data was collected by using the Internet as a tool for content analysis to see if there have been any reported incidents of classic threats on each OSNs, respectively.
- Because the Internet was used as a research tool, data for some threats could not be found. For example, no cross-site scripting (XSS) attacks were reported on any of the platforms, but it does not mean that the platforms and users have never been exposed to such attacks.

REFERENCES

- [1] D. Hiatt and Y. B., "Role of Security in Social Networking," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 2, 2016, doi: 10.14569/ijacsa.2016.070202.
- [2] K. Musiał and P. Kazienko, "Social networks on the Internet," *World Wide Web*, vol. 16, no. 1, pp. 31–72, Jan. 2012, doi: 10.1007/s11280-011-0155-z.
- [3] D. M. Boyd and N. B. Ellison, "Social Network Sites: Definition, History, and Scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, Oct. 2008, doi: 10.1111/j.1083-6101.2007.00393.x.
- [4] S. Kumari and S. Singh, "A Critical Analysis of Privacy and Security on Social Media," *2015 Fifth International Conference on Communication Systems and Network Technologies*, Apr. 2015, doi: 10.1109/csnt.2015.21.
- [5] Senthil Kumar N, Saravanakumar K, and Deepa K, "On Privacy and Security in Social Media – A Comprehensive Study," *Procedia Computer Science*, vol. 78, pp. 114–119, 2016, doi: 10.1016/j.procs.2016.02.019.
- [6] A. Ali, A. Kamran, M. Ahmed, B. Raza, and M. Ilyas, "Privacy Concerns in Online Social Networks: A Users' Perspective," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 7, 2019, doi: 10.14569/ijacsa.2019.0100780.
- [7] [1]S. Ali, N. Islam, A. Rauf, I. Din, M. Guizani, and J. Rodrigues, "Privacy and Security Issues in Online Social Networks," *Future Internet*, vol. 10, no. 12, p. 114, Nov. 2018, doi: 10.3390/fi10120114.
- [8] M. Fire, R. Goldschmidt, and Y. Elovici, "Online Social Networks: Threats and Solutions," *IEEE Communications Surveys Tutorials*, vol. 16, no. 4, pp. 2019–2036, 2014, doi: 10.1109/comst.2014.2321628.
- [9] S. Rathore, P. K. Sharma, V. Loia, Y.-S. Jeong, and J. H. Park, "Social network security: Issues, challenges, threats, and solutions," *Information Sciences*, vol. 421, pp. 43–69, Dec. 2017, doi: 10.1016/j.ins.2017.08.063.
- [10] A. K. Jain, S. R. Sahoo, and J. Kaubiya, "Online social networks security and privacy: comprehensive review and analysis," *Complex Intelligent Systems*, Jun. 2021, doi: 10.1007/s40747-021-00409-7.
- [11] DataReportal, "GLOBAL SOCIAL MEDIA STATS," *DataReportal – Global Digital Insights*, Oct. 2021. <https://datareportal.com/social-media-users>
- [12] M. Siddula, L. Li, and Y. Li, "An Empirical Study on the Privacy Preservation of Online Social Networks," *IEEE Access*, vol. 6, pp. 19912–19922, 2018, doi: 10.1109/access.2018.2822693.
- [13] Beye, M., Jeckmans, A., Erkin, Z., Hartel, P.H., Lagendijk, R.L., Tang, Q. (2010). Literature Overview - Privacy in Online Social Networks. CTIT technical report series.
- [14] S. Deliri and M. Albanese, "Security and Privacy Issues in Social Networks," *Data-Centric Systems and Applications*, pp. 195–209, 2015.
- [15] Securelca. (2020, May 20). Cyber-criminal targets facebook tags to spread malware. Medium. Retrieved from <https://medium.com/@securelca/cyber-criminal-targets-facebook-tags-to-spread-malware-2015da5a6af6>.
- [16] Hautala, L. (2020, October 1). Facebook: Malware that took over accounts and placed scammy ads a growing risk. CNET. Retrieved from <https://www.cnet.com/tech/services-and-software/facebook-malware-that-took-over-accounts-and-placed-scammy-ads-a-growing-risk/>.
- [17] Syntactics Inc., Batilong, J. (2021, September 30). Instagram virus: What you need to know and do. Syntactics Inc. Retrieved from <https://www.syntacticsinc.com/news-articles-cat/beware-instagram-virus/>.
- [18] Instagram ransoms targeting social media influencers. (2018, October 5). Tech Monitor. Retrieved from <https://techmonitor.ai/technology/cybersecurity/instagram-ransom-attacks>.
- [19] Sasnauskas, M. (2021, September 28). We uncovered a Facebook phishing campaign that tricked nearly 500,000 users in two weeks. CyberNews. Retrieved from <https://cybernews.com/security/we-uncovered-a-facebook-phishing-campaign-that-tricked-nearly-500000-users-in-two-weeks/>
- [20] Williams, R. (2018, February 20). Snapchat's phishing attack breached 50k accounts. Marketing Dive. Retrieved from <https://www.marketingdive.com/news/snapchats-phishing-attack-breached-50k-accounts/517377/>.
- [21] Bisson, D. (n.d.). Reddit clone site uses SSL certificate to lure users into handing over login credentials. Venafi. Retrieved from <https://www.venafi.com/blog/reddit-clone-site-uses-ssl-certificate-lure-users-handing-over-login-credentials>.
- [22] Kan, M. (2021, October 20). Hackers are phishing YouTube creators to steal their accounts, Google warns. PCMag. Retrieved from <https://www.pcmag.com/news/hackers-are-phishing-youtube-creators-to-steal-their-accounts-google-warns>.

TABLE II
SECURITY THREATS FOUND IN ONLINE SOCIAL NETWORKS

Threats	Facebook	YouTube	WhatsApp	Instagram	FB Messenger	TikTok	Telegram	Snapchat	Pinterest	Reddit	Twitter	Quora
Malware	✓ [15], [16]	✓ [22]	X	✓ [17]	✓ [35]	X	✓ [38]	X	X	X	✓ [23]	X
Phishing Attacks	✓ [19]	✓ [22]	✓ [34]	✓ [18]	✓ [35]	✓ [37]	✓ [39]	✓ [20]	X	✓ [21]	X	X
Spammers	✓ [24]	✓ [27]	X	✓ [25]	X	X	X	X	✓ [26]	X	✓ [23]	X
Cross-Site Scripting (XSS)	X	X	X	X	X	X	X	X	X	X	X	X
Internet Fraud	✓ [29]	✓ [30]	✓ [34]	✓ [31]	✓ [36]	✓ [37]	✓ [39]	✓ [28]	X	X	✓ [33]	X

- [23] Kerr, D. (2012, September 25). Twitter users may be victims of Direct Message malware. CNET. Retrieved from <https://www.cnet.com/tech/services-and-software/twitter-users-may-be-victims-of-direct-message-malware/>.
- [24] Varnsen, K. (2013, March 14). Types of facebook spam. Ranker. Retrieved from <https://www.ranker.com/list/types-of-facebook-spam/kel-varnsen>.
- [25] Red. (2016, September 7). Welcome to Instagram spam. Medium. Retrieved from <https://medium.com/@iamredmh/welcome-to-instagram-spam-dcbd89d692cb>.
- [26] Information environmentalism research: Fake accounts and MIS/disinformation on Pinterest. Digital Learning Inquiry (DLINQ). (2019, October 23). Retrieved November 30, 2021, from <https://dlinq.middcreate.net/blog/2018/03/14/information-environmentalism-research-fake-accounts-and-mis-disinformation-on-pinterest/>.
- [27] Breadnbeyond. (2020, August 27). Easy ways to handle spam comments on YouTube. Medium. Retrieved from <https://breadnbeyond.medium.com/easy-ways-to-handle-spam-comments-on-youtube-7b2ea20d8727>.
- [28] Snapchat Scam: This is what users need to look out for ... (2020, August 2). Retrieved from <https://www.leaderlive.co.uk/news/18621697.snapchat-scam-users-need-look/>.
- [29] Author, L. W. (2021, November 4). Top facebook scams of 2021 and how to avoid them. VPNoverview.com. Retrieved from <https://vpnoverview.com/privacy/social-media/facebook-scams/>.
- [30] Newman, L.H., 2021. How hackers hijacked thousands of high-profile YouTube accounts. Wired. Available at: <https://www.wired.com/story/youtube-bitcoin-scam-account-hijacking-google-phishing/>.
- [31] Author, L. W. (2021, October 20). Top Instagram scams of 2021 and how to avoid them: Vpnoverview. VPNoverview.com. Retrieved from <https://vpnoverview.com/privacy/social-media/instagram-scams/>.
- [32] Hall, G. E. (2019, August 12). Remove twitter virus (removal guide) - updated Aug 2019. Remove. Retrieved from <https://www.2-spyware.com/remove-twitter-virus.html>
- [33] Bisson, D. (2016, January 7). A guide on 5 common twitter scams. The State of Security. Retrieved from <https://www.tripwire.com/state-of-security/security-awareness/a-guide-on-5-common-twitter-scams/>.
- [34] Giordano, C. (2021, May 21). Whatsapp warning over scam that allows criminals to lock account and access messages. The Independent. Retrieved from <https://www.independent.co.uk/life-style/gadgets-and-tech/whatsapp-scam-hack-warning-news-b1851552.html>.
- [35] Palmer, D. (2017, August 24). Facebook Messenger user? watch out for fake messages rigged with malware. ZDNet. Retrieved from <https://www.zdnet.com/article/facebook-messenger-user-watch-out-for-fake-messages-rigged-with-malware/>.
- [36] Blanco, O. (2020, February 11). How to avoid Facebook Messenger Scams. Consumer Reports. Retrieved from <https://www.consumerreports.org/scams-fraud/facebook-messenger-scams-how-to-avoid/>
- [37] Wong, Q. (2019, August 14). TikTok is filled with adult-dating scams and fake accounts, report says. CNET. Retrieved from <https://www.cnet.com/news/privacy/tiktok-is-filled-with-adult-dating-scams-and-fake-accounts-report-says/>
- [38] Montalbano, E. (2021, April 22). Telegram platform abused in 'Toxiceye' malware campaigns. Threatpost English Global threatpostcom. Retrieved from <https://threatpost.com/telegram-toxiceye-malware/165543/>.
- [39] Ricle, J. (2019, October 2). What is "scam" label next to telegram username? how to report telegram scammers? Telegram Adviser. Retrieved from <https://www.telegramadviser.com/scammers-in-telegram-and-how-to-report/>.

Accelerating data compression using general purpose GPUs

Kristijan Ristovski

Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering
 Skopje, North Macedonia
 kristijan.ristovski.1@students.finki.ukim.mk

Vladimir Zdraveski

Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering
 Skopje, North Macedonia
 vladimir.zdraveski@finki.ukim.mk

Abstract—The amount of data has been exponentially rising over the years and data centers have invested heavily in research for solutions to efficiently transport and store the data. One of the fields that is crucial to solving this problem is compression of the generated data to reduce the amount of hardware needed for storage of said data. Multiple solutions have already been explored and proposed, and in this paper we experiment with GPU parallelization of compression algorithms to reduce the time required to compress data while maintaining the maximal possible compression ratio. The experiments in this paper explore the viability of using general purpose GPUs for accelerating at home data compression, with the possibility of scaling up to large data centers i.e. using GPU parallelization to reduce compression times.

Index Terms—GPU, CUDA, Parallelization, Compression

I. INTRODUCTION

Data compression has been around for a long time and is a widely researched topic for many years. It's a crucial part of data storage and transfer, especially in systems that utilize large amounts of data transmission. We divide data compression into lossy and lossless approaches. Audio and image data can tolerate some data loss and therefore can utilize lossy compression, but more sensitive data, such as scientific calculations or private user data, cannot tolerate losses during compression.

II. RELATED WORK

Previous research in the field of lossless compression proposed the use of GPU based lossless compression for specific types of data. Fang et al. found that Run-Length encoding based compression yields a significant increase in performance for compressing databases [1], while O'Neil et al. suggested the GFC algorithm that's specifically targeted towards double-precision floating point data [2].

Balevic also worked on RLE using GPUs, but only got increase in performance compared to the serial implementation on the CPU [3]. The parallel algorithm for variable-length encoding used atomic operations and the shared memory for computing codeword offsets in a compressed data stream and cached the codeword in a look-up table in the low latency memory.

Huffman encoding has also been studied by Cloud et al. and showing that GPUs cannot significantly impact performance using the Huffman algorithm because of the algorithms deficiencies [4]. It's proposed that full optimal use of the GPUs SMID cores requires the complete elimination of if statements from the GPU sub-routine. Rahmani et al. presents an algorithm that managed to get a pretty sizable performance increase by parallel implementation of the Huffman algorithm by bypassing the constraints of the maximum codeword

length and entropy [5]. With the amount of new research being conducted every day and with technology always improving, resulting data from studies can be extremely precise and can only be represented by the floating-point data type. Ratanaworabhan et al. compare a number of compression algorithms for scientific floating-point data and conclude that real-time compression and decompression is possible on modern hardware [6].

Works in the lossless family of algorithms, such as Lempel-Ziv family of lossless compression algorithms, suggest that its variations can benefit from parallelisation, as shown in Zu and Hua's work where they improve the LZSS algorithm by using their own approach called GLZSS and managed to get double speedup over the existing LZSS algorithm [7]. Concerning the efficiency of these algorithms there are works showing that efficiency is improved using multiple processors, as seen in Erdodi's paper where he implements the LZO algorithm using a CUDA enabled GPU which exploits the number of cores of the GPU and exceeds the CPUs efficiency when using the same algorithm [8].

As previously stated, multimedia like pictures, audio and video, is a data type that can tolerate some data loss. Utilising the weakness of some of the human senses and the human ability to fill in gaps given a context, lossy data can afford to be highly compressed, even at the cost of losing some of the nonessential data bits. For this type of data compression Patel et al. find that performance can be gained when using GPUs for lossy compression when algorithms are written with data level parallelism in mind [9].

There have been instances where algorithms intended for compressing lossless data have been used to compress lossy data. Funasaka et al. implement the LZW (Lempel-Ziv-Welch) compression algorithm to show that compressing TIFF images using a GPU yield much better performance compared to using the same algorithm on a CPU [10].

We also have data that can be compressed using both methodologies, and has pros and cons in each one. For example we have Franklin et al. that propose the HD-ODETLAP algorithm for lossy compression on high-dimensional arrays of data which tackle a pretty significant problem, compressing GIS (Geographic Information System) data [11]. GIS data can be highly compressed at the expense of reconstruction accuracy or be compressed using a lossless algorithm and maintain full accuracy.

III. SOLUTION ARCHITECTURE

In this paper we will be comparing the execution times of three different compression algorithms on different processing units i.e. on the CPU, where we will be running the

algorithm in sequential mode, and on the GPU, where we will be running the algorithm in parallel mode. Figure 1 presents a general overview of both of these scenarios.

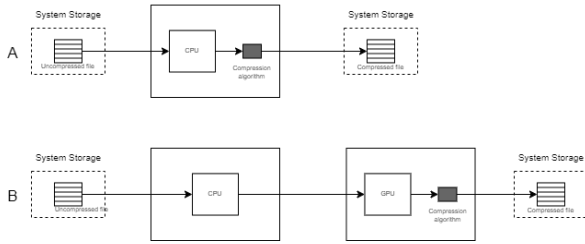


Fig. 1. Scenarios A and B

Suppose we have a large data set comprised of different types of data (audio, video, image or text data) stored in system storage (HDD or SSD). If we are running any of the compression algorithms on a file from the data set using Scenario A, the file first has to be loaded from the system storage to the main memory i.e. the RAM. After the file is put into RAM, the CPU reads it line by line and passes it through the compression algorithm which compresses the data according to the specified compression algorithm and loads the compressed file into RAM, then the resulting file is written to the system storage for the user to access.

Using Scenario B we attempt to improve upon scenario A by using the GPU as the processor for the compression algorithm to utilise to improve compression times. The process starts similarly as scenario A i.e. the desired file is loaded from system storage to RAM where the CPU can send the file using PCIe connections to the GPU. The GPU then runs the received data through the compression algorithm and generates a compressed file which is sent back to the RAM and written to the system storage.

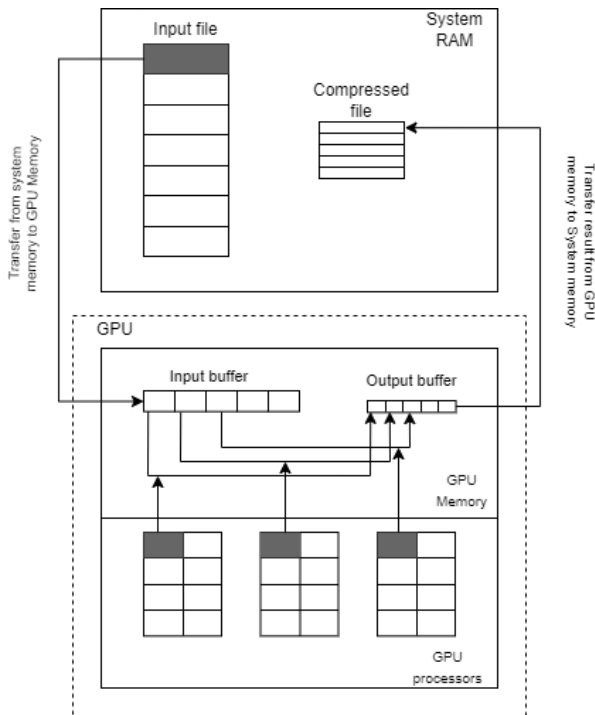


Fig. 2. GPU thread assignment

In Figure 2 we observe, in detail, the way that GPU assigns work to its multiprocessors. The desired file is loaded

into RAM and gets copied to the GPU's input buffer where the work is split into threads. The individual threads and are assigned a GPU processing core which executes the compression algorithm and sends the result to the output buffer which collects all of the individual outputs and sends them to the RAM where the compressed file is temporarily stored before it gets permanently stored in the system storage.

In both of these scenarios we evaluate the time it takes to get a compressed file from the moment the command is executed using the *Measure-command* in the Windows terminal (equivalent to *time* in UNIX based systems) and the compression ratio of each file. All of the tests will be run on the same machine with an AMD Ryzen 5 3600 CPU with 6 cores and SMT (Simultaneous Multi-Threading) enabled, an Nvidia GTX 750Ti with 2GB of memory and 640 CUDA cores, a HDD which spins at 7200RPM and 16GB of RAM.

The compression algorithms which we will be using for these benchmarks are:

- BZip2, an open-source program that uses the Burrows-Wheeler algorithm, Run-length encoding, Huffman coding and Delta encoding to compress a desired file. It's much more efficient than traditional compression algorithms but it's noticeably slower.
- LZMA aka the Lempel-Ziv-Markov chain algorithm is mainly used for lossless data compression that uses a dictionary compression algorithm whose output is encoded with a range encoder utilising a complex model to make a probability prediction for each bit. This algorithm is also noticeably slower than traditional algorithms, but can compress much larger data inputs and is also characterised with a fast decompression time.
- PPMd i.e. the Prediction by Partial Matching algorithm, is an algorithm that chooses how to compress the data further in the receiving stream by comparing it with the data it has most recently encountered. This algorithm is also lossless and is effective at compressing text files that contain natural language text but at the cost of a relatively large amount of RAM consumption because of the needs of the prediction model.

The data set on which we will be running the test consists of a large text file that contains generated Lorem-ipsum text, an image with the RAW format and a regular JPEG image.

IV. RESULTS

In this section we evaluate how quickly the files are compressed on our machine using different compression algorithms and see which compression algorithm and/or file type benefits the most from GPU acceleration.

As previously mentioned, our data set consists of a large text file with a size of 953 MB, an image with the RAW file format which has a size of 27.2 MB i.e. a large uncompressed image that contains all of the data captured by the DLSR camera and an uncompressed BMP image with a size of 116 KB. All of the tests are repeated 4 times and an average time is calculated.

Figure 3 illustrates the time needed to compress the large text file. As we can see from the graph all of the algorithms execute faster when running GPU acceleration, but the time delta is minimal in most cases. The LZMA algorithm benefits the most from GPU acceleration which reduces execution time by 27.2 seconds but is slowest out of the three by quite some margin and reduces the file size to 203 MB. The

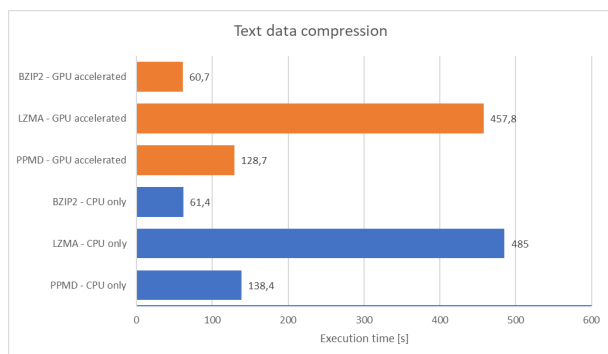


Fig. 3. Execution time of text compression

BZIP2 algorithm has the fastest execution time but also has the worst compression, only managing to reduce file size to 242 MB and the resulting file is 42MB larger than the most compressed file, i.e the resulting file from the PPMD algorithm with a resulting file size of 200 MB, but has the lowest benefit from GPU acceleration. The PPMD algorithm speeds up by 9.3 seconds when using the GPU and is the best out of the three algorithms when compressing textual data.

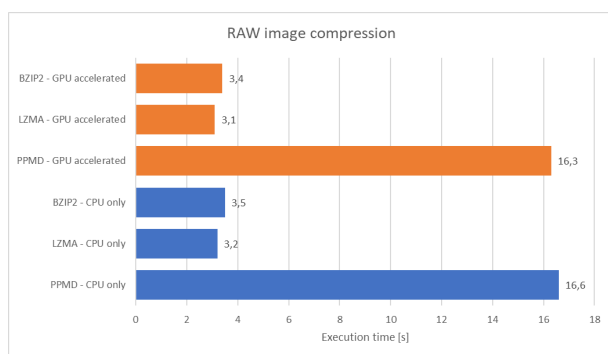


Fig. 4. Execution time of RAW image compression

In the graph in Figure 4 we observe negligible speedup when using the GPU and in the table we can see that all of the algorithms achieve the same amount of compression. This type of data i.e RAW images, is minimally compressible due to the nature of the contained data and results in all of the algorithms having the same compression result of 26.3MB. The purpose of these files is to edit the image without losing any crucial data like colour, ISO etc., so compressing the file RAW only compresses the overhead and metadata in the picture which results in minimal overall compression.

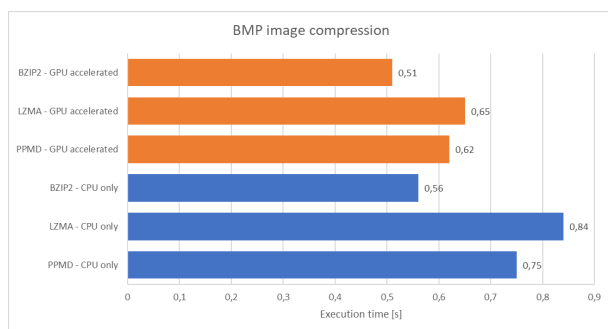


Fig. 5. Execution time of BMP image compression

Figure 5 illustrates the execution time of the compression of a BMP image. The very short execution times are a

result of the overall small file size of the image and the fact that BMP images are highly compressible. The compression algorithms all gain from the GPU acceleration but it's very negligible, but the compressed file sizes are half of the original file size. The LZMA algorithm achieves the best compression with a resulting file size of 69.2 MB, while the PPMD algorithm achieves the worst compression with a resulting file size of 77.2 MB and BZip2 algorithm generates a file size of 74.8 MB, which shows that the PPMD algorithm is better optimized for compressing textual data.

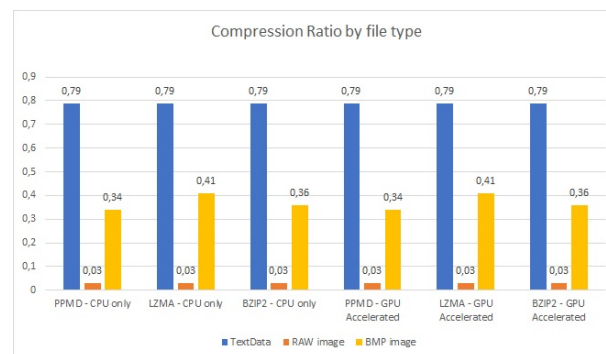


Fig. 6. Resulting compression ratios by file type

Figure 6 depicts the compression ratios of all of the files from the testing data set. We can see that the text data sample and the RAW image have a constant compression ratio that isn't affected neither by the algorithm nor whether the GPU is used or not. The only difference in the compression ratio is observed in the BMP image compression and it only differs based on the algorithm used, not the hardware used to compress the file.

V. CONCLUSION

In this paper, we compared the execution times and the compression ratios of different file types using three different compression algorithms and executing those algorithms on two different processing units, the CPU and the GPU. We observed a small speedup in execution when running the algorithms on the GPU versus running the same algorithms on the CPU, showing that some algorithms and file types benefit from running on parallel processors. This speedup is almost negligible in most user cases and doesn't make it viable in a personal working scenario, but may prove useful in professional work when compressing very large data sets, like user data bases or data resulting from a large experiment, which take a very long time to complete.

The results from our experiments confirm the results from already existing work i.e. it is possible to accelerate the execution process of compression algorithms when running them on parallel processors.

REFERENCES

- [1] W. Fang, B. He, and Q. Luo, "Database compression on graphics processors," Proceedings of the VLDB Endowment, vol. 3, no. 1-2, pp. 670–680, 2010.
- [2] M. A. O'Neil and M. Burtscher, "Floating-point data compression at 75 gb/s on a gpu," in Proceedings of the Fourth Workshop on General Purpose Processing on Graphics Processing Units, 2011, pp. 1–7.
- [3] A. Balevic, "Parallel variable-length encoding on gpgpus," in European Conference on Parallel Processing. Springer, 2009, pp. 26–35.
- [4] R. L. Cloud, M. L. Curry, H. L. Ward, A. Skjellum, and P. Bangalore, "Accelerating lossless data compression with gpus," arXiv preprint arXiv:1107.1525, 2011.

- [5] H. Rahmani, C. Topal, and C. Akinlar, "A parallel huffman coder on the cuda architecture," in 2014 IEEE Visual Communications and Image Processing Conference. IEEE, 2014, pp. 311–314.
- [6] P. Ratanaworabhan, J. Ke, and M. Burtscher, "Fast lossless compression of scientific floating-point data," in Data Compression Conference (DCC'06). IEEE, 2006, pp. 133–142.
- [7] Y. Zu and B. Hua, "Glzss: Lzss lossless data compression can be faster," in Proceedings of Workshop on General Purpose Processing Using GPUs, 2014, pp. 46–53.
- [8] L. Erdodi, "File compression with lzo algorithm using nvidia cuda architecture," in 2012 4th IEEE International Symposium on Logistics and Industrial Informatics. IEEE, 2012, pp.251–254.
- [9] P. Patel, J. Wong, M. Tatikonda, and J. Marczewski, "Jpeg compression algorithm using cuda," Department of Computer Engineering, University of Toronto, Course Project for ECE, vol. 1724, 2009.
- [10] S. Funasaka, K. Nakano, and Y. Ito, "Fast lzw compression using a gpu," in 2015 Third International Symposium on Computing and Networking (CANDAR). IEEE, 2015, pp. 303–308.
- [11] W. R. Franklin, Y. Li, T.-Y. Lau, and P. Fox, "Cuda-accelerated hd-odetlap: lossy high dimensional gridded data compression," in Modern Accelerator Technologies for Geographic Information Science. Springer, 2013, pp. 95–111.

Implementation Of The Viola Jones Algorithm For Face Detection And Tracking On Robots With Limited Computation Power

Dimitar Bezhanovski

Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering
 Skopje, North Macedonia
 dimitar.bezhanovski@students.finki.ukim.mk

Aleksandar Stankovski

Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering
 Skopje, North Macedonia
 aleksandar.stankovski@students.finki.ukim.mk

Abstract—This paper aims at presenting how to implement face detection and tracking on a low powered robot that uses hardware such as the Raspberry Pi using the Viola Jones algorithm. With face detection and tracking implemented, the robot can follow the movement of a person, and move accordingly.

I. INTRODUCTION

This paper aims to explain how the challenges of face detection and tracking with robots that have low processing power can be overcome. Face detection in image processing is a computer technology that is able to identify the presence of people's faces within digital images or videos, and it is not to be confused with face recognition, which refers to the ability to distinguish and recognize the person in the image. The main challenge that we face is the face detection part. For a mobile robot (robot with wheels or caterpillars) to be able to move towards a person and also to orient towards them, first we need a method by which we will detect the person's face. Even though the whole human figure can be detected, in this research paper, we focus on face detection because the face of a person always has the same order of characteristics. In contrast, the whole human figure can change depending of the movement or pose that the person is in, such as stretched arms or bent knees.

When it comes to face detection, we do not have a huge choice to solve this problem. A widely used approach to overcome this challenge is by applying deep learning technique, but in our case, there is a problem with that. Deep learning is too demanding in terms of computing power for devices like the Raspberry Pi. Deep learning would be an adequate approach if the embedded computer used has built-in CUDA cores. Such embedded computers can cost more than 100 USD, but this paper focuses on devices that cost less than 50 USD, as per example Raspberry Pi devices and similar variants. The purpose of this is to demonstrate a way for students to be able to implement tracking functionality in robots that are built with low cost. This research explains an implementation that is a good compromise between performance and functionality using Viola-Jones face detection algorithm. Once the face is detected we apply an algorithm to calculate the distance and location of a person in relation to the robot. The code for this paper running on the Ricardo robot can be found on the following repository <https://github.com/dimitarbez/Ricardo>

II. RELATED WORK

Many existing works use neural networks to implement the feature of user following [4]. Other works use a Bayesian

based approach for facial recognition and following using SIFT descriptors [5]. This is effective however it is a problem when implementing such functionality on a device that doesn't have a dedicated graphics processor such as the Raspberry Pi.

III. VIOLA JONES ALGORITHM

The Viola-Jones object detection algorithm [1] is an object detection algorithm that was proposed by Paul Viola and Michael Jones in 2001. Although it can be trained on various classes of objects, the main motivation of the algorithm was detection of faces.

The Viola-Jones algorithm uses a machine learning approach to detect objects in pictures. The classifier works with very simple image features. One part of the image is subtracted from another one. That alone is not very good. But if there are thousand such features, then we are well on our way to detection.

When we have a black and white image, the eye area is darker than the forehead area. When we subtract the lighter part from the darker part, we get classification. The algorithm works on this simple principle. Viola-Jones requires a full front-facing view. So in order for the face to be detected it will have to be aimed at the camera and it should not be tilted to one side. While this sounds like something that would reduce the usefulness of the algorithm, in practice these limitations are quite acceptable.

The features found by the algorithm involve a sum of pixels within the boundaries of rectangular regions. As such, they have similarities to the Haar basis functions, which have been used to detect objects in an image. But because the features used in Viola-Jones depend on more than one rectangular region, they are generally more complex.

This algorithm consists of four stages:

- 1) Haar features.
- 2) Create an integral image
- 3) AdaBoost training
- 4) Cascading

A. Haar features

The detection procedure classifies based on values from simple features. There are several motivations for using features instead of working with pixels directly. The most common reason is because with features knowledge can be encoded in an ad hoc domain. There is another critical reason and that is that a feature-based system is much faster than a pixel-based system. The simple features used are reminiscent

of Haar basis functions. The algorithm uses three types of features. The value of a two-rectangular feature is the difference between the sum of the pixels in the two regions. The regions have the same size and form and are vertically or horizontally adjacent. Feature with three rectangles calculates the sum of the two outer rectangles subtracted from the middle rectangle. Finally, a feature with four rectangles calculates the difference between diagonal pairs of rectangles.

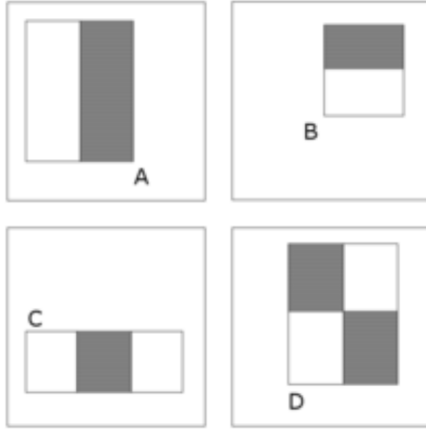


Fig. 1. Haar features [1].

B. Integral image

Rectangular features can be computed very quickly using an image that is called an integral image. The integral image at location x, y contains all the pixels above and left, inclusively.

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y')$$

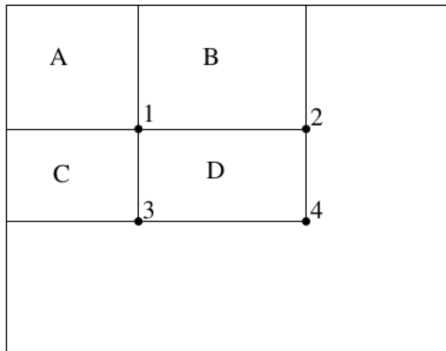


Fig. 2. Calculating the area with the integral image [1].

$ii(x, y)$ is the integral image and $i(x, y)$ is the original image. Using the following formulas:

$$\begin{aligned} s(x, y) &= s(x, y - 1) + i(x, y) \\ ii(x, y) &= ii(x - 1, y) + s(x, y) \end{aligned}$$

(where $s(x, y)$ is the cumulative sum of the row, $s(x, -1) = 0$) the integral image can be calculated with just one iteration across the pixels of the original image. Using

the integral image, any rectangular sum can be calculated using four references to an array. The difference between two rectangular sums can be calculated in eight references. Because two rectangular features above involve neighboring rectangles, they can be calculated in six references to an array, eight in case of three rectangle features and nine in four rectangle features. With this, the calculation complexity of the sum of the rectangular regions becomes of $O(1)$. Calculation of the integral image has to be done only once per new frame and its complexity is $O(n)$. This is way better than having complexity of $O(n^2)$ for the entire image.

C. Cascades with attention

This section explains the algorithm for constructing a cascade of classifiers that manages to increase detection performance but drastically reduce the computation time. A smaller and more effective classifier can be constructed that will reject many of the negative sub-windows while at the same time it will detect almost all positive instances. Simpler classifiers are used to reject most of the sub-windows, before more complex classifiers are used in order to obtain a low negative rate. The general form of the detection process is that of the degenerative decision tree, which we call cascade. A positive result from the first classifier provokes the evaluation of the second classifier which is also set to manage to reach very high rates of detection. A positive result from the second classifier causes the third classifier and so on. A negative result in any step is followed with the rejection of sub-window.

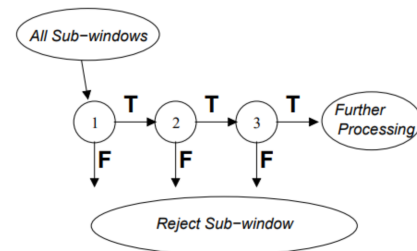


Fig. 3. Processing of sub-windows [1].

The cascade tracks are constructed by training a classifier using AdaBoost and then changing the threshold values to minimize false negatives. The default threshold values is designed to have a low error rate for data training. In general, a lower threshold leads to higher installments of false positives. For example, an excellent first-stage classifier can be constructed from two functional strong classifiers with threshold reduction to minimize false negatives. Measured against validation data, the threshold can be adjusted to detect 100% of persons with a false positive rate of 40%. The calculation of the classifier of the two features is about 60 instructions for a microprocessor. It seems hard to imagine that any simpler filter can achieve higher bounce rates. For comparison, to scan a simple template per image or a single layer of perceptron, 20 times more sub-window operations are required.

D. Training of a cascade of classifiers

The cascade training process involves two types of exchanges. In most cases, multi-feature classifiers will achieve higher detection rates and lower false positive rates. At the same time, the classifiers with more features require more time to calculate. Optimization framework can be defined in which the number of classification stages, the number of features in each phase and the threshold of each phase is traded in order to minimize the expected number of rated features. Unfortunately, finding this optimum is an incredibly difficult task.

In practice, a very simple framework is used to produce an effective classifier which is highly efficient. Each phase of the cascade reduces the false positive rate. Target is selected to minimize false positives and to maximize detection. Each stage is trained by adding characteristics until the targets are met, the stages are tested with testing the detector on the validation set.



Fig. 4. Example of frontal upright face images used for training [1].

IV. PRACTICAL IMPLEMENTATION ON THE ROBOT

The goal of the implementation is to give the robot the ability to follow users it sees. It does this by detecting their faces [2]. This can be implemented on any kind of robot that uses wheels or tank tracks to allow it to move. For this implementation the Ricardo robot is used. A 3D printable robot that uses tank tracks powered by DC motors. The robot has four DC motor that drive the two tank tracks.

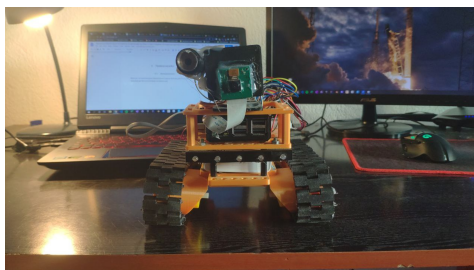


Fig. 5. Frontal view of Ricardo.

The movement of the robot to follow the user is being calculated based on the size (width) of the user's face bounding

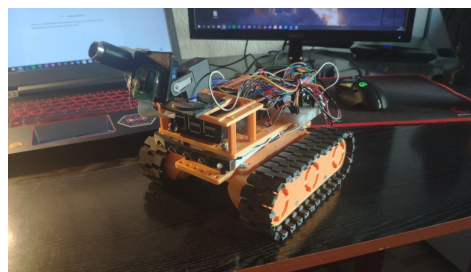


Fig. 6. Angle view of Ricardo.

box on the screen and its position accordingly [2]. The robot's movement can be divided into two categories: translational and rotational. The translational movement is being calculated based on the user's face size, meaning whether the user is close or far away from the robot, the robot should move forward or backward. There are two thresholds for calculating at which distance the robot needs to be from the user. The value we are working with is the width (size) of the users face. The higher threshold is for detecting when the user is too close. It means that the size of the user's face on the screen has passed the upper threshold. In that case the robot has to move backwards to get to a far enough distance. The lower threshold means that the size of the face has become too small on the screen. This is when the user gets too far from the robot. In this case the robot has to move towards the user to get close enough. The robots stops moving when the face size reaches a value in between the two thresholds [3].

These thresholds have to be set accordingly to the resolution of the image. For the current implementation which runs on a resolution of 400x300, the upper threshold works best at a value off 150 (pixels wide) and the lower threshold should be a value of 120 (pixels wide) for a robot placed on a table in front of you. These thresholds can be increased or decreased depending on what distance we like to keep the robot at and also keeping in mind the resolution of the image. The rotational movement is calculated based on the position of the face on the image. There are two borders that trigger the rotational movement. If the face bounding box enters the left border then the robot starts to rotate towards the left. If the face enters the right border then the robot proceeds to turn right. If the robot is in neither of the borders then just the translational movement is being calculated. If the face has entered both borders then the translational movement is being calculated only.

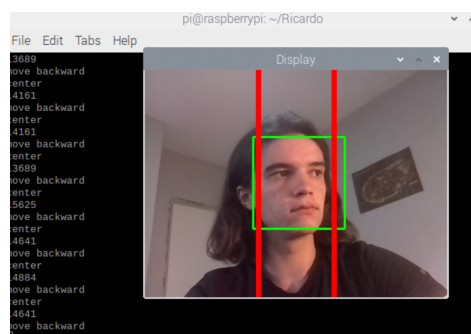


Fig. 7. Dimitar Bezhanovski.

If the face has entered both borders then the user is too

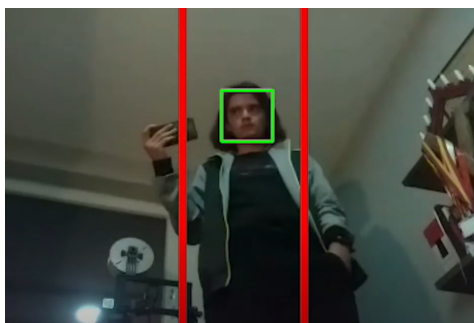


Fig. 8. Dimitar Bezhanovski.

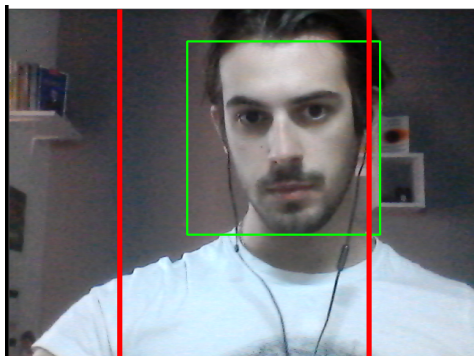


Fig. 9. Aleksandar Stankovski.

close and the robot has to back off. The rotation speed increases linearly as the user's face enters deeper into the border that triggers rotation.

V. CONCLUSION

The optimal way to implement face detection and tracking on a robot with low powered embedded computer is:

- Detect faces using the Viola-Jones algorithm
- Implement translational movement based on the person's face size in the image
- Implement rotational movement based on the person's face position in the image

This implementation is for devices that do not have a dedicated graphics processor. An argument could be made that the Nvidia Jetson Nano which has a dedicated graphics processor falls into the low powered embedded computer category and that we can run a neural network on it. But devices such as the Jetson Nano are not as available and abundant compared to devices like the Raspberry Pi. Especially due to the ongoing chip shortage, devices with graphics processors are hard to find to and are very expensive. With this Viola Jones implementation we can effectively implement a basic robot following system of readily available and cost effective hardware.

REFERENCES

- [1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 2001, pp. I-I, doi: 10.1109/CVPR.2001.990517.
- [2] M. D. Putro and K. Jo, "Real-time Face Tracking for Human-Robot Interaction," *2018 International Conference on Information and Communication Technology Robotics (ICT-ROBOT)*, 2018, pp. 1-4, doi: 10.1109/ICT-ROBOT.2018.8549902.
- [3] K. Goyal, K. Agarwal and R. Kumar, "Face detection and tracking: Using OpenCV," *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, 2017, pp. 474-478, doi: 10.1109/ICECA.2017.8203730. *Technology Robotics (ICT-ROBOT)*, 2018, pp. 1-4, doi: 10.1109/ICT-ROBOT.2018.8549902.
- [4] Hanjing Ye, Jieting Zhao, Yaling Pan, Weinan Chen, Hong Zhang, "Following Closely: A Robust Monocular Person Following System for Mobile Robot" Under review in *2022 IEEE/RSJ International Conference on Intelligent Robotics and Systems (IROS 2022)*, arXiv:2204.10540
- [5] Cruz, Claudia Sucar, Luis Morales, Eduardo. (2008). Real-Time face recognition for human-robot interaction. *Proceedings of the 8th IEEE International Conference on Automatic Face and Gesture Recognition*. 1 - 6. 10.1109/AFGR.2008.4813386.

Running Semantic Search Over Complete English Wikipedia on a Local Computer

Stojancho Tadjarski

Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
stojancho.tudjarski@students.finki.ukim.mk

Ana Madevska Bogdanova

Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
ana.madevska@finki.ukim.mk

Abstract—We implement a system that allows providing human-like answers to human-like questions extracted from a considerable amount of data in a reasonable time measured in seconds. To prove that the volume of the data used as a knowledge base where the answers to the questions are searched for, we used a complete English Wikipedia dump running on a local laptop under Windows10 OS, exposed to a software that receives questions and provides the three most relevant solutions. The entire technology stack of the implementation is the subject of this research.

The main conclusion of this research is that it is possible to implement semantic search over a vast amount of text data on a local computer with an average hardware specifications, which is of outermost importance in developing different NLP systems.

Keywords—semantic search, deep learning, transformers, BERT

I. INTRODUCTION

The semantic search is where the semantic meaning of the question is inlined with the provided search results. According to Fazainga et al.[1], there is no unique definition of the notion of semantic search on the Web. However, the most common use is the one as an improved form of search on the Web, where meaning and structure are extracted from both the user's Web search queries and different forms of Web content.

According to Berners-Lee at al.[2], The Semantic Web is a vision about an extension of the existing World Wide Web, that provides software programs with machine-interpretable metadata of the published information and data. In this way, the contextual meaning of the content of the web page would be embedded into the HTML with appropriate extensions, easily recognizable by a software. As a final result, this would expose them to an advanced search engines to be easily recognized as potential candidates for search queries where matching of the contexts of the question and the answers would play a significant role.

Therefore, it goes beyond looking for exact or similar matches of the words used in the search queries in the answers provided at the end of the search process. That leads us to the possibility of structuring questions less formally and more humanly, still recognizable by the software, which is capable of extracting answers that do not necessarily contain the exact words used in the questions, still the meaning of the answers is with (at least partially) preserved correctness from a semantic perspective, as recognized by the humans.

Unfortunately, this ideas never gain wide adoption, due to the several factors. The most significant one is the fact that is hard to design a taxonomy of terms that would cover

a wide area of topics that would be both detailed enough to cover all necessary subjects of the humans' interests, and to be acceptable for a wide population of naive users, i.e., ordinary end users who are not necessarily familiar with domain specific semantic data and just read news, and on the other side, highly specialized experts that look for very specific terms, context-wise.

The problem of the semantic search have been a subject of various research studies and applied with variable success using various approaches. Still, the breakthrough in this field is put on the scene by introducing the Transformers neural network architecture and with the BERT language model as the first helpful result of it [Attention is all you need], both applied by Google's researchers. In the following months, BERT became the fastest adopted new technology in the field, on such a scale as Google's search engine is.

The rest of the paper is organized as follows. Section 2 presents an overview on similar papers or related researches. The implemented methods, including the data set and the model selection, is described in Section 3. The results and the discussion are presented in Section 4. Section 5 presents the conclusion. The following text will provide technical details on how this technology can be applied on an average-performing home computer/laptop, still using the same technology running on Google's servers.

II. RELATED WORK

Semantic search engine for XML documents is elaborated in the work of Cohen at al. [3], and the same approach can be used for HTML content as well.

The fact that searching by keyword and do not understand polysemy and synonymy are some reasons the traditional search engine is not suitable anymore, is well documented by Kasim at al.[4]. He successfully pinpoints the need of semantic search engine that is born of traditional search engine, but with a goal to overcome this limitation.

One implementation of semantic search is SemSearch engine, described by Lei at al.[5] that pays a special attention to hiding the complexity of semantic search from end users and making it easy to use and effective.

GeneView, described by Thomas at al.[6] is built upon a comprehensively annotated version of PubMed abstracts and openly available PubMed Central full texts. This semi-structured representation of biomedical texts enables a number of features extending classical search engines. For instance, users may search for entities using unique database identifiers

or they may rank documents by the number of specific mentions they contain. Annotation is performed by a multitude of state-of-the-art text-mining tools for recognizing mentions from 10 entity classes and for identifying protein–protein interactions.

A plausible approach to the semantic search problem is presented by Nagarajan et al.[7]. He presents an idea of giving semantic to a web page so a system can understand the semantic behind the web page which automatically increases the efficiency of information search.

Dietze et al.[8] introduces GoWeb, which combines classical keyword-based Web search with text-mining and ontologies to navigate large results sets and facilitate question answering.

The need of organizing data in easily findable, accessible, interoperable and reusable way, exposed to effective searches without exact text/word matching is recognized by Sadeeq et al.[9]. In his paper, he provides a wide review of Semantic Search and Semantic Web techniques. Various types of semantic search engines are investigated and the differences between the Semantic Search and Semantic Search keywords are determined. Additionally, the benefits of using Semantic Search were highlighted.

III. IMPLEMENTATION METHODS

A. Datasets

We used a complete English Wikipedia dump, available on <https://dumps.wikimedia.org>. Various dump files are available:

- Database backup dumps, in a form of MySQL DB dumps
- Static HTML dumps
- Backup dumps of wikis which no longer exist
- Analytical data files
- Titles and descriptions only
- URLs to the pages only
- Complete articles, all the revisions included
- Complete articles, the last revision only

We used the last one, with complete articles included, but without the revisions history, so only the latest versions of the pages were used. The dump file is compressed in a format that allows parallelized decompressing, optimized for speed.

B. Data Cleanup

The data dump, decompressed, is in the format of one big XML file, 42 GB big. When XML tags are completely removed from it, still there is a lot of unusable text that acts like a noise to the search procedure, increasing the data volume, search time, and lowering the accuracy of the results and the quality of the provided data as the result.

, with many tags and metadata surrounding the leading articles that appear on the wikipedia.org site when some specific page is visited. For example, when an image is embedded in the page, the content contains the author and description of the image that is not directly connected to the page's content.

Some of the most valuable cleanup steps that did effective cleanup of the most of unwanted content are the following ones executed:

- Removing XML tags. What remains after this step is a content in Wiki syntax, that has to be cleaned further.
- Removing non-XML tags that provide information about the content in just a few words, but are not suitable for presenting to the user as a search result text.
- Removing the names of the main categories and sub-categories. Reason: again, does not provide plausible information to the user in a form of human-readable plain text.
- Removing all extra characters that act like a noise. Examples: '—', '*' and '='.

We cleaned up the XML tags metadata too short to contain any semantically valuable information. The cleanup was implemented with a Linux sh script containing 62 sed commands. This procedure took 22 hours to run on a Lenovo ThinkPad laptop with 16 GB of RAM and an Intel i7 processor, running Windows 10 as the primary OS but utilizing his Linux subsystem because of its fast 'sed' command.

C. Data Split

The texts were intended to be processed by the Hugging-Face Transformers library with the BERT (or his variation) fine-tuned with SQUAD dataset for the Q&A downstream task. One limitation of the BERT and similar models is that the input can be 512 tokens long. That means that we had to split the content of a big single-file Wikipedia dump into many small pieces of text files with content that does not exceed 512 tokens when tokenized. Otherwise, either we will produce error during the searches, or all of the tokens that exceeds 512 would be ignored.

In addition to that, we implemented text split with overlapping, meaning that the following text piece of the same article will contain the last 150 tokens of the previous text. The reason for this is the intent not to split texts in a way that possible answers to some questions would be split into two texts.

The side effect of this approach with a sliding window with padding is data redundancy. Still, since the final selection of the answer to a given question was made according to the biggest confidence score, this did not affect the user experience.

Also, a positive drawback of this padding is a fact that it is possible that a same answer to the same question might appear in two different texts. Since we are ranking the provided answers accordingly to the relevance score, and it is calculated within the provided single document only, during the question extraction phase, this approach give additional benefit of regularizing relevance score, because the bigger one will be chosen, pointing to the same answer, therefore, the chances of neglecting valuable but omitted search result due to accidental low relevance score with having contextual surrounding, will are effectively lower.

One technical detail on splitting text into chunks: the total number of files was too big to be stored in a single folder. Both Windows and Linux OSs showed problems handling folders with more than 60.000 files. We created subfolders in three levels to solve this issue, each containing no more than 1000 files. The complete text split duration: 19 hours.

D. Data Ingestion

Q&A pipeline consists of implementing a full-text search over provided texts, with the user question provided as an input. To perform fast searches over more than 19276500 files, we used Elastic Search. Its concept of inverted indexes showed its potential on a significant scale, having an average duration of the searches of less than a second. Since the texts were physically present on the disk in many small files, we used FSCrawler [<https://fscrawler.readthedocs.io/en/fscrawler-2.9/>]. Time spent on this task: 8.5 hours.

E. BERT Model Selection

We used BERT-like models available on the Huggingface models' repository, fine-tuned to Q&A downstream tasks. There were 45 models at that time, and we were doing this in April 2020. All of the models were fine-tuned using SQuAD 1.1 or 2 dataset. SQuAD stands for [Stanford Question and Answering Dataset](#). Original SQuAD 1.1 contains 107,785 question-answer pairs on 536 articles. The models trained with it showed SOTA results in finding the answers to the given questions in provided texts that should contain the answer. As a result, such models effectively answer any question in any provided text, no matter whether the text contains a valid answer to a question or not. These models were not trained to recognize an absence of a valid answer in a given text.

SQuAD 2.0 was introduced to overcome this drawback, with an additional 50.000 examples of questions and provided texts where no answer is labeled. As a result, BERT-like models trained with this dataset showed slightly lower accuracy in providing valid answers but can effectively recognize situations with no answer to a given question in a provided text.

Beyond this variation of the available modes, the following classifications were also considered:

Model size:

- small
- medium
- large

Model variation:

- original BERT
- RoBERTa
- AlBerta

There are combinations of models with all the possible sizes and variations, trained both with SQuAD 1.1 and SQuAD 2.0 datasets, and in the end we ended with experimenting with 48 models

The bigger models showed better results in finding answers but were significantly slower, and that was unacceptable having the hardware we had in our possession when we were running the experiments. Therefore, we had to find the right balance with a reasonable accuracy vs. speed tradeoff. The resulting model was a small RoBERTa model that was fine-tuned with the SQuAD 2.0 dataset chosen among 45 applicable ones available on Huggingface.

IV. RESULTS AND DISCUSSION

The semantic searches were enabled after performing all of the described data preprocessing and preparation steps. The interaction with the proposed Q&A engine is enabled by two different approaches: a slight application build using Streamlit Python library and a Flask-based website.

In Fig. 1 is presented one example of the Q&A engine in action. There might be (and usually there are) multiple answers to a certain question. Answers are ordered by the relevance score, in descending order, and the multiplicity of answers is expected when the relevance scores are close with their values.

One possibly confusing fact is that the relevance scores are calculated for the given question having only provided text. Consequently, the relevance scores are not necessarily comparable between two different texts, but the differences are not significant and reflect the human reasoning and the context of the question and the texts quite well. Normalizing texts, i.e., moving over long texts with overlapping sliding windows, as described above, makes this issue less recognizable.

V. CONCLUSION

In this paper we implemented a system that provides human-like answers to human-like questions. The knowledge base is extracted from a big amount of data in several seconds, using only a local PC/laptop hardware environment. The used data knowledge is the complete English Wikipedia dump. The total volume of data, after uncompressing and before data cleanup was at about 42 GB. The main conclusion of this research is that it is possible to implement semantic search over a vast volume of text, in a way that Google does, on a local computer with an average hardware specifications.

REFERENCES

- [1] B. Fazzinga and T. Lukasiewicz, "Semantic search on the web," *Semantic Web*, vol. 1, no. 1-2, pp. 89–96, 2010.
- [2] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 34–43, 2001.
- [3] S. Cohen, J. Mamou, Y. Kanza, and Y. Sagiv, "Xsearch: A semantic search engine for xml," in *Proceedings 2003 VLDB Conference*. Elsevier, 2003, pp. 45–56.
- [4] J. M. Kassim and M. Rahmany, "Introduction to semantic search engine," in *2009 International Conference on Electrical Engineering and Informatics*, vol. 2. IEEE, 2009, pp. 380–386.
- [5] Y. Lei, V. Uren, and E. Motta, "Semsearch: A search engine for the semantic web," in *International conference on knowledge engineering and knowledge management*. Springer, 2006, pp. 238–245.
- [6] P. Thomas, J. Starlinger, A. Vowinkel, S. Arzt, and U. Leser, "Geneview: a comprehensive semantic search engine for pubmed," *Nucleic acids research*, vol. 40, no. W1, pp. W585–W591, 2012.
- [7] G. Nagarajan and K. Thyagarajan, "A machine learning technique for semantic search engine," *Procedia engineering*, vol. 38, pp. 2164–2171, 2012.
- [8] H. Dietze and M. Schroeder, "Goweb: a semantic search engine for the life science web," *BMC bioinformatics*, vol. 10, no. 10, pp. 1–13, 2009.
- [9] M. J. Sadeeq, S. R. Zeebaree *et al.*, "Semantic search engine optimisation (sseo) for dynamic websites: A review," *International Journal of Science and Business*, vol. 5, no. 3, pp. 148–158, 2021.

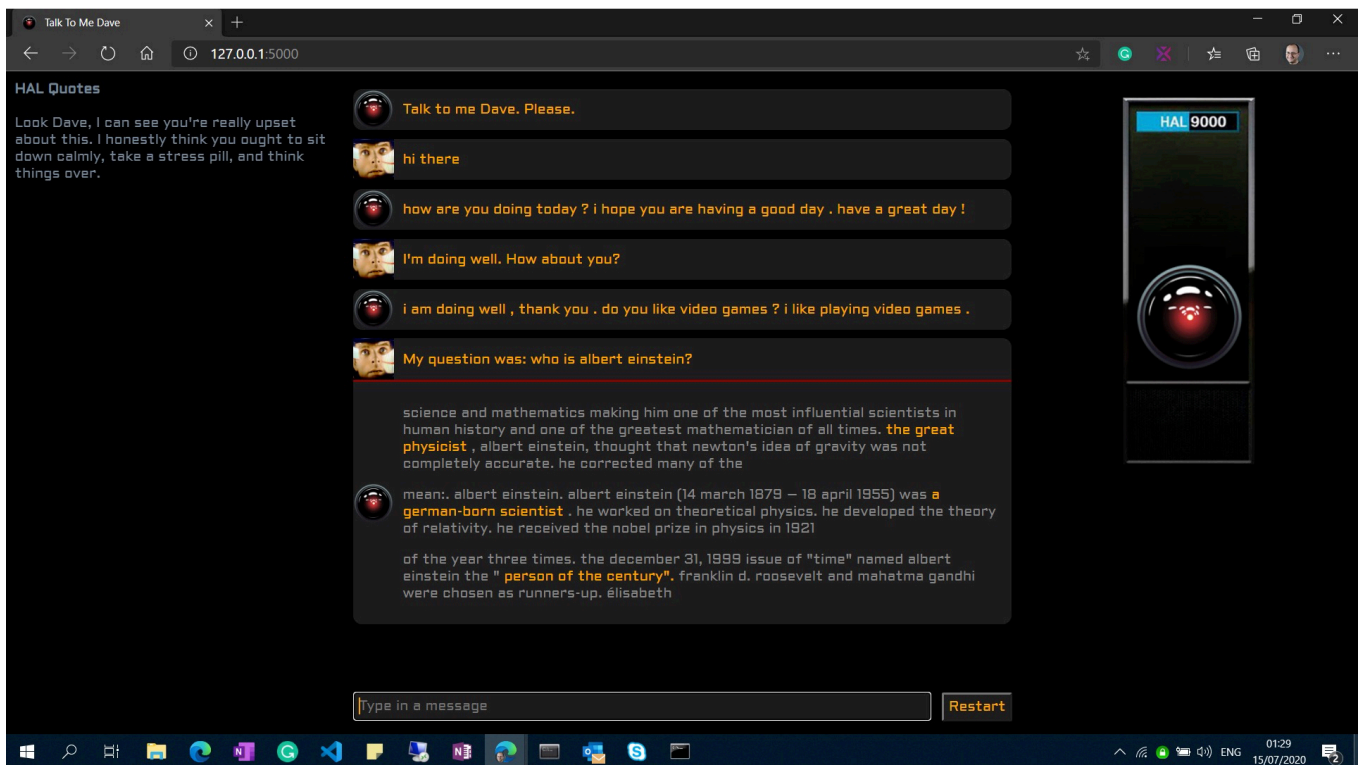


Fig. 1. Q&A in action

An Overview of Robotic Process Automation

Bojana Naumoska

*Faculty of computer science and engineering
Ss. Cyril and Methodius University in Skopje
Skopje, Macedonia
bojana.naumoska@students.finki.ukim.mk*

Georgina Mirceva

*Faculty of computer science and engineering
Ss. Cyril and Methodius University in Skopje
Skopje, Macedonia
georgina.mirceva@finki.ukim.mk*

Abstract—There are some segments of the business that require simple and everyday tasks, high-risk steps, or processes that include multiple people and steps, so implementing these working requirements can be automated, and it can leave more time for the employees to dedicate themselves to more important business goals. Process automation is simplifying complex business processes with the help of the technology. Robotic process automation (RPA) is one of the ways how a company can imitate human tasks through robots (bots). RPA is focused on the automation of repetitive actions, aiming to bring organizations an easier way of handling these actions in the most efficient way. RPA is helping companies in achieving maximum results in an effective manner. The purpose of this paper is to highlight the advantages of robotic process automation and the positive influence that this process is having over the business sector. This paper shows different tools for RPA and how they can be used to solve some challenges that businesses are facing. We describe how RPA could be used in various application areas, and also we present several case studies.

Keywords—*robotic process automation, business processes, business process automation.*

I. INTRODUCTION

The global economy today is driven by new emerging technologies that are satisfying the new needs and demands set by the customers. Companies nowadays are looking for new ways to be more efficient, competitive, save costs, and be more agile, so they use their maximum and deliver the best value. One of the emerging technologies that is placed as a solution to this challenge is robotic process automation (RPA).

RPA is replacing the repetitive tasks done by the employees, leaving them a place to work on more complicated tasks that can bring better performance to the organization. RPA software can automate some simple tasks such as online transactions, but also can simplify more complicated tasks. The main aim is to improve the overall operations of the business, lower the operations costs and bring value to the organization. There have been successfully implemented processes in different projects in banking, customer service, finance, digital services and commerce [1]. Different reports are showing a very positive impact that RPA is having in decreasing the operational costs from 30% to 50% [2]. RPA can be extended with artificial intelligence, text mining or natural language processing, and it can help in forecasting scenarios and improving the operational and business processes.

There are different techniques for automation processes that have been used so the efficiency in the companies has been increasing through the years [3]. Customer software is one of them, precisely custom-made software for performing repetitive tasks has helped companies to increase their effectiveness. Runbooks are used for sets of tasks and IT based operations, and they can be used offline as well. Batches are commands that can be run with a single click. Batch can be even scheduled so it can run at specific times.

Wrappers are watchdogs of the host client applications, and they can validate a mainframe of applications with the help of a hummingbird. There is a browser and desktop automation, so the website can be saved to a database. The data and web service integration are a process where the reading and writing into the client database is done directly.

In this paper, we give an overview of the area of RPA. Different tools for RPA are presented, as well as the capabilities that they have. We give evidence about the advantages of using RPA. Also, we give details on how RPA could be used in various areas, along with the types of problems that could be solved in these areas. Several case studies are also presented that show how RPA is used in different companies.

The rest of this paper is organized in the following manner. In section 2 we define what RPA is, while section 3 gives an overview of several existing tools for RPA. Section 4 highlights the advantages that could be earned using RPA. Section 5 shows how various areas and business processes that occur in the companies from these areas could gain benefit from RPA, and in section 6 we present several case studies. Finally, section 7 concludes the paper.

II. ROBOTIC PROCESS AUTOMATION

RPA can be defined as tools that perform operations connected with user interface in a way that people can do it. The information system does not have any changes, rather RPA is concentrated on the interactions with user interfaces and their connection with APIs to the client services. The process is mapped in the RPA tools, so software robots can follow and execute the scripts (can be referred to as "bots") [4]. The bots can mimic or emulate certain or business processes such as manipulation of data, exchanging data, answering predefined questions etc. [5].

RPA process is done with the help of artificial intelligence employees and the automation services, where the variables are set up by the developer. RPA is using software that can do human activities, interact with different applications and implement rule-based tasks.

RPA is not a traditional method, and it is not part of the information structure, rather it is on the top. Through RPA, different actions can be completed, even complicated ones such as decision-making that is usually based on data and predefined objectives. There is an opportunity to use state activity that consists of three transition sectors such as trigger, condition and action, which are giving you the opportunity to set up a trigger or a new condition for the next state. This is a huge advantage that RPA is allowing to the users. Also, the RPA platforms have a trained software robot that is using the instructions, and even a person with little programming experience can be responsible for automating a simple or complex process with proper training [3].

III. TOOLS FOR ROBOTIC PROCESS AUTOMATION

In the past years, artificial intelligence and machine learning approaches have been stimulating the process of dealing more efficiently with data, reflecting different approaches such as statistics, business cases, automatic discovery, data transformations etc. We can make extractions from information by identifying patterns, grouping, optimizing, predicting, and categorizing, but there are a lot of challenges that are arising nowadays [6]. Studies done by Deloitte [7] are suggesting that AI algorithms should be implemented in strategic processes such as reducing fraud, optimizing routine procedures, improving customer satisfaction, procedures defined by the government and other processes, which can be considered as stable and mature ones. Since there are a lot of potentials and confrontations of the automation processes, we will present some of the open-source tools that are implemented with RPA, as well as AI techniques or algorithms, and what you can expect from each of them.

UiPath [8] is a platform that consists of three modules UiPath Studio, UiPath Robot and UiPath Orchestrator, and it allows the users to create and execute programming scripts with an option for customized plug-ins for business processes. UiPath Studio can help users in transferring and managing packages, executing flows, storage of records that are linked with Microsoft's Information Services Server and SQL Server, as well as with Elasticsearch, which is an open-source search engine. There is a Kibana data visualization plugin, which is mainly oriented towards analytics. This software standouts with many advantages such as recognition, classifications, optimization, extracting information with the usage of images or character recognition [9].

Automation Anywhere [10] is a tool that is oriented towards artificial intelligence algorithms, providing information about how AI techniques can be implemented and is applicable in areas such as CRM, supply chain, HR etc. Automation Anywhere can be connected to SAP and Oracle, but there is an opportunity to connect it with other software as well. The most used process is "Digital Workers", which allows incorporation of cognitive automation and analytical data analysis tools with different functionalities in the configuration, operation and implementation process with fuzzy logic and other AI tools in order to extract information with efficiency in document validation. Automation Anywhere has an IQ Bot platform [11].

Kofax [12] is offering automation services that are oriented towards data analyses and document recognition (through Optical Character Recognition - OCR processes). This tool can extract the data from different sources such as email, files etc. into various formats, so it can optimize the tasks with ERP and allows the flows between the different applications. This application can classify and recognize the data content, as well as analyze and validate the content. Furthermore, reaching the full potential of Kofax, clients can manage, monitor and deploy different processes. Having good analytics is also an important segment with unsupervised analyses through Cognitive Document Automation module, Automation platform and information clustering [13].

IBM is combining RPA and AI so it can unlock new levels of automation processes by extending in areas where it has not been used for automation and increasing the yield of robotics. IBM is offering an automation process strategy and design that includes a virtual workspace, 30000 cognitive consultants, and setting up an automation center for implementation. There is an option for implementing control methodologies, so a systematic approach for automation programs can be on the highest level. IBM is implementing analytics techniques for extracting data and performing statistical analyses so it can discover and identify relationships, test correlations in order to enable the most effective root cause analyses. IBM is using Watson Analytics so it can discover sustainable improvements, new opportunities, and increased performance. On the other hand, IBM can make correcting actions and find the error within a robot data log, and in this way it can improve the processing by 50%. IBM is offering an option to the clients, so they can connect with Automation Anywhere and other well-ranked providers as well [2].

IV. ADVANTAGES OF ROBOTIC PROCESS AUTOMATION

There are many advantages that RPA is gaining on its side and many of them are attracting companies to implement this process [14]. One of the most important segments for companies is data security [15]. We are facing a lot of data leaks in the security systems and on the other hand, we have strict data regulations [16], so selecting a well-maintained solution is contributing to the confidence in the company [17].

RPA is processing faster since there is no human factor involved, so customer satisfaction levels in some cases are getting on a higher scale. We can expect better feedback from customers and increased trust and loyalty [18].

There are a lot of ways to handle the metrics from the company, but the RPA system can continuously track the activities since the system is automated and even the smallest improvement can be tracked. These metrics are precise, and even though the process can become more complex, the RPA system can grow in a more intelligent way.

On the other hand, it is easier to track the errors made in RPA since the process is automated and it points out and detects the errors easier. Each of the steps is followed and recorded so removing the mistake is easier.

The RPA process is helping with improved analytics and the completed time can be even faster. All data have tags and metadata, so it is easier for better insights. On the other hand, there are reduced costs, and it is said that one robot can handle the work of at least three humans and can work 24 hours [19]. This PRA process is making better performance, increased productivity, better decision-making and multitasking results that are of high quality [3].

V. APPLICATION IN DIFFERENT AREAS

The RPA processes are applicable in many areas, and they can solve different problems that have risen for the companies. Besides some commonly used RPA cases, there are a lot of business departments and industries that can have an implementation of RPA practices. When we are speaking about sales, RPA can offer test cases with creating and delivering invoices since the records are updated and the invoices can be delivered [20]. CRM system can be also updated and the channel for communication with the

customers can be constantly corrected. Also, the linkage between HR department and CRM can help in tracking the progress of the employees in sales. Automation of sales is helping the companies to spend more time on building the relations with the customers and making bigger sales rather than spending time on operational activities.

RPA techniques can also be implemented in the financial sphere [21]. Financial planning involves a lot of financial statements from different departments, so processing and merging are automated for some repetitive tasks. The bank statements can be easily automated as well, so the records can be reconciled and compared between company records and banks' spreadsheets. The reports are done for daily profit and lots are followed by companies such as trading businesses, and for them it is important to automate the preparation of the reports. The accuracy after automation has reduced the time by one third, from 60 minutes to 20 minutes that is important for companies in financial industries. Despite the fact that loan processing is complex business logic, it can be automated, especially the decision of who is getting an approval for a loan. Account closure on the other hand is also a time-consuming process and it can be automated with an electronic form that covers all required tasks, but without manual intervention.

The insurance process is a crucial segment when it comes to processing a claim [22]. The customer experience with claims can be unpleasant and time consuming. The customers are also not able to follow the process of their claims and get an update. RPA boots can deal with these issues with predefined rules, so various data and changes can be easily made. The appeals process can be also automated with 99% accuracy, which is a high percentage.

Healthcare is an important service that can be easily automated via bots for appointments, doctors' availability, financial statements, and insurance information [23]. RPA can be used for patient records, billing and claim management where data input, processing and evaluation are taking place [24]. RPA combined with other technologies like image recognition can be used for data extraction of imputing telecopying images into transactional systems [25].

Manufacturing is a rising industry, so automating different tasks can lead to reduced mistakes, increased production cycle and better customer satisfaction [26]. The inventory control in retail can help the companies to get a warning when the stock is low, and if this process is automated there won't be any chain or production loss. On the other hand, the customer service department needs logistic support for delivery. RPA automated boots can track when delivery occurs, and it can improve the response time. The linkage between the warehouse and delivery to the final destination can be automated and reported in the system in a manner of time.

VI. CASE STUDIES

We need to highlight that the successful stories of RPA implementation are even for companies that are worldwide known. The RPA techniques, designs and proper implementation have helped businesses in different areas.

Dell Technologies is a company that builds a digital future all over the world, transforming the IT segment and creating technologies that are delivered across 180 countries. The company has over 10.500 employees, which is big

human capital that requires a lot of management. Dell's HR processes have been high volume of repetitive tasks, so RPA has helped Dell in improving its cost-effectiveness [27]. The company has removed some processes such as notification or status updates, on-boarding process reminders and recruiter assignments, and it made HR employees freer in this segment. The RPA process in Dell has helped the HR department to focus more on personal interactions. Dell is one of the first companies that has implemented bots for HR. There was a pilot program that enabled the employees to have a hands-on experience while building and using the bots. So far, Dell has automated the following processes: Open requisition recruiter assignments, Offer status management, On-boarding progress reminders, Concur expense error reports, Talent Acquisition (TA) Leadership Team Reporting, Learning and Development reports and communications, Quarterly Performance and Compensation Review (PCR) letters, Employee travel invitation letters, and Job reference letters.

Another successful example of RPA is Coke Canada Bottling [28]. This company has created immediate values by automating the processes and it gained a strategic competitive advantage. The proof of concept was created in 48 hours and the results were outstanding. RPA has helped Coke Canada Bottling to deliver a value in 6 weeks, rather than waiting years for the new integration. In this case, RPA was standing as an immediate solution, which is a big advantage and step forward. Coke Canada Bottling's Intelligent Automation team was building boots, called digital workers. Digital workers were implementing RPA in uploading forms related to equipment orders for the Environmental Protection Agency, automating product sampling requests that were tracked for the accounting department. As an outcome of RPA, Coke Canada Bottling has saved 7.500 hours per month, all the solutions that are implemented have quick deployment and resolve in a shorter time.

Uber is another successful company that has implemented automation processes, so it had the chance to experience the growth and advantages of this digital transformation [29]. Uber is growing daily, so it is difficult to track the changes and coordinate global operations. Uber was in a need of automation in the operation processes in order to keep the track of them. One of the other challenges for Uber was managing the costs and hiring people. RPA has helped Uber to develop advanced solutions, to increase employees' satisfaction and processes to be inline with company's standards. Uber is having more than 100 processes, which makes Uber savings for about \$10 million per year. Uber has reduced errors that were made in their systems and has over 70% of invoices done through the automation system. Customers' satisfaction has increased due to the main portals that Uber has built, and team members can concentrate more on problem-solving.

VII. CONCLUSION

This paper gives an overview of RPA processes, their relations with artificial intelligence, and a set of tools such as Kofax, UiPath, IBM (Watson Analytics) and Automation Anywhere, which are part of RPA features. The integration of these tools is essential for better optimization, efficiency, extraction of data, presenting information to the customers, classifications and analytics. We also presented how RPA can be used for solving different problems in various areas.

Also, several case studies from particular companies are presented.

The combination of artificial intelligence, statistical methods, text mining, fuzzy logic, intelligent automation and devices are bringing another era of digital processes that are improving the RPA processes, and we can reach today another level of automation that has never been implemented before. The RPA concepts together with artificial intelligence techniques are essential for the companies that want to bring the automation processes on their agenda and increase companies' efficiency and profit.

RPA automation is beneficial since saves time, money, and resources. The technology is helping people to do more valuable work, improving the accuracy of work and making emphasizing of more complicated tasks. RPA is improving the data insights and it is increasing customers' and employees' satisfaction. The stimulating experience that RPA is creating, it is helping to the development capabilities and it is making an impact all over the business segments. RPA will contribute to even bigger efficiency in the future as we all expect that the technology will revolutionize even more.

ACKNOWLEDGMENT

This work was partially financed by the Faculty of computer science and engineering at the “Ss. Cyril and Methodius University in Skopje”, Skopje, Macedonia.

REFERENCES

- [1] M. Haenlein and A. Kaplan, “A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence,” *California Management Review*, vol. 61, no. 4, 2019.
- [2] D. Williams and I. Allen, “Using artificial intelligence to optimize the value of robotic process automation,” 2017.
- [3] A.M. Tripath, *Learning Robotic Process Automation*, Packt Publishing, 2018.
- [4] <https://www.gartner.com/en/documents/3835771>.
- [5] <https://www.gartner.com/en/information-technology/glossary/robotic-process-automation-rpa>.
- [6] S. Agostinelli, A. Marrella and M. Mecella, “Towards Intelligent Robotic Process Automation for BPMers,” 2020.
- [7] Deloitte, “Automation with intelligence Remaining the organization in the ‘Age of With’,” 2019.
- [8] UiPath, *UiPath Studio: introduction*.
- [9] UiPath, *Artificial Intelligence RPA Capabilities*.
- [10] Automation Anywhere, *Automate any ERP process with RPA*.
- [11] Automation Anywhere, *IQBot – Intelligent Document Processing*.
- [12] Kofax, *Product summary Kofax RPA*.
- [13] Kofax, *Maximize Your ERP with Integrated Accounts Payable Automation*.
- [14] T. Taulli, *The Robotic Process Automation Handbook: A Guide to Implementing RPA Systems*, Apress. 1st ed., 2020.
- [15] D. Willson and H. Dalziel, *Cyber Security Awareness for CEOs and Management*, Elsevier Inc., 2016.
- [16] J. A. T. Fairfield, *Runaway Technology: Can Law Keep Up?*, Cambridge University Press, 2021.
- [17] A. E. Waldman, *Industry Unbound: The Inside Story of Privacy, Data, and Corporate Power*, Cambridge University Press, 2021.
- [18] <https://customerthink.com/the-role-of-intelligent-automation-in-improving-customer-experience/>.
- [19] <https://botpath.com/how-rpa-is-changing-the-way-we-work>.
- [20] <https://www.helpsystems.com/resources/videos/why-your-crm-system-needs-rpa>.
- [21] <https://cdn2.assets-servd.host/lively-jackal/production/uploads/resources/case-studies/blue-prism-coop-bank-manual-efficiency-case-study.pdf>.
- [22] <https://aithority.com/guest-authors/top-11-use-cases-of-rpa-in-the-insurance-industry/>.
- [23] <https://www.uipath.com/resources/covid-automations/covid19-health-tracking-alerting-automation-for-state-government>.
- [24] <https://emergeflow.com/RPA-in-healthcare.html>.
- [25] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181/>.
- [26] <https://www.birlasoft.com/industries/manufacturing/case-studies/automotive-leader-implements-rpa-solution-to-de-risk-its-inventory-management>.
- [27] <https://www.automationanywhere.com/resources/customer-stories/dell-uses-rpa-for-hr-processes>.
- [28] <https://www.uipath.com/resources/automation-case-studies/coke-canada-bottling>.
- [29] <https://www.uipath.com/resources/automation-case-studies/uber-maintains-global-infrastructure-built-on-rpa>.

Algorithms for Process Mining – An Overview

Sandra Nikoloska

*Faculty of computer science and
engineering*

*Ss. Cyril and Methodius University in
Skopje*

Skopje, Macedonia

sandra.nikoloska@students.finki.ukim.mk

Biljana Sazdova

*Faculty of computer science and
engineering*

*Ss. Cyril and Methodius University in
Skopje*

Skopje, Macedonia

biljana.sazdova@students.finki.ukim.mk

Georgina Mirceva

*Faculty of computer science and
engineering*

*Ss. Cyril and Methodius University in
Skopje*

Skopje, Macedonia

georgina.mirceva@finki.ukim.mk

Abstract—Process mining is a research area that provides an opportunity to utilize the gathered data about business processes that occur in organizations in order to provide improved business process models. That could be made by using process mining algorithms that analyze the event logs data from the business processes. Different research groups have proposed various algorithms for process mining by employing different approaches. This paper gives an introduction in the area of process mining, as well as an overview of the algorithms used within process mining, by describing how they work and how they are applied.

Keywords— *process mining, business process, business process intelligence.*

I. INTRODUCTION

As business processes are automated with fast speed, the organizations have many data that are gathered, which could be utilized in order to improve the way how the organizations perform their work. The data from the business processes are in a form of event logs, so they present how the process executes in time for each instance (for example: a particular purchase order in the process for handling purchase orders). Since there is a huge amount and volume of data, it is not appropriate to analyze them manually. Therefore, there is a huge need for development of algorithms that could discover knowledge from these data that could be used to improve the business processes and the way how the work is done.

Process mining starts from the moment of data collection, their selection, analysis, and finally documentation. It is based on facts and is completely objective, and also it is made in automated manner thus we can save time and effort. In this way, there is no need for traditional methods for process discovery such as workshops, interviews and filtering of manual documentation, but by identifying the problems and finding solutions how they can be resolved based on the results obtained from the analyses that are made objectively and in real-time. Besides visualization of how the actual process flows, it also enables various analyses so that organizations can understand the attributes that influence the business processes and discover which activities they could improve. Process improvement, time reduction, cost reduction, increased transparency, improved customer experience, error reduction, better compliance and performance management are the main benefits of using process mining, which is a key component of business process management. It is necessary for the stakeholders to understand what their critical points are where problems arise and to rearrange the business processes for solving them [1].

There are several types of analyses that can be made in process mining, depending on the stage of the business process [2]. The most important are process discovery, compliance checking and model improvement. If these

analyses are used together, they will help organizations maintain a high level of standardization, predictability, and continuous improvement. Organizations can use process mining to optimize a single process, individual department, or even the entire organization.

In this paper we give a short introduction into process mining, and we present the most important algorithms that could be used in order to discover business process models. The rest of this paper is organized in the following way. In section 2, we give an introduction into process mining algorithms and the types of problems that could be solved using them. Section 3 presents the Alpha Miner algorithm. Then, in section 4, the Heuristic Miner algorithm is described. The Genetic Miner algorithm is presented in section 5. Finally, section 6 concludes the paper.

II. PROCESS MINING ALGORITHMS

When we mention the word business, the things we initially think of are processes that give rise to problems, which we need to solve in the simplest way, and that in business is achieved with steps that will help us achieve the desired success. For that purpose, particular algorithm could be created, which shows how its steps (activities) can be defined as a series of rules or a sequence of well-defined instructions that are used to solve a particular problem. Within the business world, over time, various changes occur that can lead to slowing down the course of activities, the occurrence of unwanted inefficiencies, bottlenecks, increased costs, and so on. These changes are tracked and detected through process mining. Process mining can reveal better models and can help to design and redesign the processes within an organization. For example, if an organization wants to reduce its total customer interactions from 5% to 3%, with the help of process mining it will be revealed which parts can be optimized in order to achieve the desired goal [3].

Every organization wants to work efficiently and effectively, to produce more in less time, but to maintain the desired quality. To achieve this, it is necessary to provide well-defined business processes, which means discovering the dependencies between activities (for each activity to discover from which activity it is dependent on), in what order they can be performed, as well as the possibility of parallelization of certain activities that are part of the business process. For the success of each innovation in the organizations, it is important in each change to review the processes, and if necessary to redesign them with the aim of maximizing the effect of introduced change [4].

The main function of the algorithms used in process mining is to provide information that is part of the current operation that plays a key role in improving existing business processes, implementing a new operating model, audit, and

analysis [5]. The key to algorithms in this regard is to provide improvement.

The algorithms used in the process discovery help analysts to better understand the business processes and the problems that are part of them through the process models that result from the input data (event log) of the system [6].

In the next sections, we present some of the most important algorithms used in process mining, i.e. Alpha Miner, Heuristic Miner and Genetic Miner.

III. ALPHA MINER

Alpha Miner is the first algorithm to bridge the gap between event logs or monitored data and the discovery of a specific process model [7]. It is an algorithm that first appears in the process of process research (play in technique). This algorithm concisely determines the purpose of the research, adequately addresses the competition, and provides an overview (model) of the relationship between the activities that are an integral part of the process that is being examined. Each algorithm has a certain input that it processes and based on which in the end it produces an output or result. As input to the Alpha Miner algorithm, an event log is given, the algorithm analyzes the input, so based on the relations between the activities results in the construction of a graph or Petri network. The input data, i.e. the event log that is given to the Alpha Miner algorithm, consist of information about a specific activity or the whole process [8].

Here are some important facts about the Alpha Miner algorithm [8]:

- Uses a basic approach to process research.
- There are a number of limitations.
- Reveals key elements of the process.
- It is very important that the algorithm is used correctly using concrete examples.

L is a log of events resulting from the set of activities. T denotes trace where the activities are in a particular order. $\alpha(L)$ is the graph that represent the model of the business process, which is obtained with the following steps [9]:

- (1) $T_L = \{t \in T \mid \exists \sigma \in L \ t \in \sigma\}$
- (2) $T_I = \{t \in T \mid \exists \sigma \in L \ t = \text{first}(\sigma)\}$
- (3) $T_O = \{t \in T \mid \exists \sigma \in L \ t = \text{last}(\sigma)\}$
- (4) $X_L = \{(A, B) \mid A \subseteq T_L \wedge A = \emptyset \wedge B \subseteq T_L \wedge B = \emptyset \wedge \forall a \in A \forall b \in B \ a \rightarrow_L b \wedge \forall a_1, a_2 \in A \ a_1 \#_L a_2 \wedge \forall b_1, b_2 \in B \ b_1 \#_L b_2\}$
- (5) $Y_L = \{(A, B) \in X_L \mid \forall (A', B') \in X_L, A \subseteq A' \wedge B \subseteq B' \Rightarrow (A, B) = (A', B')\}$
- (6) $P_L = \{p_{(A, B)} \mid (A, B) \in Y_L\} \cup \{i_L, o_L\}$
- (7) $F_L = \{(a, p_{(A, B)}) \mid (A, B) \in Y_L \wedge a \in A\} \cup \{(p_{(A, B)}, b) \mid (A, B) \in Y_L \wedge b \in B\} \cup \{(i_L, t) \mid t \in T_I\} \cup \{(t, o_L) \mid t \in T_O\}$
- (8) $\alpha(L) = (P_L, T_L, F_L)$.

- **Step 1:** The first step is to determine all the events, i.e. to determine all the activities that are part of the event log (T_L).
- **Step 2:** Defining - discovering all the start activities (T_I), i.e. all the activities that appear first within the events.
- **Step 3:** Defining - discovering a set of end activities (T_O), i.e. all activities that appear at the end or are the last within the events.
- **Step 4 and Step 5:** Define the basics of these algorithm. The main challenge in these two steps is to discover the set of activities and their connection, i.e. to discover their interrelationships. Within these steps, two sets A and B are created, set A represents input, while B represents output transitions. All activities that are part of the set A should be conditionally dependent on those that are part of set C. That is, $(a, b) \in A \times B: a \rightarrow_L^b$, the elements that are part of set A should not depend on each other, the same applies to those that are part of set C. In the 5th step, non-maximal pairs are removed, i.e. this step is also known as creating a relational matrix called Footprint matrix.
- **Step 6:** If the analysis of the relation matrix is done, it will be seen that for each element, i.e. pair there is a certain relation ($\rightarrow, \#, \parallel$). Basically, the basic goal of all these steps and globally the Alpha Miner algorithm is to determine the position and relation of the overall process.
- **Steps 7 and 8:** In these steps, the arcs are created - the relations between the activities, i.e. all the starting activities with the final transitions as an exit points.

The Alpha Miner algorithm can detect four types of relations [10], i.e. direct succession, causality, parallel (temporal independence), choice (no direct relation). These relations are described below.

• Direct Succession

Initial, basic and direct relation, on which the setting and execution of the other relations depend. This relation is denoted by the sign greater ($>$), which essentially means - "directly follows". $(a > b)$ means that a is directly followed by b (the performance of activity b is always preceded by the performance of activity a). For example: from the event log $[(a, b, c, d), (e, d, a, b), (a, e)]$, we can see that a and b are always in the same order, it is initially a, then b follows.

• Causality

This relation represents the temporal dependence between the activities. The relation is represented by the arrow sign ($\rightarrow$). More specifically, if we take the activities a and b, $(a \rightarrow b)$ – in translation "b conditionally depends on a", i.e. this relation is valid if and only if $(a > b)$ - "b directly follows a" and if $(a > b)$ - "a never follows b". Basically, b depends on a.

• Parallel – temporal independence

Relation when certain activities are performed in parallel. This relation is denoted by two parallel vertical lines ($\parallel$), which essentially denote parallel execution. If there are two activities a and b and $(a > b)$ - "b directly follows a" and $(b >$

a) - "a directly follows b", it can be concluded that these two activities are parallel, i.e. performed both activities in any order. For example, from the event log [(a, b, c, d), (e, d, b, a), (a, e)] we can see that a and b are activities that are directly monitored with each other, that means that they are temporally independent and can be performed in parallel [11].

- **Choice – No direct relation**

Independence, i.e. when a certain activity is independent. This relation is represented by the hash sign (#). If we take the activities a and b, and neither b directly follows a neither a directly follows b, then the activities are completely independent. There is no direct relation between a and b (a # b), if neither $a > b$ nor $b > a$. For example, from the event log [(a, d, c, b), (a, c, d, b), (a, e)] we can see that the activities a and b are independent [11].

IV. HEURISTIC MINER

The Heuristic Miner algorithm uses a similar representation as causal nets. It is important for this algorithm to note that event frequencies and sequences are taken into account when creating a process model. The basic idea of algorithm development is to minimize paths, i.e. those that are very rare to be completely omitted [12]. The Heuristic Miner algorithm is essentially an improvement of the Alpha Miner algorithm, the difference is that the frequencies of the activities are taken into account, and as a result it gives a model with complex dependencies in which the rare ones are left out. This algorithm is used in the research part of process discovery, it is very interesting and tempting to use due to its wide range of applications [4]. The main goal of the Heuristic Miner algorithm is to detect the causal net from the input event log [5].

The main improvements of Heuristic Miner over Alpha Miner are [3]:

- Frequencies are taken into account, which allows filtering of rare cases.
- Detection of short loops.
- Detection of skipping activities.

Heuristic Mining Algorithm uses dependency graph and causal matrix [13].

Dependency graph is a graph that shows the interdependencies of different activities and events. To build this model, first, you need a matrix in which the dependencies are displayed, and then it is necessary to determine the loops of length one and the loops of length two.

When analyzing event logs, it is difficult to determine whether certain activities or processes are parallel or sequential. A Heuristic Miner is an algorithm that builds a **causal matrix** in order to build an accurate and realistic process model.

There are two types of raw activities, AND and XOR. AND represents parallel activity, while XOR represents sequential activity.

When creating a graph dependency, thresholds could be set for the events and activities that would be modeled.

1. *Threshold* – is the minimum dependence between activities.
2. *Positive observation threshold* – is the minimum value of the frequencies between activities.
3. *Relative to the best threshold* – is the minimal difference between the value of the activity dependency and the maximum value (dependency value).
4. *Length-one threshold* – is the minimum dependency value of the same event.
5. *Length-two threshold* – is the minimum value of the loop dependency.

V. GENETIC MINER

Alpha Miner and Heuristic Miner create process models in a direct and determining way. On the other side, evolutionary approaches use iterative procedures to create and mimic the process of natural evolution. These types of approaches are not deterministic and depend on randomization in order to find new alternatives [14].

Same as the well-known genetic algorithms, Genetic Miner consists of four steps [14]: initialization, selection, reproduction, and termination. These steps are described below.

Initialization: Within the first step, the initial population is created. Process models are created randomly using the names of the activities that occur in the input event log. Process models that are part of the initial population do not have to initially correspond to the event log, i.e. the names of the activities are the same, but the behavior of the initial models will not be consistent with the input event log.

Selection: In this step, the "fitness" of each individual (model in this case) is calculated. The fitness function represents the quality of the model based on the event log. It is important to note that there are different ways to calculate the quality of a particular model. A simple criterion for determining the quality of a given model is to consider the proportion of traces in the event log that can be performed or repeated by the model. But this feature is not good because it is very likely that none of the models will be able to play any of the directions (paths) specified in the log. Also, if this type of criterion is used over universal models, then the fitness function will be too high. It is therefore necessary to use a more refined fitness function that rewards the partial correctness of the model and takes into account all four specified criteria. The best models of this generation, which are the models with highest fitness values, are passed on to the next phase, which is called elitism. In essence, only the best models go ahead and "survive", while the individuals (models) that do not have high fitness value are not retained, i.e. are considered as "dead" individuals.

Reproduction: As the name of the reproduction step itself states, it refers to the creation of new generation from the selected "parent" process models. As part of this step, two genetic operators are used, i.e.: cross-over and mutation. With cross-over two parent models are taken and used to create new ones which end up in a set of models - children. Children inherit from their parents, but they change to some extent as a result of mutations.

Termination: Reproduction and selection are repeated with the aspiration that each newly created generation will be

better than the previous one. The actual process of evolution ends at the moment when the created model is considered as a satisfactory solution. However, due to the event log, it will likely take a long time to determine the desired model, but it is also possible that the desired fitness value will not be found, so other criteria can be added and the reproduction for up to predefined number of generations can be made, so at the termination, the model with the best fitness value is returned.

VI. CONCLUSION

In this paper we gave a short introduction into process mining, as well as the benefits that it could provide to the organizations in order to rearrange how the work is performed by improving their business processes. We presented the most important algorithms that could be used to mine business process data that are gathered as event logs.

As described, the Alpha Miner algorithm discovers the temporal relations between the activities. Heuristic Miner extends this by considering the frequencies of the activities. This way, Alpha Miner and Heuristic Miner discover the process model in deterministic way. On the other side, the Genetic Miner uses randomization to randomly select initial population, as well as randomization to make reproduction (cross-over and mutation), thus it could give as an output different models in different trials.

ACKNOWLEDGMENT

This work was partially financed by the Faculty of computer science and engineering at the “Ss. Cyril and Methodius University in Skopje”, Skopje, Macedonia.

REFERENCES

- [1] M.Eisner, “What is Process Mining and Why Your Organization Needs It,” 2020.
- [2] “Why Use Process Mining,” 2021, IBM Cloud Education.
- [3] J.C.A.M. Buijs, Heuristics Miner, Technical Universiteit Eindhoven.
- [4] A.P Kurniati, G.P. Kusuma and G.A.A. Wisudawati, “Implementing Heuristic Miner for Different Types of Event Logs,” *International Journal of Applied Engineering Research*, vol. 11, no. 8, pp. 5523-5529, 2016.
- [5] W.V.D Aalst, *Process Mining: Data Science in Action*, 2nd ed., Springer Berlin, Heidelberg, 2016, pp. 205.
- [6] W.V.D Aalst, *Process Mining: Data Science in Action*, 2nd ed., Springer Berlin, Heidelberg, 2016, pp. 125.
- [7] H. Simsek, “What are the Top 5 Process Mining Algorithms in 2022?,” 2021.
- [8] W.V.D. Aalst, “Alpha Algorithm: A Process Discovery Algorithm”.
- [9] W.V.D. Aalst, *Process Mining: Data Science in Action*, 2nd ed., Springer Berlin, Heidelberg, 2016, pp. 171-172.
- [10] A. Rozinat “ProM Tips – Which Mining Algorithm Should You Use?,” 2010.
- [11] S. Conquest, “Alpha Algorithm (Process Discovery Method),” 2018.
- [12] W.V.D. Aalst, *Process Mining: Data Science in Action*, 2nd ed., Springer Berlin, Heidelberg, 2016, pp. 201.
- [13] A.J.M.M. Weijters, W.M.P van der Aalst, and A.K. Alves de Medeiros, *Process Mining with The HeuristicsMiner Algorithm*, Technische Universiteit Eindhoven, 2013.
- [14] W.V.D. Aalst, *Process Mining: Data Science in Action*, 2nd ed., Springer Berlin, Heidelberg, 2016, pp. 207-210.

