



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



2020

Proceedings of the 17th International Conference for Informatics and Information Technology

Held Online
8-9 May, 2020

Editors:
Eftim Zdravevski
Petre Lameski

ISBN 978-608-4699-10-1

Conference for Informatics and Information Technology 2020

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:

Eftim Zdravevski

Petre Lameski

Technical editing: Filip Markoski and Dushica Jankovikj

Cover page: Vangel Ajanovski

Total print run: 150

Printed in Skopje, N. Macedonia, 2020

ISBN: 978-608-4699-10-1

CIP - каталогизација на публикација

Народна и универзитетска библиотека „Св.Климент Охридски“, Скопје

004.7:621.39(062)

004(062)

PROCEEDINGS of the 17th Conference for Informatics and Information Technology (16; 2020; Mavrovo) Proceedings of the 17th Conference for Informatics and Information Technology: CIIT 2020, May, 8-9 / editors Eftim Zdravevski and Petre Lameski. - Skopje : Faculty of Computer Science and Engineering, 2020. - 164 стр. : граф. прикази ; 30 см

Библиографија кон трудовите

ISBN 978-608-4699-10-1

1. Zdravevski, Eftim [уредник] 2. Lameski, Petre [уредник]

Preface

This volume contains the papers presented at the 17th International Conference for Informatics and Information Technology (CIIT 20120) held on May 7-8, 2020 online. 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 whole world came to a stop due to the COVID-19 outbreak, our conference was not spared either. Anyone who had attended any of its previous editions knows what they are missing. Anyhow, the following editions of the conference will be back with plenty of exciting and in-person content to complement the scientific program. As unfortunate as this situation is, this opened an opportunity to provide access to the conference to a much wider audience in this virtual format. Namely, the conference now is available to all co-authors, and not only the presenters. Likewise, we have other students and staff of the Faculty of Computer Science and Engineering, as well as attendees from other companies and institutions. Additionally, we made the participation at the conference available to authors from the far corners of the world (if you consider Macedonia the center of it). For the first time, we have two high-quality contributions from Mexico and India. We also have other amazing submissions from the country and the region.

In the seventeenth edition, the aim of the CIIT conference remained to provide an opportunity for young researchers to present their work to a wider research community, but also facilitate multidisciplinary and regional collaboration. Despite the participation of scientists from the country, a substantial number of participants from abroad attended the conference as well. Building on the success of the past sixteen conferences, this year the conference attracted a large number of submissions resulting in presentations of 23 full papers and 17 short papers, which were presented in seven sessions. The 17th Conference on Informatics and Information Technologies was organized as online conference for the first time in its history. We received 48 paper submissions of which 23 accepted as full papers and 17 short papers. With this we had 48% acceptance rate on full papers. During the conference we had 39 presentations in 7 sessions and 2 keynote lectures attended by up to 63 attendees which is among the highest attendance number in the history of the conference.

This year, we performed a rigorous grading that considered several aspects of each paper: technical quality, scientific contribution and presentation. The session chairs and conference chairs graded each paper with regards to those qualities, and we also considered the reviewer grades and comments. Finally, three best student papers were awarded. The online format of the conference allowed the participants to attend 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. We had the pleasure to host two invited speakers.

Our first keynote speaker was Dr Andrzej Janusz, an Assistant Professor at the University of Warsaw and an R&D manager at QED Poland. He has a PhD in Computer science and Data mining and a Master's degree in mathematics. He's a practitioner and researcher, and has lead and participated in a verity of data science projects that resulted in many industrial applications and numerous impactful scientific articles. The topic that he talked about is a significant one from a practical and scientific point of view - "Solving Real-life Problems by Data Mining Competitions".

Our other keynote speaker was Dr. Hazim Kemal Ekenel, a Full Professor at the Department of Computer Engineering in Istanbul Technical University in Turkey. Dr. Ekenel has a vast experience in Computer Vision especially in the fields of face recognition and facial image

processing and analysis. He has founded the Facial Image Processing and Analysis group at the Department of Computer Science in Karlsruhe Institute of Technology. He was the task leader for face recognition in several large-scale European projects. His face analysis technology has been used by several research labs and companies and has taken part in several demo and press events. For his work, he has received many rewards. Since January 2011, he has been coordinating the Benchmarking Facial Image Analysis Technologies (BeFIT) initiative. Dr. Ekenel presented an overview of facial image processing and analysis research activities, as well as his group's recent work on deep learning @ SiMiT.

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.

September, 2019
Skopje

Eftim Zdravevski
Petre Lameski

Organization

Conference Chairs

Eftim Zdravevski	Assistant Professor - Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Petre Lameski	Assistant Professor - Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia

Organizing Committee

Ilinka Ivanovska	Teaching Assistant - Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Frosina Stojanovska	Teaching Assistant - Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Martina Toshevska	Teaching Assistant - Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Jovan Kalajdzieski	Teaching Assistant - Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Stefan Andonov	Teaching Assistant - Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia

Program Committee

Ackovska Nevena	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Ajanovski Vangel	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Angelova Mihaela	INSERM, France
Antovski Ljupcho	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Armenski Goce	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Baicheva Tsonka	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria

Bakeva Verica	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Balaz Antun	Institute of Physics, University of Belgrade, Serbia
Basnarkov Lasko	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Bogdanova Galina	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Borissov Yuri	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Boinovski Adrijan	School of Computer Science and Information Technology, University American College Skopje, N. Macedonia
Chorbev Ivan	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Corizzo Roberto	American University, Washington DC, USA
Davcev Danco	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Delchev Konstantin	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Delibai Boris	Faculty of Organizational Sciences, University of Belgrade, Serbia
Dimitrievska Ristovska Vesna	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Dimitrova Vesna	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Dimitrovski Ivica	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Donchev Ivaylo	St. Cyril and St. Methodius University of Veliko Turnovo, Bulgaria
Eftimov Tome	Joef Stefan Institute, Slovenia
Ekenel Hazim Kemal	Department of Computer Engineering in Istanbul Technical University, Turkey
Gievska Sonja	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Gjoreski Hristijan	Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Gjorgjevikj Dejan	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Gligoroski Danilo	Norwegian University of Science and Technology, Norway
Gramatikov Sasho	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Gusev Marjan	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Haller Stephan	Bern University of Applied Sciences, Switzerland
Ilievska Natasha	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Jakimovski Boro	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia

Janeska-Sarkanjac Smilka	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Janev Valentina	Mihajlo Pupin Institute, Serbia
Janusz Andrzej	University of Warsaw and QED, Poland
Jovanov Mile	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Jovanovik Milos	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Stoyan Kapralov	Technical University of Gabrovo, Bulgaria
Kalajdziski Slobodan	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Kitanovski Ivan	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Kon-Popovska Margita	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Kostoska Magdalena	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Koteska Bojana	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Kulakov Andrea	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Lameski Petre	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Loshkovska Suzana	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Madevska Bogdanova Ana	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Madjarov Gjorgji	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Marinkovic Bojan	Mathematical Institute of the Serbian Academy of Sciences and Arts, Serbia
Markovski Smile	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Pirez Ivan Miguel	Universidade da Beira Interior and Polytechnic Institute of Viseu, Portugal
Mihajloska Hristina	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Mihova Marija	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Mileva Aleksandra	Faculty of Computer Science, Goce Delchev University in Shtip, N. Macedonia
Mirceva Georgina	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Mirchev Miroslav	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Mishkovski Igor	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia

Naumoski Andreja	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Pachovski Veno	School of Computer Science and Information Technology, University American College Skopje, N. Macedonia
Pombo Nuno	Universidade da Beira Interior, Portugal
Papachristodoulou Louiza	Radboud University, Netherlands
Paprzycki Marcin	Systems Research Institute, Polish Academy of Sciences, Poland
Pepik Bojan	Max Planck Institute for Informatics, Germany
Popeska Zaneta	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Popovska-Mitrovikj Aleksandra	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Prckovska Vesna	QMENTA, Spain
Ribarski Panche	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Ristanoski Goce	Data61, Commonwealth Scientific and Industrial Research Organisation, Australia
Ristov Sasko	University of Innsbruck, Austria
Samardjiska Simona	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Sedigh-Sarvestani Sahra	Missouri University of Science and Technology, USA
Sherif Mohamed Ahmed	Data Science Group, Paderborn University, Germany
Shtrakov Slavcho	South West University, Bulgaria
Shurbevski Aleksandar	Kyoto University, Japan
Simjanoska Monika	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Slavkovik Marija	University of Bergen, Norway
Spasov Dejan	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Stoimenova Eugenia	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Stojanov Riste	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Stojkoska Biljana	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Stojkovikj Natasha	Faculty of Computer Science, Goce Delchev University in Shtip, N. Macedonia
Tojtovska Biljana	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Trajanov Dimitar	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Trajanovski Stojan	University of Amsterdam, Netherlands
Trajkovik Vladimir	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Trivodaliev Kire	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia

Trojacanec Katarina	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Varbanov Zlatko	St. Cyril and St. Methodius University of Veliko Turnovo, Bulgaria
Zdraveski Vladimir	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Zdravevski Eftim	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Zdravkova Katerina	Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, N. Macedonia
Zhelezova Stela	Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria

Table of Contents

Full Papers

Precision Apiculture – IoT System for remote monitoring of honeybee colonies	1
<i>Riste Poposki and Dejan Gjorgjevikj</i>	
Blockchain-based model for authentication, authorization, and immutability of healthcare data in the referrals process	7
<i>Goce Gavrilov, Orce Simov and Vladimir Trajkovik</i>	
Mapping of Auto Salons in Skopje	12
<i>Sasho Nikudinoski, Andreja Naumoski and Elena M. Jovanovska</i>	
Recent Advances in SQL Query Generation: A Survey	16
<i>Jovan Kalajdjieski, Martina Toshevska and Frosina Stojanovska</i>	
How Simple Predictive Analysis of Health Care Claims Data can Detect Fraud, Waste and Abuse Threats in Health Care Insurance - The Case Study of United Arab Emirates	22
<i>Kristijan Jankoski, Kiril Milev and Gjorgji Madjarov</i>	
Advanced analytics of Big Data using Power BI: Credit Registry Use Case	28
<i>Fisnik Doko and Igor Miskovski</i>	
The challenges of key-value stores	32
<i>Gjorgjina Cenikj, Dushica Jankovikj and Oliver Dimitriov</i>	
Analysis of Feature Selection Algorithms on High Dimensional Data	36
<i>Sowmya Sanagavarapu, Mariam Jamilah and Barathkumar V.</i>	
Evaluation of Recurrent Neural Network architectures for abusive language detection in cyberbullying contexts	42
<i>Filip Markoski, Eftim Zdravevski, Nikola Ljubescic and Sonja Gievska</i>	
Protein classification by using four approaches for extraction of the protein ray-based descriptor	47
<i>Georgina Mirceva and Andrea Kulakov</i>	
Link Prediction on Bitcoin OTC Network	51
<i>Oliver Tanevski, Igor Mishkovski and Miroslav Mirchev</i>	
Single RNA Secondary Structure Prediction based Dynamical programming algorithms: to parallelize or not?	57
<i>Bisera Chauleva, Ljubinka Sandjakoska and Atanas Hristov</i>	
Segment Labeling Method for ML-based AFIB Detection	63
<i>Dimitri Dojchinovski and Marjan Gusev</i>	
Correlating the Cholesterol Levels to Glucose for Men and Women	68
<i>Ilija Vishinov, Marjan Gushev, Lidija Poposka and Marija Vavlukis</i>	

Using Educational Escape Room to Increase Students' Engagement in Learning Computer Science.....	72
<i>Georgina Dimova, Maja Videnovik and Vladimir Trajkovik</i>	
Development of educational game for children with dyslexia.....	78
<i>Aleksandra Sholdova</i>	
The effects of flexible work in the IT industry.....	82
<i>Mile Davitkovski, Smilka Janeska Sarkanjac and Branislav Sarkanjac</i>	
Transition from the classroom to online educational environment: First Impressions.....	88
<i>Petre Lameski, Boro Jakimovski, Vladislav Bidikov, Kiril Kjiroski, Eftim Zdravevski, Ivan Chorbev and Vladimir Trajkovik</i>	
Design optimization of Rectifier Transformers.....	93
<i>Rasim Salkovski and Ivan Chorbev</i>	
Design optimization of Earthing Transformers based on Differential Evolution Algorithms.....	97
<i>Rasim Salkovski and Ivan Chorbev</i>	
Framework for Efficient Resource Planning in Pandemic Crisis.....	101
<i>Nenad Petrovic and Djordje Kocic</i>	
A Generalization of the Convolutional Codes.....	107
<i>Dejan Spasov</i>	
A meshfree formulation for the simulation of mould filling processes in casting111	
<i>Felix R. Saucedo-Zendejo</i>	

Short Papers

Smart City: Public Parking Dashboard.....	117
<i>Ivan Klandev, Marta Tolevska and Dimitar Trajanov</i>	
Home security system based on drone automation - IoT approach.....	121
<i>Darko Kostadinov, Veno Pachovski and Irena Stojmenovska</i>	
Exploratory data analysis and statistical inference for student's results on several extensive courses at Faculty of computer science and engineering.....	125
<i>Lenche Jovova</i>	
Quality of Online Teaching in Higher Education – the Case of South East European University (SEEU), North Macedonia.....	129
<i>Veronika Kareva and Daniela Kirovska-Simjanoska</i>	
Object detection and semantic segmentation of fashion images.....	133
<i>Sandra Treneska and Sonja Gievska</i>	
A novel platform for sharing and renting clothing to reduce environmental pollution.....	137
<i>Ana Todorovska, Evgenija Krajchevska, Dimitar Trajanov and Sasho Gramatikov</i>	
Educational robots in preschool education.....	141
<i>Kristina Todorovska and Ana Madevska Bogdanova</i>	

Security Situation in Republic of Macedonia Using Semantic Algorithms for Open Data .	144
<i>Zivka Jovevska, Daniel Jovevski and Leonid Djinevski</i>	
Weekly Analysis of Moodle Log Data in RStudio for Future Use in Prediction	148
<i>Neslihan Ademi and Suzana Loshkovska</i>	
Artificial Intelligence: Simulating Human Emotion and Surpassing Human Intelligence ..	152
<i>Filemon Jankuloski, Adrijan Bozinovski and Veno Pacovski</i>	
A survey of covert channels: Benign and malicious usage, conditions for creation and countermeasures	158
<i>Ema Stamenkovska and Vesna Dimitrova</i>	
A Note on a Successful WEP Attack	163
<i>Vesna Dimitrova and Stefan Pavlov</i>	
Fog Necessity Over Cloud Computing For Healthcare Applications	168
<i>Beyza Ali, Natasa Paunkoska Dimoska and Ninoslav Marina</i>	
Secure ECash Payment Method Based on Pseudo-Random Functions in Centralized and Decentralized Systems	172
<i>Lina Lumburovska, Vesna Dimitrova, Stefan Andonov, Keti Isajloska and Jovana Dobрева</i>	
Scanning of services based on E-Governance Macedonia 2020177	177
<i>Boshko Kitanov, Gzim Ibraimi and Marjan Gushev</i>	
PubSub implementation in Haskell with formal verification in Coq	181
<i>Boro Sitnikovski, Biljana Stojcevska, Lidija Goracinova-Ilieva and Irena Stojmenovska</i>	

FULL PAPERS

Precision Apiculture – IoT System for Remote Monitoring of Honeybee Colonies

Riste Poposki
Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering
Skopje, Macedonia
riste.poposki@students.finki.ukim.mk

Dejan Gjorgjevikj
Ss. Cyril and Methodius University
Faculty of Computer Science and Engineering
Skopje, Macedonia
dejan.gjorgjevikj@finki.ukim.mk

Abstract— Beekeeping practice, being very environmentally dependent, requires the temperature and the humidity in the hive to be in some regular ranges for optimal beehive health and productivity. Since most of the plants and flowers required for beehive prosperity and honey production are usually outside inhabited areas, the beekeeper must travel to the bee colonies to check them, which can be time and resource consuming. In this paper, an end to end remote monitoring and control system for a bee colony is presented. The system is consisted of a web-based system for monitoring and control of the conditions of the hives and IoT system for collecting the sensor measurements and transferring the data. The IoT system is composed of hardware units that are mounted on the beehives, containing temperature, humidity, weight sensors, actuators, and a microcontroller responsible for collecting the measurements and sending the data to the web system. The communication between the hardware unit and the web system uses WiFi or LoraWAN technology, that enables running the device on batteries. The system enables remote monitoring of multiple beehives and can be configured to alert the user via email or push notification if some sensor value is outside of predefined range. The system also enables sending commands to the unit controlling the actuators that can intervene on the beehive closing or opening a ventilation lid.

Keywords—beekeeping, IoT, monitoring, LoraWAN, WiFi

I. INTRODUCTION

The constant progress in the area of internet of things contributes to creating various systems for remote monitoring and control in different areas like agriculture [1], the health sector, smart cities, smart homes, transport etc. The apiculture or beekeeping is organized care and management of bee colonies in human made hives [2].

For the beehive prosperity and successful production of honey, it is very important the hive's internal and external environmental conditions to be in some regular ranges. According to [3] and [4], the hive's internal temperature and humidity are one of the main factors for increased honey production. Honeybees have some natural ways of regulating the temperature and humidity but despite that, sometimes an intervention from the beekeeper is required in order to balance the values. On the other hand, the bee colonies are usually placed outside of areas inhabited with people, preferably in the rural areas and most of the time the beekeeper is not living near the beehives. The threats of animal attack on the hives, as well as possibility of other people stealing the hives are common. Also, when the period for harvesting the honey comes, the beekeeper must manually inspect all the hives and check which ones are ready.

Several monitoring systems that would contribute to the beekeeping process and increase the honey production have been suggested by different researchers. Balta and Dogan in [5] show example of a software architecture for a system that

monitors the internal conditions (temperature, humidity) in a beehive. In a similar study [6] Dineva and Atanasova propose a system for monitoring the conditions with focus on different communication technologies for transferring the sensor values. Despite monitoring the internal conditions of the beehive, the authors in [7] propose adding a weight scale to periodically measure the beehive weigh. This weight values help in recognizing which hive has honey to be gathered, or if the hive has been stolen or destroyed by animals.

The aim in this paper was to develop an end to end system for remote monitoring and control of a beehive colonies. An embedded hardware unit with connected sensors would be mounted on the beehives and will periodically send data to a web system. The beehive state including the temperature, humidity and weight of the beehive can be monitored using a web application. The system also provides an option for sending commands to the beehive like changing the data transfer interval. The system can be configured to send alerts to the user if some sensor value is out of range. Special focus will be given to low powered wide area communication technologies like LoraWAN [8] since the device should be able to operate on batteries.

The remainder of the paper is organized as follows: in Section 2 the system architecture is described, with subsections about the general structure, the detailed structure, as well as explanation of the communication between components. In Section 3 the initial implementation of the web system and the prototype of the hardware unit is presented and discussed. Finally, in Section 4 a conclusion and remarks on the future work is given.

II. SYSTEM ARCHITECTURE

The architecture and intercommunication of the components of the beehive monitoring system should be carefully thought out since the whole system should meet some important requirements such as:

- The user should access the system from any device and from anywhere where internet access is available
- The data from the hives should be available near real time, with possibility to notify the user of some critical conditions
- Many areas where the beehives reside do not have electricity so the hive monitoring system should be operational in these areas
- Operation of the hive monitoring system should not be dependent on constant internet access
- Monitoring the conditions in the hives without altering the natural workflow of the bees, meaning the system should be minimally invasive

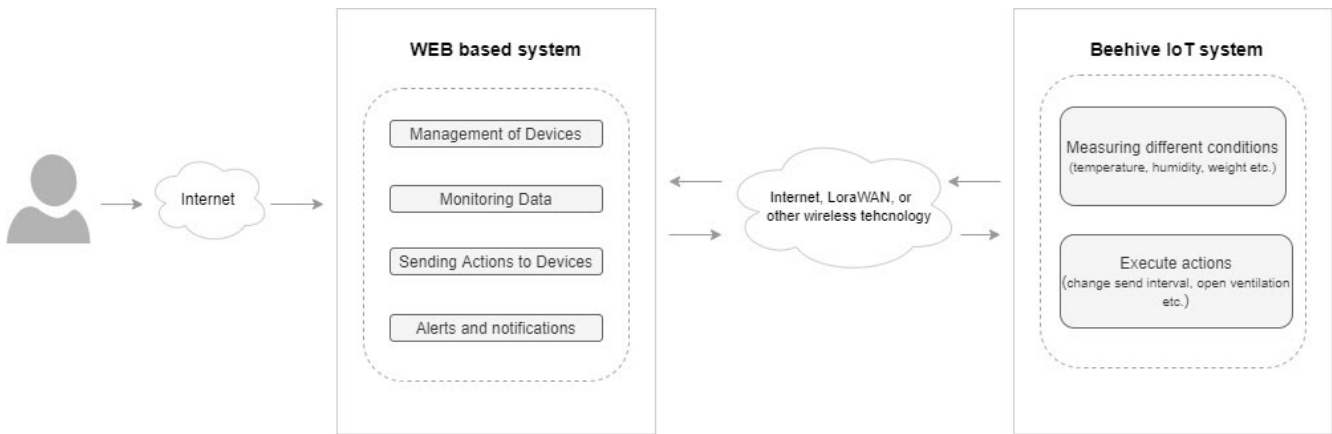


Fig. 1. General structure of the system

A. General Structure

Based on these requirements, the general system architecture is presented on Fig. 1. The beehive monitoring and control system is composed of two subsystems: a web-based system and an IoT system. The web-based system is hosted on a web server, while the IoT system is placed onsite where the actual beehives are located.

The web system is the main entry point for user interaction. To access and use this system the user only needs to have a device with internet connection and a web browser. Generally, it can be perceived like a traditional rich web application hosted on a web server. Upon visiting the application, the user can interact with the beehive monitoring and control system. Some of the functionalities and modules are:

- Monitoring the beehives' data. This includes viewing the sensor values sent from the devices mounted on the beehives. A tabular and graphical presentation of the data is available.
- Listing and managing all the devices mounted on the beehives. This includes registering new devices, editing their basic information, and adding the device on a dashboard.
- Sending commands to the devices. This is a module where the user can send commands to a device mounted on a particular beehive. The commands can be used to configure the sampling rate and the data transfer interval, or to control actuators as servo motors to increase the ventilation in the beehive. There is also a history view where the status of whether the device executed the command is presented.
- Setting rules and alerts. In this module the user can define rules regarding the sensor values of each beehive. If the rule is valid, the system can send an alert to the user via email or push notification.

The IoT system is composed of the beehives that have smart device mounted on it. The smart device contains a general-purpose microcontroller (MCU) connected to various sensors for measuring some physical parameters of the environment. Some of the sensors monitor the internal conditions inside the beehive and some of them the external conditions around the beehive.

The sensor values are periodically sent to the web-based system on a predefined period. The primary set of parameters that are measured are:

- temperature
- humidity
- pressure
- beehive weight

If the device has WiFi access to the internet, the data can be sent directly to the web system using the HTTP transfer protocol. However, considering the mentioned requirements, constant power supply or internet connection are not always guaranteed. LPWA (Low Power Wide Area) technology like LoraWAN can be very suitable for transferring slow changing data on long ranges using very low power. Devices with LoRa connectivity can generally send long range data to gateways in a 2-5km radius while running on batteries for an extended period of couple of months. Currently we use the LoraWAN wireless technology for transferring the data to the web-based system, but other LPWA technologies can be integrated to the system in the future.

Despite only sending data to the web-based system, the IoT System can also receive some predefined commands from the user. Some of the commands are:

- Changing the interval between taking measurements and data transfers. Most of the time the smart device can remain in low power mode saving the batteries. The most power is used while reading the sensors and transferring the data back to the web-based system.
- Controlling actuators like servo motors to physically move some parts of the beehive. An example can be opening a lid in order to decrease the humidity in the beehive.

Each command is followed by an acknowledgement offering an insight in the commands that have not been delivered, are not executed yet on the device, or have failed.

B. Detailed Structure

The detailed view of all the components, the external systems and communication between them is given on Fig. 2.

The internal components of the beehive monitoring and control system is composed of web client application, web

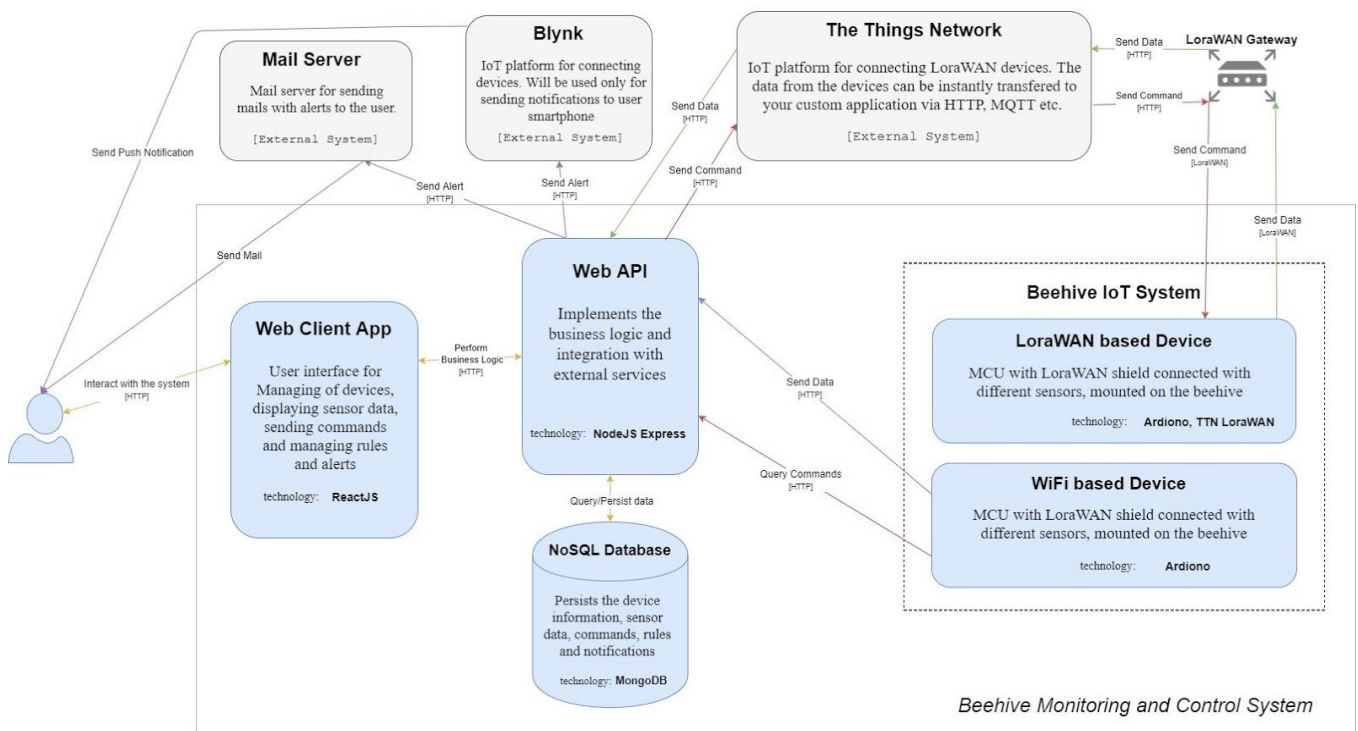


Fig. 2. Detailed structure of the system

API, NoSQL database and IoT devices mounted on the actual beehives.

To implement the LoRaWAN support, the system is integrated with the “The Things Network” cloud service [9] while for the alerting module, integration with a mail server and the Blynk IoT platform [10] is used. These are third party external services that provide some functionalities that do not need to be implemented inside the beehive monitoring and control system.

1) Web Client App

This is a single page web application developed using HTML, CSS and the ReactJS framework. It implements the user interface of the system, and exposes the following modules and functionalities:

- **Beehive Dashboard** - a module where the user can view summary information for the beehives. It includes a section showing the corresponding beehive’s minimum, maximum and average sensor values. Underneath this section there is a beehive section where each beehive is represented as a donut chart displaying some sensor values. The last section is a detailed view for the selected beehive revealing all sensor values in detail and displaying a line chart with historical data.
- **Control Dashboard** - a module where the user can view compact information for some beehives of interest. The view consists of section displaying the basic information of the beehive, section displaying the sensor data as a table and a line chart and section displaying the commands that can be sent to the device.
- **Devices Module** - a module where all registered devices (beehives) in the system are listed. Some functionalities in this module include creating new

devices or adding existing devices to the control dashboard.

- **TTN Integration Module** - a module where the integration with The Things Network is configured. The user can view all the registered devices on the network and can add new ones. For every TTN device the corresponding beehive device is presented, or if no device is assigned a connection to a device can be configured.
- **Alerts Module** - a module that is shown in the device detailed view is used for defining rules for each sensor. The predefined rules are:
 - value greater than - the system will send alert if the sensor value is greater than user specified value.
 - value lower than - the system will send alert if the sensor value is lower than user specified value.
 - maximum time without data transfer - the system will send alert if the device has failed to send data for more than prespecified amount of time.

2) Web API

This is a backend RESTful web API developed using the NodeJS express framework. It is a central place where the business logic of the system resides. Each independent module is written in a separate endpoint. The following endpoints are available:

- **Device Endpoint** - This is an endpoint for managing devices. It has CRUD methods for the beehive devices. There is also a method for changing the connected TTN device.
- **Data Endpoint** - This is an endpoint for the sensor data. It has methods for posting sensor values from the WiFi enabled devices as well as methods for getting and filtering sensor data for some specified device.

- Command Endpoint - This is an endpoint for the device commands. It has methods for sending a selected command to a specified device as well as getting the already sent commands (history). There is also a method for getting only the not acknowledged commands for a particular device. This method is used by the beehive devices that pool the server on a predefined interval.
- Alerts Endpoint - This is an endpoint for managing the rules and alerts for each device. There are methods for creating alerts for certain sensors, as well as defining the rules when this alert will be activated. There is also alerts history endpoint where the last sent alerts with the triggered rules can be retrieved.
- Beehive Summary Dashboard Endpoint - This endpoint offers data summary retrieval for the beehive dashboard.
- TTN Endpoint - This is an endpoint for managing the devices registered on The Things Network. There are CRUD methods for managing each TTN device and a method for getting the TTN application info.

Beside the REST endpoints, the Web API has integration with The Things Network's data endpoint via the MQTT (Message Queue Telemetry Transport). This is the place where actual communication with the LoRaWAN devices mounted on the beehives is happening. When a beehive transfers data to the TTN cloud service, the same data gets decrypted and sent to the Web API to be stored in the database.

There is also another integration with the Blynk IoT cloud for sending alerts via push notification and with the google mailing server for sending alerts via e-mail.

3) NoSQL Database

The database used by the Web API is NoSQL Mongo database. The data is persisted and queried using the mongoose library for NodeJS which is an ORM (Object Relational Mapper) that eases the whole data related process. The models (entities) created for the system are device, data, command, alert, alert history, and beehive summary dashboard.

4) Beehive IoT Devices

The beehive IoT devices are consisted of a microcontroller and sensors connected to it, programmed on the Arduino platform. These devices are mounted on the beehive and measure the conditions inside and outside the hive. Two types of devices have been used:

- Lora32U4 - a development board based on ATmega32u4 microcontroller with a built in Lora 868 MHz Radio module. This module has support for battery-based power supply which can last for a couple of months. On this board, the following sensors and actuators are connected: DHT11 (temperature and humidity sensor), SG90 Micro Servo (minimal servo motor), HX711 (analog to digital converter module connected to 4 strain gauge sensors composing a weight scale). The MCU is periodically reading the sensor values and is transfers the readings using the LoRaWAN technology. The data transfer is a broadcast to all accessible TTN gateways nearby. The gateways are transferring the data to the TTN cloud service via the internet. The TTN cloud then sends this data to the

web API which saves it in the database making it available to the user. After each transmission there is a small window where the TTN cloud can send some downlink message to the device. This window is used for sending commands to the device.

- NodeMCU ESP8266 - a development board based on the ESP8266 WiFi module. On this board, a BMP280 temperature and pressure sensor is connected. The sensor values are transferred to the web API periodically on a certain interval via HTTP to the data endpoint. The commands are queried via pooling i.e. sending HTTP request on a certain period and checking the command endpoint for commands waiting to be executed on that device.

5) External Systems

The external systems used in this solution include The Things Network cloud service, the Blynk IoT platform and a mail server.

The Things Network is a global public network offering easy integration for the devices that support the LoRaWAN protocol. In order to be used, one needs to create a TTN application in the TTN service, after which the registration of the devices can be performed. These devices are called TTN devices, and as described in the previous sections, they can be connected to the already registered devices in the beehive system. For the full communication to work, a gateway is required to be in the reach of the device. A gateway is a hardware device which has permanent internet and power connection equipped with software for capturing the incoming LoRaWAN messages and dispatching them via internet to the TTN Network.

The Blynk IoT platform is a large and multi feature platform for developing and implementing different IoT solutions, but in our application only the notification module is used. The user needs to have the Blynk mobile app installed, which enables receiving push notifications from the web application. Whenever an alert is triggered in the system, a notification including information about the alert that was triggered and the device that triggered it is received on the user's smartphone.

For the mail integration, any SMTP server can be used. We have used the Google's Gmail SMTP server, that upon triggering an alert, an email message including the information about the device and the rules that triggered the event is composes and sent to the configured e-mail address.

C. Communication between components

The user starts the interaction with the system from his web browser visiting the web client application. The client application requests resources from the web API using the HTTP protocol. These resources include information for the devices, the actual sensor values, commands history etc. The web API validates the requests and executes some business logic to load the data from the MongoDB database. The payload is returned in JSON format.

The IoT devices can use different communication protocols based on the actual device. The WiFi based devices send their data and query the available commands with HTTP directly with the web API. The LoRaWAN based devices use LoRaWAN protocol and communicate with the nearby gateways. The gateways than transfer the data to the TTN cloud service, that sends that data further to the web API using

MQTT protocol. The commands for the LoRaWAN devices are first sent from the web API to the TTN cloud service that queues all commands and sends them to the appropriate gateways that forward them to the actual LoRaWAN devices on their next data transfer.

III. PROTOTYPE AND INITIAL IMPLEMENTATION

The beehive monitoring and control system described in the previous section has been partially implemented and prototyped.

A. Software System

The beehive dashboard module is shown on Figure 3. The first section shows the total number of beehives monitored and the user can select which beehives and which sensors are shown on the page. Next to it is the Min-Max-Avg section showing the information for each sensor. Underneath this is the devices section showing donut gauges for each device. The data shown here can be selected from the Min-Max-Avg section. At the bottom of the screen the device detail section is shown where all the data values for the selected device are shown, as well as line chart showing the variation of the values for the selected sensor.

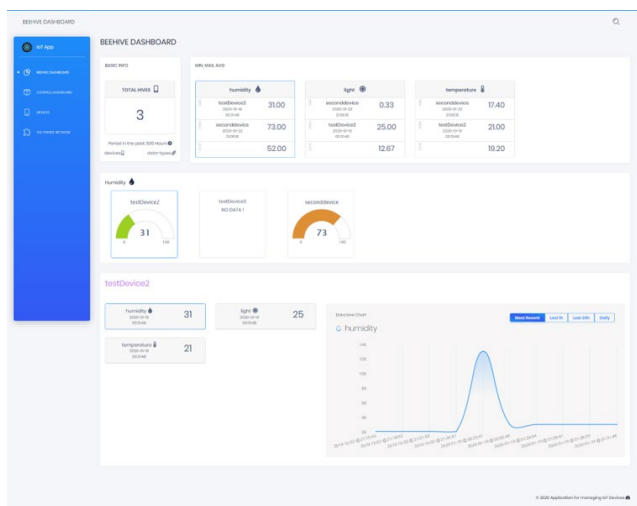


Fig. 3. Beehive dashboard from the Beehive System

B. IoT System

A photo of the principal prototype of the beehive device is shown on Figure 4. It shows the Lora32U4 development board connected to the sensors and actuators on a breadboard. The connected sensors are:

- DHT 11 temperature sensor
- LDR light sensor
- HX711 module connected to a weight sensors
- SG90 Micro Servo actuator
- LED diode for status signaling

The MCU is programmed using the Arduino platform and its main cycle is reading the values from the sensors and sending them to the TTN network. The data transfer interval is initially set to 3 minutes based on the TTN fair access policy. After each data transfer the device checks for possible messages sent back from the TTN network in the downlink window. If downlink message has been received, the MCU decodes the command and executes it. It can be for example

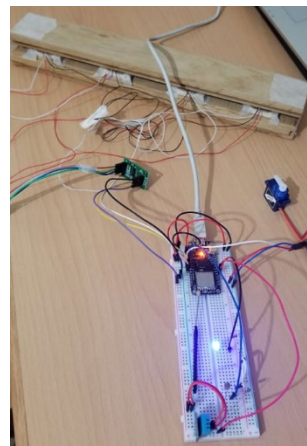


Fig. 4. Prototype of the beehive sensor device

moving the servo actuator to open/close a ventilation lid or to control another device.

IV. CONCLUSION

In this paper, an end to end system for monitoring and control of a beehive colonies is presented. It contains a web-based solution that exposes a web application where the user monitors and controls the conditions of the bee colonies. The application shows different graphical and tabular representations of the sensor values for each beehive, beehive groups, summary view of the minimum, maximum and average values across all available beehives. There is also possibility to define rules for beehive parameters that enable sending alerts via mail or push notification. The other part of the system is the hardware IoT units containing sensors. The sensors measure some crucial parameters for the beehive like the internal and external temperature and humidity and the beehive weight. These values are sent via the LoRaWAN or WiFi connection to the web system where they are persisted. Such system will enable the beekeeper to have near real-time information about the state of his bee colonies, reducing the need for travel to physically inspect the beehives.

For the future work, this prototype will be mounted on a real beehive adequately protected from the external influences (rain, sun) and we plan to perform experiments with real beehive colonies. Use of additional sensors that can give some insight of intensity of the bee activity, like microphones and light sensors that can count the bees that leave or enter the hive will also be considered. Extending the system to support other LPWAN technologies like NB-IoT or SigFox and even the GPRS network communication is also planned. As for the web-based system, a standalone mobile application can be developed and integrated with the web API. This will eliminate the need for the integration with the Blynk platform and will also provide easier way for the user to interact with the system through its mobile device.

ACKNOWLEDGMENT

This research was partially funded by the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje.

REFERENCES

- [1] Z. K. Aldein Mohammeda and E. S. Ali Ahmed, "Internet of Things Applications, Challenges and Related Future Technologies", *World Scientific News*, vol. 67, No. 2, pp. 126-148, 2017.
- [2] S. E. McGregor, "Beekeeping", *Encyclopaedia Britannica*, <https://www.britannica.com/topic/beekeeping>, [Accessed March 2020].
- [3] E. E. Southwick, G. Heldmaier, "Temperature Control in Honey Bee Colonies", *BioScience*, vol. 37, no. 6, pp. 395-399, June 1987.
- [4] H. F. Abou-Shaara, A. A. Al-Ghamdi, A. A. Mohamed, "Tolerance of two honey bee races to various temperature and relative humidity gradients", *Environmental and Experimental Biology*, vol. 10, pp. 133-138, 2012.
- [5] A. Balta, S. Dogan, G. O. KOCA and E. Akbal, "Software Modeling of Remote Controlled Beehive Design", *International Conference on Advances and Innovations in Engineering (ICAIE)*, Elazig, Turkey, May 2017.
- [6] K. Dineva and T. Atanasova, "Model of Modular IoT-based Bee-Keeping System", *The 2017 European Simulation and Modelling Conference*, Lisbon, Portugal, October 2017.
- [7] S. Gil-Lebrero, F.J. Quiles-Latorre, M. Ortiz-López, V. Sánchez-Ruiz, V. Gámiz-López, J.J. Luna-Rodríguez, "Honey Bee Colonies Remote Monitoring System", *Sensors*, vol. 17, no. 55, 2017.
- [8] "LoRaWAN 1.1 Specification", *LoRa Alliance Technical Committee*, October 11, 2017.
- [9] "The Things Network", [Online] <https://www.thethingsnetwork.org/>, Accessed: March 2020.
- [10] "Blynk IoT", [Online], <https://blynk.io/>, Accessed: March 2020.

Blockchain-based model for authentication, authorization, and immutability of healthcare data in the referrals process

1st Goce Gavrilov
University American College Skopje
School of Computer Science and
Information Technology
Skopje, Macedonia
gavrilovgoce@yahoo.com

2nd Orce Simov
International Slavic University
"Gavrilo Romanovich Derzhavin"
Faculty of Computer Science
Sv. Nikole, Macedonia
osimov@yahoo.com

3rd Vladimir Trajkovik
University "Ss. Cyril and Methodus"
Faculty of Computer Science and
Engineering
Skopje, Macedonia
trvlado@finki.ukim.mk

Abstract— The healthcare industry is continuously reforming and adopting innovative technologies that allow the digitalization of health information and automation of clinical processes. Some of the crucial requirements in these adaptations and implementations are interoperability across different departments and the security of a patient's sensitive data, mainly when data exchange.

The provisioning of confidentiality, integrity, consistency of data and data quality management is vital to ensure the best healthcare service delivery. The blockchain technology is a revolutionary invention, which ensures data integrity and confidentiality inside any system. The blockchain technology with application layers built on it, promise a mechanism that provides data integrity and privacy, most privacy, and security in healthcare services. In this paper, we propose an e-referrals model with the main focus on supporting authentication and authorization of entities in the process of issuing referrals to provide data integrity and confidentiality. The proposed model offers a framework for managing patients' referral data between doctors on the primary, secondary, and tertiary levels of healthcare.

Keywords— *blockchain, authentication, authorization, e-referral, healthcare*

I. INTRODUCTION

Today, citizens' health care is of prime importance in most world countries, as the number of patients and diseases continues to increase. Maintaining health records for every patient is a necessity for managing future health needs for citizens. Computer-aided medical support terms, such as e-Health, e-health record, e-referrals, e-scheduling, and e-prescription have appeared as a result of the evolution of information and communication technologies (ICT).

The existence of diverse health devices and apps with a combinations of the Internet of Things (IoT) have contributed to transfer of a large amount of medical data daily. Access and sharing patient's data before, during and after the treatment process are every day needs of doctors and other healthcare staff [1]. This extensive connected but distributes database of citizens' health information creates significant privacy, security, and availability issues.

Electronic availability of these data online makes it easier for hackers and other malicious attackers to access this confidential information. Also, the openness of patient's data via the internet, exposes this information to more hostile attacks compared to the paper-based records [2].

Referrals represent the link and interface between healthcare providers in primary and secondary healthcare [3]. According to [4], [5], [6], the referral process is defined as transferring (including data sharing) of responsibility of patient healthcare from referral provider to another physician

or provider. This transfer of patient healthcare should be reversed back in an appropriate time. Consequently, linking healthcare service levels is essential.

Paper-based referrals sending by fax, still the standard process in many practices, creates referral delays due to incomplete or missing information such as patient data, clinical laboratories, etc [7]. Paper-based referral processes can lead to inadequate information exchange, lost or misplaced paper records, as well as medication errors resulting from illegible handwriting with limited standardization [8].

E-referral is an electronically transmitted message such as XML documents or PDF documents that can be received and viewed by the reviewer [9], [10]. According to that, e-referrals represent a new mechanism for the integration of the different levels of health care [11]. E-referrals allude to the automation of the referral process in which appointments and other information regarding the consultation and review are transferred between two or more healthcare providers. E-referral systems have been designed to improve wait times and efficiency by electronically standardizing information and communication within the referral process.

Healthcare data are highly sensitive data, and the process of their sharing or transferring them from one institution to another (from primary healthcare to secondary, secondary to tertiary or horizontally between clinics on the same level of healthcare) has always had privacy and confidentiality concerns [12]. According to EU legislation, [13] a patient has to provide consent when someone wants to access his/her data. The law enforcement and other specific public agencies may legally access health information, according to the Health Insurance Portability and Accountability Act [14]. Medical referrals are transmitted between primary and secondary, secondary, and tertiary healthcare institutions or horizontally between clinics on the same level of healthcare daily. Medical referrals are subject to potential attacks from unauthorized persons.

The referrals are prone to attack by intruders and may be intercepted, modified, or fabricated. Even if data are safely shared between different doctors the integrity of health records [15] remains a significant issue. Data privacy [16] is also under threat, and health and medical data are prone to safety crises. Due to the sensitivity of the information contained in the referrals, especially the medical findings and doctors' opinion, the possibility of intercepting this information is a risk that should not be ignored.

Because of heavy regulation and bureaucratic inefficiency, the e-referral system's innovation is not on a high-speed line. At the same time, many healthcare facilities have a critical need for such new innovation, especially in the area of data privacy and security [17]. According to findings presented in [17], the patient's data contains data that is highly valued to cybercriminals. Healthcare facilities must introduce new security measures to address these threats or be subject to the all-out of failure to do so. Blockchain technology represents a useful mechanism for securing and protecting vulnerable patient data. The application of blockchain in the healthcare area represents an important challenge for solving privacy and security concerns [18]. Blockchain technology has the potential to address the interoperability challenges [19] in healthcare information systems and to be the technical standard that enables healthcare entities to share electronic health data securely.

The main goal of this paper is to propose a framework model for e-referrals that can be used by doctors, patients and different entities involved in e-referral processes. Our solution solves privacy, security, availability, data integrity and confidentiality problems, and access control over e-referral data. The remainder of the paper is organizing as follows. Section 2 highlights the literature review and work related to the blockchain and healthcare system. Section 3 presents the model for the e-referral blockchain system. We discuss the proposed system model and future plans in section 4 and conclude the paper in section 5.

II. LITERATURE REVIEW - BLOCKCHAIN IN HEALTHCARE

When we are trying to research and review available literature concerning blockchain in an e-referrals information system or general in healthcare, it is crucial to understand the context of blockchain technology and the proposition her usage as a means for providing patient's data security mechanism. To do this, first, we need to give a brief description of blockchain technology and then analyze the previous work in an academic environment in a relationship with the use of blockchain to address privacy and security concerns in healthcare.

The blockchain is one of the most often used phrases of the last couple of years, so that Gartner has suggested that blockchain would reach the peak of inflated expectations very soon [20]. The idea of utilization of blockchain in healthcare comes out of the need for security and interoperability in healthcare. The existence of diverse health devices and apps with a combinations of the Internet of Things (IoT) have contributed to a transfer of a large amount of medical data daily. This data traffic and exchange need management regarding privacy and security. Blockchain technology can offer a solution that not only helps to securely store and sharing of medical and healthcare data but also to assure the confidentiality of each patient's data by giving the patients, as well as their medical and health data ownership [17]. Blockchain technology can redefined the data modeling and governance deployed in many healthcare applications. This is mainly due to its adaptability and ability to segment, secure, and exchange medical data and services in an unprecedented way. Many current development projects in healthcare have blockchain technology in the center of their development [21].

With the progress in electronic health and medical-related data, data store in healthcare cloud, the promotion of regulations for patient data privacy protection, new opportunities are appearing for health data management, as well as patients' convenience to access and share their health's data [22].

The blockchain is a distributed ledger technology based on the principles of a peer-to-peer network and cryptographic notions (such as hash, asymmetric encryption, and digital signature). Blockchain, as a concept of a distributed database, was for the first time described by Nakamoto in 2008 [23]. This technology provides a transparent, decentralized, authenticated platform that applies a consensus-driven approach to facilitate the interactions of multiple entities in the network through the use of a shared ledger.

The blockchain technology consists of blocks, with each block representing a set of transactions. Like a structure of data [24], a blockchain has several significant properties described below. First, blocks are provably immutable - this means each block contains a hash, or numeric digest of its content, that verifies the integrity of the containing transactions. The hash of the next block in the blockchain network is dependent on the hash of the current block, the hash of the current block is dependent on the hash of the previous block. This effectively makes the entire blockchain history immutable, as changing the hash of any block "n - i" would also change the hash of block n [25]. The functioning blockchain does not depend on a central, trusted authority, rather than, the responsibility of functioning is distributed to all nodes which participate in the network. Because is missing central authority that will verify the validity of the blockchain, a mechanism for reaching network consensus must be employed [26]. The concept of decentralized trusted authority comes as an opposite solution to almost every system that was built using the client-server architecture. Removing the central trusted authority outside of the system means there is no longer a mediator processing the actions and the data.

Several techniques used to ensure network consensus are Proof of Work function in Bitcoin [23], Proof of Stake [27], and Proof of Activity [28]. Firstly, Blockchain technology was originally designed for the financial sector, and it has the potential to change the healthcare system for the better. By providing a mechanism for the controlled exchange of sensitive data for healthcare professionals, blockchain technology can improve the transparency and data sharing between clinical and research data systems [29].

Current identity tools and mechanisms do not support this modern approach to authentication and authorization. Instead of checking on physical identity documents, the processes, and methods that are required expensive and tedious counter visits, they must change to the direction on simplifying the checks but, at the same time, maintaining the level of security with using new technology like the blockchain.

There are a lot of different approaches in the literature for supporting healthcare authentication. Trust management is tied to authentication mechanisms as the means to identify the trustee. Recent work from Zyskind et al. [30] shows the interest of blockchain technology as a personal data management platform focused on privacy. According to Zyskind et al. [30], the blockchain helps to leverage user control over data in the context of social networks and big

data. Blockchain technology may offer a way to bypass the problem of the central government body of identity control by delivering a secure solution without the need for a trusted, central authority. Blockchain-based identity authentication is particularly salient in the last few years of internet penetration.

III. SYSTEM DESIGN FOR E-REFERRAL

For the implementation of blockchain technology in the healthcare system, especially in the e-referrals process, we have first to understand how blockchain ledger works under the hood. Blockchain technology owns a built-in identity mechanism, a cryptographically secure key pair. So each participant with a specific activity on the network is assigned these keys. All participants knew each other by these keys because the original identities of participants in the network are not visible [31].

The smart contract plays a vital role in performing the agreement among various stakeholders involved in the system when implementing the blockchain in the e-referrals system. By developing the codes can be created a smart contract and these codes define the agreement signed by the various stakeholders such as a patient, doctor, or physician. A smart contract is an integral and inseparable part of the blockchain-based applications. A smart contract represents a computer protocol that follows specific rules, codes and constraints agreed by all participants in the network. It is an agreement made among various involved stakeholders in the defined system. The referrals data can be encrypted and shared with the whole ledger available within the respective network.

To implement the e-referral system in our case, we propose decentralized identity management built on top of an Ethereum (or multi-block) consortium network. Our proposed identity management system is using the decentralized smart-contract standard that defines the method for ownership and transferability of the referrals. With these features are enabled:

- Eliminate the possibility of the existence of counterfeit referrals
- Enable regulatory insight into the number of issued and realized references for every citizen
- Enable regulatory insight into the number of unfulfilled referrals for every citizen
- Enable regulatory insight into the number of specialist reports, findings, assessments and opinions
- Create an immutable record for issued referrals, specialist reports, findings, assessments, and opinions

The patient referral process workflow starts when a patient is registered to receive health services, and medical personnel determines the diagnosis of the patient. Referral form has to complete with the name of the referred healthcare facility and other required information when the patient needs to be referred to. There are two types of referrals which be issued, an emergency referral and an outpatient referral. Depending on the referral type, the patient will be served in the designated healthcare facility by either emergency or outpatient measures. If a referral is returned, then the referred healthcare facilities have to fill in the referral form. After that, the referral's institution receives referral data.

Figure 1 represents a logical architecture of the proposed e-referral management system model. Application layer, a Database, and an authentication and authorization server are the main components of the proposed model. Authentication and authorization of the users, system make by validating transactions in the blockchain network. The user (the patient, doctor, physician) communicates with the system through the e-referral application. This e-referral application can be web, mobile, or a standalone application that interacts with the application server via integration services to perform the desired functions. For delivering a requested medical and health data, the application server needs to communicate with the database and the authentication server. To accomplish these activities, users must be authenticated and authorized.

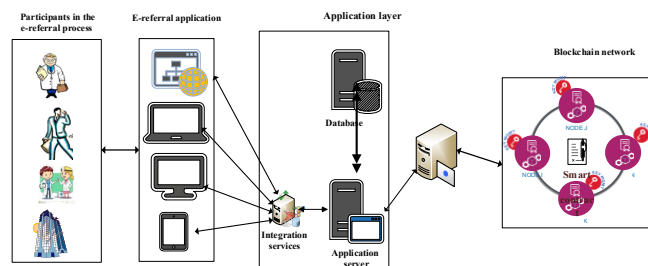


Fig. 1: Logical architecture of the e-referral system model

The application server assigns work to the authentication server to check the authenticity of the user and to authorize access to the database. A standalone or cloud-based database may be used [32], and it is used in the system to limit the amount of data in the blockchain as much as possible. After receiving a permission validation from the authentication server, access to the database from the application server is allowed. The authentication server has an intermediary role between the application server and the blockchain network. It authenticates and authorizes the user and able to interact with the blockchain network. The blockchain network records data operations and various data access requests for immutability and integrity protection. The nodes in the blockchain network keep the network in running state and maintain the ledger. Nodes follow the rules in the smart contract, validate and broadcast transactions and run the consensus protocol.

The patient first must registers in the trusted services offices by providing personal details (such as ID, Biometrics, and PIN), together with the public key of the patient. Also, to enable a referral needs the public key of the doctor. Since the registration process is a one-time process, patients provide their details using their mobile devices or web application.

The public key(s) of the doctor(s) responsible for referrals are added to the information file of the patient. The doctors together with their public keys, must be already registered in the trusted services. After the patient's registration process, the constructed information file about the patient will be sent to the blockchain. Next, the patient goes to the doctor to gets the needed referral. By using the mobile application on a mobile device, the patient generates ID and a secret key pair (private/public key pair) to authenticate him/herself. After that, if the doctor wants to issue referral/s, he/she will send a request to the blockchain network using his/her key material. Upon receiving the referral request by the blockchain network, it will check the validity of the doctor and whether the patient has granted the update permission to that

particular doctor. If the check is successful, it performs the referral issuing operation. When filling the data for specialist reports, findings, assessments and opinions, similar kind of steps are taken.

Figure 2 shows the steps of healthcare data management workflow in blockchain.

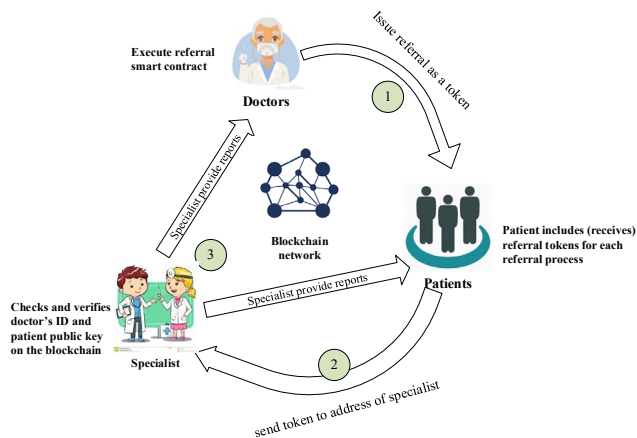


Fig. 2: E-referral process with blockchain

Three leading roles of participants are mainly existing in our proposed system:

1. Primary healthcare doctor - Issues referrals by executing a smart contract that tokenizes a valid prescription. During the process of issuing referrals are used metadata such as doctor's ID, patient's ID, quantity, healthcare institution where the patient is referred, the type of review to be performed, physician (see Figure 2).

2. Patient - Receives a token representing a valid referral prescribed by a doctor. The patient receives the necessary health service at an authorized healthcare facility by sending token to the facility's public wallet (responsible for storing the users' private and public cryptographic keys).

3. Specialist (doctor on the next level of healthcare) - After receiving a token from a patient, the doctor makes the necessary examinations and filled the specialist reports, findings, assessments and opinions, prescribe medication, etc. After these actions specialist sends information to the doctor and patient. Specialist checks and verifies a valid referral by checking the permission blockchain for a signature between the patient and doctor.

Healthcare authorities with read-only access to the ledger may be considered to be a "fourth role".

IV. DISCUSSION AND FUTURE WORK

The rapid developments in Information Communication Technology (ICT) causes increased digitalization in the healthcare sector. One of the top priorities in the healthcare sector is to provide safer manners of accessing the patients' health and medical information throughout the whole process of referrals process. Blockchain technology assumed to be among one of the suitable ways of authentication, authorization, and sharing the health and medical information. The potential of blockchain in the e-referrals process is being realized by many involved stakeholders (patients, doctors, physicians, healthcare management staff) and its immense impact on improving the healthcare

interoperability and collaboration and enhanced healthcare economy and revenues.

One of the vital on-going obstacles in the current e-referrals systems is the lack of a mechanism for authentication and authorization, and blockchain offers the possibility to handle this issue. The blockchain's future in healthcare sector seems to be quite pronounced and visionary. However, the practicality of the healthcare applications and especially application for e-referral using blockchain is mostly untested yet. From these reasons, the future development includes implementation a functional prototype of the proposed architecture, shown in Section 3. A proposed system model is based on an open-source community blockchain framework called Hyperledger Fabric. It is also permitted instead of using Hyperledger Fabric, to use other cloud-based blockchain services. Future work includes implementation and testing of the proposed system in a closed environment, development of most of the components of the system, connection with some outsourced components, demonstration of the up-scaling of the system and goes to real implementation.

V. CONCLUSION

Blockchain technology in healthcare information systems has brought immense opportunities in terms of not only providing secure and efficient data storage but also sharing and control access to the data. System model for identity and access management in the e-referrals process using blockchain technology are proposed in our paper. The core focus in the paper is pointed to the theoretical design of a secure and efficient data access mechanism for current referrals systems using the blockchain technology. We have also proposed the potential smart contract agreement considering this e-referrals scenario.

Blockchain implementation in the e-referrals system and in general in healthcare systems is a significant challenge in a rapidly evolving era of privacy and security concerns. With the progress in electronic health and interoperability, healthcare data store in the cloud and patient data privacy protection regulations, new opportunities are appearing for health data management, as well as patients' convenience to access and share their health data.

REFERENCES

- [1] S. Alla, L. Soltanisehat, U. Tatar, and O. Keskin, "Blockchain Technology in Electronic Healthcare System", Proceedings of the 2018 IISE Annual Conference, In K. Barker, D. Berry, C. Rainwater, eds. ISBN: 978-1-5108-6935-6, 2018 May 19-22, pp.754-760.
- [2] G. Gavrillov, O. Simov, and S. Manasov, "Blockchain technology for authentication, authorization and immutability of healthcare data in process of recipes prescriptions", *Scientific Journal «INTERNATIONAL DIALOGUE: EAST-WEST»* (ISSN print:1857-9299, ISSN online: 1857-9302), pp. 319-326, 2019
- [3] A. Mehrotra, C.B. Forrest, and C.Y. Lin, "Dropping the baton: specialty referrals in the United States", *Milbank Quarterly*, vol. 89(1), pp.39-68, 2011.
- [4] J. Warren, S. White, K.J Day, Y. Gu Y, and M. Pollock, "Introduction of electronic referral from community associated with more timely review by secondary services", *Applied Clinical Informatics*, vol. 2(4): pp. 546-564, 2011.
- [5] A. Esquivel, "Characterizing, Assessing and Improving Healthcare Referral Communication", PhD Thesis, The University of Texas School of Health Information Sciences at Houston, 2008.
- [6] T.M. Akande, "Referral system in Nigeria: study of a tertiary health facility", *Annals of African Medicine*, vol. 3, No. 3, pp. 130 – 133, 2004.

- [7] C. Hughes, P. Allen, M. Bentley, “e-Referrals: why we are still faxing”, *Aust Fam Physician*, vol. 47(1–2), pp. 50–57, 2018.
- [8] F.K. Thiong’o, “Framework for the Implementation of a Patient Electronic Referral System: Case Study of Nairobi Province”, MSc Thesis, University of Nairobi, School of Computing and Informatics Scientific 2011.
- [9] L. Tian, “Improving knowledge management between primary and secondary healthcare: an e-referral project”, *Health Care Inform Rev Online*, vol. 15, pp. 31-37, 2011
- [10] A. Coleman, “Developing an e-health framework through electronic healthcare readiness assessment”, PhD Thesis, Nelson Mandela Metropolitan University, 2010.
- [11] A.H. Chen, E.J. Murphy, and H.F Jr Yee, “eReferral - A New Model for Integrated Care”, *The New England Journal of Medicine*, vol. 368(26), pp. 2450-2453, 2013 Jun 27.
- [13] G. Gavrilov, E. Vlahu-Gjorgievska, and V. Trajkovik, “Healthcare data warehouse system supporting cross-border interoperability”, *Health Informatics Journal*, pp. 1-12, <https://doi.org/10.1177/1460458219876793>, 2019.
- [13] GDPR, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).
- [14] M.L. Johns, “HIPAA privacy and security: A practical course of action”, *Topics in Health Information Management*, vol. 22(4), pp. 40-48, 2002.
- [15] M. Herlihy, M. Moir, “Enhancing accountability and trust in distributed ledgers. arXiv preprint arXiv:1606.07490, 2016.
- [16] J. Zou, Y. Wang, M.A Orgun, “A dispute arbitration protocol based on a peer-to-peer service contract management scheme”, In: 2016 IEEE international conference on web services (ICWS). IEEE, June 2016, pp. 41–48.
- [17] K. Peterson, R. Deeduvanu, P. Kanjamala, and K. Boles, “A Blockchain-Based Approach to Health Information Exchange Networks”, Available at: <https://www.healthit.gov/sites/default/files/12-55-blockchain-based-approach-final.pdf> [Accessed January 25, 2020].
- [18] K.J. Smith, G. Dhilon, “Blockchain for Digital Crime Prevention: The Case of Health Informatics”, *Blockchain for Digital Crime Prevention: Health Informatics, Twenty-third Americas Conference on Information Systems, Boston, 2017*.
- [19] N. Mountford, K. Threase, M. Quinlan, R. Maher, R. Smolders, P. Van Royen, I. Todorovic et al. “Connected Health in Europe: Where are we today?”, University College Dublin, 2016.
- [20] Gartner (2016). “Hype Cycle for Emerging Technologies”, Stamford, CT, USA: The Gartner Group. Available at: <http://www.gartner.com/newsroom/id/3412017>. [Accessed January 2, 2020].
- [21] S. Khezr, M. D. Moniruzzaman, A. Yassine, and R. Benlamri. “Blockchain Technology in Healthcare: A Comprehensive Review and Directions for Future Research”, *Appl. Sci.*, vol. 9, pp. 1736-1764, 2019, <https://doi.org/10.3390/app9091736>.
- [22] D. V. Dimitrov, “Blockchain Applications for Healthcare Data Management”, *Healthc. Inform. Res.*, vol. 25, pp. 51–56, 2019.
- [23] S. Nakamoto, “Bitcoin: A peer-to-peer electronic cash system”, Available from: <https://bitcoin.org/bitcoin.pdf>. [Accessed January 12, 2020].
- [24] B. Singhal, G. Dhameja, and P. S. Panda, “Beginning Blockchain: A Beginner’s Guide to Building Blockchain Solutions”, ISBN-13 (pbk): 978-1-4842-3443-3. ISBN-13 (electronic): 978-1-4842-3444-0. <https://doi.org/10.1007/978-1-4842-3444-0>, Apress.
- [25] E. Karafiloski, “Blockchain Solutions for Big Data Challenges- A literature review”, *IEEE EUROCON 2017, 6–8 JULY 2017, OHRID, R. MACEDONIA*, 978-1-5090-3843-5/17/\$31.00 ©2017 IEEE.
- [26] G. Zyskind, O. Nathan, A.S. Pentland, “Decentralizing Privacy: Using Blockchain to Protect Personal Data”, 2015 IEEE CS Security and Privacy Workshops. DOI 10.1109/SPW, 2015, pp. 180-185.
- [27] S. King and S. Nadal, “PPCoin: Peer-to-peer crypto-currency with proof-of-stake”, Self-Published Paper, August, 19, 2012.
- [28] I. Bentov, C. Lee, A. Mizrahi and M. Rosenfeld, “Proof of activity: Extending bitcoin’s proof of work via proof of stake”, *ACM SIGMETRICS Performance Evaluation Review*, vol. 42(3), pp. 34-37, 2014.
- [29] A.C. Marek, “Blockchain as a Foundation for Sharing Healthcare Data”, *Blockchain in Healthcare Today™* ISSN 2573-8240 online <https://doi.org/10.30953/bhty.v1.13>. [Accessed February 12, 2020].
- [30] G. Zyskind, O. Nathan, and A.S. Pentland, “Decentralizing privacy: Using blockchain to protect personal data”, *Proceedings - 2015 IEEE Security and Privacy Workshops, SPW 2015*, pp. 180–184.
- [31] Blockchain in Healthcare. Available: <https://www.hyperledger.org/wpcontent/uploads/2016/10/ey-blockchain-in-health.pdf>. [Accessed January 10 2010].
- [32] A. Dimitrievski, E. Zdravevski, P. Lameski, and V. Trajkovik, “Addressing Privacy and Security in Connected Health with Fog Computing”, In *Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good (GoodTechs '19)*. Association for Computing Machinery, New York, NY, USA, pp. 255–260. DOI: <https://doi.org/10.1145/332428.3342654>

Mapping of Automobile Dealership Outlets in Skopje

Sasho Nikudinoski
Faculty of Computer Science and
Engineering
Ss. Cyril and Methodius University
Skopje, N. Macedonia
joce.nexus@gmail.com

Andreja Naumoski
Faculty of Computer Science and
Engineering
Ss. Cyril and Methodius University
Skopje, N. Macedonia
andreja.naumoski@finki.ukim.mk

Elena M. Jovanovska
Faculty of Computer Science and
Engineering
Ss. Cyril and Methodius University
Skopje, N. Macedonia
jovanovska.elena14@gmail.com

Abstract—The purpose of this project is to provide an easier visual access to most of the automobile dealership outlets in Skopje. All car dealership outlets in Macedonia have websites that give detailed information, however sometimes information is not frequently updated. This project will allow people who search for a new car to look at their contact details, address, rating information, opening hours and website info more easily using the visual benefits of ArcGIS. The above data is marked and shown on a geographical map, through a simple overview with visually tagged symbols for each automobile dealership outlet.

Keywords—car, automobile dealership outlets, customers, ArcGIS, interpolation

I. INTRODUCTION

Although the number of imported secondhand cars in our country is increasing every year, new car sales also have a decent share in the local automobile market. All car dealership outlets in Macedonia have websites that present detailed information, however sometimes information is not frequently updated. This project includes information concerning the years of warranty that certain auto dealership offer and the number of cars each brand had sold in Europe. The purpose of this project is to provide information about the automobile dealership outlets (36), which are general importers in Macedonia as well as two others which are not general importers. Additional information is provided concerning their proximity and comparisons by rating, warranty, and cars sold in Europe categorized by brand.

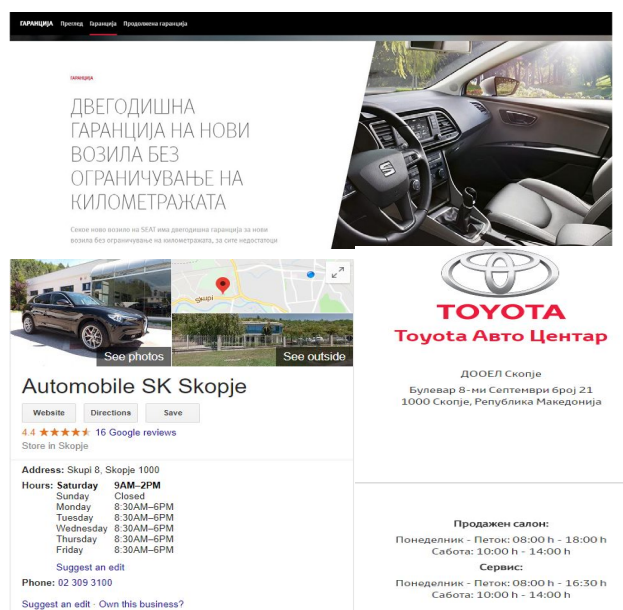
GIS was used for many business related studies, service opportunities and planning [1], and in these studies the researchers used the population data to create a geographical relationship with the locations where new businesses were located. Furthermore, the long-term trends in many businesses have been studied, and one of them is the retail trend in UK [2]. Many studies emphasize the use of geographical modeling redistribution in various businesses [3]. In other studies, like [4], the researchers pointed out at large geographically diversified data set of registered car-dealerships in the US. In that study, the authors made spatial cluster analysis, trade area analysis and regression models to identify the determinants of brand origin affinity based on socio-economic attributes across the trade areas. Their results show that specific group of people tend to buy cars made in certain places, but the affinity decreases with the income, and the strength of this relationship is weaker than expected.

The rest of the paper is organized as follows: Section II presents the data we obtained to form the geographical distribution of all the automobile dealership outlets in Skopje, as well as the methods used in this research. In Section III, we give a visual representation of the model's results

obtained from our GIS data analysis, while Section IV concludes the paper and outlines the direction for further work.

II. DATASETS AND METHODS

The purpose of this project is to show the locations of all automobile dealership outlets in Skopje, as well as some additional car information from the brands they offer, like, working hours, phone, website info, ratings (retrieved from Google Maps), years of warranty and a sum of sold cars by each brand in Europe. This information could be of great importance when people are looking to buy a new car. The data used in this project is taken from every automobile dealership outlet official website and from information provided by the Golden Book and ABC contact editions. Ratings for all car outlets are gathered from Google Maps (Fig. 1) while warranty information is downloaded from the automobile dealership outlet websites. Information on car sales for each car brand in Europe is taken from a relevant website source and refers to the year 2019 sales [5]. After we gathered information for all automobile dealership outlets in Skopje, we created a new Geodatabase called AutoDealers.gdb where two Feature classes MainDealership and Dealership were added. You can create a new Geodatabase with a right-click on the Catalog menu on the folder, then New -> File Geodatabase.



The image shows a screenshot of a website advertisement for Toyota. The top part features a car interior and text in Macedonian: "ДВЕГОДИШНА ГАРАНЦИЈА НА НОВИ ВОЗИЛА БЕЗ ОГРАНИЧУВАЊЕ НА КИЛОМЕТРАЖАТА". Below this is a map showing the location of "Automobile SK Skopje" with a red pin. The map includes a "See photos" button and a "See outside" button. To the right of the map is the Toyota logo and the text "TOYOTA Toyota Авто Центар". Below the map, there is a section for "Automobile SK Skopje" with a 4.4 star rating and 16 Google reviews. It lists the address as "Skupi 8, Skopje 1000" and provides a table of hours: Saturday 9AM-2PM, Sunday Closed, Monday 8:30AM-6PM, Tuesday 8:30AM-6PM, Wednesday 8:30AM-6PM, Thursday 8:30AM-6PM, Friday 8:30AM-6PM. There are also buttons for "Suggest an edit" and "Own this business?". To the right of the hours table, there is a section for "Продажен салон:" (Sales Salon) with hours: Monday-Friday 08:00 h - 18:00 h, Saturday 10:00 h - 14:00 h. Below that is a section for "Сервис:" (Service) with hours: Monday-Friday 08:00 h - 16:30 h, Saturday 10:00 h - 14:00 h.

Fig. 1. Example of a data source (Web page source)

We added the new class by right-clicking on the newly created database, then New -> Feature Class.

Each class contains the following attributes:

- ObjectID – automobile dealership outlet unique identifier
- Name – automobile dealership outlet name
- Address – address where the automobile dealership outlet is located
- Phone – automobile dealership outlet phone number
- Make – car brands that the automobile dealership outlet sells
- Website – automobile dealership outlet website
- Hours – automobile dealership outlet working hours
- Rating – automobile dealership outlet rating
- Warranty – warranty offered when buying a car
- SoldAutomobiles – number of cars sold in Europe for a particular brand.

After creating the classes, all automobile dealership outlets were marked on the map and filled with the appropriate information.

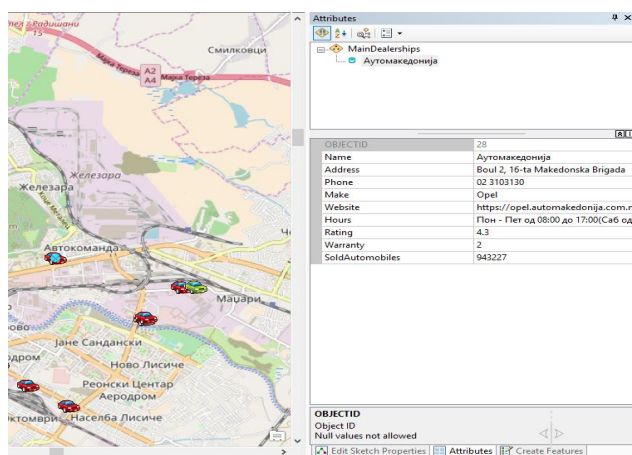


Fig. 2. Data entry for an automobile dealership outlet

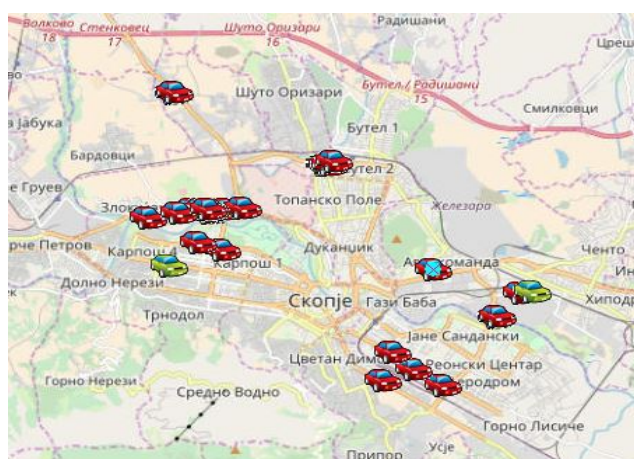


Fig. 3. The final preview of all marked automobile dealership outlets in Skopje

Marking an instance of a given class on the map is done by selecting the class from the Create Features menu and then marking the given coordinates on the map. Attributes were entered for each instance in the Attributes section (Fig. 2). The official map of Macedonia was used provided in our GIS

course. The final preview of all marked automobile dealership outlets in Skopje is shown on Fig. 3.

III. VISUAL MODELS IN GIS

The predefined automobile dealership outlet display was not sufficient to present the information, so we made a few changes to adapt this visual look. First, the automobile dealership outlet instances symbols were changed. To distinguish main importers from the rest of the automobile dealership outlets we marked the main importers with a red car symbol while the other importers were marked with a green car symbol. This can be done by clicking on the class symbol in the Table of Contents menu, which opens the Symbol Selector window.

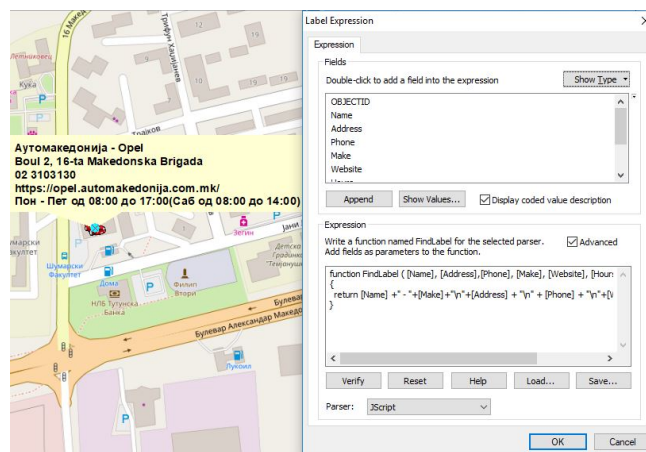


Fig. 4. Changing Expression by adding additional label display info

In addition to changing the symbols, the label displayed for each salon was changed, in Properties -> Label a yellow rectangle was chosen. In Label Expression, the label display information was added concerning the name, address, phone, website and working time of the outlets (see Fig. 4). Since many of the labels overlapped, we created a minimum scale on which the labels appear to be 1 : 5000.

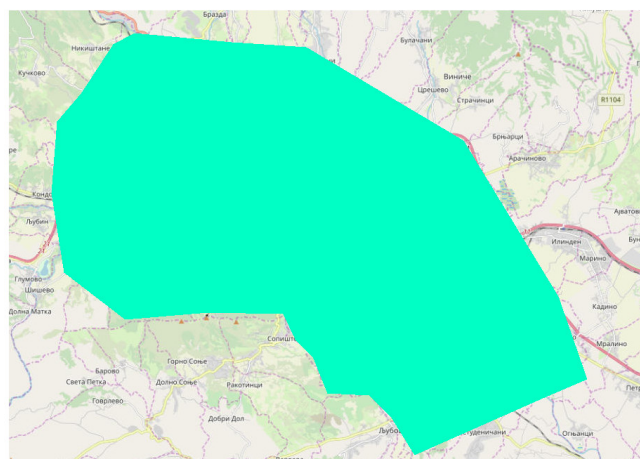


Fig. 5. Marking the territory of Skopje and its surroundings

After this phase where marking all the automobile dealership outlets was finished, marking of the polygon that will represent the territory of the city of Skopje was performed. So, from the Catalog menu, a new Sharp File called Map was created. In this file, the territory of the city of Skopje and the surrounding area was marked, as represented on Fig. 5. We will need the administrative board map later for the interpolation analysis.

A. Point Distance

Moreover, we also added information about the distance of each auto salon in Skopje from the city center. This functionality was done by choosing ArcToolbox - Analysis Tool - Proximity - Point Distance. The following window appears when selecting this functionality (Fig. 6).

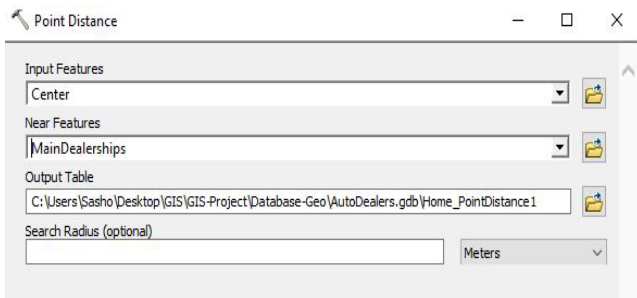


Fig. 6. Creating a Point Distance Table

On this window, Input Features is the point where the distance is measured, and Near Points are the points that will measure the distance in our case the automobile dealership outlets (Fig. 7 and Fig. 8).

OBJECTID *	INPUT_FID	NEAR_FID	DISTANCE
1	1	3	4490.056216
2	1	1	3263.790532

Fig. 7. The distance of the non-official auto dealers from the city center

OBJECTID *	INPUT_FID	NEAR_FID	DISTANCE
1	1	30	2351.9616
2	1	20	3309.2074
3	1	24	2601.0432
4	1	35	1996.4179
5	1	28	3769.4448
6	1	25	3745.3241
7	1	31	4277.0900
8	1	28	2514.7379
9	1	32	2885.7504
10	1	17	2370.5331
11	1	15	4106.2437
12	1	19	3619.7228
13	1	8	3093.5851
14	1	7	3039.9278
15	1	6	3448.9078
16	1	37	3188.4871
17	1	36	3144.5421
18	1	16	3287.5563
19	1	5	3303.9930
20	1	4	3239.3703
21	1	2	3132.4192
22	1	1	3188.1008
23	1	12	2913.6948
24	1	11	2947.6931
25	1	10	2964.1703
26	1	9	3022.1558
27	1	3	3078.5838
28	1	23	2727.0722
29	1	21	2739.8330
30	1	42	3319.0021
31	1	41	3208.4911
32	1	40	3229.9672
33	1	39	3223.5798
34	1	38	3220.6287
35	1	34	5823.0478
36	1	33	5807.0907

Fig. 8. Distances to the main auto importers from the city center

B. Interpolation Results

Interpolation aims to display data measured from one place and it is referred to large areas. In this project, it was used by interpolating the parameters: rating, warranty and number of sold cars.

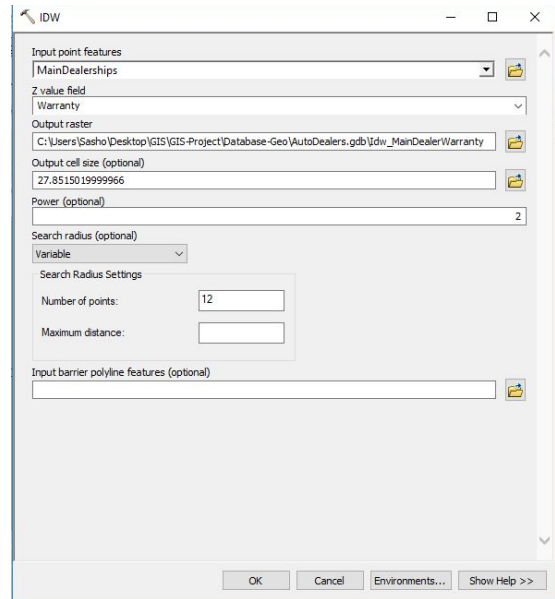


Fig. 9. Creating the interpolation

In order to interpolate, first we need to go to Geoprocessing -> Environment Settings and set the coordinate system to match with the one on the map of R. Macedonia. Moreover, in the Raster Analysis menu for the Mask option, it is necessary to limit the interpolation to the surface of the city of Skopje, in this case, the Map file. Restricting the interpolation surface will produce more accurate results than interpolating the whole surface.

Interpolation is done by selecting ArcToolbox -> Spatial Analyst Tool -> Interpolation -> IDW. Selecting this option is shown on Fig. 9.

After the interpolation was created, in Properties - Symbology the color was changed in order to present each interpolation with different colors and get a better visual representation (Fig.10).

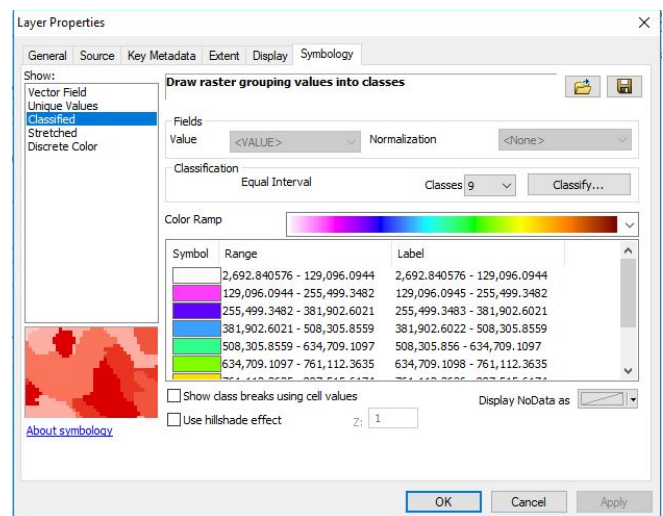


Fig. 10. Interpolation color change

On Fig.11, we present a preview from the interpolation of sold cars in correlation with the car dealer who is the main importer and sells them in Skopje. As we can see from the

interpolation analysis, in the northern part of the city car dealers sold more cars than the ones located in the center of the town. Compared to the rest of the city, the car dealers in the western part of the city sold little below the average sale per year.

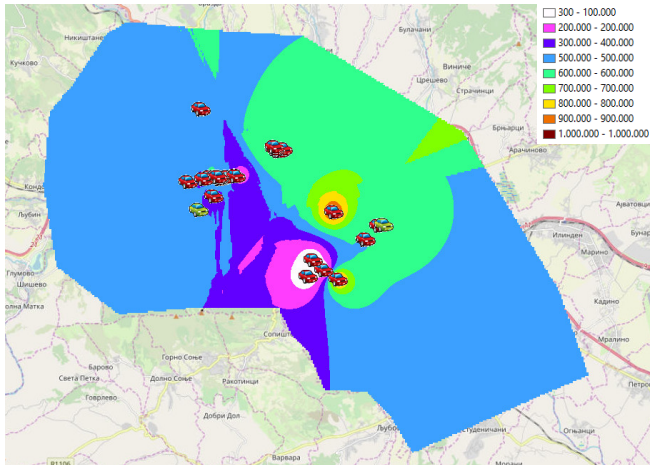


Fig. 11. Interpolation of sold cars in the city of Skopje

On Fig. 12, we present a preview from the interpolation of the warranty offered by the major auto importers in Skopje. If we analyze the warranty years given to the customer, it is no coincidence that the number of sold cars correlates with the number of warranty years that the customers gets. Therefore, sold cars with 4 or more years of warranty are sold the most in the car dealership outlets in city of Skopje.

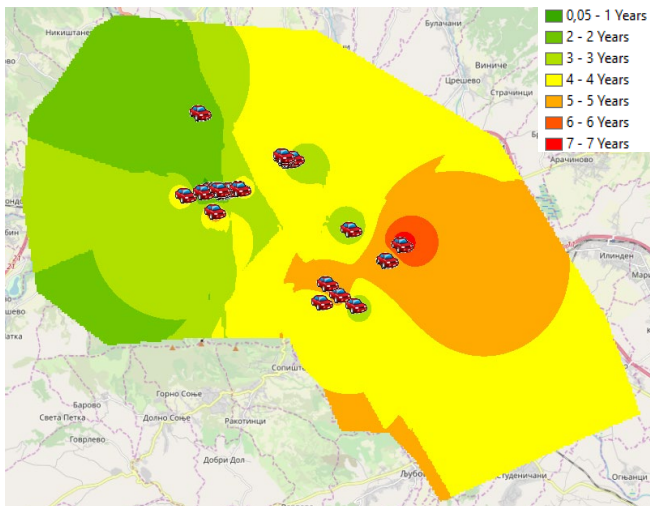


Fig. 12. Interpolation of the offered warranty (in years)

On Fig. 13, we present a preview from the interpolation of ratings of the major auto importers in Skopje, gathered from Google Maps. The rating scale ranges from 3 to 5 stars. The average star ratings of the major auto importers are presented on this map.

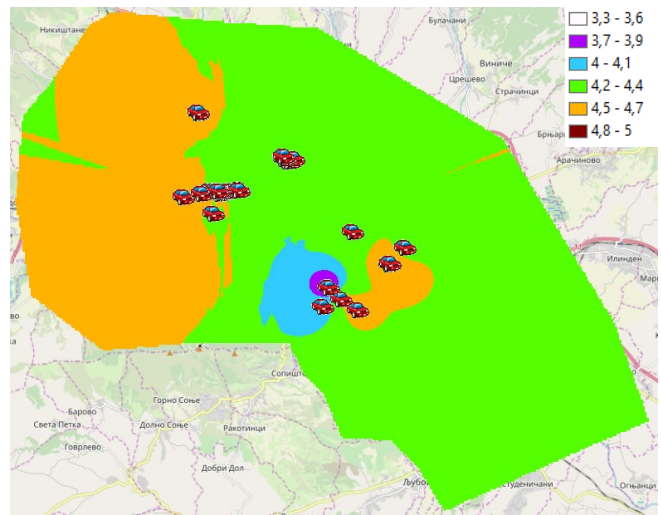


Fig. 13. Interpolation of auto importer ratings from Google Maps

We can conclude that most of the car dealership outlets in Skopje received an average of 4 stars and some of them above 4.5.

IV.CONCLUSION

This project might help customers with their next car choice. Customers can easily obtain information about automobile dealership outlets in Skopje and find out which of them are major importers and which are not. By clicking on the other layers, they will be able to get a visual representation of the areas where automobile dealership outlets offer better services. Moreover, they will be able to see what areas and outlets offer longer car warranty, which is of great importance when buying a new car. Finally, with one click, people can see where the best car sales were conducted and which brands made lower sales. In addition to this, with the click of a button they can see the distance of each automobile dealership outlet from the city center.

ACKNOWLEDGMENT

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

REFERENCES

1. M. Birkin, G. Clarke, and M. Clarke. "GIS for business and service planning." *Geographical Information Systems*, vol. 2, pp. 709-722, 1999.
2. M.T. Goodwin, and Y. Gong. "Mapping long-term retail trends in London." *Journal of Targeting, Measurement and Analysis for Marketing*, vol. 13, no. 3, pp. 220-233, 2005.
3. L. Azaz, "The use of geographic information systems (GIS) in business." In: *Int. Conf. Humanit*, pp. 299-303, 2011.
4. F. Wang, T Xu, and J. Li, "Designment of Power Mobile Repair System Based on Embedded GIS." *Agricultural Science&Technology and Equipment* vol. 2, pp. 34, 2011.
5. X. Zhao, "Ethnicity in Car Purchase Decisions." *Journal of Marketing Management*, vol. 5, no. 2, pp. 1-14, 2017.
6. H. Bekker, "2019 (Full Year) Europe: Best-Selling Car Manufacturers and Brands", online: <https://www.best-selling-cars.com/europe/2019-full-year-europe-best-selling-car-manufacturers-and-brands/>, January 2020.

Recent Advances in SQL Query Generation: A Survey

Jovan Kalajdzieski
*Faculty of Computer Science
and Engineering*
Ss. Cyril and Methodius University
Skopje, North Macedonia
jovan.kalajdzieski@finki.ukim.mk

Martina Toshevska
*Faculty of Computer Science
and Engineering*
Ss. Cyril and Methodius University
Skopje, North Macedonia
martina.toshevska@finki.ukim.mk

Frosina Stojanovska
*Faculty of Computer Science
and Engineering*
Ss. Cyril and Methodius University
Skopje, North Macedonia
frosina.stojanovska@finki.ukim.mk

Abstract—Natural language is hypothetically the best user interface for many domains. However, general models that provide an interface between natural language and any other domain still do not exist. Providing natural language interface to relational databases could possibly attract a vast majority of users that are or are not proficient with query languages. With the rise of deep learning techniques, there is extensive ongoing research in designing a suitable natural language interface to relational databases.

This survey aims to overview some of the latest methods and models proposed in the area of SQL query generation from natural language. We describe models with various architectures such as convolutional neural networks, recurrent neural networks, pointer networks, reinforcement learning, etc. Several datasets intended to address the problem of SQL query generation are interpreted and briefly overviewed. In the end, evaluation metrics utilized in the field are presented mainly as a combination of execution accuracy and logical form accuracy.

Keywords—SQL Query Generation, Text-to-SQL, Deep Learning, Semantic Parsing

I. INTRODUCTION

The possibility to use a natural language statement to query a database has the potential to attract a vast majority of users that are not proficient in using query languages such as the Structured Query Language (SQL). This language is the main query language for relational databases currently in use. The problem of text to SQL mapping could be viewed as a Semantic Parsing problem [1], which is defined as transforming a natural language input into a machine-interpretable representation. Semantic parsing is a long-standing question and is a well-studied problem in Natural Language Processing (NLP). As such, it has attracted much attention both from academia and from the industry, especially translating natural language into SQL queries. A large amount of the data in today's age is stored in relational databases for applications ranging from financial and e-commerce domains to medical domains. Therefore, it comes as no surprise that querying a database using natural language has many different applications. It also opens up the prospects of having self-serving dashboards and dynamic analytics, where people not accustomed to the SQL language could use it to get the most relevant information for their business. The task of translating natural language to SQL has many related tasks such as code generation and schema generation. All these tasks could be ultimately combined

to form a general task of translating natural language to a complete application.

There have been a variety of methods proposed to tackle the semantic parsing problem such as: rule-based [2], unsupervised [3], supervised [4], response based [5] and many others. However, the problem of generating SQL is more challenging than the traditional semantic parsing problem. A short natural language question could require joining multiple tables or having multiple filtering conditions. This requires more context based approaches.

For that purpose, in recent years, with the extensive development of deep learning techniques, especially convolutional and recurrent neural networks, the results are drastically improving. There have been quite a few researches attempting to generate data processing results by directly linking records in the tables to the semantic meaning of the natural language input, such as [6] and [7]. However, these attempts are not scalable to big tables and are not reusable when the database schema is changed. More recent approaches use only the natural language input and the database schema and metadata to generate the queries. We review the most recent approaches in our research. Furthermore, the release of large annotated datasets containing questions and the corresponding database queries has additionally enhanced the ability to use deep learning or supervised techniques to tackle this problem. This has enabled the problem to evolve into a more complex task where the approaches should be domain independent and involving multiple tables with complex queries.

In this paper, we provide an extensive research on the most used datasets, as well as the most recent approaches applied on these datasets to handle the problem of generating SQL queries from natural language input. The main motivation of this paper is to provide a comprehensive explanation and analysis of the most recent methods to handle the task of generating SQL using natural language as well as the different datasets and evaluation techniques used. The rest of the paper is organized as follows. Section II describes the different datasets that have been used in the approaches described. A comprehensive explanation of the different methods for generating SQL queries from text is presented in Section III. The different evaluation methods used for this problem are outlined in Section IV. Finally, Section V concludes the paper.

II. TEXT-TO-SQL DATASETS

The datasets designed for semantic parsing of natural language sentences to SQL queries are composed of annotated complex questions and SQL queries. The sentences are questions for a specific domain, and the answers for these questions are derived from existing databases. Therefore, the particular question is connected with an SQL query. The execution of the SQL query extracts the answer from the existing database/s.

Nowadays, there are several semantic parsing datasets developed for SQL query mapping. All of the different datasets vary in several aspects. Table I provides detailed statistics of the most used datasets among researchers. The early developed datasets concentrate on one domain and one database: ATIS [8], GeoQuery [9], Restaurants [8], [10], Academic [11], Scholar [12], Yelp [13], IMDB [13] and Advising¹ [14].

The newest datasets, WikiSQL² [15] and Spider³ [16], are cross-domain context-independent with a larger size. Also, newer datasets have a greater number of questions and more comprehensive queries. The size of the datasets is crucial for the proper model evaluation. Unseen complex queries in the test sets can evaluate the model generalization ability. Authors in [14] show that the generalisability of the systems is overstated by the traditional data splits. The WikiSQL dataset contains a large number of questions and SQL queries, yet these SQL queries are simple and concentrated on single tables [16]. The Spider dataset contains a more modest number of questions and SQL queries than WikiSQL. However, these questions are more complex, and the SQL queries include different SQL clauses such as join of tables and nested query [16].

The SParC [17] and CoSQL [18] are the extension of the Spider dataset that are created for contextual cross-domain semantical parsing and conversational dialog text-to-SQL system. These new aspects open new and significant challenges for future research in this domain.

III. METHODS

With the rise of deep learning techniques, there is extensive ongoing research in designing a suitable natural language interface to relational databases. Mostly, the models in this area rely upon the encoder-decoder framework that is widely used in the field of natural language processing. The following subsections present some of the models utilized in the field. Some of the models described in this paper are publicly available which enables other researchers to evaluate or build other models upon them.

A. SQLNet

The order of two constraints in the WHERE clause of an SQL query does not matter, but syntactically, two queries with a different order of constraints are considered as different queries. This can affect the performance of a sequence-to-sequence model. That is what SQLNet⁴ [20] attempts to

overcome. SQLNet is a novel approach for generating SQL queries from a natural language using a sketch based approach on the WikiSQL task. The sketch is generically designed to express all the SQL queries of interest. The sketch separates the query into two different token types: keywords and slots to be filled. The slots belong to either the SELECT clause or to the WHERE clause.

The WHERE clause is the most complex structure to predict and consists of three types of slots: column, op and value. All of these types can appear multiple times, as in real queries where we can have multiple filter conditions. When predicting the WHERE clause, the authors firstly need to predict which columns to include in the conditions. For that purpose, they generate the probability of a column name col appearing in the natural language query Q which is computed as $P_{wherocol}(col|Q) = \sigma(u_c^T E_{col} + u_q^T E_{Q|col})$ where σ is the sigmoid function, E_{col} and $E_{Q|col}$ are the embeddings of the column name and the natural language question respectively, and u_c and u_q are two column vectors of trainable variables.

The embeddings E_{col} and $E_{Q|col}$ are computed as hidden states of a bidirectional LSTM (Long-short term memory introduced in [21]) which do not share their weights which enables the decision whether to include a particular column to be independent of other columns. $E_{Q|col}$ has an additional column attention mechanism to be able to remember the particular information useful in predicting a particular column name. After $P_{wherocol}(col|Q)$ is computed, the next step is predicting which columns need to be included in the WHERE clause. To be more precise, the authors use a network to predict the number of column slots K by translating it into a $N + 1$ classification problem where N is an upper bound of the number of columns to choose.

After selecting the top- K columns in the where clause, a prediction is done to predict one of the three possible operands $\{=, >, <\}$. This prediction again uses the column attention embedding $E_{Q|col}$, which clearly shows the logical connection between the operand and the particular column upon which the operand will be used. For every column, the method also needs to predict the value slot. Unlike the column slots, the order of the tokens matters in the value slot, so a sequence-to-sequence structure is used to predict a substring from the natural language question. The encoder phase still employs a bidirectional LSTM, while the decoder phase computes the distribution of the next token using a pointer network [22], [23] with the column attention mechanism.

On the other side, the SELECT clause consists of two types of slots: an aggregator and a column name. Another difference is that there is only one column name in the SELECT clause, unlike multiple column name slots in the WHERE clause. However, the prediction is the same as the one done in the WHERE clause, keeping in mind that the model is only trying to predict one column. After predicting the column, the probability of an aggregator is also predicted, which shares a similar structure as the prediction done for the operation slot in the WHERE clause, since there are five possible aggregators to choose from.

¹<https://github.com/jkkummerfeld/text2sql-data>, last visited: 05.05.2020

²<https://github.com/salesforce/WikiSQL>, last visited: 05.05.2020

³<https://github.com/taoyds/spider>, last visited: 05.05.2020

⁴<https://github.com/xiaojunxu/SQLNet>, last visited: 05.05.2020

TABLE I
TEXT-TO-SQL DATASETS

Dataset	Year	Domain(s)	Databases	Tables	Questions	Queries
ATIS [8]	1994	air travel information	1	32	5,280	947
GeoQuery [9]	2001	US geography	1	8	877	247
Restaurants [10], [19]	2003	restaurants, food types, locations	1	3	378	378
Academic [11]	2014	Microsoft Academic Search (MAS)	1	15	196	185
Scholar [12]	2017	academic publications	1	7	817	193
Yelp [13]	2017	Yelp website	1	7	128	110
IMDB [13]	2017	Internet Movie Database	1	16	131	89
WikiSQL [15]	2017	/	26,521	24,241	80,654	77,840
Advising [14]	2018	student course information	1	10	4,570	211
Spider [16]	2018	138 different domains	200	645	10,181	5,693

B. Bidirectional Attention

The Bidirectional Attention model⁵ [24], much like SQLNet employs the sketch based approach for generating an SQL query. The model consists of four separate modules: character-level and word-level embedding module, the COLUMN-SELECT module, the AGGREGATOR-SELECT module and the WHERE module.

The character embeddings in the first module are initialized using the pre-trained character-level GloVe model with 300 dimensions and then leverage convolutional neural networks with three kernels to get the next representation of the embedding. The word embeddings are initialized using the pre-trained word-level GloVe model with size 300. The words not present in the GloVe model are initialized to 0 and not to a random value because the authors have inferred that using a random value and making it trainable makes the results decrease. Because a column may contain several words, the words of one column are encoded after applying an LSTM network.

In the COLUMN-select module, each token of the questions and the column names is represented as a one-hot vector, and then is an input to a bidirectional LSTM. Using this approach, the attention information of questions and column names is captured and then used to make a prediction over the column names.

On the other hand, the authors infer that there are five types of aggregation keywords in SQL: *MAX*, *MIN*, *COUNT*, *SUM*, *AVG*. They also conclude that the column name does not impact the prediction result, so the AGGREGATION-SELECT module only needs to predict the type of aggregation using the question as input. This would translate the problem to a text classification problem, where the input text is the encoded question.

The last and most challenging part is the WHERE module. Because the order of conditions does not matter, this model employs a very similar approach like SQLNet. Firstly, the number of conditions K is predicted. The prediction once again can be viewed as a $(N+1)$ classification problem. After the number of columns is predicted, taking questions and column names as input and leveraging the bi-attention information from the inputs, it predicts the column slots, which is the same computation like in the COLUMN-SELECT module

with the only difference that in this part the top- K columns are selected for the column slots. For each column slot predicted, the model then needs to select the operator from the set of three possible operators. Again, it uses the bi-attention info from the question and the column names, but now with the addition of the prediction for the column slot. The last part is the value part where leveraging the predicted columns info, a sequence-to-sequence structure is used to generate the values by taking the predicted columns info as input.

C. Encoder-Decoder Framework

The grammatical structure of a language can be described using Backus Normal Form (BNF), which is a set of derivation rules, consisting of a group of symbols and expressions. The BNF specification consists of two types of symbols: terminal and non-terminal symbols. Non-terminal symbols can be substituted by a sequence of expressions. There can be more than one sequence for a non terminal symbol, divided by a vertical bar meaning that one of them needs to be selected. On the other side, as the name suggests, the terminal symbols are not substituted. The terminal symbols are usually SQL keywords, operators or a concrete value expression. The encoder-decoder framework [25] leverages the BNF for the purpose of translating natural language inputs to SQL queries. As the name states, it consists of two phases: encoder phase and decoder phase.

This approach firstly starts with the encoder phase with an objective of digesting the natural language input and putting the most important information in the memory before proceeding to the next phase. For this purpose, the authors propose extracting additional semantic features that link the original words to the semantics of the SQL language. The semantic features are in fact group of labels, where each label corresponds to a terminal symbol in the BNF. In the BNF of SQL-92⁶ there are four terminal symbols: derived column, table reference, value expression and string expression. The authors manually label a small group of samples with these four label types and employ conditional random fields (CRFs) [26] to build effective classifiers for these labels.

The decoder phase employs two different techniques: including the embedding of grammar state in the hidden layer and the masking of word outputs. The first technique is used for state transition. The authors state that given a particular

⁵<https://github.com/guotong1988/NL2SQL>, last visited: 05.05.2020

⁶<https://en.wikipedia.org/wiki/SQL-92>, last visited: 05.05.2020

word in the output sequence, the grammar state of the word is the last expression of BNF this word fits in. To facilitate grammar state tracking, a binary vector structure is used to represent all possible states. The length of the vector is identical to the number of expressions in the BNF.

The second technique is used to filter out invalid words for outputting, based on short term and long term rule dependencies. At each step, the decoder chooses one rule from the candidate short-term dependencies, and one or more rules from the candidate long-term dependencies. These rules are used for rule matching, and once the decoder identifies a matching rule it generates a mask on the dictionary to block the output of words not allowed by the rule. The short-term dependency is updated according to the current grammar state as well as the last output word from the decoder. Long-term dependencies on the other hand, are updated based on the active symbols chosen by the SQL parser, maintained in the grammar state vector.

D. Seq2SQL

Seq2SQL⁷ [15] method consists of two parts: augmented pointer generator network and main Seq2SQL model. The augmented pointer network generates the content of the SQL query token-by-token by copying from the input sequence. The input sequence x is composed of the following tokens: words in the question, column names in the database tables and SQL clauses. The network encodes x with two-layer bidirectional LSTM network using the embeddings of its words. Next, a pointer network [22] is applied. The decoder is a two-layer unidirectional LSTM that generates one token at each timestep using the token generated in the previous step. It produces scalar attention score for each position of the input sequence. The token with the highest score is selected as next token. The second part, Seq2SQL, is composed of three different parts: Aggregation Operation, SELECT Column and WHERE Clause.

The first part, Aggregation Operation, classifies aggregation operation of the query, if any. First, scalar attention score is computed for each token in the input sequence. The vector of scores is then normalized in order to produce a distribution over the input tokens. It is computed with a Multilayer Perceptron (MLP) with cross-entropy loss. The second part, SELECT Column, points to a column in the input table. Each column name is first encoded with LSTM network such that the last encoded state of the LSTM is assumed to be representation of the specific column. With the same architecture, representation for the input question is calculated. MLP with cross-entropy loss is applied to compute score for each column conditioned on the input representation. The last part, WHERE Clause, generates the conditions for the query. For this part, reinforcement learning is applied to optimize the expected correctness of the execution result. Next token is sampled from the output distribution. When the complete query is generated, it is executed against the database. The reward is: (1) -2 if the generated query is not a valid SQL

query, (2) -1 if the generated query is a valid SQL query but executes to an incorrect result, and (3) +1 if the generated query is a valid SQL query and executes to the correct result. The loss is the negative expected reward over possible WHERE clauses.

The overall model is trained using gradient descent to minimize the objective function that is the combination of the objective functions of its composing parts. However, this method does not incorporate complex SQL queries such joining tables and nested queries.

E. STAMP

Syntax- and Table-Aware seMantic Parser (STAMP) [27] is a model based on Pointer Networks [22]. It is composed of two separate bidirectional Gated Recurrent Unit (GRU) networks as encoder and decoder. An additional bidirectional RNN is used to encode the column names. The STAMP model is composed of three different channels, that are attentional neural networks: (1) SQL channel - predict SQL clause, (2) Column channel - predict column name and (3) Value channel - predict table cell. For SQL and Value channel, the input is the decoder hidden state and representation of the SQL clause. Column channel has an additional input that is the representation of the question. Feed-forward neural network is used as a switching gate for the channels.

Column-cell relation is incorporated into the model in order to improve the prediction of SELECT column and WHERE value. The representation of the column name is enhanced with cell information. The importance of a cell is measured with the number of cell words occurring in the question and then the final importance of the cell is normalized with softmax function. The vector representing the column is concatenated with weighted average of the cell vectors that belong to that column. An additional global variable to memorize the last predicted column name is added. When the switching gate selects the Value channel, the cell distribution is only calculated over the cells belonging to the last predicted column name.

The model is trained in different ways: with standard cross-entropy loss over the pairs of question and SQL query, and with reinforcement learning with policy gradient as in [15].

F. One-Shot Learning for Text-to-SQL Generation

A method for SQL query generation composed of template classification and slot filling is presented in [28]. The first phase, template classification consists of two networks: Candidate Search Network and Matching Network. The first network, Candidate Search Network, chooses n most relevant templates. The network is a convolutional neural network and is trained to classify a natural language question where the classes represent SQL templates. For a given question, features from the layer before the final classification layer are extracted. Then, n most similar vectors with the question vector are obtained using cosine similarity. The second network, Matching Network, predicts the SQL template. First, an encoder is used to embed the question. The encoder is convolutional neural network consisted of convolutional layers with different

⁷<https://github.com/salesforce/WikiSQL>, last visited: 05.05.2020

window sizes with max-pooling. The final representation of the question is a concatenation of each pooled feature. An attention-based classifier predicts the template label based on the feature vectors obtained with the Candidate Search Network.

The second phase, slot filling, is a Pointer Network [22] that fills the slots of the predicted SQL template. The encoder is bidirectional LSTM network, while the decoder is unidirectional LSTM network. The network determines the tokens by maximizing the log-likelihood of the predicted token for the given natural language question and list of variables in the SQL template. The decoder generates one token at each timestep using attention over its previous hidden state and the encoder states.

G. Relation-Aware Self-Attention for Text-to-SQL Parsers

This approach explained in [29]⁸ is an improvement on the already existing methods so that it overcomes some crucial limitations such as: working in only one domain, working with one database schema, working with only one table or overlapping training and test sets. The main improvement of this approach focuses on the encoder part of the encoder-decoder framework already seen in previous approaches. To incorporate the relationships between schema elements in the encoder, the database schema is translated to a directed graph where each node represents either a table or a column and the edge represents the relationship between the elements. The label in the node represents the name of the table or column appropriately. The columns additionally have their type prepended. All the edges between the nodes are labeled as well to represent the exact relationship they represent. The relationships can be: (1) Column X Column Y relationship where X and Y belong to the same table or X is a foreign key for Y (or vice versa), (2) Column X Table T relationship (or vice versa) where X is the primary key of T or X is a column of T (but not a primary key), and (3) Table T Table R relationship where T has a foreign key column in R (or vice versa) or T and R have foreign keys in both directions.

After obtaining the initial graph representation, bidirectional LSTM is applied to the labels in the nodes and the output of the initial and final timesteps of the LSTM is concatenated to obtain the embedding for the node. For the input question, bidirectional LSTM is also used. Until this point, these initial representations are independent of one another in the sense that they do not have any information which other columns or tables are present. For that purpose, a relation-aware self-attention transformation is applied to all the elements to encode the relationship between two elements. For brevity, we omit the mathematical model used to represent the relationships.

The formulation of the relation-aware self-attention is the same as the one used in Shaw et al. [30]. However, in this approach, it is shown that relation-aware self-attention can effectively encode more complex relationships that exist within an unordered sets of elements compared with relationships

between two words. After applying this transformation, the final encoding of the columns, tables and the input question is used in the decoder. The decoder used is the same as in [31]. The decoder generates the query using a depth-first traversal order in an abstract syntax tree. It outputs a sequence of production rules that expand the last generated node in the tree. The decoder does not output the FROM clause. It is recovered afterwards using hand-written rules where only the columns referred to in the remainder of the query are used. Small modifications have been made to this decoder, namely: (1) when the decoder needs to output a column a pointer network based on scaled dot-product attention is used, and (2) at each step, the decoder accesses the encoder outputs using multi-head attention.

IV. EVALUATION

There is no single metric for evaluation of the text-to-SQL model. One strategy is to estimate the correctness of the result for the question. This metric is called *execution accuracy* [15]. It compares the result from the generated SQL query and the result from the ground truth query. Then it returns the number of correct matches divided by the total number of examples in the dataset. One shortcoming of this approach is that it does not eliminate the cases when a completely different query is giving the same result as the expected, for example, the NULL result.

The second metric is the *logical form accuracy* [15]. This approach calculates the exact matches of the synthesized query and the ground truth query. The queries are represented as strings, and the method for comparison is the exact string match of the queries. The weakness of this approach is the penalization of the queries that are correct but do not achieve a complete string match with the ground truth query; for example, different order of the returning columns or different queries for the same purpose. To partially address this issue, the authors in [20] introduce the *query match accuracy*. The predicted and ground truth queries are represented in a canonical form to perform the matching of the queries. This approach only solves the false negatives due to the ordering issue. SQL canonicalization is an approach used to eliminate the problem of the different writing style by ordering the columns and tables and using standardized aliases [14].

The evaluation metric in [16] includes *component* and *exact matching* of the queries. Each query is divided into components: SELECT, WHERE, GROUP BY, ORDER BY and KEYWORDS. The predicted and ground truth queries are divided and represented as subsets for each of the components, and these subsets are then compared with exact matching. However, the problem of the novel synthesized syntax for the identical logic of the SQL query is not eliminated, so the execution accuracy is needed for a comprehensive evaluation. The approach in [16] also incorporates one novelty in the evaluation process, the difficulty of the SQL query. Dividing the results by the hardness criteria can be more informative of the general ability of the model.

None of the current metrics can be used as a standalone evaluation metric for exact evaluation and comparison of the

⁸<https://github.com/rshin/seq2struct>, last visited: 05.05.2020

models, so the combination of the existing metrics is essential. It is critical for the future work in this domain to incorporate the evaluation question.

V. CONCLUSION

The translation of a natural language to SQL queries is a problem of semantic parsing. There are several text-to-SQL datasets developed that include natural language questions that can be answered by executing an SQL query from a database. The progression of the datasets introduces a combination of different domains with multiple databases and tables. The increase in the size of the datasets is apparent. Also, the questions are becoming more complex and in more extensive number.

The progressions in the NLP area are reflected in the designed models of this problem. The encoder-decoder framework is incorporated to translate the natural language into an SQL query. The encoder serves for natural language processing, whereas the decoder predicts the BNF representation of the SQL output. The sketch-based approach is introduced for SQL representation for eliminating the ordering effect of sequence generation. Additional efforts incorporate attention to the bidirectional LSTM network with the sketch-based method. The augmented pointer network is also combined in the novel models. The Relation-Aware Self-Attention approach is an improvement on already existing methods to overcome several limitations. It includes a relationship graph of the database schema and self attention to encode more complex relationships.

To evaluate the models, several approaches combine the execution accuracy and logical form accuracy. The latest approaches divide the accuracy metric into component and exact matching with the additional information of the difficulty of the SQL query.

REFERENCES

- [1] J. Andreas, A. Vlachos, and S. Clark, "Semantic parsing as mach. transl.," in *Proc. of the 51st Annu. Meeting of the Assoc. for Comput. Linguistics (Vol. 2: Short Papers)*, pp. 47–52, 2013.
- [2] A.-M. Popescu, O. Etzioni, and H. Kautz, "Towards a theory of natural lang. interfaces to databases," in *Proc. of the 8th Int. Conf. on Intell. User Interfaces*, pp. 149–157, 2003.
- [3] D. Goldwasser, R. Reichart, J. Clarke, and D. Roth, "Confidence driven unsupervised semantic parsing," in *Proc. of the 49th Annu. Meeting of the Assoc. for Comput. Linguistics: Human Lang. Technol.*, pp. 1486–1495, 2011.
- [4] J. M. Zelle, *Using inductive logic program. to automate the construction of natural lang. parsers*. PhD thesis, 1995.
- [5] P. Liang, M. I. Jordan, and D. Klein, "Learning dependency-based compositional semantics," in *Proc. of the 49th Annu. Meeting of the Assoc. for Comput. Linguistics: Human Lang. Technol.-Vol. 1*, pp. 590–599, Assoc. for Comput. Linguistics, 2011.
- [6] A. Neelakantan, Q. V. Le, M. Abadi, A. McCallum, and D. Amodei, "Learning a natural lang. interface with neural programmer," *ICLR*, 2017.
- [7] P. Yin, Z. Lu, H. Li, and B. Kao, "Neural enquirer: learning to query tables in natural lang.," in *Proc. of the 25th Int. Joint Conf. on Artif. Intell.*, pp. 2308–2314, 2016.
- [8] D. A. Dahl, M. Bates, M. Brown, W. Fisher, K. Hunnicke-Smith, D. Pallett, C. Pao, A. Rudnicki, and E. Shriberg, "Expanding the scope of the atis task: The atis-3 corpus," in *Proc. of the workshop on Human Lang. Technol.*, pp. 43–48, Assoc. for Comput. Linguistics, 1994.
- [9] L. R. Tang and R. J. Mooney, "Using multiple clause constructors in inductive logic program. for semantic parsing," in *Eur. Conf. on Mach. Learn.*, pp. 466–477, Springer, 2001.
- [10] L. R. Tang and R. J. Mooney, "Automated construction of database interfaces: Intergrating statistical and relational learning for semantic parsing," in *2000 Joint SIGDAT Conf. on Empirical Methods in Natural Lang. Process. and Very Large Corpora*, pp. 133–141, 2000.
- [11] F. Li and H. V. Jagadish, "Constructing an interactive natural lang. interface for relational databases," *Proc. of the VLDB Endowment*, vol. 8, pp. 73–84, September 2014.
- [12] A. C. J. K. Srinivasan Iyer, Ioannis Konstas and L. Zettlemoyer, "Learning a neural semantic parser from user feedback," in *Proc. of the 55th Annu. Meeting of the Assoc. for Comput. Linguistics (Vol. 1: Long Papers)*, pp. 963–973, 2017.
- [13] I. D. Navid Yaghmazadeh, Yuepeng Wang and T. Dillig, "Sqlizer: Query synthesis from natural lang.," in *Int. Conf. on Object-Oriented Program., Syst., Languages, and Appl., ACM*, pp. 63:1–63:26, October 2017.
- [14] L. Z. K. R. S. R. Z. Catherine Finegan-Dollak, Jonathan K. Kummerfeld and D. Radev, "Improving text-to-sql eval. methodology," in *Proc. of the 56th Annu. Meeting of the Assoc. for Comput. Linguistics (Vol. 1: Long Papers)*, pp. 351–360, July 2018.
- [15] V. Zhong, C. Xiong, and R. Socher, "Seq2sql: Generating struct. queries from natural lang. using reinforcement learning," *CoRR*, vol. abs/1709.00103, 2017.
- [16] T. Yu, R. Zhang, K. Yang, M. Yasunaga, D. Wang, Z. Li, J. Ma, I. Li, Q. Yao, S. Roman, *et al.*, "Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task," in *Proc. of the 2018 Conf. on Empirical Methods in Natural Lang. Process.*, pp. 3911–3921, 2018.
- [17] T. Yu, R. Zhang, M. Yasunaga, Y. C. Tan, X. V. Lin, S. Li, H. Er, I. Li, B. Pang, T. Chen, *et al.*, "Sparc: Cross-domain semantic parsing in context," in *Proc. of the 57th Annu. Meeting of the Assoc. for Comput. Linguistics*, pp. 4511–4523, 2019.
- [18] T. Yu, R. Zhang, H. Er, S. Li, E. Xue, B. Pang, X. V. Lin, Y. C. Tan, T. Shi, Z. Li, *et al.*, "Cosql: A conversational text-to-sql challenge towards cross-domain natural language interfaces to databases," in *Proc. of the 2019 Conf. on Empirical Methods in Natural Lang. Process. and the 9th Int. Joint Conf. on Natural Lang. Process.*, pp. 1962–1979, 2019.
- [19] O. E. Ana-Maria Popescu and H. Kautz, "Towards a theory of natural lang. interfaces to databases," in *Proc. of the 8th Int. Conf. on Intell. User Interfaces*, pp. 149–157, 2003.
- [20] X. Xu, C. Liu, and D. Song, "Sqlnet: Generating structured queries from natural lang. without reinforcement learning," *CoRR*, vol. abs/1711.04436, 2017.
- [21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [22] O. Vinyals, M. Fortunato, and N. Jaitly, "Pointer networks," in *Advances in Neural Inf. Process. Syst.*, pp. 2692–2700, 2015.
- [23] Z. Yang, P. Blunsom, C. Dyer, and W. Ling, "Reference-aware lang. models," in *Proc. of the 2017 Conf. on Empirical Methods in Natural Lang. Process.*, pp. 1850–1859, 2017.
- [24] G. Huilin, G. Tong, W. Fan, and M. Chao, "Bidirectional attention for sql gener.," in *2019 IEEE 4th Int. Conf. on Cloud Comput. and Big Data Anal. (ICCCBDA)*, pp. 676–682, IEEE, 2019.
- [25] R. Cai, B. Xu, Z. Zhang, X. Yang, Z. Li, and Z. Liang, "An encoder-decoder framework translating natural lang. to database queries," in *Proc. of the 27th Int. Joint Conf. on Artif. Intell.*, pp. 3977–3983, 2018.
- [26] J. D. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. of the Eighteenth Int. Conf. on Mach. Learn.*, pp. 282–289, Morgan Kaufmann Publishers Inc., 2001.
- [27] Y. Sun, D. Tang, N. Duan, J. Ji, G. Cao, X. Feng, B. Qin, T. Liu, and M. Zhou, "Semantic parsing with syntax-and table-aware sql gener.," in *Proc. of the 56th Annu. Meeting of the Assoc. for Comput. Linguistics (Vol. 1: Long Papers)*, pp. 361–372, 2018.
- [28] D. Lee, J. Yoon, J. Song, S. Lee, and S. Yoon, "One-shot learning for text-to-sql gener.," *CoRR*, 2019.
- [29] R. Shin, "Encoding database schemas with relation-aware self-attention for text-to-sql parsers," *CoRR*, vol. abs/1906.11790, 2019.
- [30] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *Proc. of the 2018 Conf. of the North Amer. Chapter of the Assoc. for Comput. Linguistics: Human Lang. Technol., Vol. 2 (Short Papers)*, pp. 464–468, 2018.
- [31] P. Yin and G. Neubig, "A syntactic neural model for general-purpose code gener.," in *Proc. of the 55th Annu. Meeting of the Assoc. for Comput. Linguistics (Vol. 1: Long Papers)*, pp. 440–450, 2017.

How Simple Predictive Analysis of Health Care Claims Data can Detect Fraud, Waste and Abuse Threats in Health Care Insurance - The Case Study of United Arab Emirates

1st Kristijan Jankoski
Netcetera AG
Skopje, R. of N. Macedonia
kristijan.jankoski@netcetera.com

2nd Kiril Milev
Netcetera AG
Dubai, United Arab Emirates
kiril.milev@netcetera.com

3rd Gjorgji Madjarov
Faculty of Computer Science and Engineering
SS. Cyril and Methodius University,
Elevate Global LLC
Skopje, R. of N. Macedonia
gjorgji.madjarov@finki.ukim.mk

Abstract—The usage of unethical practices which does not follow prescribed clinical standards and leads to the unnecessarily high expenditure for health care (waste, abuse and fraud) is increasing day by day in the Middle East countries. Reports show that about 30% of health care companies expenditures are based on a fraudulent medical claim. The rule-based approaches and expert systems that are used traditionally for tackling the health care waste, abuse and fraud (WAF) are very limited and require experts with extensive knowledge of medicine and expertise in the domain itself. The predictive analysis can be more flexible and less susceptible to some of the problems encountered with rules-based systems by focusing on the outcomes rather than the entire decision making process. In this paper, we present how simple predictive analysis and unsupervised learning on health care claims data can be used for detecting waste, abuse, and fraud threats in health care insurance in UAE. Our focus is to detect abnormal behavior of the clinicians from different specialties from different medical providers using the patterns made on the diagnosis and activity level prescription. The results obtained from the experiments performed on over 370K medical claims showed that only 0.007% of the clinicians caused potentially over 10% of the WAF marked claims. 27 clinicians marked with the analysis and scored as being most suspicious by the auditors made total of 4.929 claims.

Index Terms—health care, waste, abuse, fraud, machine learning, unsupervised learning

I. INTRODUCTION

Health care insurance is a multifaceted industry that brings together care providers, insurance companies and patients. As the industry is expected to create social benefits, there is constant pressure to contain costs while providing security and improving the health of the general population.

Misuse of the health insurance is an ongoing issue. Motivated by the financial incentives, different stakeholders are creating waste, abuse the market or even commit fraud. The volume of waste, abuse and fraud (WAF) is estimated to be in the range of 5-10% of the yearly health care expenditure [1]. This makes WAF a significant contributor to the medical

inflation. Insurance claims are under continuous scrutiny by the health care payers for being one of the key tools to control health care spending.

Looking from an insurance fraud detection technique perspective, WAF is being generated by both health care providers and insured members in the Middle East. In the worst cases, a conspiracy-type of fraud involves several parties colluding in the misuse. When looking at the complexity levels, we can categorize WAF in seven levels. This starts with single transaction as the simplest one and goes up all the way to multi-party, criminal conspiracies [2].

The most prevalent types of fraud carried out by policy holders are: gaining access to or being reimbursed for services typically not covered by the policy. For clinicians and health care providers, financial gain is the main motivation with up-coding, service unbundling, and billing for unnecessary or even not rendered services. Another problem behind the seemingly rampant issue of health insurance fraud is that many businesses and their employees do not know how it looks like action. The truth is that this scam exists in many forms and it is often horribly difficult to notice. According to a study by D. Thortnton et. al. [3] some of the most common types of frauds that are the most prevalent are:

Payment for more expensive treatments than necessary - otherwise known as 'upcoding'. This includes hyperbolizing the diagnosis in a much more serious condition, in order to increase the cost of the claim.

Counterfeiting of the diagnosis: essentially examining multiple diagnoses, in order to collect procedures that are not medically needed.

Refunds for services that have not been made: this type of fraud can be achieved by falsifying claims by using real patient information, creating a false claim from scratch or by supplementing any of the actual claims with procedures that are never happened.

Misrepresentation of unnecessary treatments: this is a request for an unnecessary procedure, for example scanning

the brain with magnetic resonance, as part of a medically necessary procedure for heart surgery.

Payment procedures: perhaps the most courageous form of fraud in health insurance, whereby medical professionals perform treatments over healthy patients exclusively for the purpose of submitting a claim.

Upcoding is one of the most expensive and most sophisticated types of waste, abuse and fraud in health care. Between 2002 and 2012, this method cost publicly funded medical assistance programs around \$11 billion. These are not victims of crime because they put unnecessary efforts on the social security network over which millions of citizens rely on their basic medical needs.

The fraud itself is difficult to detect. Medical diagnoses, length of work visits and complexity of treatment are left at the discretion of the health care provider. Individual cases of this criminal behavior, or even small groups of them, may be almost impossible to find or prove. Scams may be even more prestigious or harder to discover in large institutions, such as laboratories or hospitals, which have a wide range of procedural options and where insurance recovery tends to be very loose.

In the United Arab Emirates currently premiums amount to over \$9 billion dollars, and with the introduction of compulsory health insurance for expatriates in Dubai (95 percent of the workforce), that figure is constantly increasing. But there is additional factor that influence the increase of insurance premiums, the waste, abuse and fraud.

Insurance waste, abuse and fraud is nothing new, but levels that are currently thought to be taking place globally are shocking. The data from the Health Insurance Counter Fraud Group, a group of 32 health insurance companies, with a mandate to detect and prevent fraud in the health sector, suggest that global losses as a result of fraudulent claims are hundreds of billions of dollars.

In this paper, we show how simple predictive analysis and unsupervised learning on health care claims data can be used for detecting waste, abuse, and fraud threats in health care insurance in UAE. Our focus is to detect abnormal behavior of the clinicians from different specialties from different medical providers using the patterns made on the diagnosis and activity level prescription. In the following section, the most relevant related work is reviewed first, and then the use case and the health care insurance system of UAE is described. The analysis that was performed on the obtained data for indicating the potential threats of waste, abuse and fraud of the medical personnel is presented in section 3. This section also provides visualization of the obtained results and their interpretation. The concluding remarks are given in section 4.

II. RELATED WORK AND PROBLEM STATEMENT

Traditionally, detection of health care waste, abuse and fraud is based on archaic methods that is very limited and requires experts with extensive knowledge of medicine and expertise in the domain itself [4] [5]. It is based on the work of auditors who need to manually review and identify

suspected medical claims, which is a very costly and time-consuming process. They have quite limited time to process each claim by following predefined rules and procedures on certain characteristics of a claim without paying attention of a provider's behavior. This process is supported by the rule-based engines and expert systems based on the information disclosed from the past events and findings in the research.

But with the development of electronic health records and advances in artificial intelligence, other opportunities for automatic dealing with fraud have been made. Machine learning can help with the knowledge ex-traction process and create models from thousands of medical claims that can identify a much smaller subgroup for further assessment by the auditors, which significantly increases the time that any expert can dedicate to claims while he inspects. In turn, this leads to a higher rate of detected fraud.

There are different data mining approaches that are used for addressing the problem of health care waste, abuse and fraud. The most common and well-accepted categorization that is used in this domain divides the approaches into 'supervised' and 'unsupervised' [6] [7].

Supervised approaches use samples of previously known fraudulent and non-fraudulent records. They are quite successful in detecting already known patterns of fraud and abuse. Taking this into account, these models should be regularly updated to reflect new types of fraudulent behaviors and changes in the regulations and settings [4]. Examples of the supervised methods that have been applied to health care fraud and abuse detection include neural networks [8] [9], decision tree [10] [8], and Support Vector Machine (SVM) [11].

Unsupervised methods typically compare claim's attributes to other claims and determine the level of difference by measuring the distance from a concrete distribution of claims. They are able to select anomaly records or group of similar records. Examples of the unsupervised methods that have been applied to health care fraud and abuse are clustering [12] and outlier detection [13].

A. *Shafafiya standard and data description*

Abu Dhabi has implemented a standard for data exchange within the health sector called Shafafiya. Any electronic transaction between 2 health care institutions must be structured and exchanged according to the imposed standard. The purpose of such provisions from 'Shafafiya' is to establish a national standard for improving the efficiency and effectiveness of the health care system by simplifying the processes themselves. The various definitions, formats and data that health systems have used for decades are now united in one standard format (Figure 1) and thus the electronic transfer of information is far more efficient and reliable. This has improved the overall quality of data and significantly reduced the administrative burden. In this way, information management methods are simplified, giving health care professionals more opportunities to provide better care and reduce the costs of patients themselves.

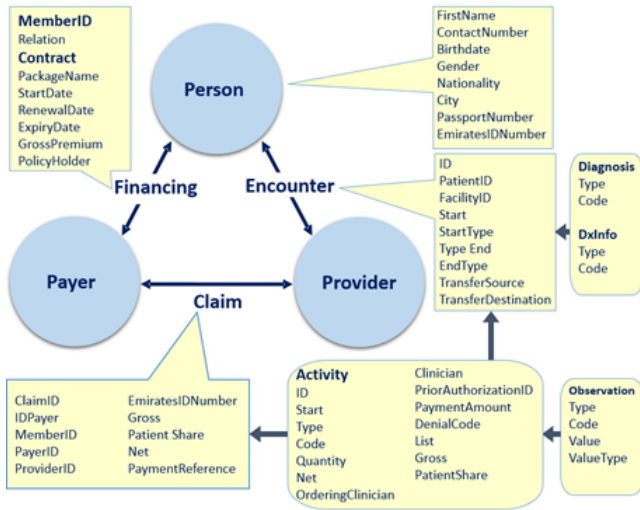


Fig. 1. Entities defined by the Shafafiya standard © Property of Department of Health (DOH)

The data that were used in this research satisfied the Shafafiya standard. We received 368.541 unique claims obtained from 1.567 medical providers for 2 years. The total number of prescribed activities on all claims were 1.224.259. The total number of investigated clinicians were 9.566 from different specialties (not given by the stakeholder). On all claims 12.384 diagnoses were detected. 35% of them were primary diagnoses, while the rest 65% were secondary diagnoses. All diagnoses were coded according to the ICD10 standard.

III. ANALYSIS OF WASTE, ABUSE, AND FRAUD THREATS

Health care claims contain cleanly structured data elements that can be used as input for predictive modeling. These elements include information about the insured member with their medical condition, the medical procedures and services performed on the patient, the prescribed medications, time, date and location of the services, and others.

One representation of the problem space for waste, abuse and fraud in health insurance could be the multi-layer graph representation (Figure 2). Each data element in the transaction can be represented by one or more nodes in the graph. Nodes are linked to one or more of the other nodes, from the same or from a different type. Each edge in the graph has an assigned weight based on the relationship of the specific nodes. After determining the nodes and edges of the initial graph, additional layers can be added that represent different abstractions for the transactions. Once the problem is defined in this way, one should represent the nodes and edges using descriptive attributes and start the learning process of the machine learning based approach.

This iterative process of knowledge discovery uses a combination of different data analysis and visualizations. Conveying the information and the gained knowledge to an individual during the analysis, even when working with highly skilled actuaries, is one of the key tasks. Therefore, good data visualization is just as important as the data analysis.

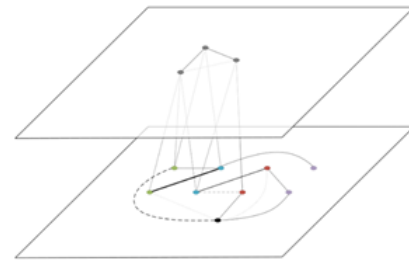


Fig. 2. Multi-layer graph representation of the entities and their interactions

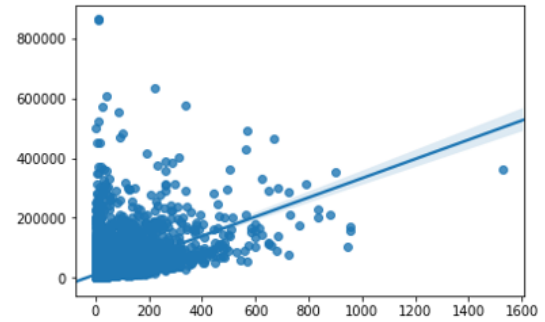


Fig. 3. Comparison

Defining the baseline in behavior analysis mandates proper analysis and definition of peer groups first. For example, pharmaceutical transactions (eRX or PBM) have very different features than inpatient visits (hospital stay). Depending on the transaction types, this classification can be straightforward. In absence of the elements containing the required information (due to the private and personal data limitation), however, a data-driven approach should be applied. Aggregations on a clinician level is another case. Figure 3 indicates the importance of identification of peer groups among the clinicians. It shows a comparison on all clinicians (represented as blue dots on the figure) on the number of unique activities that they have prescribed and the net amount on the claims that they have made. The big bang approach for solving this problem could lead us to many wrong assumptions and conclusions. Identifying the correct peer group for comparison is a critical step in the process. To address these kinds of problems, we use unsupervised data-driven approaches that try to find hidden patterns in the data based on a similarity measure using unlabeled data.

A. Peer groups identification using unsupervised learning

One of the most common methods for unsupervised learning is cluster analysis, which is used to analyze data previews to find hidden patterns or groupings in the data. Clusters are modeled using a similarity measure that is defined on a metric, such as Euclidean distance, probable distance, etc. Therefore, clustering can be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings depend on the individual set of data and the intended use of the results. The cluster analysis as

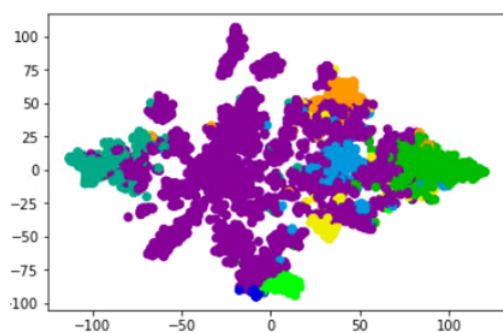


Fig. 4. t-SNE visualization of the results obtained from data-driven clustering

such is not an automated task, but an iterative process of knowledge discovery or interactive multi-target optimization that involves trial and failure. It is often necessary to pre-process the data and define the parameters of the model, until we achieve the desired result. One of the most commonly used clustering algorithms are: K-means and Hierarchical Agglomerative Clustering.

In our context, the use of such groups avoids the problem of comparing specialists in a medical field and doctors who are less professional. Therefore, each doctor who appears in the data set is represented as a rare vector of the frequency of the diagnoses he has prescribed, that is, the columns in such a vector are all unique diagnoses that occur in the set of data.

Then, the vectors from all doctors are concatenated into a single matrix, over which the two clustering algorithms applied. The best results are obtained using hierarchical clustering, and they are presented on Figure 4 in two dimensional space by reducing the dimensionality with t-distributed stochastic neighbor embedding (t-SNE) [14].

B. Univariate analysis

Univariate models are one of the simplest forms of data analysis. They don't deal with causes or relationships and the main goal is to describe the data. With the help of this type of model, we want to answer one of the following questions:

- How many times did the doctor prescribe a certain activity compared to other similar doctors?
- How much did a doctor earn by prescribing a particular activity compared to other similar doctors?
- How much activity has been prescribed for a given diagnosis compared to other prescribed activities?
- How much do the doctor prescribe for a diagnosis compared to other similar doctors?
- How many doctors diagnose a diagnosis compared to other similar doctors?

An example answer to some of these questions is presented on the Figures 5 and 6, where it can be clearly seen which doctors don't follow the trend of their group.

C. Bivariate analysis

Bivariate analysis is a simultaneous analysis of two variables. Here, the concept of the relationship between these

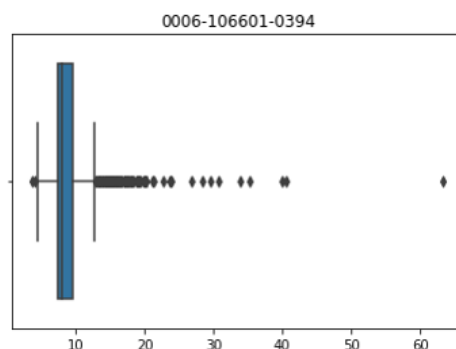


Fig. 5. Box-plot visualizations of the univariate models. Each dot represents a doctor and the amount of prescriptions he made on the 0042-114504-2481 activity

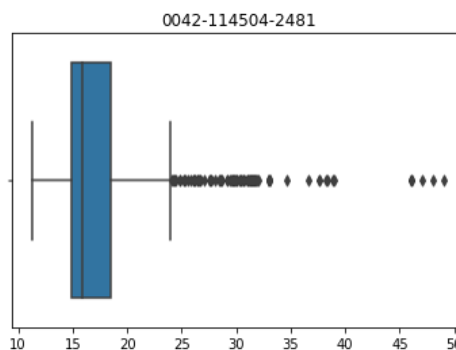


Fig. 6. Box-plot visualizations of the univariate models. Each dot represents a doctor and the amount of prescriptions he made on the 0006-106601-0394 activity

variables is being investigated, regardless of whether there is any association between them, as well as the strength of that association, or whether there are differences between those variables and their significance. It is important to note that in our context, this type of analysis can detect a trend only if the data are well grouped. Using these models, the trends in 3 types of models were analyzed:

- The number of prescribed activities against the quantity of activities (Figure 7)
- The number of prescribed activities against the net amount of claims (Figure 8)
- The number of claims against the net amount of claims (Figure 9)

Then the relationship between these two variables is modeled using a linear predictive function whose unknown parameters are estimated from the data. Finally, we define a metric for the abnormality of the doctor according to this model as the z-score of the distances of doctors from the linear function Figure 10.

This predictive analysis leads us to over 36K suspicious claims. 27 clinicians marked with the analysis and scored as being most suspicious by the insurer made total of 4,929 claims. 0.007% of the clinicians caused potentially over 10% of the WAF marked claims.

IV. CONCLUSION

Health fraud significantly affects the ability of insurance companies to provide effective health care. Utilizing the power of machine learning can help detect fraudulent events, revealing perpetrators and reducing health care costs.

In this paper, we investigated various methods of machine learning to detect doctors who made claims for payment of activities that were not needed in the particular case. The results of the models applied in the practice of real data show that the data-driven methodologies used are very successful in dealing with this type of problem. They showed: over 36K suspicious out of 370K evaluated medical claims; 0.007% of the clinicians caused potentially over 10% of the WAF marked claims; 27 clinicians marked with the analysis and scored as being most suspicious by the insurer made total of 4.929 claims. It is also worth noting that predictive analysis is a much more efficient strategy than analyzing individual medical claims because it allows real-time detection over a large number of data and does not require supervision from a human auditor.

Current and further research will include improving the performance of existing models by using a larger number of data as well as using new methods based on supervised learning with a limited number of labels received by the auditors themselves. In addition, future research will dive deeper into assessing specific techniques for detection of anomalies based on other types of fraud in order to achieve a more automatic ranking and greater adaptability of the model. We hope this research will advance the state-of-the-art detection of fraud in health care and help address this important social challenge.

REFERENCES

- [1] J. Gee and M. Button, *The Financial Cost of Healthcare Fraud 2014: What Data from Around the World Shows*. BDO, 2014.
- [2] D. Thornton, R. M. Mueller, P. Schoutsen, and J. van Hillegerberg, "Predicting healthcare fraud in medicaid: A multidimensional data model and analysis techniques for fraud detection," *Procedia Technology*, vol. 9, pp. 1252 – 1264, 2013. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S2212017313002946>
- [3] D. Thornton, M. Brinkhuis, C. Amrit, and R. Aly, "Categorizing and describing the types of fraud in healthcare," *Procedia Computer Science*, vol. 64, pp. 713 – 720, 2015. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1877050915027295>
- [4] R. A., J. H., and V. T., "No evidence of the effect of the interventions to combat health care fraud and abuse: a systematic review of literature," *PLoS One*, vol. 7, pp. 2579–2605, 2012.
- [5] J. Li, K.-Y. Huang, J. Jin, and J. Shi, "A survey on statistical methods for health care fraud detection," *Health Care Management Science*, vol. 11, no. 3, pp. 275–287, 2008. [Online]. Available: <https://doi.org/10.1007/s10729-007-9045-4>
- [6] H. Joudaki, A. Rashidian, B. Minaei, M. Mahmoud, B. Geraili, and M. Nasiri, "Using data mining to detect health care fraud and abuse: A review of literature," *Global journal of health science*, vol. 7, p. 37879, 01 2015.
- [7] R. Bauder, T. M. Khoshgoftaar, and N. Seliya, "A survey on the state of healthcare upcoding fraud analysis and detection," *Health Services and Outcomes Research Methodology*, vol. 17, no. 1, pp. 31–55, 2017. [Online]. Available: <https://doi.org/10.1007/s10742-016-0154-8>
- [8] F.-M. Liou, Y.-C. Tang, and J.-Y. Chen, "Detecting hospital fraud and claim abuse through diabetic outpatient services," *Health Care Management Science*, vol. 11, no. 4, pp. 353–358, 2008. [Online]. Available: <https://doi.org/10.1007/s10729-008-9054-y>

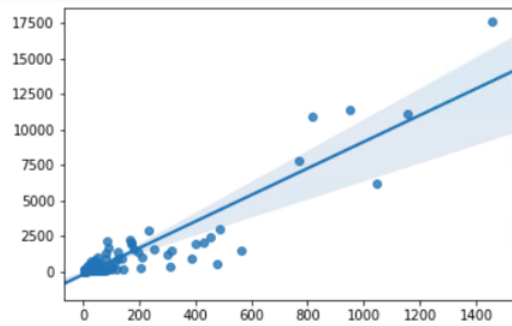


Fig. 7. Number of activities vs quantity of activities

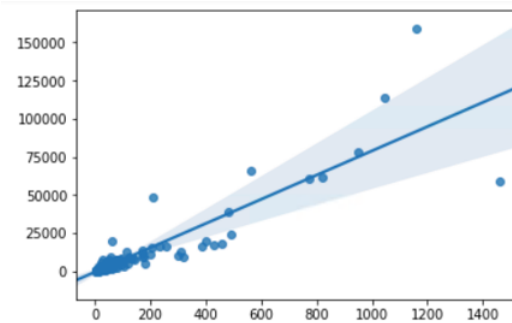


Fig. 8. Number of activities vs net amount of claims

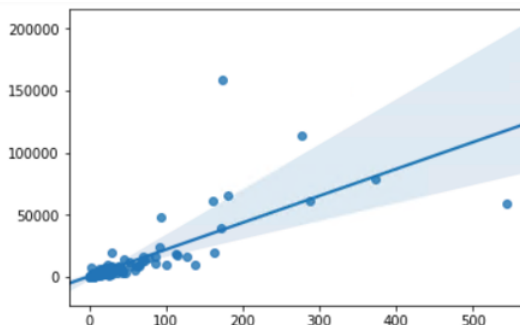


Fig. 9. Number of claims vs net amount of claims

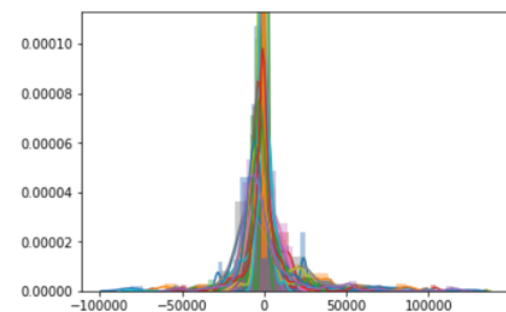


Fig. 10. Modified z-score distribution per peer group

- [9] P. Ortega, C. Figueroa, and G. Ruz, "A medical claim fraud/abuse detection system based on data mining: A case study in chile," vol. 6, 01 2006, pp. 224–231.
- [10] H. Shin, H. Park, J. Lee, and W. C. Jhee, "A scoring model to detect abusive billing patterns in health insurance claims," *Expert Systems with Applications*, vol. 39, no. 8, pp. 7441 – 7450, 2012. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0957417412001236>
- [11] M. Kirlidog and C. Asuk, "A fraud detection approach with data mining in health insurance," *Procedia - Social and Behavioral Sciences*, vol. 62, pp. 989 – 994, 2012. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1877042812036099>
- [12] T. Ekin, "Application of bayesian methods in detection of healthcare fraud," *Chemical Engineering Transactions*, vol. 33, 01 2013.
- [13] D. Thornton, G. van Capelleveen, M. Poel, J. Hillegersberg, and R. Mueller, "Outlier-based health insurance fraud detection for u.s. medicaid data," *ICEIS 2014 - Proceedings of the 16th International Conference on Enterprise Information Systems*, vol. 2, pp. 684–694, 01 2014.
- [14] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

Advanced analytics of Big Data using Power BI: Credit Registry Use Case

Fisnik Doko
Faculty of Informatics
“Mother Teresa” University
Skopje, Macedonia
fisnik.doko@unt.edu.mk

Igor Miskovski
Faculty of Computer Science and Engineering
“SS. Cyril and Methodious” University - Skopje
Skopje, Macedonia
igor.mishkovski@finki.ukim.mk

Abstract—Big Data is emerging trend which brings and the need for more effective data analysis and visualization to get new knowledge and to leverage the benefits of advanced analytics of the volume of data they collect without IT knowledge. Analyzing and visualizing large volumes of data in financial services often suffers from performances in traditional systems with traditional tools. For understanding the data through visualization, we have tried various approaches but this described in the paper with Power BI was the most efficient. This paper aims to provide use case of effective implementation of Power BI tools in banking, more specifically in Credit Registry database, using the methodology of Big Data analytics and the features of Power BI tool.

Keywords—Big Data, Power BI, Credit Registry

I. INTRODUCTION

With the exponential increasing of the amount of data, raises the need to understand trends in business and to gain important insights from the existing data. Different businesses need to understand analytical concepts using statistical methods, data prediction and machine learning [1]. These operations in the past were done just by developers, but now, these modern tools provide these abilities directly to people from the business. Microsoft Power BI is a tool for advanced analytics which enables normal users to use it and to leverage its capabilities for extracting useful knowledge from data, data visualizations and integration with R. Power BI empowers predictive analyzes by using machine learning without any previous programming.

Credit Registry is dataset that captures and persists financial information about credit borrowers and persists their history to contribute in improving the quality of loans and maintain the stability of banking system. This is one of the biggest datasets in central banks and one of the favorites for Big Data in central banks [2] [3]. In our case we have subset of the Credit Risk dataset of the Republic of North Macedonia. The data in the dataset is submitted by banks and saving houses and contains credit exposure data for the purposes of credit risk management. The Credit Registry is a collection of personal data, controlled by the central banks, which in our use case is anonymized and we work with extracted subset of the dataset [4].

This paper is organized as follows. In Section 2 we overview the Power BI features and our decision to use it. Then in Section 3 we describe the methodology and the process of performing the analysis workflow starting from the data source and ending with visualization. In Section 4 we describe our model created in Power BI Desktop, then in Section 5 analytics done on Power BI Service. The last

Section reviews the benefits and future work which needs to be done.

II. POWER BI

Power BI [5] is a business analytics tool that provides insights to enable fast and informed decisions. Power BI is a set of software services, applications, and connectors that work together to turn unrelated data sources into coherent, visual, and interactive displays. Data can be an Excel spreadsheet or a cloud-based hybrid data warehouse or local database. Power BI makes it easy to connect to data sources, visualize and discover important information, and share it through the web and mobile applications.

After reviewing multiple tools, we decided to use Power BI because of the high optimization and speed of data manipulation and analysis. In our example the 13GB database in SQL on import has been reduced to 330mb in Power BI .pbix format, since it has its own format that is adapted to handle big data [6]. Power BI supports around 115 types of data sources [7].

Our decision was done also according to Gartner [8], where Power BI is an application leader for business intelligence that has quickly become the most well-known and valuable to large corporations, surpassing Click and Tableau.

Power BI consists of:

- **Power BI Desktop** – Application for Windows desktop
- **Power BI Service** – Internet SaaS application named Power BI Service
- **Power BI Mobile** – Mobile application of Power Bi which works Windows, iOS and Android.

Power BI integrates with on-premises or cloud data sources, supporting all the data sources of Azure and many others.

When the data source changes not in a real-time manner, it is feasible to import the data in Power BI and then manipulate with top performances and easily analyze and visualize.

Large datasets which are more real-time data, can be leveraged by using Direct Query which get only needed data from the data source and efficiently display and refresh avoiding performance impacts.

Our data for this project is a subset of Big Data dataset which contains relational data, where the extraction is done in SQL Server.

Processing with Power BI starts with using Power BI Desktop when importing data, setting up relationships and visualization. Once the analysis is completed locally, it can be published on the SaaS version of Power BI on the cloud, where users can be granted access through Azure Active Directory user accounts. At the same time when the dataset, model and reports are published on SaaS, data, reports and displays are also available through the official mobile application. For a native mobile display, you need to edit your mobile layout via the desktop application.

Using this approach one can quickly and efficiently view and analyze analytics via desktop, web and mobile applications. Power BI can integrate reports into an external application through program code. To be able to embed external applications you need to have a Power BI Pro account with Azure Active Directory. The application needs to be registered in Azure Active Directory before it can access the Power BI REST API. Registering an application allows you to submit application identities. Power BI applications and their connections is shown in Fig. 1

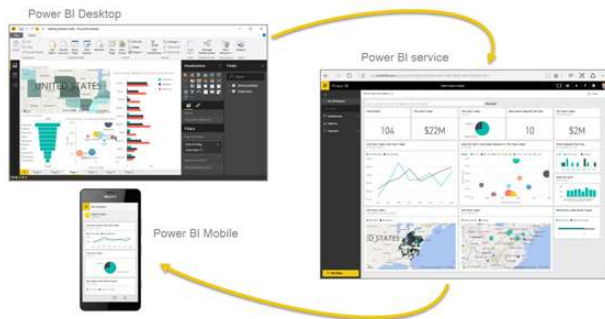


Fig. 1. Power BI applications and their connection

III. PROCESS OF PERFORMING THE ANALYSIS

For performing the analysis, we have performed several stages of data processing to achieve the final results. Initially the data is imported from multiple sources, edited and updated in Power BI Desktop.

After creating models, reports and dashboards they are finally published and made available on the web and mobile application.

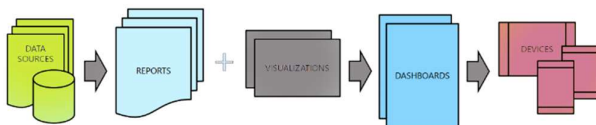


Fig. 2. Process of performing the analyses

The following is a detailed process from the data source to the final analysis.

A. SQL Server process

- **Analysis** - the SQL data source originally had more than 50 attributes. The minimum and maximum values are analyzed. It is important to note that due to legal changes over the years, not all fields have data.

- **Missing values** [9] - Missing values are initially located through SQL code, those that are few in a column are filled with the mean of the others for the corresponding attribute.
- **Feature selection** [10] - Locating the most important attributes is done with the help of fellow economists who are specialized in credit registry management. With their help, about 15 most important attributes are logged.
- **Feature engineering** – For improving the model efficiency we have derived new attributes from existing data [11].
 - **SizeOfBank** - derived column according to bank size code, which represents the size of the bank as a factor of three values.
 - **NumberOfLoans** – for every person using the SQL code we derived the number of loans in the current reporting period.
 - **DateStartLoan** - It has been found that the column for the first cash flow date has a rather illogical date due to the prior lack of control for that column, so a new column has been drawn that sets the credit start date at the time that credit party first appears in the database for the respective client.
 - **LoanDurationYears** - Derived column using the loan due date column.
 - **Age** - Age is derived for individuals through their identification number.

B. Excel process - create additional data sources

To improve the model and to set up and analyze with external factors, several sets of public data, which are an important economic factor, have been downloaded from the Internet.

- USD, EUR exchange rate list, to see how it affects the number of loans and their exposure.
- GDP data expressed in dollars. The purpose is to see how GDP is linked to the issued loans.
- Table of the municipalities of the Republic of North Macedonia, which has the number of municipalities required for merging and displaying the municipality's name for the loans.

C. Power BI Desktop process

- **Importing Data from SQL & Excel** - The data is imported into Power BI, a process that initially takes up to 20 minutes. The imported data is no longer dependent on the SQL and Excel sources. When changing data sources, Power BI only refreshes the data and retrieves it from scratch.
- **Creating Calculations** - Calculations are derived columns that are additionally calculated at the level of each row.
- **Creating Measures** - Measures are amounts at the table level, not at the level of each row. To create new information from data we used Data Analysis Expressions (DAX) which is a set of functions,

operators and constants that can be used in a formula or expression.

- **Creating hierarchies** - To enable drill down functionality, hierarchies of attributes have been created.
- **Model Design and Relationships** - In order all columns to behave as if they were in the same table, the model was created by merging all source tables through relationships.
- **Create Visualization Reports** - Reports are created using ready-made visualizations, then the corresponding columns displayed during the analysis are configured. For visualizing numeric attributes outliers, we installed additional extension from the AppSource to plot outliers with Box Plot.
- **Forecasting** - Power BI has built-in analysis using Linear Regression.
- **Creating views for mobile** - For neat display of mobile devices, the views for viewing from mobile phones are also designed.
- **Publishing Power BI Service** - After completing all the analysis and modeling locally, the dataset along with all visualizations are published and uploaded on the online cloud version of Power BI named Power BI Service.

D. Power BI Service process

- **Dashboard** - By selecting important visuals and displaying them on a new page or adding most important ones to dashboards.
- **Get insights** - Power BI Service has the ability to search on its own to learn new knowledge and visualizations of data. This knowledge is not always helpful. The Quick Insights feature automatically analyzes all the relations of data applying sophisticated algorithms to provide insights.
- **Ask question** – This is an option that allows writing queries in English and strives to produce results according to source columns and data.
- **Web publishing** - Ability to place graphs on external sites, but such access will also require external sites to have the user sign in details.

E. Power BI Mobile process

- **Mobile application testing** - You need to install the native Power BI Mobile application, and by logging into your Power BI account, all visualizations and views are displayed appropriately for mobile view.

IV. ANALYSIS AND PROCESSING WITH POWER BI DESKTOP

The full analysis is done with Power BI Desktop, the main application of the set of the Power BI tools. An important part is pre-processing process and then designing the model. Our model which is kind of a Star schema (see Fig. 3) and all the source charts are merged with the main credit registry dataset [12]. For having time series and grouping by months, quartiles and years we have used additional excel sheet which is imported and linked in the model.

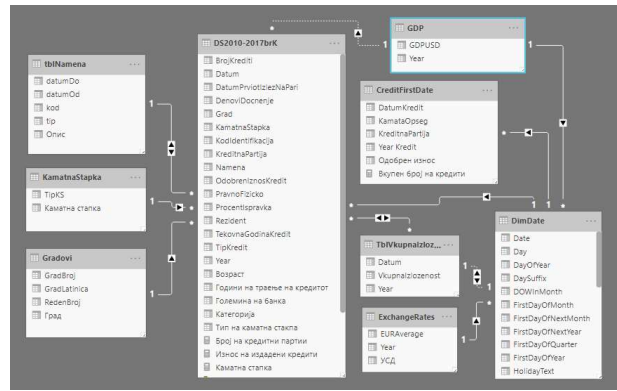


Fig. 3. Design of the model

After designing the model and creating new calculations and measures, the next step is to place visualizations on each side (report) and adjust them separately. It is beneficial that after configuring all visualizations separately, when selecting some information in one visualization (chart) it is updated in all visualizations in the level of report for the selected feature or range. Also for some reports with time series, is enabled and drilldown functionality.

The tool is powerful for finding dependencies between different attributes [13]. The following is an analysis of how the risk category depends on the day-to-day installment payment delay, and on the percentage of impairment entered by the banks themselves. The tool through Influencer visualization finds with high accuracy the range of days in which the client should be for every risk category, and also with more accuracy it detects the dependency for percentage of impairment. Figure 4 displays example that category is more likely to be B when the percentage of impairment is 8.9-25 with accuracy of 73.63%, and also when day-to-day installment payment delay is between 30-105 days.

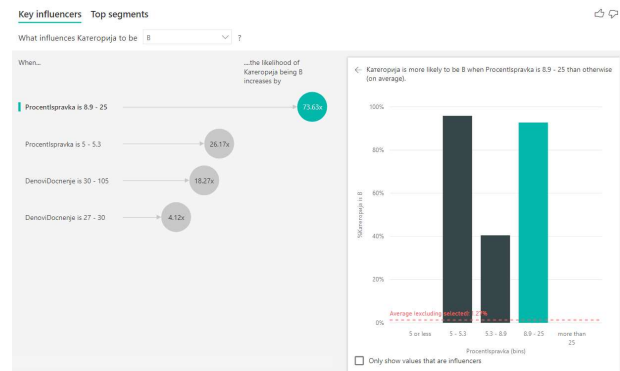


Fig. 4. Visualizing category dependencies through Influencer visualization

Business leaders can easily understand and work with data when they are in visualized. You can provide available visualizations in Power BI gallery or add custom visualizations including and R custom visualizations.

V. ANALYTICS ON POWER BI SERVICE

The Power BI Service enables full data upload and visualization of the Power BI cloud and its availability everywhere and at all times.

In the following there are some Figures of the solution and their description. Figure 4 shows analysis by age, number of loans by age of people and distribution of amount per age.

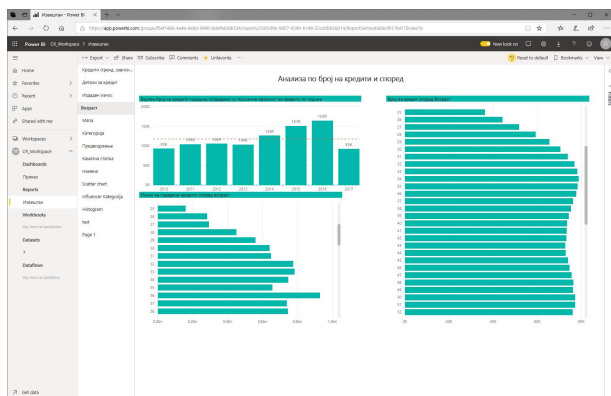


Fig. 5. Visualization by age, number of loans by age, and amount of loan by age

Figure 6 shows the distribution of number of loans per municipality of the country. The map is by Bing maps and using Power BI are displayed the sum of number of loans using green circles with appropriate size.

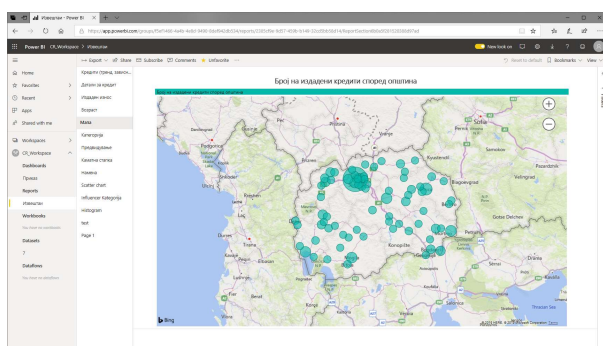


Fig. 6. Map of the municipalities in the Republic of Macedonia by number of loans

Figure 7 shows predicted values for period 2018-2020 using the data from previous layers. The prediction is integrated in Power BI and uses linear regression, which can be additionally configured.

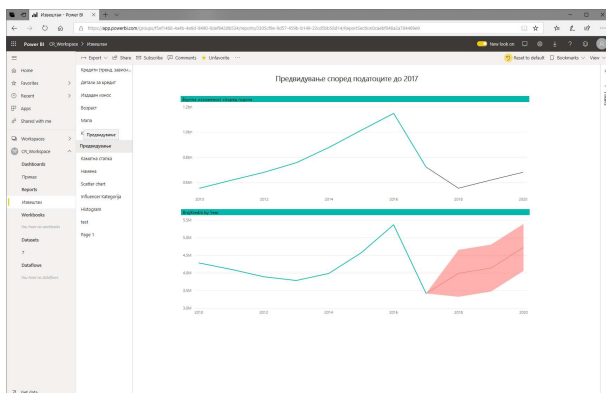


Fig. 7. Forecast visualization using Power BI feature, total exposure and number of credits per year.

VI. CONCLUSION

Power BI business analytics makes easier the analytics and visualization for all users of the company, gaining new insights and making effective advanced analytics. Using Power BI tool for Big Data Analytics is an effective way of discovering and visualizing knowledge in a very fast way, which unlike traditional tools is incomparably fast and powerful with modern capabilities for visualization, dependency and prediction. Dashboards in Power BI can transform the way people guide the business by supporting monitoring of social media, video streaming and real-time data.

With the help of Power BI one can achieve detailed insight and dependency analysis of the attributes themselves and correlation with external factors.

The project helped a lot to gain complete knowledge to the big data on the credit registry. The next step with this datasheet is to implement state-of-the-art machine learning algorithms for predicting and evaluating the results.

VII. REFERENCES

- [1] B. a. P. Z. Fang, "Big data in finance," *Big data concepts, theories, and applications.*, no. Springer, Cham, pp. 391-412, 2016.
- [2] BearingPoint, "Big data in central banks: 2017 survey," 2017.
- [3] C. B. M. P. J. L. & S. Altavilla, "Banking supervision, monetary policy and risk-taking: Big data evidence from 15 credit registers.," 2020.
- [4] Microsoft, "Advanced Analytics with Power BI White Paper,"
- [5] A. Aspin, *Pro Power BI Desktop*, Apress, 2016.
- [6] S. M. Ali, "Big data visualization: Tools and challenges," in *2nd International Conference on Contemporary Computing and Informatics (IC3I)*, 2016.
- [7] "Power BI data sources," Microsoft, 10 03 2020. [Online]. Available: <https://docs.microsoft.com/en-us/power-bi/power-bi-data-sources>. [Accessed 13 03 2020].
- [8] Microsoft, "Microsoft Power BI Blog," [Online]. Available: <https://powerbi.microsoft.com/en-us/blog/microsoft-named-a-leader-in-gartners-2020-magic-quadrant-for-analytics-and-bi-platforms/>. [Accessed 09 03 2020].
- [9] A. C. Acock, "Working with missing values," *Journal of Marriage and family*, vol. 67.4, pp. 1012-1028, 2005.
- [10] J. Li, "Feature selection: A data perspective," *ACM Computing Surveys*, vol. 50(6), pp. 1-45.
- [11] M. R. a. M. C. Anderson, "Input selection for fast feature engineering," in *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, 2016.
- [12] P. a. P. G. Cerchiello, "Big data analysis for financial risk management," *Journal of Big Data (2016)*, vol. 3.1, 2016.
- [13] C. S. C. a. L. G. Cantú, "How Do Bank-Specific Characteristics Affect Lending? New Evidence Based on Credit Registry Data from Latin America," in *Bank for International Settlements (BIS)*, 2019.

The challenges of key-value stores

Gjorgjina Cenikj

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

Dushica Jankovikj

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

Oliver Dimitriov

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

Abstract— The escalation of flexibility, scalability, and elasticity demands for data storage solutions has implored the quest of finding an alternative to the traditional relational databases, which in turn, lead to the popularization of NoSQL databases. This paper takes a deeper look into the key-value stores, and accentuates their strengths and weaknesses. The performance of Redis, LevelDB, and Oracle NoSQL is compared to that of the PostgreSQL database. The results affirm the inability of key-value stores to achieve comparable performance on queries that are typically performed on relational databases, such as inserting and deleting records, sorting and querying that involves several join operations. The results indicate that LevelDB outperforms Redis and Oracle NoSQL in most scenarios. The evaluation of the modeling capabilities of each database reveals additional challenges that one needs to be aware of when choosing which key-value store is best suited for their requirements.

Keywords— NoSQL databases, key-value, Redis, LevelDB, Oracle

I. INTRODUCTION

Key-value stores are one example of NoSQL databases, which use method based on key-value pairs, where each key uniquely identifies the value. Both the keys and the values can be of different data types, from simple strings, to complex BLOBs, depending on the concrete implementation.

The main advantage of key-value databases is their high partitioning and horizontal scaling potential. Due to the fact that most of them do not use a predefined schema, they are considered to be one of the most flexible NoSQL database types, which give the application a complete control over the stored data, with next to no restrictions. Furthermore, since no placeholder values are assigned for the optional attributes, they are more efficient as far as the memory needed to store the data is concerned.

One of the major disadvantages of the typical key-value databases is the inefficiency of querying through the values. The standard implementations do not use a query language, but instead merely provide options for key-value pair addition and removal. Extending the possibilities of manipulating the data using conventional SQL queries imparts the benefits of speed, availability, and scalability, while keeping the familiar relational language.

According to the CAP theorem, it is impossible for a distributed computer system to simultaneously provide consistency, availability and partition tolerance. While relational databases guarantee consistency and availability, the key-value databases guarantee availability and partitioning tolerance, while sacrificing consistency.

Some of the most well known key-value databases are: Dynamo, Hazelcast, LevelDB, Riak, Redis, Oracle, Voldemort и RocksDB. In this project, we take a look at LevelDB, Redis and Oracle NoSQL. We provide a short

overview of the capabilities of each dataset in Section 2, and present the most related research in Section 3. A detailed description of the databases, method and proposed models is provided in Section 4, the discussion of the evaluation results are presented in Section 5, while Section 6 concludes the study.

II. BACKGROUND

A. Redis

Redis is an in-memory data structure project implementing a distributed, in-memory key-value database. Redis supports different kinds of abstract data structures, such as strings, lists, maps, sets, etc. The basic functionalities it offers are put, get and delete, i.e storing, retrieving and deleting key-value pairs.

Although Redis is an in-memory database, it can persist data on disk. However, the disk storage format is not suitable for querying, as its sole purpose is reconstructing the data in memory, once the system is rebooted.

B. LevelDB

LevelDB is an embeddable, light-weight key value store developed by Google, which offers support for many programming languages such as C++, NodeJS and Java. This database stores the keys and values on-disk in a sorted fashion in the form of byte arrays. The data itself is organised by the core storage architecture which is represented by a log-structured merge tree (LSM). This type of write-optimised storage system makes levelDB appropriate for large sequential (batch) updates instead of sparse random writes.

LevelDB is not a SQL database, signifying it doesn't have a relational data model, nor does it support SQL queries or indexes. The key/value store supports a simple flat mapping from a key to a value, so the creation of data models and data relational handling is considered only in the higher abstraction levels.

LevelDB is published with respect to the New BSD License which ensures that the wider technical community can use this storage engine. This database optimises storage size and bandwidth predominantly with the utilization of compression algorithms such as Google's Snappy and LZ4. The compression showed useful for the research paper since we were dealing with raw text such as JSON and XML. A possible disadvantage of LevelDB is that it requires quite a few disk seeks for information retrieval.

C. Oracle NoSQL

Oracle NoSQL is a distributed database that is typically applied in web applications, where it is used as the primary or as an auxiliary backend database. It supports the majority of the most popular data types. The latest versions provide the option of using tabular structures with an additional level of abstraction that simplifies the modeling and querying with the use of the familiar SQL syntax.

The concept of tables in Oracle NoSQL is to some extent similar to that of SQL tables, in the sense that the tables are composed of rows with predefined, named columns. Each table must have at least one attribute that forms part of the primary key, which uniquely identifies each table row. Some of the methods that refer to multi row operations, allow or even demand the use of partial primary keys, which are keys where the values of some of the key's attributes are not specified.

Contrary to the SQL implementation, each table can have nested, indexable child tables with an unlimited number of rows. The primary key of each child table is composed of their parent's and their own primary keys, meaning that each table implicitly contains the primary keys of all of their ancestors. There are no limitations in regards to the number of child tables present, nor the amount of nesting possible, i.e. the depth of the hierarchy.

Alternatively, lower-level data can be represented with the Record data type, and it is recommended to use this approach in the case of a fixed and low number of attributes. Indexes are an alternative way of querying tables through values that are not necessarily part of the primary key. Using indexes, it is possible to query rows that have different primary keys, but share another common characteristic.

III. RELATED WORK

Redis has previously been used as a representative of the key-value family of NoSQL databases, to compare their functionalities to traditional relational databases [1]. While this work has a wider scope, and takes into account all types of non-relational databases, it is primarily concerned with their general capabilities, without going into too much detail about any particular type. Our work is completely focused on the key-value databases, and the difference between their concrete implementations.

Apart from this, Redis' capabilities have also been compared to Riak's, [2] and it has been pointed out that Redis excels in situations where the data is subjected to rapid changes.

A study that bears more similarities to ours, [3] compares the performance of Redis and Oracle NoSQL, and finds that while Oracle NoSQL provides more convenient querying methods, Redis consistently outperforms it in terms of execution time. While this paper also investigates document stores and extensible record stores, ours takes into account one additional key-value database, LevelDB.

IV. METHODOLOGY

We setup an instance of each database and consider several different data representation models, in order to be able to fully explore the modeling capabilities offered by the corresponding database. All of the models are based on the same dataset, which is necessary for the comparison of their performance on different types of queries, which is presented in the next section.

A. Dataset

The used dataset refers to completed auctions, where each row contains data about the seller, item, auction, shipping, payment types, buyer protection, and bid history. The original dataset was given in 4 XML files: '321gone', 'ebay', 'ubid' and 'yahoo'. As a preliminary step, each file was converted to JSON format, which is better suited for reading with java

libraries. Owing to the fact that all of the fields were originally of type string, additional preprocessing was required before the database insertion took place.

B. Redis

For the purpose of the paper, we made a Java application in order to import the data and run some queries on it. The project setup is straight forward, all that is required is to add a dependency for the Redis Client and make a connection with the Redis-server. In order to import the data in Redis, we transformed the given datasets from xml into key-value pairs, where the keys have the format shown below:

```
$database_name:$table_name:$primary_key_value:$attribute_name
```

The key is made of four main parts. The first two parameters are the database and table name. After them comes the primary key, and finally separated with colons are listed the names of the attributes.

C. LevelDB

For the purposes of this research paper the Java implementation of the database was utilised and tested. The api used for this project is part of the project dain/leveldb which offers the core functionalities: get(key), put(key,value), and delete(key).

The entity of interest in this research paper is Listing. For the handling of its nested structure and related attributes, three experimental models were built:

- **L_natural model** - This model is the most straightforward approach. Here, every complex object within the nested structure of the Listing entity is represented as a separate class. Even though this is an intuitive approach, it might be an overcompartmentalisation of an otherwise rather simple entity with not that many depth levels. The query execution for this model is simple, since no serialisation is included in the nested classes.
- **L_flat model** - This model is an example of an oversimplification. Here only the leaf nodes of the xml object tree are included as attributes of the Listing model. Even though this increases the accessibility of data for certain keys, some structural information is lost in the process. To an individual familiar with the structure and key characteristics, the performance of this model and the previous one is identical.
- **L_formatted model** - A model identical to the first one when talking about object structure, but different in the data that it carries. This model ensures input data goes through a formatting phase before being written in the database. This makes the database less error prone and value data types become a universally known fact. The downside here is that some data types are not primitive and might require serialization in order to be written as a byte array value for a key - such as lists, sets, maps.

Aside from the previously mentioned models, some further experimentation was executed by the inclusion of serialisation of objects of deeper levels. What this means is

that the depth of an entity is being decreased by representing the class attributes of non primitive data types as serialised java objects. This process gives a flatter structure, with a major downside - time consuming value extraction which has detrimental effects on performance.

D. Oracle NoSQL

Several models were considered for structuring the data in the OracleNoSQL database. All of the models use the appropriate data type for each field, which required some additional preprocessing effort before the database imports, since the original dataset represented all of the information as strings.

The O_flat model is obtained by moving all of the attributes that would represent leaves in the json tree of the dataset into the root element, listing. This way, the table contains only attributes of simple types, without any further nesting. The column names were created so that they represent the hierarchical organization of the original structure, so that queries can be easily executed if one is familiar with this nomenclature.

The two remaining models, named O_record and O_child, use the Record data type to represent the seller_info, bid_history and item_info fields. They only differ in the representation of the auction_info field. The O_record model relies on the use of the Record data type to represent the auction_info field, while the O_child model represents it using a child table, where all of the attributes are represented using simple data types. The keys of the parent listing table are implicitly added to the auction_info child table, as one would add foreign keys in a relational database.

V. RESULTS

For the purposes of comparison with relational databases, two baseline models were created using PostgreSQL. The P_flat model is created by placing all of the data in a single table, so that no joins are required in order to retrieve the data. The P_nested model involves placing every JSON object in its respective table. This way, the tables item_info, seller_info, high_bidder, bid_history, auction_info and finally listing_nested were created. The P_nested model has foreign keys to every other respective table, which might be redundant since all relations are one-to-one.

The experiments were conducted on machines with 16GB of RAM and i7 processors, with the difference of the machine used for the Redis experiments having 2 cores and 4 threads, as opposed to the 4 cores and 8 threads on the machine used for the LevelDB and Oracle NoSQL experiments. We measured the time needed for importing the data, and executed 8 additional queries that are standard filtering (Q1-Q6) and sorting queries (Q7,Q8) that are typically executed on relational databases. Fig.1 features an overview of the performance of the previously presented models and databases, measured as the time in milliseconds needed to import the data and execute the queries.

TABLE I. QUERY EXECUTION TIME FOR EACH PROPOSED MODEL

Query	Database			
	Oracle NoSQL	Redis	LevelDB	PostgreS QL
Import	O_flat: 3001 + 269	469	L_natural: 94.66	P_flat:

	O_record: 2983 + 293 O_child: 5115 + 414		L_flat: 91.29 L_formatted: 103.42	64 + 472 P_nested: 116 + 84
Q1	O_flat: 421 ^O , 30 ^J O_record: 189 ^J O_child: 267 ^J	48	L_natural: 29.04 L_formatted: 39.43	P_flat: 38 P_nested: 37
Q2	O_flat: 404 ^O O_record: 168 ^J O_child: 427 ^O	49	L_natural: 28.35 L_formatted: 35.27	P_flat: 47 P_nested: 32
Q3	all models: 144 ^O , 15 ^J	23	L_natural: 39.3 L_formatted: 28.59	P_flat: 44 P_nested: 31
Q4	O_flat: 441 ^O , 13 ^J O_record, O_child: 22 ^J	150	L_natural: 44.7 L_formatted: 31.23	P_flat: 31 P_nested: 25
Q5	all models: 168 ^J	39	L_natural: 31.24 L_formatted: 25.55	P_flat: 31 P_nested: 33
Q6	O_flat: 396 ^O , 39 ^J O_record: 140 ^J O_child: 391 ^O , 33 ^J	42	L_natural - 26.63 L_formatted: 26.06	P_flat: 36 P_nested: 31
Q7	O_flat: 153 ^J , 274 ^J O_record: 162 ^J O_child: 363 ^J , 630 ^O	144	L_natural: 30.67 L_formatted: 20.2	P_flat: 47 P_nested: 33
Q8	O_flat: 145 ^J , 260 ^J O_record: 20 ^J O_child: 356 ^J , 633 ^J	152	L_natural: 33.88 L_formatted: 21.53	P_flat: 47 P_nested: 32

Fig. 1. Execution times in milliseconds required for the data import and 8 executed queries, for all of the proposed models and databases. The superscripts next to the results of the Oracle models refer to the time required when using the SQL syntax (O), java (J) or previously created indexes (I).

It comes as no surprise that the relational database outperforms the non-relational ones, since key-value stores are not meant to be used when accessing the data through non-key values is required.

A. Redis

In order to get a result for a query, one needs to manipulate the data. In our project, all manipulation and actions with the data in Redis and LevelDB was made using Java code, since the databases themselves provide no operations other than basic retrieval by key, and there is no alternative way to execute the queries.

The average execution time for the queries run on Redis is 80.87ms, the lowest one is 23ms and the highest one is 152ms. It is interesting to point out that while LevelDB and Redis have roughly similar execution times for the select queries, Redis is somewhat slower when it comes to the database creation and sorting queries.

B. LevelDB

Out of the three databases explored in this paper, LevelDB has the best performance on most of the queries, and manages to outperform the baseline PostgreSQL on the sorting queries and some of the filtering queries.

As far as the different proposed models are concerned, it can be noticed that the L_natural model that does no data

preprocessing executes the database population faster than the L_formatted model, because no time is spent on serialization. On account of this, the L_formatted model performs better on the data retrieval queries, because there is no need for formatting the data on the fly.

C. Oracle NoSQL

Oracle NoSQL does support the execution of some of the queries, and therefore, its results are accompanied by a superscript, which indicates the type of approach that the specified execution time refers to.

When analyzing the time needed to import the data, one can clearly note the significantly higher amount of time required by the O_child model. This is due to the fact that, unlike the other models, O_child requires the creation of an additional table for representing the auction_info.

A slightly less obvious observation is the lack of a result referring to the performance of the O_record model using a secondary index. The absence is a consequence of the inability to create indices based on the values encapsulated in a Record field, and it is one of the main disadvantages of this model. A related issue is the inability to query through these values, which is the reason why queries Q1, Q2, Q4, and Q6 had to be executed using java code.

The O_flat model supports the execution of all of the queries using Oracle's SQL syntax, but this model sacrifices the original data structure.

As far as execution time is concerned, the O_child model seems to yield the most inferior performance out of the 3 Oracle models. However, this model would be more memory efficient in the case of missing auction_info values, since both

of the other models would store null values for all of its fields, while the O_child would waste no memory resources. This is also the only model that could represent a one-to-many relationship, if needed.

While it is rather obvious that the Table API used for the Oracle NoSQL database is slower than the other two databases, it is worth noting that the syntax the API provides is significantly more compact and convenient to anyone familiar with SQL.

VI. CONCLUSIONS

Through the review of Redis, LevelDB and Oracle NoSQL as representatives of the key-value databases, this work once again reaffirms the strengths, and especially the weaknesses of this type of non-relational databases. The obtained results indicate that the Table API used for the Oracle NoSQL database is slower than the Redis and LevelDB implementations. While relational databases outperformed on the dataset used in our study, the empirical results have noteworthy implications for future research that could involve the use of a different and larger dataset, analyzing additional queries and experimenting with different data models.

REFERENCES

- [1] M. Radoev, "A Comparison between Characteristics of NoSQL Databases and Traditional Databases," *Computer Science and Information Technology* 5.5, 2017, pp.149 - 153.
- [2] C.A. Baron, "NoSQL Key-Value DBs Riak and Redis", *Database Systems Journal*, 6, issue 4, 2015, pp. 3-10.
- [3] F. Bugiotti and L. Cabibbo, "A comparison of data models and APIs of NoSQL datastores," *21st Italian Symposium on Advanced Database Systems*, 2013, pp. 63-74.

Analysis of Feature Selection Algorithms on High Dimensional Data

Sowmya Sanagavarapu
Department of Computer Science and
Engineering
College of Engineering Guindy, Anna
University
Chennai, India
sowmya.ruby7@gmail.com

Mariam Jamilah
Department of Civil Engineering
College of Engineering Guindy, Anna
University
Chennai, India
mariamjamilah24@gmail.com

Barathkumar V
Department of Computer Science and
Engineering
College of Engineering Guindy, Anna
University
Chennai, India
barathkumarv98@gmail.com

Abstract— Dimensionality of a dataset refers to the number of attributes present in the dataset. At times, the number of attributes is greater than the number of observations, this gives rise to high dimensional data. In high dimensional data, the dimensions are so high that calculations become extremely difficult and this in turn increases the processing and training time. Thus, it is vital to reduce the dimensionality of data [1]. Dimensionality reduction means to simplify the data without affecting data integrity. For this study, we have taken the Dorothea dataset [10] from UC Irvine Machine Learning Repository. Dorothea is a drug discovery dataset. Drugs are organic molecules that bind to a target on a receptor, they are classified as active or inactive based on their ability to bind. New drugs are formed usually by identifying and isolating the receptor to which the chemical compounds have to bind. Then many small molecules are tested for their ability to bind to this receptor. The class label shows whether the molecule will bind to the drug or not. In this paper, we investigate the dimensional reduction achieved by applying three Feature Selection algorithms [2]- Filter, Wrapper and Hybrid with no loss in the integrity of the dataset. We evaluated the accuracy of the obtained data using a C4.5 Classification algorithm [6]. It is used to predict categorical class label of the dataset after training it using the training dataset. The results of each algorithm [1] have been compared and analyzed in order to arrive at the best suited algorithm.

Keywords— classifier, relief filter, hybrid, Las Vegas wrapper, test data, training data

I. INTRODUCTION

A data is said to have high dimensionality when the number of attributes is greater than the number of observations in a dataset. Such data is hard to handle and hence it is important to reduce the dimensionality of the dataset without any drop in the accuracy with which the class label is predicted. If the reduction in the dimensionality of data results in decrease in accuracy, then the decreased accuracy counterweights the reduction. “Curse of dimensionality” [8] means that large data doesn’t necessarily have to have a good accuracy in the prediction of the class label of the dataset. The dataset used in the project is from the UC Irvine Dorothea drug testing dataset [10]. There were a significant number of missing values in the dataset.

A raw dataset has dirty data [3] -that is, the data contains missing values and noise- in it. These anomalies make the data inconsistent, harder to process and give us inaccurate results upon processing. To avoid this, data preprocessing is performed to obtain a clean dataset- a dataset with little noise in it. The obtained dataset with reduced number of features after feature selection gives more accurate results than the original dataset, consequently reducing the run time. On a

clean dataset, the various feature selection algorithms are applied to obtain a dataset with reduced number of attributes. To this dimensionally reduced dataset, the classification algorithm is applied to check the accuracy with which the class label of the dataset is predicted.

The dataset used in the project is from the UC Irvine Dorothea drug testing dataset [10]. There were a significant number of missing values in the dataset. The preprocessing techniques were applied to this dataset.

Data processing [7] is first begun by removing the missing values from the dataset. There are a few methods through which this can be achieved. This fully filled dataset is subjected to further processing to smoothen out the noise in the data. Here, the binning method [8] has been employed to smoothen out the noise in the data. The data that results from the preprocessing stage is complete and consistent. Next, the Feature Selection is applied to obtain a reduced feature subset. Here the three methods of feature selection that have been applied are- the Filter Method, the Wrapper Method and the Hybrid Method ([1], [2], [8]). The preprocessing techniques were applied to this dataset. A Classification algorithm [6] is employed to find the accuracy with which the class label is predicted in the dataset. After obtaining each reduced dataset, the accuracy with which the class label of this dataset is predicted is calculated and compared with that of the original dataset.

A Classification algorithm [6] is employed to find the accuracy with which the class label is predicted in the dataset. After obtaining each reduced dataset, the accuracy with which the class label of this dataset is predicted is calculated and compared with that of the original dataset.

The paper is organized as follows. Section II, discusses the preprocessing steps performed on the dataset. In Section III, the Feature Selection algorithms are explained. Section IV, where the classification algorithm has been used. Section V displays the results obtained from each of the Feature Selection Algorithm on the dataset. Finally, we conclude the paper and discuss future work.

II. DATA PREPROCESSING

The dataset chosen initially contained dirty data -meaning incomplete, noisy and inconsistent data. The dataset had missing values: lacking attribute values or certain attributes of interest or/and noisy and inconsistent data: containing errors, discrepancies or outliers.

Missing values may be due to

- Equipment malfunctioning

- Inconsistent with other recorded data and thus deleted
- Data not entered due to misunderstanding
- Certain data may not be considered important at the time of entry

In case of missing values, we can do any one of the following

1) *Manually enter the value*

By manually entering the missing values in the data, we are giving way to data discrepancies in the data. This doesn't necessarily ensure a good accuracy of data.

2) *Assign a constant value*

The global constant of the attribute in the data is chosen to fill the missing values.

3) *Eliminate rows with missing values*

This method is not preferred as a high number of attributes may be lost thus resulting in a very much reduced feature subset.

4) *Fill the missing values with the mean of the column*

The mean value of the non-missing values of an attribute is calculated and the missing values are filled with the mean value of that attribute. To find the mean of the attribute, an inbuilt mean function is used.

5) *Fill the missing values with the median of the column*

The median value of the non-missing values of an attribute is calculated and the missing values are filled with the median value of that attribute. To find the median of the attribute, an inbuilt median function is used.

6) *Fill the missing values with the mode of the column*

The mode value of the non-missing values of an attribute is calculated and the missing values are filled with the mode value of that attribute. To find the mode of the attribute, a mode function was coded and used to find the mode value of an attribute.

Noisy and inconsistent data [9] maybe due to

- faulty data collection instruments
- data entry problems
- data transmission problems
- technology limitation

For reducing the noise in the dataset binning method [8] is used. In binning method, the data values are placed into equal width bins. The bins are set to accept data values between two limits. The limit is decided depending upon the range of the values in that attribute.

The limits of the bins are to be set such that the lowest and the highest limit are also taken into consideration. The noisy data is then smoothened by applying the mean, median or boundary of the bins to the data values in the bin.

1. Binning by mean

Here the mean of the bin limits is applied to the data values in the bin to smoothen out the noise.

2. Binning by median

Here the median of the bin limits is applied to the data values in the bin to smoothen out the noise.

3. Binning by boundary

Here the upper or the lower boundary of the bin limits is applied to the data values in the bin to smoothen out the noise.

The missing values in the Dorothea dataset were handled by filling them with binning by median value in the column. This ensured that the extreme noises observed in some features didn't decrease the integrity of the dataset, as it would have happened with using the mean of a feature.

III. FEATURE SELECTION ALGORITHMS

Feature Selection [5] is the process of selecting a subset of relevant features for further processing. The main concern when using a feature selection technique is that the data contains some features that are either redundant or irrelevant, and can thus be removed without incurring much loss of information.

A. Filter method

Filter feature selection methods apply a statistical measure to assign a score to each feature. The features are ranked by the score and either selected to be kept or removed from the dataset. The methods are often univariate and consider the feature independently, or with regard to the dependent variable.

Here, the Relief Filter Algorithm [1] is applied to the dataset. Filter method selects the feature subset on the basis of the intrinsic characters of the data, independent of the machine learning algorithm. Among that, the Relief algorithm is considered one of the most successful algorithms used for assessing the quality of features due to its simplicity and effectiveness. The key idea of Relief Algorithm [1] is to iteratively estimate the feature weights according to their ability to discriminate between neighboring patterns. In this algorithm, a random object is selected from the dataset and the nearest neighboring sample with the same class label (nearHit) and different class label (nearMiss) are identified. The nearest neighboring sample of an object means the object with the greatest number of features having same value. The weight of each feature is updated by:

$$W_j = W_j + |x(i) - nM(i)(x)| - |x(i) - nH(i)(x)| \quad (1)$$

After running the for loop for the number of iterations, the weight threshold value is calculated by taking the mean of weight of all the features. Then select the features which are having the weight greater than the threshold value.



Fig. 1. Flow of the Relief Filter Algorithm

```

Algorithm 1: Relief Filter Algorithm
input : O, the set of all objects; C, the set of all conditional
        features; its, the no. of iterations;  $\epsilon$ , weight threshold value
output: R, the feature subset
R = {} for  $W_s$  where  $W_s=0$  do
  for  $i$  in its do
    choose an object  $x$  in O randomly;
    calculate  $x$ 's nearHit (nH) and nearMiss (nM) ;
    for  $j$  in range(1, |C|) do
       $W_j = W_j + |x(i)-nM(i)(x)| - |x(i)-nH(i)(x)|$  ;
    for  $j$  in range(1, |C|) do
      if  $W_j \geq \epsilon$  then
         $R = R \cup j$  ;

```

Features of Filter Method:

- 1) Relief method considers all attributes into consideration.
- 2) It doesn't consider the relationship between two attributes.
- 3) Eliminates attributes by comparing with threshold value.
- 4) The subset selection is done only a single time.

B. Wrapper Method

Wrapper methods [2] consider the selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared to other combinations. A predictive model is used to evaluate a combination of features and assign a score based on model accuracy. As wrapper methods train a new model for each subset, they are very computationally intensive, but usually provide the best performing feature set for that particular type of model. At times, Filter methods have been used as a preprocessing step for wrapper methods, allowing a wrapper to be used on larger problems. Here, the Las Vegas Wrapper Algorithm ([1], [2]) is applied to the dataset.

The Las Vegas Wrapper algorithm uses random subset creation that guarantees given enough time, the optimal solution will be found. It produces intermediate solutions while working towards better ones that result in a lower classification error.

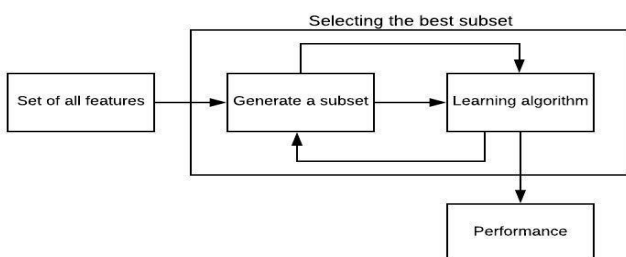


Fig. 2. Flow of the Las Vegas Wrapper Algorithm

```

Algorithm 2: Las Vegas Wrapper Algorithm
input : C - the set of conditional features; K - update threshold;  $\epsilon$  -
        error threshold O
output: R, the feature subset
R = C;
k = 0;
while  $\epsilon$  not updated for K times do
  T = randomFeatureSubset();
   $\epsilon_t = \text{learn}(T)$ ;
  if  $\epsilon_t < \epsilon$  or  $\epsilon_t == \epsilon$  and  $|T| < |R|$  then
    return T;
    k = 0;
     $\epsilon = \epsilon_t$ ;
    R = T;
  k=k+1;
   $\epsilon = \text{learn}(R)$ ;
return R;

```

In this algorithm, initially the full set of conditional features is taken as the best subset. The algorithm then generates a random feature subset and evaluates the error threshold ϵ_t using an inductive learning algorithm, here C4.5 [6] is used. It compares ϵ_t with ϵ ,

- if $\epsilon_t < \epsilon$ or $\epsilon_t == \epsilon$ and $|T| < |R|$, then ϵ_t becomes the new ϵ , T becomes the new R and k becomes 0 and then the algorithm continues to generate random subsets.
- if $\epsilon_t > \epsilon$ or $\epsilon_t == \epsilon$ and $|T| > |R|$, the algorithm continues to generate random subsets until K times.

This algorithm requires two threshold values to be supplied: ϵ , the classification error threshold and the value K, used to determine when to exit the algorithm due to there being no recent updates to the best subset encountered so far.

Features of Wrapper Method:

- 1) Not all attributes are considered.
- 2) A varying number of attributes are randomly chosen and the mean absolute error is compared for the different subsets.
- 3) The relationship between attributes is considered here.
- 4) The subset is selected a number of times as per requirement.

C. Hybrid Method

The Hybrid Algorithm [1] is defined as a combination of both Iterative Relief Filter method [2] and Las Vegas Wrapper method [2]. The complete and consistent dataset is considered and the Iterative filter method algorithm is applied to it. With the feature subset obtained thus, we apply the Las Vegas Wrapper method [2] to it to get a further reduced feature subset. Thus, with the now doubly reduced feature set, we test its accuracy by applying the C4.5 Classification algorithm [6] and the results are recorded.

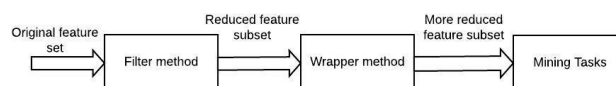


Fig. 3. Flow of the Hybrid Algorithm

IV. CLASSIFICATION ALGORITHM

A Classifier [6] is a tool in Data Mining that takes a bunch of data representing things we want to classify and attempts to predict which Class Label the test data belongs to. Classification is used to find out in which group each data instance is related within a given dataset. It is used for classifying data into different classes according to some constraints. Classification techniques in data mining are capable of processing a large amount of data. It can be used to predict categorical class labels and classifies data based on training set and class labels and it can be used for classifying newly available data. Several major kinds of classification algorithms include C4.5, ID3, K-nearest neighbor classifier, Naive Bayes, SVM, ANN etc. Here, C4.5 [6] was used as the Classification algorithm. C4.5 is used because of its quick classification and high precision. C4.5 is a Classification algorithm that is used to produce a decision tree which is an expansion of prior ID3 calculation. It enhances the ID3 algorithm. C4.5 creates decision trees from a set of a training data same way as an ID3 algorithm. Decision tree learning creates something similar to a flowchart to classify new data. C4.5 uses greedy (non-backtracking) approach in which decision trees are constructed in top – down recursive divide and conquer manner. C4.5 algorithm is a supervised learning algorithm as it cannot learn on its own. For this it was trained by using the Training dataset. The algorithm analyzes the training set and builds a decision tree and now it uses the decision tree to classify. As the decision tree is being built, the goal of each node is to decide the split attribute (feature) and the split point that best divides the training instances belonging to that leaf.

V. ANALYSIS

The accuracy of the preprocessed high-dimensional dataset was observed to be at 94.25% with 6061 features. The feature algorithms were applied on this dataset and results are analysed.

A. Iterative Relief Filter Algorithm

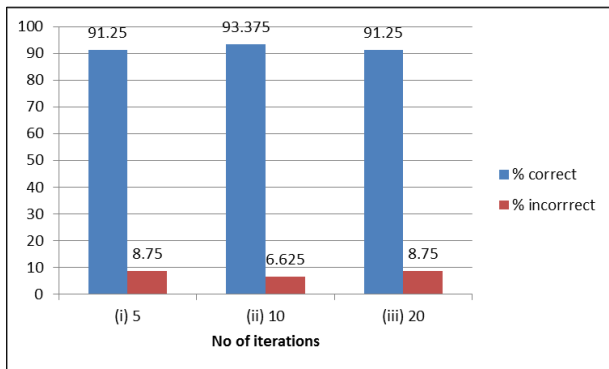


Fig. 4. Iterative Relief Filter Algorithm for varying “its”

From the Fig. 4. - (i) we infer that, after applying the Relief Filter algorithm where the number of iterations is 5, the number of attributes has reduced from 6061 to 852 and the percentage of correctly classified attributes thus obtained is 91.25%. Therefore with 852 attributes we can obtain an accuracy of 91.25%. Hence dimensionality reduction is achieved.

From the Fig. 4. - (ii) we infer that, after applying the Relief Filter algorithm where the number of iterations is 10, the number of attributes has reduced from 6061 to 1112 and

the percentage of correctly classified attributes thus obtained is 93.375%. Therefore with 1112 attributes we can obtain an accuracy of 93.375%. Hence dimensionality reduction is achieved.

From the Fig. 4. - (iii) we infer that, after applying the Relief Filter algorithm where the number of iterations is 20, the number of attributes has reduced from 6061 to 1436 and the percentage of correctly classified attributes thus obtained is 91.25%. Therefore with 1436 attributes we can obtain an accuracy of 91.25%. Hence dimensionality reduction is achieved.

B. Las Vegas Wrapper Algorithm

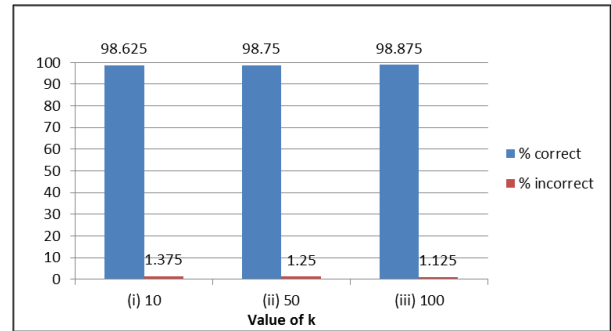


Fig. 5. Las Vegas Wrapper Algorithm for varying “k”

From the Fig. 5. - (i) we infer that, after applying the Las Vegas Wrapper algorithm where K=10, the percentage of correctly classified attributes has increased from 94.25% to 98.625% and the number of attributes has reduced from 6061 to 3034. Therefore with 3034 attributes we can obtain an accuracy of 98.625% which is higher than the accuracy of the clean dataset. Hence dimensionality reduction is achieved.

From the Fig. 5. - (ii) we infer that, after applying the Las Vegas Wrapper algorithm where K=50, the percentage of correctly classified attributes has increased from 94.25% to 98.75% and the number of attributes has reduced from 6061 to 4349. Therefore with 4349 attributes we can obtain an accuracy of 98.75% which is higher than the accuracy of the clean dataset. Hence dimensionality reduction is achieved.

From the Fig. 5. - (iii) we infer that, after applying the Las Vegas Wrapper algorithm where K=100, the percentage of correctly classified attributes has increased from 94.25% to 98.875% and the number of attributes has reduced from 6061 to 3427. Therefore with 3427 attributes we can obtain an accuracy of 98.875% which is higher than the accuracy of the clean dataset. Hence dimensionality reduction is achieved.

C. Hybrid Algorithm

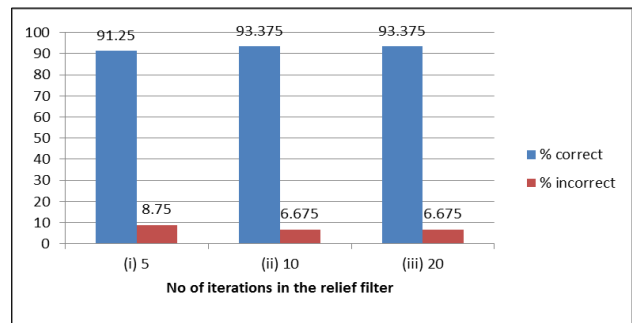


Fig. 6. Hybrid Algorithm for k=10, and varying “its”

From the Fig. 6 - (i) we infer that, after applying the Hybrid algorithm where the number of iterations is 5 and K= 5, the number of attributes has reduced from 6061 to 132 and the percentage of correctly classified attributes thus obtained is 91.25%. Therefore with 132 attributes we can obtain an accuracy of 91.25%. Hence dimensionality reduction is achieved.

From the Fig. 6. - (ii) we infer that, after applying the Hybrid algorithm where the number of iterations is 10 and K= 5, the number of attributes has reduced from 6061 to 1102 and the percentage of correctly classified attributes thus obtained is 93.375%. Therefore with 1102 attributes we can obtain an accuracy of 93.375%. Hence dimensionality reduction is achieved.

From the Fig. 6 - (iii) we infer that, after applying the Hybrid algorithm where the number of iterations is 20 and K= 5, the number of attributes has reduced from 6061 to 1436 and the percentage of correctly classified attributes thus obtained is 95.375%. Therefore with 1436 attributes we can obtain an accuracy of 95.375%. Hence dimensionality reduction is achieved.

VI. CONCLUSION

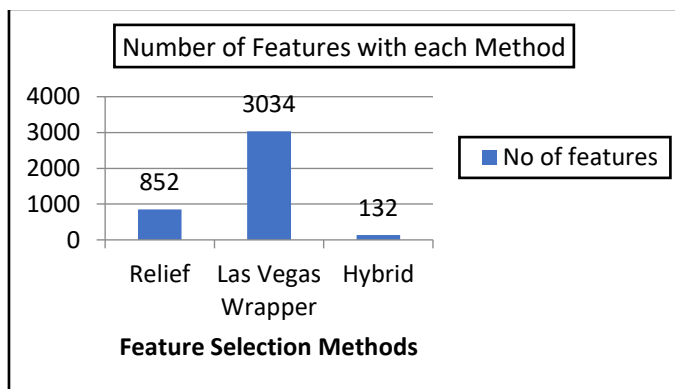


Fig.7. Comparing no. of reduced features obtained after applying the algorithms

From the Fig. 7. it is observed that although the Iterative Relief Algorithm achieves a high reduction in the number of features considered from 6061 to 852, it also results in a slight drop in the accuracy of the reduced dataset to 91.25% from 94.25%. The Hybrid algorithm achieves dimensionality reduction from 6061 features at accuracy of 94.25% to 136 with accuracy of 91.25%. The Las Vegas Wrapper Algorithm performs by reducing the number of the features to almost half with an increase in the accuracy of the dataset from 94.25% to 98.625%.

VII. FUTURE WORKS

The algorithms involved in this research and the results obtained can be used for various applications. These algorithms can be applied to reduce the dimensionality of dynamic data, consequently reducing the processing/run time. They can also aid in the reduction of bands in satellite imagery obtained through multispectral or hyperspectral sensors. In combination with these algorithms, different classifiers can be used to test the accuracy for different applications. The number of iterations for each algorithm can be varied and based on the resulting accuracy further analysis can be made.

ACKNOWLEDGEMENT

This research was supported by College of Engineering, Guindy, Anna University, India. We are grateful to Dr. T. V. Geetha, Dean of College of Engineering Guindy for giving us this golden opportunity. Our wholehearted gratitude to Dr. S. Valli, Head of Department of Computer Science, and Engineering, College of Engineering Guindy for providing us with the lab facilities and the necessary essentials. We thank our mentor Dr. A P Shanthy, Professor, Department of Computer Science and Engineering, College of Engineering Guindy who provided insight and expertise that greatly assisted the research. We thank Dr. T. Raghuvveera, Associate Professor, Department of Computer Science and Engineering, College of Engineering Guindy and Dr. D Manjula for their undying support and encouragement throughout the project, and Ms. Susi E, Research Scholar, Department of Computer Science and Engineering, College of Engineering Guindy for her guidance and constant supervision in completing the project successfully. We take full responsibility in case of any errors and this should not tarnish the reputation of these esteemed persons.

REFERENCES

- [1] K.Sutha and Dr.J.Jebamalar Tamilselvi, "A Review of Feature Selection Algorithms for Data Mining Techniques" Vol. 7 No.6 Jun 2015
- [2] Richard Jensen and Qiang Shen, "Computational Intelligence and Feature Selection Rough and Fuzzy Approaches" Aberystwyth University
- [3] Preeti Patidar and Anshu Tiwari, "Handling Missing Value in Decision Tree Algorithm" International Journal of Computer Applications (0975 – 8887) Volume 70– No.13, May 2013
- [4] WANG Hongwei, "A Method of Feature Selection for Continuous Features base on Similarity Degrees of Interval Numbers" College of Information Science and Technology, Bohai University, Jinzhou, China
- [5] S. Visalakshi and V. Radha, "A Literature Review of Feature Selection Techniques and Applications Review of Feature Selection in Data Mining" IEEE International Conference on Computational Intelligence and Computing Research 2014
- [6] Hehui Qian, Zhiwei Qiu, "Feature Selection using C4.5 Algorithm for Electricity Price Prediction" International Conference on Machine Learning and Cybernetics, Lanzhou 2014
- [7] Surekha Samsani, "An RST based Efficient Preprocessing Technique for Handling Inconsistent Data" IEEE International Conference on Computational Intelligence and Computing Research
- [8] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques" Morgan Kaufmann Publishers, Third Edition, 2011
- [9] Qing Ang, Weidong Wang, Zhiwen Liu and Kaiyuan Li, "Explored Research on Data Preprocessing and Mining Technology for Clinical Data Applications" 2nd IEEE International Conference on Information Management and Engineering 2010, Chengdu, China
- [10] UC Irvine Machine Learning, Center for Machine Learning and Intelligent Systems. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/dorothea>
- [11] K. Naidu, A. Dhenge and K. Wankhade, "Feature Selection Algorithm for Improving the Performance of Classification: A Survey," 2014 Fourth International Conference on Communication Systems and Network Technologies, Bhopal, 2014, pp. 468-471.
- [12] A. S. Abdullah, C. Ramya, V. Priyadharsini, C. Reshma and S. Selvakumar, "A survey on evolutionary techniques for feature selection," 2017 Conference on Emerging Devices and Smart Systems (ICEDSS), Tiruchengode, 2017, pp. 58-62.
- [13] B. C. Santos, M. W. Rodrigues, L. N. Cristiano Pinto, C. N. Nobre and L. E. Zárate, "Feature selection with genetic algorithm for protein function prediction," 2019 IEEE International

- Conference on Systems, Man and Cybernetics (SMC), Bari, Italy, 2019, pp. 2434-2439.
- [14] B. C. Santos, M. W. Rodrigues, L. N. Cristiano Pinto, C. N. Nobre and L. E. Zárate, "Feature selection with genetic algorithm for protein function prediction," 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy, 2019, pp. 2434-2439.
- [15] W. Pearson, C. T. Tran, M. Zhang and B. Xue, "Multi-Round Random Subspace Feature Selection for Incomplete Gene Expression Data," 2019 IEEE Congress on Evolutionary Computation (CEC), Wellington, New Zealand, 2019, pp. 2544-2551.
- [16] T. Chandak, C. Ghorpade and S. Shukla, "Effective Analysis of Feature Selection Algorithms for Network based Intrusion Detection System," 2019 IEEE Bombay Section Signature Conference (IBSSC), Mumbai, India, 2019, pp. 1-5.
- [17] F. Koumi, M. Aldasht and H. Tamimi, "Efficient Feature Selection using Particle Swarm Optimization: A hybrid filters-wrapper Approach," 2019 10th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2019, pp. 122-127.

Evaluation of Recurrent Neural Network architectures for abusive language detection in cyberbullying contexts

Filip Markoski¹, Eftim Zdravevski¹, Nikola Ljubešić², Sonja Gievska¹

¹Faculty of Computer Science and Engineering
Skopje, Macedonia

filip.markoski45@gmail.com, eftim.zdravevski@finki.ukim.mk, sonja.gievska@finki.ukim.mk

²Department of Knowledge Technologies, Jožef Stefan Institute,
nikola.ljubestic@ijs.si
Ljubljana, Slovenia

Abstract—Cyberbullying is a form of bullying that takes place over digital devices. Social media is one of the most common environments where it occurs. It can lead to serious long-lasting trauma and can lead to problems with fear, anxiety, sadness, mood, energy level, sleep, and appetite. Therefore, detection and tagging of hateful or abusive comments can help in the mitigation or prevention of the negative consequences of cyberbullying. This paper evaluates seven different architectures relying on Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) gating units for classification of comments. The evaluation is conducted on two abusive language detection tasks, on a Wikipedia data set and a Twitter data set, obtaining ROC-AUC scores of up to 0.98. The architectures incorporate various neural network mechanisms such as bi-directionality, regularization, convolutions, attention etc. The paper presents results in multiple evaluation metrics which may serve as baselines in future scientific endeavours. We conclude that the difference is extremely negligible with the GRU models marginally outperforming their LSTM counterparts whilst taking less training time.

Keywords—Deep Learning, NLP, RNN, LSTM, GRU, Abusive Language Detection, Hate Speech, Cyberbullying

I. INTRODUCTION

Cyberbullying is the use of technology to harass, threaten, embarrass, or target another person. It includes posting online threats and mean, aggressive, or rude texts, personal information, pictures, or videos designed to hurt or embarrass someone else. Online bullying can be particularly damaging and upsetting because it is usually anonymous, and therefore, hard to trace and control. Online bullying and harassment can be easier to commit than other acts of bullying because the bully does not have to confront his or her target in person. Online bullying, as any other kind of bullying, can lead to serious long-lasting problems. The stress of being in a constant state of upset or fear can lead to problems with mood, energy level, sleep, and appetite. It also can make someone feel jumpy, anxious, or sad. If someone is already depressed or anxious, cyberbullying could lead to much more serious consequences.

For these reasons, providing a systematic solution that can recognize and tag textual comments that represent some sort of hateful or abusive content can be valuable in the prevention and mitigation of their consequences. [1]

Previous works have already applied various approaches aiming to tackle this kind of tasks. Most successful approaches for this task usually employ Recurrent Neural Networks (RNNs). RNNs are Deep Neural Networks (DNN) that are adapted to sequence data, i.e. to input and output of variable length. RNNs contain loops in the hidden layer to retain information from a previous time step which will later

be used to predict the value of the current time step. This retention of information makes the neural networks extremely deep and thus makes them difficult to train to capture long-term dependencies because the gradients tend to either vanish or explode and thus the RNNs are prone to exploding or vanishing gradients. [1]

The most prominent ways to reduce the negative effects of training RNNs are either to design a better learning algorithm than stochastic gradient descent, such as a powerful second-order optimization algorithm or design an improved activation function such as the LSTM architecture, which was developed and proposed in [2] which proves to be an effective way of dealing with the vanishing gradient problem and thus became a standard, or the GRU architecture which was proposed in [3] and shares many similarities to the LSTM architecture whilst still employing different circuitry. Other ways of dealing with the problems faced by RNNs are to perform regularization of the RNN's weights that ensures that the gradient does not vanish, to entirely stop learning the recurrent weights and finally, to very carefully initialize the RNN's parameters, such as in [4] and [5].

In this paper, we attempt to evaluate the two most prevalent architectures as answers to dealing with the vanishing gradient problem whilst training on sequential data. The approaches we have chosen are LSTM and GRU in the context of other components. Our aim is to see which architecture performs better when its most defining component is an LSTM module, or a GRU module. We perform this evaluation on two abusive language detection tasks whilst situating the most defining component in architectures which incorporate various neural network mechanisms such as bi-directionality, regularization, convolutions, attention etc. We draw our conclusions on the basis of a variety of evaluation metrics, which may subsequently serve as baselines for future research.

II. RELATED WORK

There exist many empirical comparisons performed on RNN architectures, such as LSTM or GRU. In [2], the authors evaluated multiple models, namely LSTM, GRU and tanh-RNN, with all approximately the same number of parameters and trained using RMSProp on a suite of sequence modelling tasks, namely, tasks of polyphonic music modelling and speech signal modelling. The authors concluded that although GRU produced superior results to the other models overall, the difference was not too great as to lead to a firm conclusion of which model is best.

Deeming the LSTM's architecture to be 'ad-hoc', in [3], the authors perform an ablation study and an empirical evaluation of LSTM, GRU and LSTM-mutated architectures which they produced using an evolutionary architecture

search, more extensive than the architecture search conducted by [4] in which the authors performed fewer experiments with small models. From the ablation study and model results from [3], it was shown that the forget gate in the LSTM architecture is most important and that its removal results in drastically inferior performance, except in language modelling. Furthermore, the authors noted that initializing the bias of the forget gate to be a number between 1 and 2 leads the LSTM models to have very comparable results to that of the GRU models, thus closing the performance gap between the LSTM and GRU models.

In [5], it is shown that LSTM models with a large number of parameters take up a considerable amount more training time than their GRU counterparts whilst still producing similar results. Their models used ReLU as an activation function and the Adam optimization algorithm.

In comparison to [2], we utilize more appropriate parameter initialization strategies, employ the use of a more robust parameter optimization algorithm and a plethora of architectures leading to perhaps a more reliable empirical comparison between the LSTM and GRU components.

III. METHODOLOGY

The two data sets used within this study, namely the Wikipedia and Twitter data sets. Further, this section describes the data preprocessing, the appropriate evaluation metrics for both data sets and describes the generalized model architectures and the specific LSTM or GRU models which are manifestations of those architectures.

A. Toxic Wikipedia Comment Data Set

The data set, described in Table I, used to train and evaluate the models is the same one used in [6] and offered publicly as part of the Toxic Comment Classification Challenge competition. [7] The multi-labeled data set is comprised of a training set containing 159571 entries and of a testing set comprised of 153164 entries. The six labels, each presented in a separate column that are provided in the training set and need to be predicted in the testing set, are the following: 'toxic', 'severe_toxic', 'obscene', 'threat', 'insult', 'identity_hate'. From the entire training set, only 16225 entries are labeled with any of the aforementioned labels, meaning that the labels often overlap. The training set has a class-imbalance problem, in relation to this, the authors of [8] present a “real-life” distribution of abusive language use via surveying available abusive language data sets. From these data sets, one can see that they are usually comprised of an overwhelming majority of non-abusive entries. Additionally, sentence length does not seem to be a significant indicator of toxicity which is in accordance with the conclusion of [9], which is that word-length distribution features provide little to no improvement in a model’s predictive abilities.

TABLE I. LABELS FOR THE TOXIC WIKIPEDIA COMMENTS TRAINING DATA SET (MULTI-LABELED DATA SET WITH OVERLAPPING LABELS)

Label	<i>toxic</i>	<i>severe toxic</i>	<i>obscene</i>	<i>threat</i>	<i>insult</i>	<i>identity hate</i>
Count	15294	1595	8449	478	7877	1405
%	9.6	1.0	5.3	0.3	4.9	0.9

B. Twitter Data Set

We use the same data set as [10], which is comprised of approximately 100000 tweets of which only 61194 were able to be retrieved using the Twitter API. Of the ones retrieved, in a single class column (unlike the Wikipedia data set which separates each label in a separate column), 63% are annotated as ‘normal’, 19% as ‘abusive’, 14% as ‘spam’ and 4% as ‘hateful’ with exact counts shown in Table II.

Each tweet text was further processed using a tweet normalization tool¹ which directly optimizes the vocabulary and in turn the models’ power to generalize by replacing usernames with a single token ‘<user>’ and changing words such as ‘gooodd’ to ‘good’ for example.

The data set was split into a training, validation and test data set each consisting with 76%, 4% and 20% of the data respectively.

TABLE II. LABELS FOR THE TWITTER DATA SET

Label	<i>abusive</i>	<i>hateful</i>	<i>spam</i>	<i>normal</i>
Count	11766	2461	8561	38407
%	19.2	4.0	14.0	62.8

C. Data Preprocessing

Each of the comments was represented with a padded indexed representation of itself. Keras² Tokenizer was used to perform the tokenization, indexing and padding of each comment, in which the vocabulary was limited to the most frequent 20000 tokens and each comment was padded to a maximum length of 200 indices.

D. Evaluation Metric

1) Toxic Wikipedia Comments Data Set

The models had to predict a probability for each of the six possible columns, each representing one of the labels. This was evaluated using the area under the receiver operating characteristic curve (ROC-AUC) which was calculated after each epoch for each of the models to get the metrics for the training set predictions. Additionally, the ROC-AUC evaluations are also provided for the private and public Kaggle testing sets evaluated by the Kaggle platform. Although we deem the classification metrics used on the Twitter data set as more appropriate, we could not calculate the same due to not having the corresponding labels for the test set, thus we only resort to the ROC-AUC scores returned for each submission of predictions by the Kaggle platform in the form of private and public scores.

2) Twitter Data Set

The models had to predict a probability for each of the four possible labels with the prediction being a one-hot encoded vector whose only active components correspond to the label assigned with the highest probability. This was evaluated using the following metrics: accuracy, F1-micro, F1-macro, weighted precision, and weighted recall scores.

E. Models

Each of the models was constructed using the python deep learning library Keras. In total, there are seven model architectures containing a recurrent neural network layer which manifests a total of 12 models, that is, all the model architectures once with an LSTM component as the most representative component, and similarly, once with a GRU

¹ <https://github.com/cbaziotis/ekphrasis>

² <https://github.com/keras-team/keras>

component. Each of the architectures use an Embedding layer which transforms the indexed words with a vector representation of size 50. We chose not to work with pre-trained word embeddings as the primary goal of our experimentation is to isolate the benefits of different architectures. The architectures described through their dual manifestations are:

1. (unidirectional) LSTM / GRU – Unidirectional approach to the language task (Fig. 2).
2. (bidirectional) Bi-LSTM / Bi-GRU – A bidirectional variant of the first model architecture (Fig. 2).
3. (bi-then-conv) Bi-LSTM-CNN / Bi-GRU-CNN – Following the bidirectional recurrent neural network layer, the model extracts one-dimensional convolutions, performs global average pooling and global max pooling, concatenates them and used the resultant vector to infer a prediction (Fig. 3).
4. (bi-conv-uni) Bi-LSTM-CNN-LSTM / Bi-GRU-CNN-GRU – In addition to the Bi-RNN-CNN architecture we add another RNN component that attempts to learn on the convolved information and afterwards infer a prediction (Fig. 4).
5. (convolutional) Multi-CNN-Bi-LSTM / Multi-CNN-Bi-GRU – Contrary to some of the previous architectures, we use one-dimensional convolutional layers of kernel sizes 1, 2, 3 and 5 with the hopes to derive unigram, bigram, trigram and 5-gram features which shall later be used in training the RNN component of the architecture (Fig. 5).
6. (conv-attention) Conv-Att-LSTM / Conv-Att-GRU – Two pairs of a kernel-size-three convolutional layer and max-pooling layer precede an attention mechanism right before the RNN component (Fig. 6).
7. (attention) Attention-LSTM / Attention-GRU – Only an attention mechanism is added before the RNN component (Fig. 7).

In Figures 2 through 7, the shape of the tensor is described below the name of the component in the architecture.

Each of the models is trained for three epochs with a batch size of 256 and optimized with an Adam optimizer with a default learning rate of 0.001 on Google Collaboratory Tensor Processing Unit (TPU) runtime, which most likely offered a TPU v2 device with 8 GiB of high-bandwidth memory, two TPU cores and one matrix unit for each TPU core. For the Toxic Wikipedia Comments data set, as each label is presented in its separate column, the output layer utilizes a sigmoid function, thus it outputs independent probabilities for each label-column, this type of output is in tandem with a binary cross-entropy loss function. For the Twitter data set, the output layer utilizes a Softmax function which is paired with a categorical cross-entropy loss function ensuring the model can assign a probability for each possible label. The rate for any drop-out mechanism, including the recurrent drop-out, ranges from 0.1 to 0.2.

According to the example of [11], each Batch Normalization layer has been positioned before any singular Drop-Out layer. For the recurrent neural networks, the activation function is tanh and the recurrent activation function is Sigmoid. All other hidden layers have ReLu as an activation function.

Inspired from [12] and [13], the layers using the ReLu activation function are initialized with a He uniform distribution, whilst the others are initialized with a Glorot uniform distribution.

It is important to note that Keras follows the advice from [3] and initializes the bias of the LSTM forget gate to 1.

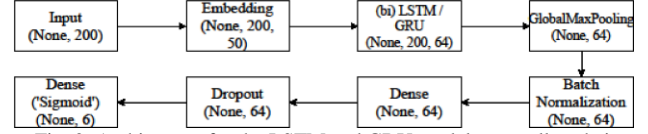


Fig. 2. Architecture for the LSTM and GRU models, as well as their Bidirectional alternatives

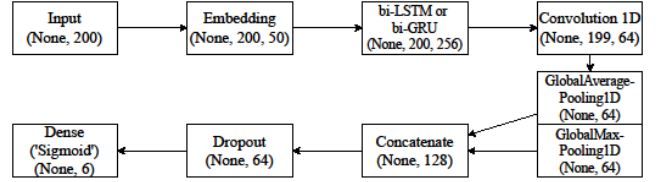


Fig. 3. Architecture for the bi-LSTM-CNN and bi-GRU-CNN models

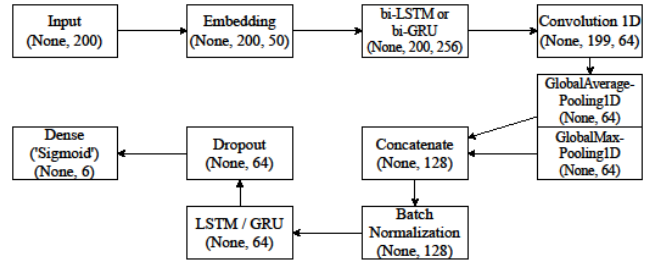


Fig. 4. Architecture for the bi-LSTM-CNN-LSTM and bi-GRU-CNN-GRU models

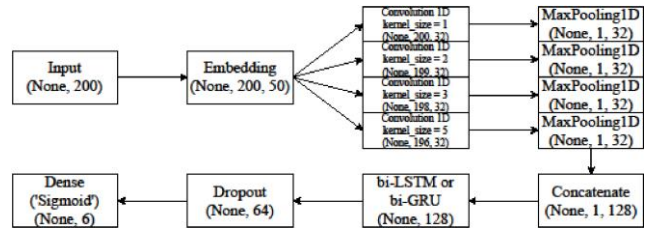


Fig. 5. Architecture for the multi-CNN-Bi-LSTM and multi-CNN-Bi-GRU models

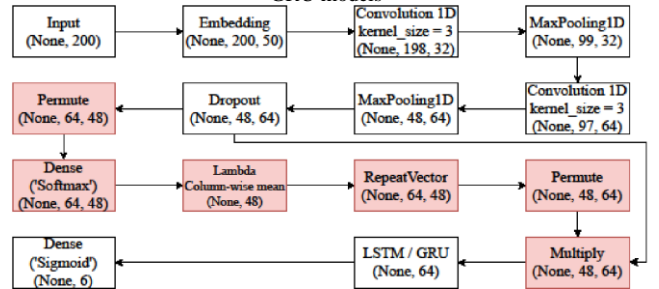


Fig. 6. Architecture for the conv-att-LSTM and conv-att-GRU models (the red sections highlight the attention mechanism)

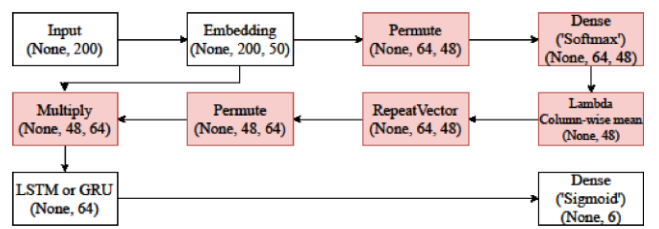


Fig. 7. Architecture for the attention-LSTM and attention-GRU models (the red sections highlight the attention mechanism)

IV. RESULTS

Concerning the Toxic Wikipedia Comments data set, in Fig. 8 we present the ROC-AUC scores for each model for each of the three epochs on a validation data set and the Kaggle private and public testing data sets. We deem the private testing data set ROC-AUC score as most representative of the model's power to generalize as performs its evaluation on a testing set not publicly provided.

Concerning the Twitter data set, in Fig. 9 we present the aforementioned metrics for each of the models. The authors of [14] and [15] note that macro metrics provide a better sense of effectiveness on the minority classes in a class-imbalanced problem, thus we deem the F1-macro score as the most relevant indicator.

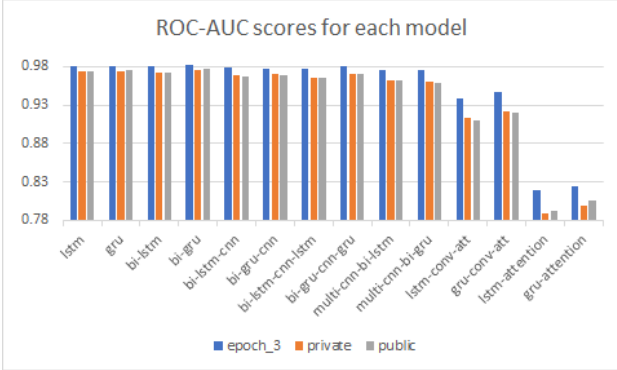


Fig. 8. ROC-AUC scores for each model on the Toxic Wikipedia Comment data set

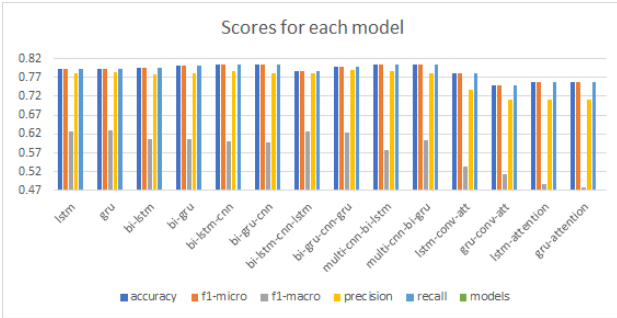


Fig. 9. Accuracy, F1-micro, F1-macro, precision and recall scores for each model on the Twitter data set

Our findings are summarized below:

- Each of the models, for both data sets, began to overfit after the second epoch of training, except for the only-attention models. Possible causes for this are presented in Section V.
- Judging the models on both data sets, looking at the private score for the Toxic Wikipedia Comments data set, and the F1-Macro score for the Twitter data set and awarding points to each recurrent neural component whenever its model scores better than its counterpart, we can say, although the differences are negligible, the GRU models outperformed the LSTM models. The scores are presented in Table III and Table IV, for the Wikipedia and Twitter data sets, respectively.
- In terms of the number of parameters, the LSTM models have significantly more trainable parameters than their GRU counterparts. Details can be found in Table VI.

- In terms of training time, which is closely related to the number of trainable parameters, the GRU models always finished their training faster than their LSTM counterparts. Details can be found in Table V.
- The simpler architectures got the best scores, with the best model for the Wikipedia data set being the Bi-GRU model achieving a private score of 0.975, closely followed by the unidirectional GRU model. For the Twitter data set the best model being GRU, with 0.631 F1-Macro score is followed by Bi-LSTM-CNN-LSTM with 0.626 F1-Macro score.
- The conv-attention and attention architectures trained for significantly less time while achieving decent results, yet still inferior to the rest of the evaluated models. Section V gives possibilities to why the attention-including architectures noticeably underperform in comparison to the other alternatives.

V. DISCUSSION AND FUTURE WORK

The authors of [6], which contributed the Wikipedia data set, managed to obtain a ROC-AUC score of 0.971 using a DNN architecture with character n-grams. To contrast this, a more traditional machine learning approach can be found in [16], which uses feature construction analogous to [17] which achieved a ROC-AUC score of 0.89 using a logistic regression classifier. Evidently, the Bi-GRU model, with a score of 0.975 here performs sufficiently well with both of these attempts.

Regarding the Twitter data set, which is differently annotated as in [14] or [15], our best model, being the unidirectional GRU, achieved 0.63 F1-Macro score. Even though it does not outperform [18], it does compare well with the reported [19] F1-Micro score of 0.827, whilst our Bi-GRU has 0.802.

Regarding the underperformance of the attention-based models, one possibility might be due to the use of a single attention vector which is shared across the input dimensions. In the computation of the attention vector is a mean operation which possibly cumulatively worsens the feature vectors.

Regarding future work, further empirical evaluations may be performed on the current model architectures using various pre-trained embeddings. Additionally, these evaluations may incorporate different types of data sets and different model architectures. Theoretical analysis also may be conducted as well as an ablation study with the goal of forming a more concrete and reliable comparison. This could be done in the form of a survey in which the various findings are collected, and the conclusion is more significant. The algorithms may also be adapted for real-time classification or abusive language detection, as well as employed by in-production platforms.

VI. CONCLUSION

In the context of the two abusive language detection tasks, the difference between the LSTM models and their GRU counterparts is extremely negligible. Still, the GRU models train faster due to the smaller number of trainable parameters and thereby, overall, outperform the LSTM models. In agreement with the results obtained by [2], we cannot make a firm conclusion on which of the gating units is better.

TABLE III. SCORES ON THE TOXIC WIKIPEDIA COMMENTS DATA SET (LSTM IS SUPERIOR IN CELLS HIGHLIGHTED WITH BOLD, GRU IS SUPERIOR IN CELLS HIGHLIGHTED WITH ITALIC)

Models	epoch 1	epoch 2	epoch 3	private	public
lstm	0.9732	0.9798	0.9815	0.9737	0.9734
gru	0.9738	0.9787	0.9812	<i>0.9738</i>	<i>0.9748</i>
bi-lstm	0.9737	0.9777	0.9810	0.9727	0.9717
bi-gru	0.9772	0.9809	<i>0.9818</i>	<i>0.9751</i>	<i>0.9775</i>
bi-lstm-cnn	0.9749	0.9773	0.9787	0.9694	0.9679
bi-gru-cnn	0.97478	0.9753	0.9766	<i>0.9706</i>	<i>0.9695</i>
bi-lstm-cnn-lstm	0.9713	0.9750	0.9772	0.9661	0.9648
bi-gru-cnn-gru	<i>0.9764</i>	<i>0.9784</i>	<i>0.9815</i>	<i>0.9710</i>	<i>0.9700</i>
multi-cnn-bi-lstm	0.9704	0.9764	0.9759	0.9624	0.9615
multi-cnn-bi-gru	0.9728	0.9750	0.9757	0.9609	0.9595
lstm-conv-att	0.9168	0.9359	0.9386	0.9139	0.9105
gru-conv-att	0.9307	<i>0.9456</i>	<i>0.9466</i>	0.9220	<i>0.9196</i>
lstm-attention	0.7872	0.8126	0.8188	0.7884	0.7924
gru-attention	0.7832	<i>0.8168</i>	<i>0.8245</i>	<i>0.7988</i>	<i>0.8065</i>

TABLE IV. SCORES ON THE TWITTER DATA SET (LSTM IS SUPERIOR IN CELLS HIGHLIGHTED WITH BOLD, GRU IS SUPERIOR IN CELLS HIGHLIGHTED WITH ITALIC)

Models	Accura-cy	F1-micro	F1-macro	Preci-sion	Recall
lstm	0.7939	0.7939	0.6259	0.7810	0.7939
gru	0.7928	0.7928	<i>0.6309</i>	<i>0.7838</i>	0.7928
bi-lstm	0.7948	0.7948	0.6062	0.7773	0.7948
bi-gru	<i>0.8022</i>	<i>0.8022</i>	<i>0.6062</i>	<i>0.7798</i>	<i>0.8022</i>
bi-lstm-cnn	0.8042	0.8042	0.6016	0.7863	0.8042
bi-gru-cnn	<i>0.8046</i>	<i>0.8046</i>	0.5981	0.7814	<i>0.8046</i>
bi-lstm-cnn-lstm	0.7880	0.7880	0.6260	0.7826	0.7880
bi-gru-cnn-gru	<i>0.7978</i>	<i>0.7978</i>	0.6231	<i>0.7890</i>	<i>0.7978</i>
multi-cnn-bi-lstm	0.8040	0.8040	0.5763	0.7879	0.8040
multi-cnn-bi-gru	0.8034	0.8034	<i>0.6044</i>	0.7798	0.8034
lstm-conv-att	0.7803	0.7803	0.5343	0.7388	0.7803
gru-conv-att	0.7494	0.7494	0.5118	0.7119	0.7494
lstm-attention	0.7580	0.7580	0.4878	0.7110	0.7580
gru-attention	<i>0.7591</i>	<i>0.7591</i>	0.4793	<i>0.7124</i>	<i>0.7591</i>

TABLE V. DIFFERENCES IN TOTAL SECONDS TAKEN TO TRAIN EACH MODEL (LSTM IS SUPERIOR IN CELLS HIGHLIGHTED WITH BOLD, WHERE THE DIFFERENCE IS NEGATIVE, GRU IS WITH ITALIC, WHERE THE DIFFERENCE IS POSITIVE)

Architecture	Wikipedia Data Set			Twitter Data Set		
	LSTM	GRU	Differ-ence	LSTM	GRU	Differ-ence
unidirectional	890	812	78	277	242	35
bidirectional	1703	1402	301	546	451	95
bi-then-conv	4681	3879	802	1519	1235	284
bi-conv-uni	5445	4474	971	1753	1444	309
convolutional	500	507	-7	160	164	-4
conv-attention	568	495	73	178	163	15
attention	1154	964	190	381	320	61

TABLE VI. DIFFERENCES IN THE TOTAL NUMBER OF TRAINABLE PARAMETERS FOR EACH MODEL (GRU HAS FEWER TRAINABLE PARAMETERS IN ALL CASES)

Architecture	LSTM	GRU	Difference
unidirectional	29440	22080	7360
bidirectional	58880	44160	14720
bi-then-conv	183296	137472	45824
bi-conv-uni	232704	174528	58176
convolutional	98816	74112	24704
conv-attention	33024	24768	8256
attention	70030	62670	7360

VII. REFERENCES

- [1] S. Hinduja and J. W. Patchin, "Bullying, Cyberbullying, and Suicide," *Archives of Suicide Research*, vol. 14, no. 3, pp. 206-221, 2010.
- [2] J. Chung, C. Gulcehre and K. Cho, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," 2014.
- [3] R. Jozefowicz, W. Zaremba and I. Sutskever, "An empirical exploration of recurrent network architectures," *International Conference on Machine Learning*, vol. 2350, p. 2342, 2015.
- [4] J. Bayer, D. Wierstra, J. Togelius and J. Schmidhuber, "Evolving Memory Cell Structures for Sequence Learning," pp. 755-764, 2009.
- [5] A. Shewalkar, D. Nyavanandi and S. Ludwig, "Performance Evaluation of Deep Neural Networks Applied to Speech Recognition: RNN, LSTM and GRU," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 9, pp. 235-245, 2019.
- [6] E. Wulczyn, N. Thain and L. Dixon, "Ex Machina: Personal Attacks Seen at Scale," 2016.
- [7] "kaggle," [Online]. Available: <https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data>.
- [8] P. Mishra, H. Yannakoudakis and E. Shutova, "Tackling Online Abuse: A Survey of Automated Abuse Detection Methods," 2019.
- [9] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," in *Proceedings of the NAACL Student Research Workshop*, San Diego, California, 2016.
- [10] A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos and N. Kourtellis, "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior," in *11th International Conference on Web and Social Media, ICWSM 2018*, 2018.
- [11] X. Li, S. Chen, X. Hu and J. Yang, "Understanding the Disharmony between Dropout and Batch Normalization by Variance Shift," 2018.
- [12] G. Xavier and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Journal of Machine Learning Research - Proceedings Track*, vol. 9, pp. 249-256, 2010.
- [13] K. He, X. Zhang, S. Ren and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," *IEEE International Conference on Computer Vision (ICCV 2015)*, vol. 1502, 2015.
- [14] P. Mishra, H. Yannakoudakis and E. Shutova, "Neural Character-based Composition Models for Abuse Detection".
- [15] Z. Zhang and L. Luo, "Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter," 2018.
- [16] M. Todosovska and S. Gievska, "Detection of Abusive Language in OnlineComments," 2018.
- [17] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad and Y. Chang, "Abusive Language Detection in Online User Content," *Proceedings of the 25th international conference on world wide web*, pp. 145-153, 2016.
- [18] P. Mishra, M. D. Tredici, H. Yannakoudakis and E. Shutova, "Abusive Language Detection with Graph Convolutional Networks," 2019.
- [19] J. H. Park and P. Fung, "One-step and Two-step Classification for Abusive Language Detection on Twitter," 2017.
- [20] Y. Bengio, P. Simard and P. Frasconi, "Learning long-term dependencies with gradient descent is," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, p. 157-166, 1994.
- [21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.
- [22] K. Cho, B. van Merriënboer, C. Gulcehre, F. Bougares, H. Schwenk and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," 2014.

Protein classification by using four approaches for extraction of the protein ray-based descriptor

Georgina Mirceva

Faculty of computer science and engineering
Ss. Cyril and Methodius University in Skopje
Skopje, Macedonia
georgina.mirceva@finki.ukim.mk

Andrea Kulakov

Faculty of computer science and engineering
Ss. Cyril and Methodius University in Skopje
Skopje, Macedonia
andrea.kulakov@finki.ukim.mk

Abstract—The knowledge about the protein molecules, and how they influence the processes in the humans is very worth, because it is really needed in order to develop new drugs for diseases. In proteomics, one of the most important tasks is solving the problem of classification of protein molecules. The literature provides plethora of methods that could be used for this task. However, it is still an open issue where still there is a need for fast computational methods that would provide accurate classification of proteins. In this paper, we focus on solving this task. For that purpose, first, we extract feature vectors that hold information about the main features of the proteins. The feature vectors that are used in this study are obtained by following the procedure for extraction of our protein ray-based descriptor that we have introduced in our former studies. For that purpose, the skeleton of the protein is interpolated with predefined number of interpolation points, and then the elements of the feature vector are extracted as Euclidean distances between the interpolation points and center of mass. Besides this approach, in this study we also use three additional approaches for extraction of the feature vectors, where we focus on the change of the Euclidean distance to the center of mass between two consecutive interpolation points. After extracting feature vectors, next we apply several well-known classification methods in order to generate classification model. We present the results obtained with these four approaches used for extraction of the feature vectors.

Keywords—protein structure, protein classification, protein ray-based descriptor

I. INTRODUCTION

Proteins are an important component of the living organisms. The understanding of protein molecules and their influence in the processes in which they participate is essential in order to be able to design new drugs for various diseases. Proteomics is an area where the focus is on discovery, analysis and understanding of the proteins. One of the most important tasks in this research area is the problem of classification of protein structures, which could help to understand these structures and to determine the functions that they make have.

The data about the protein structures that are discovered are deposited in the Protein Data Bank (PDB) [1], [2]. Nevertheless, these data are not worth if we do not determine the functions of the protein structures. It is supposed that the proteins that have common ancestor and belong to same class, also share similar functions. Therefore, the protein classification problem is among the most important tasks that should be solved regarding proteins. The literature provides plethora of methods that could be used to classify protein structures.

SCOP (Structural Classification Of Proteins) [3] is one of the most widely known methods used for protein classification. With this method, the classification of proteins is manual, thus the classification time is too long. Therefore, there is need for methods where the classification would be done in automatic manner. In the literature, also automatic and semiautomatic methods could be found. CATH (Class, Architecture, Topology and Homologous superfamily) [4] is among the semiautomatic methods, because it first tries to classify the inspected protein in automatic way, and if it is not possible for a given protein, in the next stage the protein is classified manually by human experts.

Another class of methods tries to classify protein structures by making alignment of their sequences. Needleman–Wunch [5], BLAST [6] and PSI-BLAST [7] are among the most important representatives from this category. However, the methods based on sequence alignment may not discover similarity between proteins whose structures are very similar if they do not have similar sequences. Therefore, it is better to find the similar structures by aligning the structures of the proteins, rather than making alignment of their sequences. CE [8], MAMMOTH [9] and DALI [10] are among the most broadly known methods from this category. Of course, there is also third category, where the methods perform both sequence alignment and structure alignment, as in the methods SCOPmap [11] and FastSCOP [12].

However, the methods from the three categories mentioned above, which make alignment of protein sequences or/and structures need long time to classify an inspected protein. To overcome this problem, various methods extract vectors with features for the proteins, thus later the comparison between the proteins is made by calculating the distance between their feature vectors. In this category, there are both methods that consider features of the protein sequences [13] or protein structures [14]. After extraction of the proteins' feature vectors, then a classification model could be built by using some classification method.

In this paper, we use an approach as it is used in the methods from the last category. Namely, we extract feature vectors for the proteins, and then we generate prediction model. In our previous study [15], we presented several approaches for finding similar protein structures based on the features of their tertiary structures. In [15], we focused on the task how to find similar protein structures, known as protein retrieval. In this paper, our aim is not to find the homologous proteins that are similar with the inspected proteins, but our aim here is to make decision in which class the inspected protein should be classified in.

In this study, we perform protein classification by using the protein ray-based descriptor [15]. In [15], we showed that the protein ray-based descriptor is very accurate for finding similar proteins, even though it is very simple. The extraction of the protein ray-based descriptor starts with interpolation of the protein skeleton. Then, the elements of the feature vector are extracted as Euclidean distances between the obtained interpolation points and the center of mass. Besides this approach, in this paper additionally we consider three additional approaches where we analyze how these Euclidean distances change as we traverse the skeleton of the protein from one interpolation point to its consecutive interpolation point. After extraction of the feature vectors, then we apply several classification methods in order to build classification model for making class decisions. In this paper we use the following classification methods: C4.5 [16], Naive Bayes [17], Bayesian Network [18], k-nearest neighbors (knn) [19] and Support Vector Machines (SVM) [20], [21].

The remaining of this paper is structured in this way. In Section 2, we give description of the original protein ray-based descriptor. Besides this approach, where the Euclidean distances between the interpolation points and center of mass are used as features, we also present three other approaches where the features of the feature vectors are calculated based on the difference between two consecutive elements of the protein ray-based descriptor. Section 3 is focused on the evaluation of the approaches presented in this study. This section provides results obtained by using the four approaches for feature vector extraction combined with various classification methods for model generation. The conclusions are presented in Section 4, which also contains several ideas for future work.

II. PROTEIN CLASSIFICATION BY USING THE PROTEIN RAY-BASED DESCRIPTOR

In this study, the classification of protein structures is made by using feature vectors, which contain information about the geometrical features of the proteins. For that purpose, first, for each training protein structure a corresponding feature vector is generated. In the second stage, prediction model is built by using some classification method, where the elements of the feature vectors correspond to the attributes in the data set that is used for training the model. Once the prediction model is generated, next, we can make decisions in which class a given query protein belongs to.

Protein molecules contain one or several chains. The ground true data that would be used in this study contain information about the classes of the protein chains, therefore we need to obtain a feature vector for each protein chain that would be used in the study. In this way, the samples in the data set correspond to the individual protein chains of the protein molecules.

As it was mentioned before, in this paper we use four approaches for extraction of the protein ray-based descriptor. These approaches are very close to each other, the difference is just in the last step where the final values for the elements of the feature vector are calculated. The first approach corresponds to the original version of our protein ray-based descriptor that was presented in our previous study [15]. The remaining three approaches focus on the difference between the consecutive elements in the feature vector. First, we give description of the first approach, as it is presented in [15], and

then we give description about the differences introduced with the other three approaches.

A. Protein Ray-Based Descriptor

In the extraction of the protein ray-based descriptor [15], we consider only the C_α atoms, which form the protein's skeleton. We take into account the information how the C_α atoms of the protein are positioned in the 3D space, thus forming its 3D model. First, this model is scaled thus obtaining model where the Euclidean distance between the most distant C_α atom and the center of mass is 1. In this way, scale invariance is provided.

Different protein chains have different number of C_α atoms, thus we are not able to extract some feature vector directly, by considering all these atoms, because in that way we will obtain feature vectors with different lengths. Therefore, we interpolate the skeleton of the protein backbone with interpolation points, where the number of interpolation points is predefined and is equal for each protein chain. The skeleton of the protein chain could be seen as a curve in the 3D space, where the consecutive C_α atoms are connected by this curve. The idea of the interpolation is to find predefined number of points that would be good representatives of this curve. In [15], we considered two different ways how to make interpolation of the protein skeleton. In this study, we use uniform interpolation, which showed as better choice according to the results presented in [15].

Next, we give description how the interpolation of the skeleton is made. For that purpose, we calculate the length of the protein's skeleton by summing up the Euclidean distances between each pair of two consecutive C_α atoms. Then, we need to find interpolation points that are placed over the skeleton and form segments over the skeleton with same length. In this study, we interpolate the skeleton with $N=64$ interpolation points, thus we obtain vectors with 64 features.

After finding the interpolation points, the feature vector could be extracted. When we introduced the protein ray-based descriptor, we were inspired from the ray descriptor [22] that is used for making retrieval of similar 3D objects. In the ray descriptor, for a given 3D object a mash model is obtained, and then the feature vector is extracting by calculating the distances between the center of mass and the vertices in the mash model. Similarly, here we extract the elements of the feature vector by calculating the Euclidean distances between the interpolation points that were found in the previous step and the center of mass. By calculating the feature vector's elements in this way, we provide feature vector that is invariant to both translation and rotation.

B. Four Approaches for Extraction of the Protein Ray-Based Descriptor

The description presented in the previous sub-section corresponds to the first approach used for extraction of the protein ray-based descriptor. With this approach, the interpolation points are examined individually, because each element of the feature vector focuses on one of the interpolation points. If we try to visualize this approach, the idea of the protein ray-based descriptor is to present how the skeleton of the protein goes towards the center of mass or goes towards the protein surface. If we assume that there are concentric spheres in the space where the protein is placed, then, with the protein ray-based descriptor we describe how the skeleton of the protein passes from one concentric sphere to another as we traverse along the backbone. This is

illustrated in our previous paper [23], where we build HMM (Hidden Markov Model) for classification of proteins.

In this study, we came with the same idea that we used in [23], where we tried several different ways in order to represent the Euclidean distances that are obtained. With the first approach, which is actually the original version of the protein ray-based descriptor [15], the elements of the feature vector $f_{Eucl} = [f_1, f_2, \dots, f_N]$ are calculated as Euclidean distances between the interpolation points and center of mass. The i -th elements of the feature vector is $f_i = D_i$, for $i=1, 2, \dots, N$, where D_i denotes the Euclidean distance between the i -th interpolation point and the center of mass. In the second approach, we analyze the difference between two consecutive interpolation points, therefore the feature vector is calculated as $f_{diff} = [diff_1, diff_2, \dots, diff_{N-1}]$, where $diff_i = f_i - f_{i+1} = D_i - D_{i+1}$, $i=1, 2, \dots, N-1$. With the third approach, the feature vector is calculated as $f_{abs} = [abs_1, abs_2, \dots, abs_{N-1}]$, where $abs_i = |diff_i|$, $i=1, 2, \dots, N-1$. The fourth approach just considers whether the difference between two consecutive Euclidean distances rises or declines, without considering the amount of the increase or decrease. Thus the feature vector with the fourth approach is calculated as $f_{sign} = [sign_1, sign_2, \dots, sign_{N-1}]$, where $sign_i = sign(diff_i)$, $i=1, 2, \dots, N-1$, where the function $sign(x)$ is defined as

$$sign(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{if } x = 0. \\ -1, & \text{if } x < 0 \end{cases} \quad (1)$$

C. Classification Methods

After extraction of the feature vectors for all training samples, next, we generate a model that would be later used for making decisions about the classes of the test samples. The elements of the feature vectors correspond to the descriptive attributes that are used to describe the samples, the protein chains in this case.

In this paper we use several classification methods that are commonly used for solving classification task, i.e.: C4.5 [16], Naive Bayes [17], Bayesian Network [18], k-nearest neighbors (knn) [19] and Support Vector Machines (SVM) [20], [21]. For knn, we use $k=1$.

III. EXPERIMENTAL RESULTS

The standard of truth used in this study is obtained from part of the knowledge from the SCOP database [3], which holds knowledge how the protein chains are classified by human experts. With the SCOP method [3], the classification is in hierarchical manner, where the SCOP domain level is considered as most important for protein classification. Therefore, in this study the classes correspond to the different SCOP domains at domain level. We formed a data set that holds 6145 protein chains from 150 SCOP domains. The protein chains are approximately uniformly distributed in these SCOP domains. This means that we have 6145 samples and 150 classes. This data set is divided in a ratio 90:10 into training and test data set, thus obtaining 5531 samples for training the models and 614 samples for testing the models. The prediction power of the obtained classification models is estimated using the classification accuracy evaluation measure. Additionally, we also present the results obtained for the AUC-ROC (Area under the ROC curve) evaluation measure.

We made experiments by using the four approaches for extraction of the protein ray-based descriptor in combination with the five classification methods listed before. The results for the classification accuracy are given in Table 1, while the results for AUC-ROC are given in Table 2. The bolded values correspond to the best results obtained with each of the classification methods.

As it can be seen from the results, by using the first approach, where the Euclidean distances between the interpolation points and center of mass are used as feature vectors' elements, the best results are obtained. Then, the second approach follows, where we analyze how the Euclidean distances changes from one element of the vector to another. Or described in other way, we analyze how the protein backbone goes towards its center or its surface. With the last two approaches, we misplace some of the information that is considered with the second approach. With the third approach, we evade the evidence whether the difference between two consecutive Euclidean distances is increased or decreased, while with the fourth approach we keep that information, but we lose the information about the amount of the increase or decrease. Although with the fourth approach we obtain feature vector that requires less memory to be kept, in general it showed as better choice than the third approach.

The first approach corresponds to the absolute representation in [23], the second approach corresponds to the relative representation in [23], while the fourth approach is almost as the binary representation in [23]. If we compare the results from [23] and the results from this study, it is evident that with the HMM used in [23] the best results are obtained with relative representation (corresponds to the second approach from this study), while in this study the best results are obtained with the absolute representation (corresponds to the first approach from this study). The reason for that is the type of classification method that is used. With HMM, the model is defined by a final number of states, and the next state depends on one or several previous states. This type of model is appropriate if we want to analyze sequences, in our case the sequence corresponds to the transition of the skeleton from one concentric sphere to another. In this study, we use other type of classification methods, where we do not have states that are dependent from the previous states, and therefore the results are different, as expected.

Regarding the classification methods, we can conclude that, in general, with Bayesian Network best results are obtained, then knn, SVM and Naïve Bayes follow, while C4.5 showed as the worst choice in this case.

TABLE I. THE RESULTS FOR CLASSIFICATION ACCURACY BY USING THE FOUR APPROACHES FOR EXTRACTION OF THE PROTEIN RAY-BASED DESCRIPTOR IN COMBINATION WITH THE FIVE CLASSIFICATION METHODS

Classification Method	Approach for extraction of the protein-ray based descriptor			
	<i>Eucl</i>	<i>diff</i>	<i>abs</i>	<i>sign</i>
C4.5	92.997	91.042	88.111	89.739
Naive Bayes	94.625	92.182	90.717	90.717
Bayesian Network	96.417	95.603	94.788	93.322
knn	98.534	97.883	97.231	97.394
SVM	97.557	96.743	96.254	97.231
AVERAGE	96.026	94.691	93.420	93.681

TABLE II. THE RESULTS FOR AUC-ROC BY USING THE FOUR APPROACHES FOR EXTRACTION OF THE PROTEIN RAY-BASED DESCRIPTOR IN COMBINATION WITH THE FIVE CLASSIFICATION METHODS

Classification Method	Approach for extraction of the protein-ray based descriptor			
	<i>Eucl</i>	<i>diff</i>	<i>abs</i>	<i>sign</i>
C4.5	0.971	0.964	0.947	0.958
Naive Bayes	0.996	0.996	0.994	0.995
Bayesian Network	1.000	1.000	0.999	0.999
knn	0.993	0.990	0.987	0.992
SVM	0.995	0.994	0.996	0.990
AVERAGE	0.991	0.989	0.985	0.987

IV. CONCLUSION AND FUTURE WORK

This study focused on solving a problem for protein classification based on the features of the tertiary structure of proteins. For that purpose, first we generated feature vectors for the training samples, and then we applied five well-known classification methods for generating classification models. In this paper, we used our protein ray-based descriptor as a feature vector. However, besides its original version (denoted as the first approach), where the Euclidean distances between the interpolation points and center of mass are used as features, we also considered three additional approaches that focus on the changes of the values of the previously extracted features. With the second approach, we present how the values of the features increase or decrease as we navigate along the skeleton of the protein, while the third and fourth approach consider only the amount of the change or the direction of the change, respectively.

For evaluation, we used a part of the knowledge from the SCOP database, as a ground truth. The results showed that with the classification methods used in this study, it is best to use the original version of the protein ray-based descriptor (the first approach). Next, the second approach follows because it preserves both the amount and the direction of the change. The fourth approach although requires less memory than the third approach, in general it showed as better than the third approach, meaning that it is more important on which places there are changes of the direction in which the skeleton moves (towards the center or towards the surface), while the amount of this change is a little bit less important than the direction.

Continuing the research for solving the protein classification problem, we plan to extend our studies in several directions. We believe that the choice of the feature vector is the most important factor, thus if we have better attributes, we can later make more accurate models. Therefore, we will continue our hunt for other feature vectors and other features that are the most important and relevant for this task. Besides geometrical features, also some features of the primary and secondary structure of the proteins could be considered. Of course, our effort will be also put on looking for the most appropriate classification method, that would lead to most accurate models.

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] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, “The protein data bank,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, January 2000.
- [2] RCSB Protein Data Bank, <http://www.rcsb.org>, 2019.
- [3] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, “Scop: a structural classification of proteins database for the investigation of sequences and structures,” *J. Mol. Biol.*, vol. 247, no. 4, pp. 536–540, April 1995.
- [4] C. A. Orengo, A. D. Michie, D. T. Jones, M. B. Swindells, and J. M. Thornton, “CATH – a hierarchic classification of protein domain structures,” *Structure*, vol. 5, no. 8, pp. 1093–1108, August 1997.
- [5] S. B. Needleman and C. D. Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins,” *J. Mol. Biol.*, vol. 48, no. 3, pp. 443–453, March 1970.
- [6] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, October 1990.
- [7] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs,” *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, September 1997.
- [8] I. N. Shindyalov and P. E. Bourne, “Protein structure alignment by incremental combinatorial extension (CE) of the optimal path,” *Protein Eng.*, vol. 11, no. 9, pp. 739–747, September 1998.
- [9] A. R. Ortiz, C. E. Strauss, and O. Olmea, “Mammoth: an automated method for model comparison,” *Protein Sci.*, vol. 11, no. 11, pp. 2606–2621, November 2002.
- [10] L. Holm and C. Sander, “Protein structure comparison by alignment of distance matrices,” *J. Mol. Biol.*, vol. 233, no. 1, pp. 123–138, September 1993.
- [11] S. Cheek, Y. Qi, S. S. Krishna, L. N. Kinch, and N. V. Grishin, “SCOPmap: automated assignment of protein structures to evolutionary superfamilies,” *BMC Bioinformatics*, vol. 5, pp. 197–221, December 2004.
- [12] C. H. Tung and J. M. Yang, “fastSCOP: a fast web server for recognizing protein structural domains and SCOP superfamilies,” *Nucleic Acids Res.*, vol. 35, W438–W443, July 2007.
- [13] K. Marsolo, S. Parthasarathy, and C. Ding, “A multi-level approach to SCOP fold recognition,” *IEEE Symposium on Bioinformatics and Bioeng.*, pp. 57–64, October 2005.
- [14] P. H. Chi, Efficient protein tertiary structure retrievals and classifications using content based comparison algorithms, PhD thesis, University of Missouri-Columbia, 2007.
- [15] G. Mirceva, I. Cingovska, Z. Dimov, and D. Davcev, “Efficient approaches for retrieving protein tertiary structures,” *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 9, no. 4, pp. 1166–1179, July/August 2012.
- [16] R. Quinlan, *C4.5: Programs for Machine Learning*, 1st ed. San Mateo, CA, USA: Morgan Kaufmann Publishers, 1993.
- [17] G. H. John and P. Langley, “Estimating continuous distributions in bayesian classifiers,” In: Besnard, P., Hanks, S. (Eds.), *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Francisco, CA, USA, pp. 338–345, 1995.
- [18] N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian network classifiers,” *Mach. Learn.*, vol. 29, no. 2–3, pp. 131–163, November/December 1997.
- [19] D. W. Aha, D. Kibler, and M. K. Albert, “Instance-based learning algorithms,” *Mach. Learn.*, vol. 6, no. 1, pp. 37–66, January 1991.
- [20] V. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed., Springer, New York, 1999.
- [21] C. J. C. Burges, “A tutorial on support vector machine for pattern recognition,” *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.
- [22] D. V. Vranic, *3D Model Retrieval*, Ph.D. Thesis, University of Leipzig, 2004.
- [23] G. Mirceva, M. Mirchev, and D. Davcev, “Hidden Markov Models for classifying protein secondary and tertiary structures,” *Journal of Convergence*, vol. 1, no. 1, pp. 57–64, 2010.

Link Prediction on Bitcoin OTC Network

Oliver Tanevski

*Faculty of Computer Science and
Engineering*

Skopje, North Macedonia

oliver.tanevski@students.finki.ukim.mk

Igor Mishkovski

*Faculty of Computer Science and
Engineering*

Skopje, North Macedonia

igor.mishkovski@finki.ukim.mk

Miroslav Mirchev

*Faculty of Computer Science and
Engineering*

Skopje, North Macedonia

miroslav.mirchev@finki.ukim.mk

Abstract—Link prediction is a common problem in many types of social networks, including small Weighted Signed Networks (WSN) where the edges have positive and negative weights. In this paper, we predict transactions between users in Bitcoin OTC Network, where the links represent the ratings (trust) that the users give to each other after each transaction. Before predicting, we transform the network where we convert negative weights into positive so that the feature scores, calculated by existing algorithms (such as Common Neighbours, Adamic Adar etc.) would improve the models performance in our link prediction problem. We consider two methods that will help us in our link prediction: attributes estimation based on similarity scores link prediction and link prediction as supervised learning problem. The first method can be used more as a way to determine which of the attributes (feature scores) are more important in link prediction. The second method is used for estimating attributes importance, but even more for actual prediction using the calculated feature scores as input to the machine learning and deep learning models. The predicted links can be interpreted as possible transactions between certain users.

Index Terms—link prediction, weighted signed directed graphs, network science, machine learning

I. INTRODUCTION

In the past several years, the Bitcoin and other cryptocurrencies emerged into the digital world of transactions. The ability to transfer unlimited amount of money without paying fees to a third party, like banks, and the idea of mining Bitcoin using the blockchain technology, caught the users interest. Because of that, a lot of platforms were created where people can make these transactions based on certain market policy. When a user registers on such platform and it has been authenticated, he can later use it to buy or sell bitcoins. On some platforms, when a transaction has been completed, users can rate each other. That rating indicates how trustworthy the user is and helps other users to estimate this user whenever they want to make a transaction with him in the future.

In this paper, we use the Bitcoin OTC Network¹ in order to predict whenever a possible transaction will be made between two users in the network. In order to achieve this we use

a plethora of methods and algorithms from network science and machine learning. We represent this network as a graph, where links indicate the user ratings. Since users usually rate each other after transaction or several transactions, the links also indicate that those users made a transaction, which comes down to link prediction in a graph and/or predicting future transactions in the network.

This research can be used on other similar trading markets which are based on ratings. Also it can be used in order to determine if there would be an interaction between users or companies and who would initiate the interaction.

II. RELATED WORK

There have been many studies in the field of link prediction, most of them focused on big social networks like Facebook and Twitter [1], citation networks [2] and dynamic networks (networks where edges are created over time) [3]. The problem have been also addressed for multilayer online social networks [4]. Most of these studies use methods from machine learning, but on graphs where the edges are undirected and unweighted. There also has been work on link prediction on directed weighted graphs [5], where the weights are always positive.

In this work, we will present a plethora of algorithms and methods used on our sparse network with negative weights and evaluate them using ROC Curve and Precision-Recall Curve.

III. DATASET

We used the Bitcoin OTC network dataset [6, 7] that is available in the Stanford Large Dataset Collection.²

The dataset contains source and target nodes, where the source node gives rating to the target node that is between -10 , which means that the user did not hold his end of the bargain and that he is a fraud, and 10 , which means that this user is very trustworthy.

The Bitcoin OTC Network consists of 5881 users and 35592 links between users, which is a relatively small network. The *Average Clustering Coefficient* is 0.151, which shows that this network is clustered. We wanted to see the node degree distribution (Fig. 1) in our network in order to determine if the network is scale-free and follows a Power-Law distribution. Additionally, we calculated the *Power-law exponent*: $\gamma = 1.969$ and *Tail power-law exponent*: $\gamma_t = 2.051$. Comparing the two

¹Bitcoin OTC Network: <https://www.bitcoin-otc.com/>

²<https://snap.stanford.edu/data/soc-sign-bitcoin-otc.html>

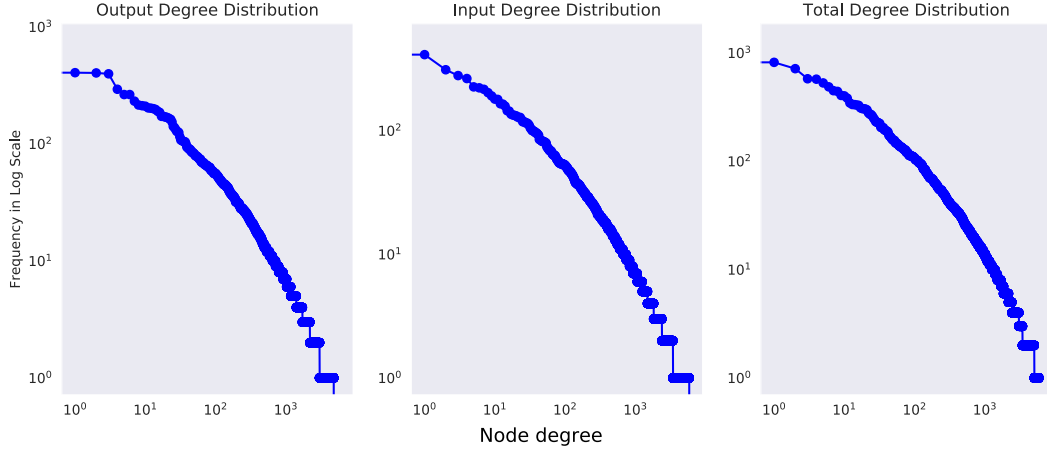


Fig. 1: Node degree distribution in the Bitcoin OTC Network.

exponents, we can see that the Bitcoin OTC Network is not a typical scale-free network, since they are not almost equal.

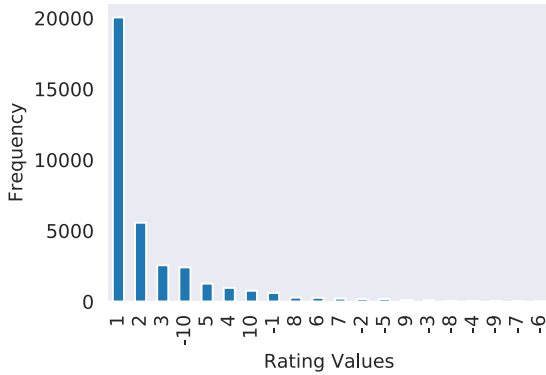


Fig. 2: Ratings given in the Bitcoin OTC Network.

We also examined what rating was most often given by the users (Fig. 2) in order to have a rough estimation of how much of the transactions were successful. We see that most of the ratings are positive, which probably indicates that most of the transactions were successful.

IV. RESEARCH METHODOLOGY

By definition, the term *Weighted Signed Network* means a directed, signed graph $G = (V, E, W)$ where V is the set of users, $E \subseteq V \times V$ is the set of edges in the graph and $W : E \rightarrow [-n, +n]$ is weight value between some $-n$ and n assigned to an edge. In our graph, a $W(u, v)$ means how much the user u "likes" or "trusts" user v . In the following, we describe the methods used for link prediction and define the attributes used by these methods.

A. Methods

In our paper, we use two methods that will help us in our link prediction. In both of these methods we consider two

graphs G_1 and G_2 where G_1 is used as a training graph and G_2 is the graph in which we predict links. The methods are:

- **Attribute estimation based on Similarity Scores link prediction**
- **Link Prediction as a Supervised Learning Problem.**

1) *Attribute estimation based on Similarity Scores link prediction:* Let u and v denote two nodes in G who are not connected in G_1 , but are connected in G_2 . Then we can assign similarity measurement $score(u, v)$ for the node pair (u, v) . Thus, we assign score value $\forall e \notin G_1, S : score(e)$ where e is the edge (node pair) that does not exist in the graph G_1 and create a list in decreasing order. Then we select the top M high scored edges and see how many of them are in G_2 . Furthermore, we will show how the accuracy changes for different sizes of M . The more the accuracy increases, the more valuable is the feature score in our link prediction.

There are mainly three groups of similarity measurement scores, where each gives different results based on the problem. Those are:

- **Local Measurement Scores** - where we calculate the score mostly based on the information depending on nodes u and v .
- **Global Measurement Scores** - where we use all paths in the graph to calculate the score. These measurement scores usually give good results, but depending on the graph it may take very long time to compute.
- **Quasi-Local Measurement Scores** - which are in-between measurement scores, that balance the trade-off between accuracy and computing complexity.

We must emphasize that in this work we focus only on local measurement scores and estimate their importance for link prediction.

2) *Link Prediction as a Supervised Learning Problem:* From the original graph G we choose node pairs (u, v) that do not exist in the graph. We choose equal number of non-existing edges and existing edges in order to solve the problem with class imbalance while training. Whenever there is a

present edge in the graph, we label it as positive, whereas the non-present edges are labeled as negative. This is crucial pre-processing step before we split the original graph in two subgraphs G_1 and G_2 .

For each sample, we use a variety of attributes, such as: topological (global), neighbourhood-based (local) etc. in order to form a dataset for training and test, which is then fed into the machine learning and deep learning models. In the next subsection, we will give an insight of the types of attributes used by our models.

B. Attributes - Feature Scores

As we mentioned before, in each of our methods we will use attributes (feature scores) that will help us build models and evaluate them later on. All of these attributes represent some score for a given pair of nodes u and v .

1) *Common Neighbors Score*: The idea of common neighbours comes from social networks where it states that if two strangers who have a friend in common a more likely to be introduced to each other than those who do not have any friends in common. For our problem, we use Laishram's [5] variation of Common Neighbours in directed weighted graphs:

$$CN_{i,weighted}(u,v) = \sum_{z \in (\Gamma_i(u) \cap \Gamma_i(v))} \frac{w(z,u) + w(z,v)}{2},$$

$$CN_{o,weighted}(u,v) = \sum_{z \in (\Gamma_o(u) \cap \Gamma_o(v))} \frac{w(u,z) + w(v,z)}{2},$$

where u and v are the nodes, Γ_i are the predecessors of a given node and Γ_o are the successors of a given node and $w(x,y)$ is the link's weight.

2) *Preferential Attachment Score*: Preferential attachment is measure to compute the closeness of two nodes based on their neighbours. If both of the nodes u and v have more neighbours, the chance is bigger for them to connect. We will use Laishram's [5] weighted variation:

$$PA_{i,weighted}(u,v) = \left(\frac{\sum_{z \in \Gamma_i(u)} w(z,u)}{|\Gamma_i(u)|} \right) * \left(\frac{\sum_{z \in \Gamma_i(v)} w(z,v)}{|\Gamma_i(v)|} \right),$$

$$PA_{o,weighted}(u,v) = \left(\frac{\sum_{z \in \Gamma_o(u)} w(z,u)}{|\Gamma_o(u)|} \right) * \left(\frac{\sum_{z \in \Gamma_o(v)} w(z,v)}{|\Gamma_o(v)|} \right).$$

3) *Adamic Adar Score*: Adamic Adar score [8] is based on a concept that if a person has a lot of friends, and he is a common friend or acquaintance of other two people, then it is less likely to introduce the two people to each other, other than when he would have very few friends [9]. We will use Laishram's [5] weighted variation of Adamic Adar:

$$AA_{i,weighted}(u,v) = \frac{1}{2} \left(\sum_{z \in (\Gamma_i(u) \cap \Gamma_i(v))} \frac{w(z,u) + w(z,v)}{\log(\sum_{x \in \Gamma_i(z)} w(x,z))} \right),$$

$$AA_{o,weighted}(u,v) = \frac{1}{2} \left(\sum_{z \in (\Gamma_o(u) \cap \Gamma_o(v))} \frac{w(u,z) + w(v,z)}{\log(\sum_{x \in \Gamma_o(z)} w(z,x))} \right).$$

4) *Shortest Path Score*: Shortest path score is a global measurement score that calculates reciprocal of the length of all shortest paths from u to v . We use Laishram's [5] weighted variation of Shortest Path:

$$SP_{weighted}(u,v) = \frac{1}{|\rho(u,v|l_{min})|} \left(\sum_{p \in \rho(u,v|l_{min})} \frac{\sum_{(x,y) \in p} w(x,y)}{l_{min}} \right),$$

where $|\rho(u,v|l_{min})|$ are number of paths between u and v with length l_{min} , and l_{min} is the length of the shortest path.

5) *Fairness and Goodness*: The Fairness and Goodness algorithm [6] is an extension of HITS [10] for directed signed graphs. The fairness attribute of a node is a measure of how fair is the node in giving ratings to other nodes (users). The goodness attribute of a node measures how much the other nodes "like" him and what is his true quality. Fairness and Goodness are calculated as follows:

$$\text{Let } f^0(u) = 1 \text{ and } g^0(u) = 1, \forall u \in V,$$

$$g^{t+1}(u) = \frac{1}{|\Gamma_i(u)|} \sum_{z \in \Gamma_i(u)} f^t(z) \times w(z,u),$$

$$f^{t+1}(u) = 1 - \frac{1}{|\Gamma_o(u)|} \sum_{z \in \Gamma_o(u)} \frac{|w(u,z) - g^{t+1}(z)|}{R},$$

where R is the maximum difference between an edge weight and goodness score [6]. The full algorithm is described in Kumar's paper [6].

6) *Jaccard Similarity*: Jaccard Similarity is a measure that shows us how similar are two nodes in a graph by calculating how many common neighbours they have from all their neighbours. The Jaccard Similarity for directed graphs will be calculated as follows:

$$JC_i(u,v) = \frac{|\Gamma_i(u) \cap \Gamma_i(v)|}{|\Gamma_i(u) \cup \Gamma_i(v)|},$$

$$JC_o(u,v) = \frac{|\Gamma_o(u) \cap \Gamma_o(v)|}{|\Gamma_o(u) \cup \Gamma_o(v)|}.$$

7) *Cosine Similarity*: Cosine Similarity in a directed graph is calculated as follows:

$$CS_i(u,v) = \frac{|\Gamma_i(u) \cap \Gamma_i(v)|}{|\Gamma_i(u)| * |\Gamma_i(v)|},$$

$$CS_o(u,v) = \frac{|\Gamma_o(u) \cap \Gamma_o(v)|}{|\Gamma_o(u)| * |\Gamma_o(v)|}.$$

8) *Katz Score*: Katz score is similar to Shortest path score, however instead of taking only the shortest paths, it takes all paths in the graph. The Katz score in directed graph is calculated like this:

$$KS(u, v) = \sum_{l=1}^{\infty} \beta^l * |\rho(u, v|l)|,$$

where $|\rho(u, v|l)|$ are the number of paths between u and v with length l and β is damping factor between 0 and 1. We used the unweighted variation of Katz score since the weighted one was computationally expensive.

9) *PageRank Score*: The PageRank Score is a node based feature and is calculated as follows:

$$\text{Let } PR(u) = \frac{1}{|V|}, \forall u \in V,$$

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)},$$

where B_u are the links that u is connected to and $L(v)$ are the number of links that node v points to.

10) *HITS*: HITS [10] or *Hubs and Authorities* was initially created for rating Web pages. Authorities in our case will be the users who participated in most of the transactions, while hubs are those users that can point us to the authoritative users. HITS is calculated as follows:

$$\text{Let } u = (1, \dots, 1),$$

$$v = A^T u, \quad u = Av,$$

where A is the adjacency matrix, u is the hub weight vector and v is the authority weight vector. Since we use Fairness and Goodness as weighted variation of HITS, we also wanted to use the unweighted variant for comparison and determine which one is more valuable in the prediction.

V. RESEARCH STUDY

In this section, we study the problem of link prediction in Weighted Signed Network. There are many studies on link prediction focused on the big online social networks, which prompted us to experiment on smaller networks with the previously mentioned approaches and attributes.

Before we explain how we used the methods, we need to point out that certain weight mapping was used in order to calculate the feature scores. For all attributes except Shortest Path Score, the mapping used is $[-10, 10] \rightarrow [1, 20]$, while in Shortest Path a mapping of $[-10, 10] \rightarrow [20, 1]$ is used. The mapping was not used for the Fairness and Goodness attribute, since the algorithm itself can handle negative weights.

In the first method (i.e. Attribute estimation based on Similarity Scores link prediction), we find the maximal strongly connected component and we split it in two subgraphs G_1 and G_2 where we take 15% of the edges in the maximal component to the subgraph G_2 and the rest to G_1 . We mentioned above that we will use different sizes for M which is the number of top high scored edges that do not exist in graph G_1 . M takes values $l_{est} 2^i, \forall i \in [0, 1, \dots, k]$, where k is a number which

satisfies the condition $l_{est} 2^k$ equals 15% of the number of all possible edges in G_1 and l_{est} is the number of edges in graph G_2 . We also mentioned that for this method we will use only local feature scores. Those are **Common Neighbours**, **Adamic Adar** and **Preferential Attachment**. With each of these measures, we calculate the score for all edges that do not exist in graph G_1 and we select the top M high scoring edges and see how much of them belong in G_2 as we change M .

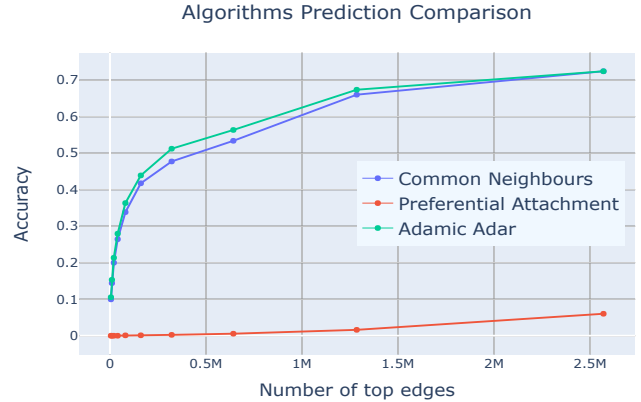


Fig. 3: Algorithms accuracy based on top M edges selected

TABLE I: Accuracy for each algorithm

Value of M	Algorithms		
	Common Neigh.	Adamic Adar	Preferential Attach.
5019	0.099	0.105	0.0
10038	0.144	0.153	0.0
20076	0.2	0.213	0.0
40152	0.264	0.28	0.0
80304	0.339	0.363	0.001
160608	0.418	0.439	0.001
321216	0.477	0.512	0.002
642432	0.534	0.564	0.006
1284864	0.66	0.674	0.016
2569728	0.724	0.724	0.06

From the results above, we can see that Adamic Adar gives the best results. Common Neighbours has similar accuracy results compared to Adamic Adar in the beginning and same accuracy in the end. However, Preferential Attachment gave very poor result. These results gives us a rough estimate about which local features are more significant for link prediction when using the second method, i.e. Link Prediction as a Supervised Learning Problem.

In the second method, we select edges whose links are not present in the original graph G . Since the graph is very sparse, the number of selected non-present edges will be very high which will produce very high class imbalance. For the edges that exist in the graph, we label them as positive, while the other as negative. Next, we split our main graph G with the generated edges into G_1 , which is used for training and has equal amount of existing and non-present edges in order to solve the problem of class imbalance while training, and G_2 ,

which is used for test and uses 15% of the existing edges and 15% of the non-present edges, which makes it highly unbalanced. For each of these graphs, we generate a dataset where the attributes are all the feature scores we previously mentioned along with their node pair whose feature scores are calculated. Afterwards, we remove the node pairs from the dataset in order to prevent the machine learning algorithms to be biased based on the edges. These datasets were used for training and evaluating several machine learning and deep learning models, which are shown in the following figures.

For evaluation metrics, we used ROC Curve and Precision-Recall Curve for each of the models. We used *5-fold Cross Validation with Grid Search* in order to find the optimal hyper-parameters for our machine learning models.

For our deep learning model, we used classical Feed Forward Network with the idea of Residual Networks [11] by using only the *skip connections*. For simplicity, we called it SkipConnNet. We used 3 hidden Linear layers, each with 100 neurons. The optimizer used was Ranger [12] with learning rate of 0.0001. We used validation split of 10% and batch size of 128. The loss function we used was Mean False Error (MFE) [13]. The model was created using Pytorch[14] and Poutyne [15].

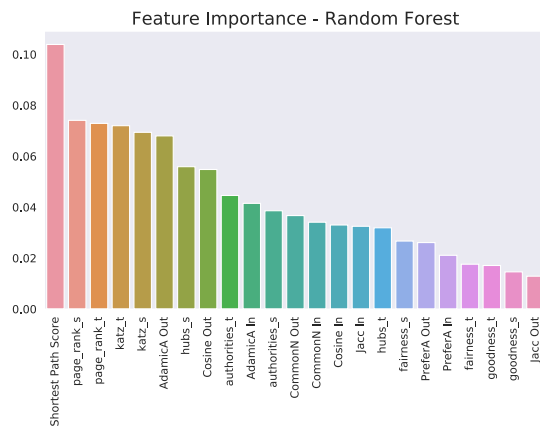


Fig. 4: Feature Importance using Random Forest

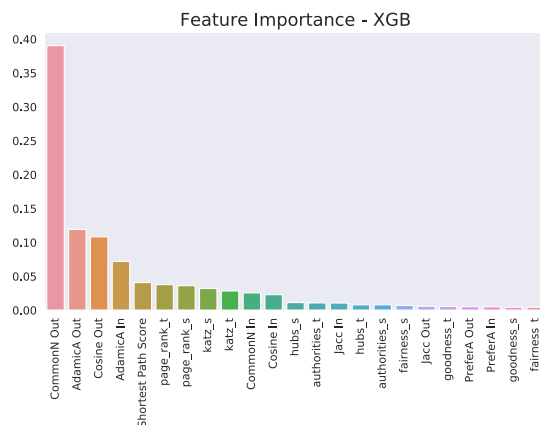


Fig. 5: Feature Importance using XGB

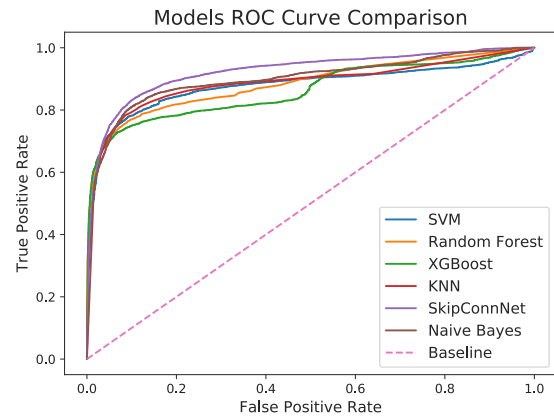


Fig. 6: Models ROC Curves

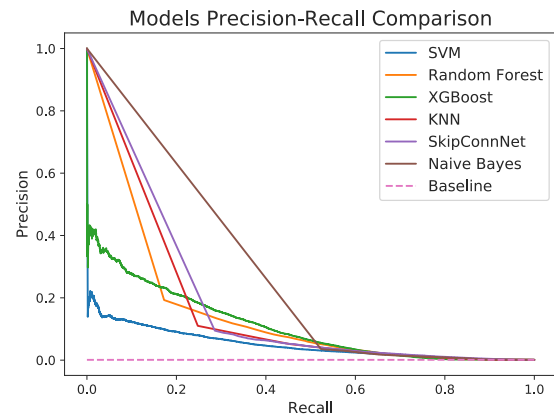


Fig. 7: Models Precision-Recall Curves

As we mentioned above, we have a highly imbalanced test set for evaluation, where the dominant class is the negative one (ratio of around 1:1000). From Fig. 6 we can see that the **SkipConnNet** has the best ROC curve than the other models. However, in Fig. 7 is hard to distinguish which model performs better. All of them vary significantly for certain precision and recall, however for recall above 0.5, all of them converge to a certain precision value.

We also used Random Forest and XGB (Fig. 4 and Fig. 5) to see which features impacted their prediction the most. On both plots, on the x-axis, we see two formats of labeling our features **xxx_s/xxx_t**, which means that this feature was calculated individually for both Source and Target nodes respectively, and **xxx Out/xxx In**, which means that this feature was calculated based on the successors and predecessors of both Source and Target nodes respectively. It can be seen that Common Neighbours and Adamic Adar have bigger impact than Preferential Attachment, which is what we roughly estimated with the first method.

VI. CONCLUSION

In our research we have shown:

- how we use the problem of link prediction to predict a possible transaction in Bitcoin OTC Network,
- that certain weight mapping from negative to positive can improve the score from Common Neighbours, Adamic Adar etc. and with that, the overall model performance,
- that link prediction using similarity scores can also be used to give a rough estimation of how much each feature is important in the prediction,
- that link prediction as supervised learning problem with machine learning models can give decent result on a sparse network with very big class imbalance like the one we used.

In the future we will extend this work by using state-of-the-art networks like Node2Vec to generate additional features and improve the performance of our models even more.

REFERENCES

- [1] Peng Wang et al. “Link Prediction in Social Networks: the State-of-the-Art”. In: *CoRR* abs/1411.5118 (2014). arXiv: 1411.5118. URL: <http://arxiv.org/abs/1411.5118>.
- [2] Mohammad Al Hasan et al. “Link prediction using supervised learning”. In: 2006.
- [3] Catherine A. Bliss et al. *An Evolutionary Algorithm Approach to Link Prediction in Dynamic Social Networks*. 2013. arXiv: 1304.6257 [physics.soc-ph].
- [4] Haris Mandal et al. “Multilayer Link Prediction in Online Social Networks”. In: *2018 26th Telecommunications Forum (TELFOR)*. IEEE. 2018, pp. 1–4.
- [5] Laishram Ricky. “Link Prediction in Dynamic Weighted and Directed Social Network using Supervised Learning”. In: (2015). Dissertations - ALL. 355. URL: <https://surface.syr.edu/etd/355/>.
- [6] Srijan Kumar et al. “Edge weight prediction in weighted signed networks”. In: *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE. 2016, pp. 221–230.
- [7] Srijan Kumar et al. “Rev2: Fraudulent user prediction in rating platforms”. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM. 2018, pp. 333–341.
- [8] Lada A. Adamic and Eytan Adar. “Friends and neighbors on the Web”. In: *Social Networks* 25 (2001), pp. 211–230.
- [9] Fei Gao et al. “Link Prediction Methods and Their Accuracy for Different Social Networks and Network Metrics”. In: *Scientific Programming* (2015), p. 172879. ISSN: 1058-9244. DOI: 10.1155/2015/172879. URL: <https://doi.org/10.1155/2015/172879>.
- [10] Jon M. Kleinberg. “Authoritative Sources in a Hyperlinked Environment”. In: *J. ACM* 46.5 (Sept. 1999), pp. 604–632. ISSN: 0004-5411. DOI: 10.1145/324133.324140. URL: <https://doi.org/10.1145/324133.324140>.
- [11] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].
- [12] Michael R. Zhang et al. *Lookahead Optimizer: k steps forward, 1 step back*. 2019. arXiv: 1907.08610 [cs.LG].
- [13] Shoujin Wang et al. “Training deep neural networks on imbalanced data sets”. In: July 2016, pp. 4368–4374. DOI: 10.1109/IJCNN.2016.7727770.
- [14] Adam Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035. URL: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- [15] Frédéric Paradis. *Poutyne: A Keras-like framework for PyTorch*. <https://poutyne.org>. 2018.

Single RNA Secondary Structure Prediction based Dynamical programming algorithms: to parallelize or not?

Bisera Chauleva¹

Ljubinka Sandjakoska¹

Atanas Hristov³

¹ Faculty of Computer Science and
Engineering at
UIST “St.Paul the Apostle”,
Ohrid, Republic of North Macedonia
bisera.cauleva@cse.uist.edu.mk

¹ Faculty of Computer Science and
Engineering at
UIST “St.Paul the Apostle”,
Ohrid, Republic of North Macedonia
ljubinka.gjergjeska@uist.edu.mk

³ Faculty of Information and
Communication Sciences at
UIST “St.Paul the Apostle”,
Ohrid, Republic of North Macedonia
atanas.hristov@uist.edu.mk

Abstract— RNA Secondary Structure Prediction has a huge importance for Bioinformatics. Over the last decade, Dynamical Algorithms used for that purpose reached performance bottlenecks, with data produced by RNA sequences. The main idea of this paper is to answer the question – to parallelize or not? We aim to achieve a better performance over different algorithms. The chronological development of algorithms is followed and we try to obtain better execution time accordingly, where we introduce comparison between serial and parallel version of the algorithm. As a performance measurements are obtained the Time Complexity and Accuracy Level with accent on the best algorithm for the purpose needed.

Keywords—Bioinformatics, Dynamical Programming, RNA Secondary Structure Prediction, Parallelization,

I. INTRODUCTION

RNA is defined as the second most important element after DNA, which is complex molecule that takes function in cellular protein synthesis. Its structure is used in encoding and decoding genes, as well as regulation of their expression in living organisms. RNA is constructed of ribose nucleotides connected with phosphodiester bonds, forming strands of varying lengths. The nitrogenous bases in RNA are adenine, guanine, cytosine, and uracil. The three-dimensional structure of RNA is critical for its stability and function, since bases could be modified in different ways by cellular enzymes and manipulation of groups to the chain. RNA could be defined as a structure which features are between linear molecule and 3-D structure. If secondary structure is taken in consideration, the RNA is composed of double stranded regions, when the single linear RNA is folded upon itself. Different structures are evaluated further with some algorithms such as: *Stem Loops or Hairpins; Bulge Loops; Interior Loops; Junctions or Multiloops; PseudoKnots* etc. RNA secondary structure was discovered by X-Ray methods, which were extremely hard and expensive to be performed for all possible RNA sequences. Therefore a new concept for faster and efficient formation of a secondary structure was invented, known as computational prediction of RNA secondary structure. One type of computational prediction is by dynamical programming (DP) which is a useful technique for complex problems like the RNA structure. With the help of DP algorithms calculation is done over one major problem when subdivided on multiple smaller problems, therefore efficient prediction of the RNA structure could be performed. There are different algorithms using this method that are going to be presented in our work, such as:

- *Nussinov-Jacobson Algorithm* – for formation of secondary structure of RNA based on folding upon itself. That base-pairs can form secondary structure. This algorithm is known as the first algorithm for that purpose, with time complexity of $O(n^3)$ [1]. Different type of implementation of this algorithm is known as Four Russians Algorithm with perfect time complexity of $O(n^3/\log(n))$ [4];
- *Minimal-free Energy Algorithm* (Zucker’s Algorithm) where the main focus is on amount of free energy expressed by each adjacent base pair. Since different RNA structure has different amount of free energy presented, this algorithm gives great accuracy for shorted RNA sequences, with time complexity of $O(n^4)$ [3];
- *Maximum Expected Accuracy* (MEA) - mostly focused on partition function calculation based on McCaskill’s Algorithm, that utilizes the free energy change with the usage of nearest-neighbor parameters. MEA predict base pair probabilities as well as probabilities of nucleotides being single-stranded, with time complexity of $O(n^3)$ [6] [7];
- *Pseudoknotted Algorithm* – simple dynamic programming algorithm for RNA secondary structure prediction with pseudoknots which is mainly based on the theoretical approach of Akutsu’s algorithm, with time complexity of $O(n^5)$ [10][11];

In this work, all of the mentioned algorithms are going to be evaluated. The evaluation is followed by parallelization with OpenMP in order to find proved answer of the question – *to parallelize or not* where we are interested in better performance for different sequence lengths. The rest of the paper is organized as follow. In the second section theoretical background of the RNA second structure prediction algorithms is given. In the third, the implementation and testing of parallelization with OpenMP is presented, followed by performance and accuracy discussion sections. Before the list of used references, in the last section, the experimental setup that include the testing platform and the description of the requirements is given.

II. RNA SECONDARY STRUCTURE PREDICTING ALGORITHMS

A. Nussinov-Jacobson Algorithm

Nussinov-Jacobson algorithm is proposed by Nussinov and Jacobson in 1978 [1]. It is defined as algorithm for secondary RNA structure prediction based on the folding principle, when the RNA strand is folding onto itself, without

taking in consideration formations like pseudoknots. Mainly it considers usage of the maximum amount of base pairs for optimization of the score. It is based on the usage of the standard 2D array. As a score values of X_i and X_j are used in form of a matrix $M[i][j]$. This algorithm, take in considerations two cases: if leftmost base is unpaired or paired with other base. We use three stages when constructing the Nussinov-Jacobson algorithm [1]:

1. Initialization step, in this step the scoring of matching elements present on the main diagonal and the diagonal below it are done, where the rules respected are: $M[i][j] = 0$ for $i=1$ to L , and $M[i][i-1] = 0$ for $i=2$ to L , where L is the length of the RNA

2. Recursive step, which considers fulfilment of the matrix, using the four major conditions like: if i is unpaired, added onto a structure for $i+1, j$, for it the matrix will follow $M[i+1][j]$ (i th residue is hanging off), if j is unpaired, added onto a structure for $i, j-1$, where the matrix follows $M[i][j-1]$ (j -th residue is hanging off), if i and j base pair are added on to a structure for formation of $i+1, j-1$, where the matrix has $M[i+1][j-1] + S(x_i, x_j)$ (i -th and j -th residues are paired and if x_i is complement of x_j , then $S(x_i, x_j) = 1$; otherwise it is 0, and finally if i and j are making a base pair but not to each other, the structure for $i..j$ adds together to a substructures, for two sub-sequences, $i..k$ and $k+1..j$ (making a bifurcation). The matrix will follow $M[i][j] = \text{MAX}_{i < k < j} (M[i][k] + M[k+1][j])$ (merging two substructures). (Fig 1.)

3. Traceback step, considers the formation of actual secondary structure of RNA sequence based on the trace-back from the given scores in the matrix, which are filled in by the previous steps. This step is missed in the algorithm used in this project, it can be formed by the scores obtained from the matrix.

In order to implement these steps practically must use computation of M matrix by diagonals and within a diagonal from top to bottom. Calculated run time is $O(n^3)$. Even if the algorithm is written using two-dimensional array notation for M , we need only the upper triangle of M . Therefore if we want to have memory efficient implementation, with the usage of either mapping of the upper triangle into a 1D array or dynamically allocated 2D array with variable size rows, must be used. In both cases it is expected to have memory for $n*(n+1)/2$ elements of M [2].

For better time complexity we consider another type of implementation known as Four Russian Algorithm [4]. The contribution is both theoretical and practical, since the basic RNA-folding problem is often solved multiple times in the inner-loop of more complex algorithms and for long RNA.

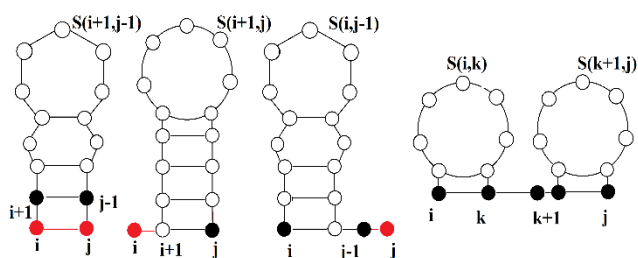


Fig. 1. Recursive step possible cases with pairing

As RNA input for this folding problem we take a string K of length n as representative of nucleotides, and an element d

to show the maximum distance between two sites of a match. In order to obtain matching we use a set M which holds the pairs that are disjoint to a set K which hold other sides. If pair (i, j) is in M , then the nucleotide i will match to the one on site of j . We obtain a permitted match if the nucleotides at sites i and j are complimentary, and $|i - j| > d$. M is non-crossing or nested if and only if it does not contain any four sites $i < i' < j' < j$ where (i, j) and (i', j') are matched in M . If we place the sites of K in a circular order, and draw a straight line between them, then in each pair in M , will have a non-crossing pair if and only if no two straight lines cross. Finally, a permitted matching M is a matching that is non-crossing, where each match in M is a permitted match [4]. From the following we obtain a cubic-time algorithm is we consider the work over three nested loops, for j, i and k that make increments of $O(n)$ times when entered. The speed-up can be obtained where instead of incrementing k through each value from $j-1$ down to $i+1$, it is practical to make a combination into groups of size q , which gives constant amount of time per group. This modification is done with the introduction of a vector V_g . [4]

B. Minimum Free Energy (Zuker's) Algorithm

In order to calculate the minimum energy we are adding experimentally predetermined values for each base pair, which is found in the dynamic programming matrix. The free energy depends on the sequence part of actual segment and the most adjacent base pairs. The total free energy is the sum of all increments. This concept is implemented as algorithm for RNA secondary structure prediction, available as MFold. The approach is also known as Minimum Free Energy (MFE) and was developed by M. Zuker [5]. There are certain limitations to MFE method such as that within the method, energies of bulge loops and single non-canonical pairs are not predicted. Zuker's Algorithm [9] mainly uses approach that divides a secondary structure, such that loops also known as graphs are used, and the free energy value is given based upon those graphs. Calculation of the lowest free energy structure, gives us the optimal structure of RNA molecule with consideration of maximum base pair amount. Zuker's algorithm takes in account different energies for calculation.

Zuker's algorithm defines two matrices $W(i, j)$ and $V(i, j)$, where $W(i, j)$ is the total free energy of subsequence i to j . For the $V(i, j)$ is defined as the total free energy of subsequence i to j if i and j pairs, otherwise, $V(i, j) = \infty$ consecutively $FH(i, j)$ is the energy of hairpin loop $i..j$. Whereas $FL(i, j, h, k)$ is the energy of 2nd order loop such as stack region, bulge loop and interior loop $i...h...k...j$. The last item is the energy for bifurcation loop, where item repeats over $i+1 < k < j-1$ because i and j must be a base pair, otherwise $V(i, j) = \infty$. From where $W(1, L)$ gives the final total minimum free energy (Fig. 2).

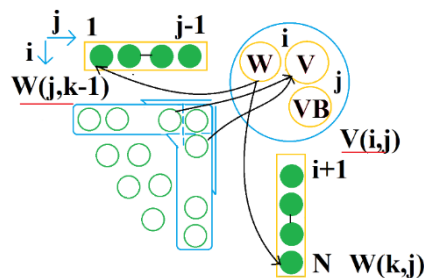


Fig. 2. Internal Loops Wand V matrix calculation depicted

C. Maximum Expected Accuracy (MEA) Algorithm

The concept of maximum expected accuracy algorithm, is a concept where in order to predict a secondary structure of RNA we need to use the technique of partition function calculation [10], with which we predict base-pair probabilities. These probabilities are then used by dynamic programming algorithm, which we can find in the RNA structure online service where the native C code for development and further improvement could be obtained.

MFE method finds one specific best guess for secondary structure of RNA, due to that it has drawbacks such as, when we use only one conformation at equilibrium instead of more. Compared when using MEA, assumed conformation from more conformations has probability of base pair based on a partitioning McCaskill's function algorithm. With this method usually high probability base pairs are chosen, and they have higher accuracy level. Additionally the base pair probability is less prone to change than it is in the thermodynamic measures of the MFE, which proves that some errors are overcome. MEA method uses derivation temperature of 37°C which is optimal, even if we make testing over RNA sequences which are from different organisms functioning on different temperatures. There are certain methods that have modification over this parameter too. Another characteristic of this method is the prediction of MEA structure to use only independent base pairs, which is not right when considering structures like helix one, where cooperation between pairs is present. Still there are methods for advancing this basis of algorithm that uses this characteristics to obtain higher accuracy level.

Constructing blocks of MEA method are:

a) Nearest Neighbor Parameters (NNP), which are the one used by Watson-Crick helices that go under the partitioning function developed by McCaskill;

b) McCaskill's partitioning function, which was developed by McCaskill in 1990 [10], and is mainly based on original statistical problem where the calculation of the partition function must be made first, in order to obtain further specific quantities of thermodynamic interest. The most useful part of the equilibrium ensemble of structures is in the binding probabilities between base pairs. Here the notice is made on the probabilities which are not locally determined by the sequence, but instead each probability has some effect over the equilibrium sum of the structures. From here, the formed matrix gives information about the global ensemble of structures in equilibrium. We need to establish that the equilibrium ensemble made of summed probabilities from bounded bases propose a direct comparison to enzymatic and chemical modification experiments, with goal to detect any modifications in bases exposition.

McCaskill's partition function algorithm is composed of two parts. First partitioning function scores are obtained and after that the probabilities of base pairs are calculated. Also known as folding and backtracking steps. The folding part is corresponding to the Zuker's algorithm, whereas the backtracking is completely authentic.

QB_{ij} , is defined as the partition function of the substrings i and j which are paired and Q_{ij} is taken for the unconstrained partition function. From there the partition function of whole molecule would be given as $Q = Q_{1n}$. If we end up with i and

j which are paired, then we can form a hairpin loop or interior loop i^*j or h^*l or eventually a loop with multiple components.

The variables Q_m and Q_{ml} are used for handling of the multiloop formations. Also we have variable of Q_a which is used for the size of the internal loops which makes modification in the time from $O(n^4)$ to $O(n^3)$. In the backtracking of the algorithm for the pairing probabilities P_{ij} we get value from the partitioning function QB_{ij} and Q_{ij} mentioned before.

For the implementation purposes we use the simplified version of the Nussinov-like energy scoring scheme, where each pair formed in a structure has a contribution to a fixed energy term E_{bp} which is aside of its context. From here we form two dynamic programming tables Q and Q_{bp} . The partition function for a sub-sequence from position i to position j is provided by Q_{ij} . Array Q_{bp} holds the partition function of the sub-sequences, which form a base pair or 0 if base pairing is not possible.

Recursive functions are used to compute Q and Q_{bp} . The input data are RNA sequence S as a chain of nucleotides. We have specification of minimal loop length l (also defined as minimal number of enclosed positions), energy weight of base pair E_{bp} and normalized temperature RT . The memory complexity of the arrays is $O(n^2)$, while the time complexity of a direct implementation of this algorithm is $O(n^3)$ in the sequence of length n .

D. Pseudoknotted Algorithm

When choosing which algorithm we can use it is difficult to choose upon high amount of algorithms suggested because some algorithms lack the accuracy of prediction in the pseudoknots considering the following characteristics: based on lack of knowledge in area of energy models we have more difficulty in discovery of secondary structures, the folding principle in formation of a structure can be affected by kinetic energies, ligand-binding, interactions in transition, and finally small amount of experiments result in small knowledge in pseudoknot formation, therefore only H-type of pseudoknots are found with most of the algorithms.

For the suggested dynamic programming algorithm in first step we use the algorithm for prediction of MFE structure which will have up to 100 suboptimal structures predicted. Additionally this algorithm will be able to generate dot plot which will provide nucleotides i and j , with MFE structure containing i - j base pair. The ΔG° values will be calculated using the current Turner nearest-neighbor parameters but with the multi-branch loop [13]. Pseudoknot helix list, H , follows with corresponding helix energies. There is some criteria that needs to be pleased to get into helix list, such as sequence size restriction to be longer than 100 nucleotides, due to which examined sequences are of 200, 500 and 1000 nucleotides, also ΔG° must be 25% of the free energy of MFE. The ΔG° of H_i will be obtained from the nearest-neighbor AU/GU pairs.

Filtration of helices H is done in some particular steps. A helix H_i is accepted into H if it has more than 3 base pairs. Helices are going through comparison with the MFE structure. If they have more than 50% base pairs paired in the MFE they are discarded. For each H_i , a new set of structures, taken lowest of all MFE structures and up to 99 suboptimal structures, is generated by the dynamic programming algorithm, with H_i prohibited from pairing. [11]

Consecutively, base pairs from H_i are restored to the structures. The ΔG° of each structure is incremented by the free energy of the corresponding helix H_i . All unique structures are added to S . In S , an entropic cost of the pseudoknot formation is generated by the ΔG°_{PK} . In order to be a pseudoknot it must have at least two helices in a formation, one side of a helix to match with base pairs in the second helix. (Fig.3) Formed structures in pseudoknot can be: structure (SS) with single-stranded nucleotides inside the pseudoknot and (NE) representing number of nested helices inside the pseudoknot, also the IL (N) defined as the amount of in-line helices of length N . [12]

Before the intervening structures are calculated, the pseudoknot are in advance calculated with filling single and grouped mismatches with base pairs and removing isolated pairs.[15] Helices containing a single bulged nucleotide are counted as a single helix. Terms e and f give the values of entropic penalty by the distance between carbons of neighboring unpaired nucleotides and across a single base pair. In-line helix frequencies $P1$ and $P2$ are constant energy parameters that include Boltzmann constants and temperature terms and must be determined empirically. ΔG°_{PK} is added to the total ΔG° of each pseudoknot-containing structure. (Fig. 3) [15]

The sorting of S is done based on the total energy that each structure has. As the rule says, the first 20 that have lowest MFE are discarded. The one which will be executed are put in the Window parameter which separates them from other structures. In order to be part of this, the structure must have equal amount or more than the window base pairs amount. Usually a default value for the window parameter is given based on a length of a sequence. Finally, structures that are higher in folding MFE are discarded with the help of a parameter which accounts the maximum percentage of energy difference. By default the value is 10%.

Coaxial stacking of helices stabilizes pseudoknot formation and is included indirectly in the energy function. [11] Coaxial stacking has effect over the helices which are chose in order to assemble the pseudoknots.

III. PARALLELIZATION WITH OPENMP

A. Implementation

For the purpose of implementation we are going to evaluate each algorithm with corresponding parallelization segment of code considering OpenMP rules [18].

- For the Nussinov-Jacobson Algorithm, we have considered parallelization over the main recursive segment where the scoring matrix over diagonals and fullfilment of it is made;
- For the Zuker's MFE Algorithm the practise would be the same, since we have one starting recursive function taken for initialization and calculations over two matrices and different structures;
- MEA Algorithm taken advantage of the partitioning McCaskill's function, therefore the parallelization would be performed over it;

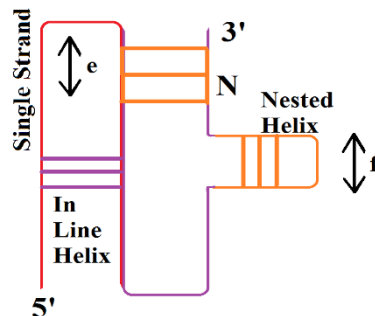


Fig. 3. Formation of pseudoknots with given penalties for in-line helix, single stranded helix and nested helix

- Finally, for the Pseudoknotted Algorithm we can take advantage of the structure S . The structures need to go under sorting based on total energy which takes a lot of time and takes advantage of parallelization.

B. Testing and Results

We are going to provide results from testing over serial implementation of algorithms.

For testing purpose we are using HSBGPG Human gene for bone gla-protein (BGP) in FASTA form. [20]

The time complexities of the represented algorithms are: Nussinov-Jacobson Algorithm defines as $O(n^3)$, Nussinov's Four Russian Algorithm defined as $O(n^3/\log(n))$, Zuker's MFE and the time complexity of $O(n^4)$, MEA algorithm with the time complexity of $O(n^3)$, the Pseudoknotted algorithm with the highest time complexity of $O(n^5)$. Obtained results and comparison are given in the Fig. 4). From the provided results we can make comparison and obtain number of times we have obtained speed up. We can test the speed up with the help of Amdahal's Law, formula given (1).

$$Speedup = \frac{1}{(1-p) + \frac{p}{N}} \quad (1)$$

According to the given formula obtained results of speed up are given on table (Fig 5). From calculated speedup we can make conclusion that parallelization can decrease execution time for at least 2.67 times.

IV. PERFORMANCE AND ACCURACY

For the purpose of accuracy level comparison of the algorithm we use the classical benchmark specification known as PPV or positive-predictive value, which is founded on the base-pair prediction accuracy. How sensitive is this benchmark value is given with the percentage of obtained base pairs that are correct, also the PPV value can be defined as a value of a structure that provides the amount of predicted pair.

These characteristics are calculated with the following formula (2) and (3).

$$PPV = \frac{\text{Number of Correct Predicted pairs}}{\text{Total Number of Predicted Base pairs}} \quad (2)$$

$$Sensitivity = \frac{\text{Number of Correct Predicted pairs}}{\text{Total Number of known Pairs}} \quad (3)$$

Win. Size	Nussinov Four Russian		Zuker's MFE		MEA		ProbKnot	
	Serial	Parallel	Serial	Parallel	Serial	Parallel	Serial	Parallel
200	18.89	08.11	1289.36	451.31	122.03	46.28	1535.03	575.40
500	72.39	34.79	4744.84	1658.88	413.12	133.69	5951.44	2635.30
1000	333.35	96.09	9354.40	3769.50	1147.09	419.21	11700.10	4460.13

Fig. 4. Execution Time of proposed Algorithms tested for Serial and Parallel Implementation of Code over different Window Size

After calculation performed over higher training set, conclusion made are based on average results for the sensitivity of algorithms and PPV:

a) Nussinov-Jacobson's Algorithm has Sensitivity average score of 0.65 or 65% and PPV of 0,48 or 48%

b) Zuker's (MFE) Algorithm has Sensitivity average score of 0.73 or 73% and PPV of 0.66 or 66%

c) MEA Algorithm has Sensitivity average score of 0.72 or 72% and PPV of 0.67 or 67%

d) Pseudoknotted Algorithm has Sensitivity average score of 0.72 or 72% and PPV of 0.76 or 76%

From where we can conclude that the smallest sensitivity level and PPV has the Nussinov algorithm as oldest method, and the best results has the Pseudoknotted algorithm as expected since it is the newest and predicts all kinds of structures.

V. DISCUSSIONS AND CONCLUSION

From the parallelization of introduced secondary structure predicting RNA algorithms, a couple of conclusions can be made. First discussion of the most important algorithms in chronological order of occurring in dynamical programming branch was given.

After the serial implementation and introduction of parallelized version, comparison in the time complexity and accuracy level was performed, from where a couple of assumptions were made.

First of all, the fastest algorithm of all was the Four Russian's algorithm or so known as the new generation of the Nussinov's algorithm, but not always the fastest algorithm means the most accurate algorithm. Considering the PPV and Sensitivity tests based on benchmarks, calculations show that the most accurate of all is the pseudoknotted algorithm, as the most advanced which is in capability to predict any kind of structure. But as a cost for this advantage, this algorithm is the slowest.

Win. Size	Nussinov	Zuker's	MEA	ProbKnot
200	2.32 Times	2.86 Times	2.63 Times	2.67 Times
500	2.08 Times	2.80 Times	3.09 Times	2.26 Times
1000	3.50 Times	2.41 Times	2.73 Times	2.63 Times

Fig. 5. SpeedUp for each tested Algorithm calculated according Amdahl's Law Formula

Depend on that how prediction of secondary structure should be done from sequence of RNA, due to gene expression, encoding and decoding of genes, it could be chosen across lots of different approaches.

In the branch of dynamical programming in bioinformatics, from the most popular and stable were chosen Zuker's as the basis for all other newly created algorithms, until reaching the last one which predicts all kinds of secondary structure formations, such as the pseudoknots. According to that, as the time goes by, the algorithms become more complex but also more capable of predicting any kind of RNA secondary structure formation from a given sequence, which again gives the chance for lots of advancements in the bioinformatics field.

In order to decrease the time complexity, we use the parallelization option with OpenMP, which showed a high amount of improvement in the execution time for about 3 times per each algorithm that has huge importance when working with longer sequences.

VI. TESTING PLATFORM AND REQUIREMENTS

For the purpose of testing of the introduced dynamical algorithms was used the platform CPU Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz, where were tested serial and parallelized version of the algorithms.

For the algorithms: Nussinov-Jacobson, Zuker's and Four Russian implementation of Nussinov, were used publically available source codes with elementary changes implied.

For the algorithms MEA and pseudoknotted prediction were used source codes taken from the RNA structure developers package where small improvements were introduced and tested on our platform. Source code in all algorithms tested was C/C++ develop and tested in the Visual Studio 2017 Package with additional support of OpenMP for the purpose of parallelization.

Characteristics of building the RNA structure algorithms MEA with McCaskill's function and pseudoknotted algorithm are the following:

a) C++ class libraries encapsulated in the I/O functions of RNA structure and also the secondary structure prediction and analysis methods were used. The classes are designed to be easily included in C++ projects;

b) Text interfaces;

c) Thermodynamic parameters, for nearest neighbor parameters used for the purpose of prediction of the stability in secondary structures. Here we include the change of free energy parameters at 37°C and the change of enthalpy parameters which are used for the conformational stability and the arbitration of temperature. These are taken from the Turner's group.

Finally in order to access the code and build we need Windows GUI: Microsoft Foundation Classes (MFC) as found in Microsoft Visual Studio 2005 or later and the Intel C++ compiler.

For the purpose of parallelization of the code, we need to enable the OpenMP option in the Visual Studio package.

As RNA input sequence we use the FASTA file format, saved as .txt outside the algorithm. Input module is defined as one which takes the FASTA file and extracts the sequence into a string.

REFERENCES

- [1] Palkowski, M., & Bielecki, W. (2017). Parallel tiled Nussinov RNA folding loop nest generated using both dependence graph transitive closure and loop skewing. *BMC bioinformatics*, 18(1), 1-10.
- [2] Zhao, C., & Sahni, S. (2017). Cache and energy efficient algorithms for nussinov's rna folding. *BMC bioinformatics*, 18(15), 518.
- [3] Zhao, C., & Sahni, S. (2017, October). Efficient RNA folding using Zuker's method. In *2017 IEEE 7th International Conference on Computational Advances in Bio and Medical Sciences (ICCBMS)* (pp. 1-6). IEEE.
- [4] Venkatachalam, B., Gusfield, D. & Frid, Y. "Faster algorithms for RNA-folding using the Four-Russians method. Algorithms"2014, Mol Biol, 2014, pp.9-5.
- [5] Will, S., Jabbari, H. "Sparse RNA folding revisited: space-efficient minimum free energy structure prediction. Algorithms" 2016, Mol Biol, 2016, pp.11-7.
- [6] Clote, P., Lou, F. & Lorenz, W.A. "Maximum expected accuracy structural neighbors of an RNA secondary structure." 2017 BMC Bioinformatics, 2017, pp.13- S6
- [7] Aghaeepour, N., Hoos, H.H. "Ensemble-based prediction of RNA secondary structures." 2013 BMC Bioinformatics, pp.14-139.
- [8] Zhang, B., Yehdego, D.T., Johnson, K.L. et al. "Enhancement of accuracy and efficiency for RNA secondary structure prediction by sequence segmentation and MapReduce."2013 BMC Struct Biol, 2013, pp.13-S3
- [9] Wu, Y., Shi, B., Ding, X., Liu, T., Hu, X., Yip, K. Y., ... & Lu, Z. J. (2015). Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data. *Nucleic acids research*, 43(15), 7247-7259.
- [10] Palkowski, M., & Bielecki, W. (2019, September). Parallel cache-efficient code for computing the McCaskill partition functions. In *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)* (pp. 207-210). IEEE.
- [11] Sharma, D., Singh, S., & Chand, T. (2015, January). RNA Pseudoknot: Topology and prediction. In *2015 International Conference on Computer and Computational Sciences (ICCCS)* (pp. 244-248). IEEE.
- [12] Song, Y., Liu, C., & Li, Y. (2015). A new parameterized algorithm for predicting the secondary structure of rna sequences including pseudoknots. *The Computer Journal*, 58(11), 3114-3125.
- [13] Liu, Z., Zhu, D., & Dai, Q. (2015, December). Predicting Algorithm of RNA Folding Structure with Pseudoknots. In *2015 11th International Conference on Computational Intelligence and Security (CIS)* (pp. 34-37). IEEE.
- [14] Liu, Z., Kong, Q., Fu, Y., Ye, H., Zhao, S., Su, X., ... & Wang, Y. (2017, December). The Algorithm and Scheme of Prediction in RNA Folding Structure with Pseudoknots. In *2017 13th International Conference on Computational Intelligence and Security (CIS)* (pp. 469-474). IEEE.
- [15] Liu, Z., Liu, F., Kong, Q., Hao, F., & Zhao, H. (2018, November). Algorithm and Scheme in RNA Structure Prediction Including Pseudoknots. In *2018 14th International Conference on Computational Intelligence and Security (CIS)* (pp. 196-200). IEEE.
- [16] Stern, H.A., Mathews, D.H, "Accelerating calculations of RNA secondary structure partition functions using GPUs,"2013 Algorithms Mol Biol 2013, pp.8-29.
- [17] Zhang, H., Zhang, C., Li, Z., Li, C., Wei, X., Zhang, B., & Liu, Y. (2019). A new method of RNA secondary structure prediction based on convolutional neural network and dynamic programming. *Frontiers in genetics*, 10.
- [18] OpenMP API C/C++ Syntax Reference Guide, "OpenMP application program interface version 4.5,"2015.[Online].Available: <https://www.openmp.org/wp-content/uploads/OpenMP-4.5>
- [19] Microsoft Visual Studio Software for editing and code Evaluation, <https://visualstudio.microsoft.com/vs/whatsnew/>
- [20] Human gene for bone protein – Nucleotide - Retrieved on 20.12.2019 from <http://www.cbs.dtu.dk/services/NetAspGene/fasta.php>
- [21] Kong, Q., Liu, Z., Tang, X., Yang, Z., Fu, Y., & Li, H. (2018). The Computation of the Barrier Tree for BHG of RNA Folding Structure. *2018 14th International Conference on Computational Intelligence and Security (CIS)*, 6-9.

Segment Labeling Method for ML-based AFIB Detection

Dimitri Dojchinovski
Innovation Dooel

Skopje, North Macedonia
dimitri.dojchinovski@innovation.com.mk

Marjan Gusev
St. Cyril and Methodius University
Computer Science and Engineering
Skopje, North Macedonia
marjan.gushev@finki.ukim.mk

Abstract—Atrial Fibrillation is one of the most common and mortal types of heart rhythm problems - arrhythmias. Therefore, early and accurate detection is important in detecting heart diseases and prescribing appropriate treatment therapy. Developing a technology of this kind is of pivotal importance and a challenging problem for noninvasive tools for patient monitoring and analysis.

Electrocardiography provides comprehensive information that can be efficiently used in the management of the patients heart health. Detecting and classifying episodes of the different types of heart diseases is a subject of continuous research and immediately with new technological advances. Machine learning methods emerged as frequently used technology recently and become acknowledged for their relevance and results in this field.

Developing an effective model for detecting and classifying Atrial Fibrillation in ECG recordings requires the right data and adequate feature engineering. For this purpose we propose two methods, majority and pure segment labeling method used in the performed segmentation for feature engineering using the most popular ECG database and by integrating them in three machine learning algorithms, Support Vector Machines, Decision Trees and Random Fores.

The research concluded that the majority method trained on the Random Forest algorithm gives the highest results in the defined research space.

Index Terms—Atrial fibrillation, Machine learning, ECG.

I. INTRODUCTION

Detecting and classifying episodes of the different types of heart diseases is a subject of continuous research and immediately with new technological advances. Machine learning (ML) methods emerged as frequently used technology recently and become acknowledged for their relevance and results in this field.

Atrial Fibrillation (AFib) is one of the most common and mortal types of heart rhythm problems - arrhythmias. Therefore, early and accurate detection is important in detecting heart diseases and prescribing appropriate treatment therapy. Developing a technology of this kind is of pivotal importance and a challenging problem for noninvasive tools for patient monitoring and analysis.

The dangerous consequences from AFib, are described in countless research papers. Wolf et al. [1] concluded that AFib,

by itself, has a huge contribution in developing a heart attack. Other cardiovascular abnormalities diminish with age, but AFib prevail in older humans. Heeringa et al. [2] concluded that patients in age between 55 and 75 years, the risk of developing AFib is 23.8% for males and 22.2% for females.

The main goal in this research is choosing the right method for labeling the featured engineered segments from the LTAfDB ECG database used to develop a model for detecting and classifying AFib episodes in long term ECG recording based on binary classification by observing the anomalies in the RR intervals.

Three types of machine learning algorithms are implemented with in the search for the optimal model:

- Support Vector Machines - SVM [3],
- Decision Tree - DT [4],
- Random Forest - RF [5].

which are broadly used in the domain of detecting arrhythmias [6] [7] [8] [9].

The Long Term AF (LTAfDB) [10] ECG database from Physionet was used for training, validating and testing as one of the most used in many researches for decades in the field of cardiovascular diseases.

The analyzed ECG database features a unique set of ECG recordings upon which a set of feature engineering methods were applied:

- calculating RR intervals and labeling them,
- feature extraction by grouping subsequent RR intervals in segments with a certain length,
- labeling the generating segments.

When labeling the segments, two methods are proposed for assigning a rhythm annotation to the segment, majority and pure method. Due to the unique nature of each of the used ECG databases it is of great importance to pick the best method for the given problem.

The performance metrics of Sensitivity, Positive Predictive Value and F-score are used for evaluation of the algorithms alongside Improvement Factor (IF) and Duration Method [11].

Section II presents the basic background in the field of cardiovascular medicine required to understand the AFib diseases. Feature engineering process and segment labeling method are analyzed in Section III and the evaluation methodology in

Corresponding author: Dimitri Dojchinovski
email: dimitri.dojcinovski@innovation.com.mk

Section IV. Section V presents the obtained results discussed in Section VI. Final remarks and future work are presented in Section VII.

II. BACKGROUND

Electrocardiography (electrocardiogram – ECG) is a graphic method for measuring the electrical activity of the heart by tracing the electric current generated by the heart muscle during a heartbeat and it provides information of the current condition of the heart.

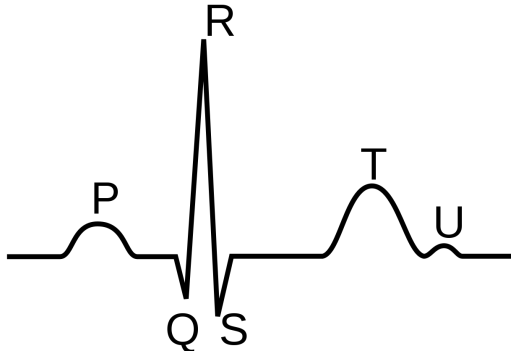


Fig. 1. QRS complex

The QRS complex is the most visible spike on the ECG line (Figure 1). Is made up of three main waves, which indicate the changing direction of the electrical impulse as it passes through the heart. The largest is the R wave (R peak), which represents the electrical impulse as it passes through the main portion of the ventricular walls. For detecting AFib there are two indices, the absence of non synchronized appearance of the P wave and irregularities in the heart rhythm. The P wave is very hard to detect, for that reason the more convenient method, irregularities in the heart rhythm, is being used. The time calculated by subtracting two consecutive R waves is labeled as RR interval. Irregularity in these intervals are considered as one of the most important indicators for AFib detection. Figure 2 pinpoints the difference between the intervals with irregularities in the heart rhythm and normal sinus rhythm.

Arrhythmia occurs as a variation from the normal heartbeat, usually as a result from irregularities within the specialized cardiac muscle cells that control the signals sent to the heart muscles (conduction system). There are several types of arrhythmias, but the most sever can appear in the form of AFib episode.

The main challenge in diagnosing heart disease using ECG is that the normal ECG may still differ for each person and sometimes one disease has dissimilar signs on different patient’s ECG signals. Also, two distinct diseases may have approximately identical effects on normal ECG signals. These challenges complicate the heart disease diagnose. So adequate feature engineering is important for pattern recognition.

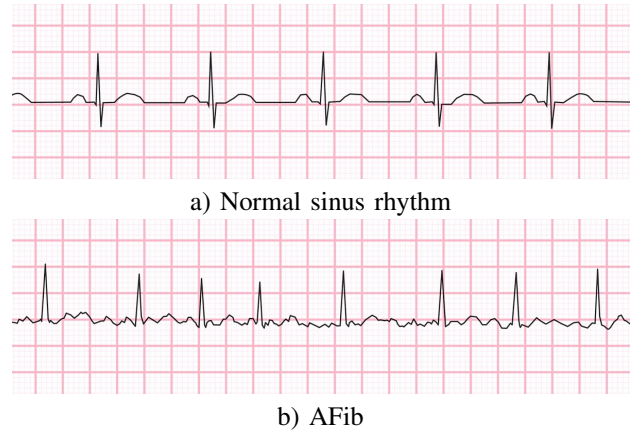


Fig. 2. Analyzed ECG rhythm episodes

III. FEATURE ENGINEERING

In this section we describe how and with what features we use the ECG databases in the ML algorithms. The procedure is realized according to the following activities:

- RR intervals are calculated for the annotation records,
- the calculated RR intervals are labeled according to the type of arrhythmia of the second R beat for each RR interval,
- Using the sliding window technique we generate segments of multiple consecutive RR intervals, and label whether the corresponding interval belongs to the AFib rhythm episode.

The recordings in the ECG databases are transformed into data containing calculated RR intervals of each adjacent R beat, annotated with a rhythm annotation of the second of the two adjacent R heartbeats.

The labeled RR intervals are used to generated segments of successive RR intervals with the sliding window technique where each segment represents a sample as part of a research set used with machine learning algorithms.

Each of the segments belongs to one of two classes, AFib (positive class, denoted by 1) or nAFib (negative class denoted by 0). The negative class includes samples of other arrhythmias that appear in ECG data bases, and the positive class consists of samples with Atrial Fibrillation arrhythmia. In the formed segments, the sliding window can encompass samples of both classes, with the segment class depending on the number of samples of both classes. Two methods are introduced for this problem:

- labeling the segments with majority vote (*majority method*),
- labeling pure segments (*pure method*).

For example, if we are creating a segment with sliding window size 5, the created segment has 5 features: f1, f2, f3, f4, f5, where each feature represents the length of the RR intervals of the 5 consecutive samples covered by the sliding window.

Since the sliding window in each iteration contains a certain number of samples, the class of the newly added segment is

EKF samples			method M					method P							
Sample #	Type	Rhythm	RR	f1	f2	f3	f4	f5	class	f1	f2	f3	f4	f5	class
83664	N	(N	#	#	#	#	#	#	#	#	#	#	#	#	#
83774	N	(N	110	110	#	#	#	#	#	110	#	#	#	#	#
83885	N	(N	111	111	110	#	#	#	#	111	110	#	#	#	#
83993	N	(N	108	108	111	110	#	#	#	108	111	110	#	#	#
84103	N	(N	110	110	108	111	110	#	#	110	108	111	110	#	#
84214	N	(N	111	111	110	108	111	110	0	111	110	108	111	110	0
84324	N	(N	110	110	111	110	108	111	0	110	111	110	108	111	0
84396	N	(AFIB	72	72	110	111	110	108	0	72	110	111	110	108	#
84484	N	(AFIB	88	88	72	110	111	110	0	88	72	110	111	110	#
84548	N	(AFIB	64	64	88	72	110	111	1	64	88	72	110	111	#
84643	N	(AFIB	95	95	64	88	72	110	1	95	64	88	72	110	#
84783	N	(AFIB	140	140	95	64	88	72	1	140	95	64	88	72	1
84863	N	(AFIB	80	80	140	95	64	88	1	80	140	95	64	88	1
84974	N	(N	111	111	80	140	95	64	1	111	80	140	95	64	#
85082	N	(N	108	108	111	80	140	95	1	108	111	80	140	95	#
85192	N	(N	110	110	108	111	80	140	0	110	108	111	80	140	#
85303	N	(N	111	111	110	108	111	80	0	111	110	108	111	80	#
85413	N	(N	110	110	111	110	108	111	0	110	111	110	108	111	0
85522	N	(N	109	109	110	111	110	108	0	109	110	111	110	108	0
85630	N	(N	108	108	109	110	111	110	0	108	109	110	111	110	0
85740	N	(N	110	110	108	109	110	111	0	110	108	109	110	111	0
85848	N	(N	108	108	110	108	109	110	0	108	110	108	109	110	0
85958	N	(N	110	110	108	110	108	109	0	110	108	110	108	109	0

Fig. 3. Forming of segment with 5 segment length

determined by the class of the dominant rhythm annotation that most of the RR intervals that make the segment belong to.

Method P (pure segments) annotates the segment with one of the classes if all the samples that make up the segment belong to one of the classes. With this method, the segments that contain samples of both classes are not added to the dataset for research. This kind of segments every time the rhythm annotations change, which means they take a very small part of the total number of segments.

Method M (majority segments) annotates the segment with one of the classes depending of the dominant rhythm annotation that most of the RR intervals that make the segment belong to. From the example with five segment length (Figure 3), the segments that contain three or more samples of the same class, it labels the whole segment with that class.

The occurrence of this segments is the reason we use odd numbers in the segment lengths, so that we can easily determine the class of the segment with the majority method (Figure 3).

An example of applying the majority method is presented in Figure 3. Six segments are assigned to the AFib class (labelled with 1), and 15 segments with the normal class (labeled with 0). Note that three segments can not be determined (labeled with #) due to missing data to apply the majority vote rule in the segment.

Figure 3 presents also the result of applying the pure method. Only two segments are assigned to the AFib class (labeled with 1) and 9 segments to the nAfib class (labeled with 0). Note 13 segments can not be determined since the segments are with mixed rhythm episodes.

After extracting the features, the data generated from each ECG data base is divided by a ratio of 80/10/10, where 80% of the data is for training, 10% for validation and 10% for testing. The training dataset is used exclusively to train the machine

learning algorithms. A validation set checks the performance of training algorithms to see if there is room for improvement and to serve as an indicator when optimizing algorithms. The test dataset shows the final performance of the algorithms after being optimized, such a dataset is hidden from the algorithms to the end to see how they respond to new unseen samples. During the training procedure we used 3-fold cross validation.

Since each of the data in the ECG databases represents individual patients, actually dividing them with such a ratio does not amount to merging all the patients and randomly dividing them. Since the ratios of classes in each data set are different, the data are grouped in such a way that the ratios of classes in the training, validation and testing dataset are almost the same as the ratios of classes in the respective ECG database.

IV. EVALUATION METHODS

Since the goal is to develop a binary classifier, the rhythm annotations are divided into a positive class AFib and a negative class nAFib where all other rhythm annotations are, including normal and other abnormal heart rhythm annotations.

LTAfDB includes 83 recordings from 83 individual patients each 24 hours long with a total of 8.903.169 annotated beats sampled at 128 Hz.

To evaluate the performances of the trained models, in the validation and testing phase the models process the data for validation or testing in the same input shape as the training data used so far, but only the segment class in the feature set is not exposed to the models, they have to predict it. Once the models have predicted the classes of the appropriate segments, the results are processed so that the data returns to its original state as ECG recordings labeled with the newly predicted classes. The predicted results are then compared with the original reference data set for validation and testing and the model performance are calculated. Namely, the segments are subdivided into the constituent RR intervals assigned to the corresponding class, her one RR interval may belong to the segments of the two classes, thus applying the majority method that assigns the class with the most RR interval in the segments.

Sensitivity (*SEN*), also known as hit rate or true positive rate (TPR), measures the proportion of correctly identified positive cases (sequences correctly classified as AFib) in regards to the actual number of positive cases.

Positive Predictivity Value (*PPV*), also known as precision, is the proportion of positive results that are true positives (sequences correctly classified as AFib) in regards to the total number of positive results

A statistical measure of a test's accuracy that combines *SEN* and *PPV* is known as F1 score. The F1 score, also called F score or F Measure, is calculated by the harmonic mean of the precision and sensitivity.

Due to the large research dimension and the large number of ML algorithms with different feature engineering and segment labeling during training, when presenting and comparing

the results, an Improvement Factor IF (*Improvement Factor*) metric has been introduced which is computed by comparing the F1-score values of the performance of the analyzed and reference algorithm.

V. RESULTS

Training was conducted on LTAfDB with RR intervals as features. The results are presented in Table 4 for each odd length of segments from 5 to 49 (23 segments) and for both majority and pure segment labeling methods.

Segment length	majority			pure		
	SVM	DT	RF	SVM	DT	RF
5	63.66%	84.94%	85.28%	69.04%	85.02%	85.32%
7	65.71%	86.11%	86.17%	49.65%	85.96%	86.22%
9	69.60%	86.50%	86.56%	69.57%	86.36%	86.40%
11	61.43%	86.84%	86.74%	65.59%	86.43%	86.72%
13	69.08%	87.04%	86.89%	60.36%	86.47%	86.90%
15	70.64%	87.22%	87.30%	64.13%	86.54%	87.14%
17	62.00%	87.19%	87.53%	60.06%	86.69%	87.54%
19	43.71%	87.08%	87.87%	68.27%	86.85%	87.76%
21	64.63%	87.37%	87.94%	65.53%	86.94%	88.05%
23	61.45%	87.11%	88.26%	64.79%	86.86%	87.78%
25	53.25%	87.47%	88.24%	64.81%	86.87%	88.16%
27	50.64%	87.69%	88.50%	56.87%	86.85%	88.11%
29	66.74%	87.72%	88.43%	57.72%	87.01%	88.05%
31	44.65%	87.71%	88.55%	65.49%	87.16%	88.37%
33	61.66%	87.82%	88.58%	65.68%	87.21%	88.44%
35	65.82%	87.82%	88.86%	65.56%	87.40%	88.57%
37	58.65%	87.95%	88.82%	65.40%	87.36%	88.75%
39	69.54%	88.00%	88.79%	65.55%	87.44%	88.58%
41	65.32%	88.42%	89.05%	65.37%	87.53%	88.68%
43	65.50%	88.49%	89.10%	65.40%	87.52%	88.99%
45	64.19%	88.30%	89.02%	65.32%	87.58%	88.86%
47	65.75%	88.31%	89.13%	65.25%	87.66%	88.82%
49	65.62%	88.40%	89.10%	65.27%	87.84%	88.90%

Fig. 4. F1 score of ML based AFib detection algorithms with majority and pure methods

VI. DISCUSSION

Table 5 shows the improvement factor values of all tested ML algorithms and feature engineering segment labeling methods for different segment length. The average values of these results lead to a conclusion that the improvement factor is positive in DT and RF algorithms which means that the majority method is better, while for SVM it is worse.

Since RF outperformed DT and SVM we conclude that the majority method for segment labeling in feature extraction process is better than the method of pure segments. Note that the majority method includes segments containing features of both the AFib and non-AFib classes, they are one type of transition from one rhythmic episode to another making their number in the whole dataset insignificant, but good to include so that algorithms would know how to handle such situations.

VII. CONCLUSION

In this paper, we experiment in determining the optimal method for labeling segments with a certain length of consecutive RR intervals for training the models used for detecting

segment length	ratio majority/pure validation		
	SVM	DT	RF
5	-7.78%	-0.09%	-0.05%
7	32.34%	0.17%	-0.05%
9	0.05%	0.16%	0.18%
11	-6.34%	0.47%	0.03%
13	14.44%	0.65%	-0.01%
15	10.15%	0.79%	0.18%
17	3.24%	0.58%	-0.01%
19	-35.97%	0.26%	0.13%
21	-1.38%	0.49%	-0.12%
23	-5.15%	0.28%	0.55%
25	-17.84%	0.69%	0.09%
27	-10.95%	0.96%	0.43%
29	15.62%	0.82%	0.43%
31	-31.83%	0.62%	0.21%
33	-6.12%	0.70%	0.15%
35	0.39%	0.48%	0.32%
37	-10.32%	0.67%	0.07%
39	6.08%	0.64%	0.24%
41	-0.07%	1.01%	0.41%
43	0.15%	1.11%	0.13%
45	-1.73%	0.83%	0.18%
47	0.77%	0.75%	0.35%
49	0.53%	0.64%	0.22%
Average	-2.25%	0.60%	0.18%

Fig. 5. Improvement factor of majority versus pure method

and AFib in ECG recordings. The majority method proves to be adequate for this problem.

Future work aims at developing an optimal model for ML-based AFib detection in ECG recording, and find the optimal ML algorithm, segment length, features sets, segments labeling method and adequate ECG database.

REFERENCES

- [1] P. A. Wolf, R. D. Abbott, and W. B. Kannel, "Atrial fibrillation as an independent risk factor for stroke: the framingham study.," *Stroke*, vol. 22, no. 8, pp. 983–988, 1991.
- [2] J. Heeringa, D. A. van der Kuip, A. Hofman, J. A. Kors, G. van Herpen, B. H. C. Stricker, T. Stijnen, G. Y. Lip, and J. C. Witteman, "Prevalence, incidence and lifetime risk of atrial fibrillation: the rotterdam study," *European heart journal*, vol. 27, no. 8, pp. 949–953, 2006.
- [3] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152, 1992.
- [4] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.
- [5] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1, pp. 278–282, IEEE, 1995.
- [6] S. H. Jambukia, V. K. Dabhi, and H. B. Prajapati, "Classification of ecg signals using machine learning techniques: A survey," in *2015 International Conference on Advances in Computer Engineering and Applications*, pp. 714–721, IEEE, 2015.
- [7] R. Colloca, A. E. Johnson, L. Mainardi, and G. D. Clifford, "A support vector machine approach for reliable detection of atrial fibrillation events," in *Computing in Cardiology 2013*, pp. 1047–1050, IEEE, 2013.
- [8] S. Datta, C. Puri, A. Mukherjee, R. Banerjee, A. D. Choudhury, R. Singh, A. Ukil, S. Bandyopadhyay, A. Pal, and S. Khandelwal, "Identifying normal, af and other abnormal ecg rhythms using a cascaded binary classifier," in *2017 Computing in Cardiology (CinC)*, pp. 1–4, IEEE, 2017.
- [9] J. Hu, W. Zhao, Y. Xu, D. Jia, C. Yan, H. Wang, and T. You, "A robust detection method of atrial fibrillation," in *2018 Computing in Cardiology Conference (CinC)*, vol. 45, pp. 1–4, IEEE, 2018.

- [10] S. Petrutiu, A. V. Sahakian, and S. Swiryn, "Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans," *Europace*, vol. 9, no. 7, pp. 466–470, 2007.
- [11] M. Gusev and M. Boshkovska, "Performance evaluation of atrial fibrillation detection," in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 342–347, IEEE, 2019.

Correlating the Cholesterol Levels to Glucose for Men and Women

Ilija Vishinov, Marjan Gusev
Ss. Cyril and Methodius University,
Faculty of Computer Sciences and Engineering,
Skopje, North Macedonia

Lidija Poposka, Marija Vavlukis
Ss. Cyril and Methodius University,
University Clinic of Cardiology
Skopje, North Macedonia

Abstract—Objectives: This paper explores the correlation between multiple cholesterol levels of the lipid profiles of patients and their diabetes regulation abilities in men and women.

Methodology: The methodology includes the following techniques: i) Pearson correlation ii) Spearman rank correlation and iii) setting thresholds for certainty of class assumption.

Data: The methods were applied on data from 161 patients of which 110 male and 41 female, analyzing the variables about patients' age, height, weight, BMI, lipid profile (total cholesterol, HDL cholesterol, LDL cholesterol, triglycerides), glycated hemoglobin levels with respective glucose regulation and diabetes classes, history of heart, diabetes and other chronic illnesses, habitual behaviors (smoking, alcohol consumption, physical activity), and medications intake (calcium channel blockers, BETA blockers, anti-arrhythmics, ACE/ARB inhibitors, diuretics, statins anti-aggregation medication and anticoagulants).

Conclusion: Analyzing the correlations between the lipid profile and glucose regulation in patients led to different results when the analysis was done separately on men and women. Thus, better predictions and insights can be made dependent on gender. The research found no strong stand-alone correlation when analyzing all data, but when the data was segmented in male and female records, a strong negative linear ($r=-0.52$, $p=0.001$) and non-linear ($r=-0.55$, $p=0.001$) correlation was found for the HDL-C and glucose levels in female patients. In men, statistically significant negative correlations with HbA1c were assessed for Chol ($r=-0.27$, $p=0.009$), LDL-C ($r=-0.33$, $p=0.002$) and HDL-C ($r=-0.23$, $p=0.026$).

Index Terms—cholesterol, diabetes, glucose regulation, correlation, lipid profile

I. INTRODUCTION

This research is part of the Glyco project [1] aiming to detect blood glucose level out of an electrocardiogram (ECG) measurement. In addition to the main research goal of this project, we have noticed possible correlations between the lipid profile and glucose regulation ability of the patients. Therefore, in this paper, we present the research and findings analyzing all possible correlations between measured biochemical parameters that describe a complex health condition of a patient, and especially analyzing the gender differences in the correlations.

The dataset within the realized project contains a total of 161 patients (110 male and 41 female) and among all medical records we focused on patients' age, height, weight, BMI, lipid profile (total cholesterol, HDL cholesterol, LDL cholesterol, triglycerides), history of heart, chronic and diabetes illnesses, habitual behaviors (smoking, alcohol consumption, physical

activity), medications intake (calcium channel blockers, beta blockers, anti-arrhythmics, ace/arb inhibitors, diuretics, statins, anti-aggregation medication and anticoagulants) and glycated hemoglobin levels with respective glucose regulation and diabetes classes.

Particularly, the research question in this paper focuses on determination if the measured cholesterol levels are correlated to blood glucose levels and whether these correlations are different for men and women.

II. RELATED WORK

Researchers typically have been investigating various correlations between patient's lipid profile and glucose regulation ability. We classify their findings according to the diabetes category (prediabetes, type1 and type 2 diabetes).

A. Prediabetes patients

Calanna et al. [2] concluded that prediabetes patients exhibited lower HDL and higher Triglycerides levels.

B. Type 1 diabetes

Prado et al. [3] found that HbA1c levels are positive correlated to Total Cholesterol, Triglycerides and LDL cholesterol, but found that there is no significant correlation between HDL cholesterol and HbA1c. Kim et al. [4] had a similar conclusion except for the LDL cholesterol, and concluded that for LDL cholesterol there is no a statistically significant correlation.

C. Type 2 diabetes

Triglycerides are positively and HDL cholesterol is negatively correlated with HbA1c in diabetic patients according to [5] and [2].

A significant positive correlation is found for Total Cholesterol, Triglycerides and LDL cholesterol when compared to HbA1c by several studies [6], [7], [8], [9], [10], although no significant correlation of HDL cholesterol is reported for worsened diabetes condition.

Similar conclusion for positive correlation of non-HDL (Total Cholesterol, Triglycerides and LDL) cholesterol is reported by [11], [12], [13], [14], [15], [16] and a significant negative correlation with HDL cholesterol.

TABLE I
GENDER STATISTICS OF THE PROPRIETARY DATASET

Gender	Quantity	Percentage (%)
Male	110	68.3
Female	41	31.7

TABLE II
NUMERICAL FEATURES STATISTICS OF THE PROPRIETARY DATASET
(INNOVATION DOOEL)

Feature	mean \pm std		
	All	Male	Female
Age	60.23 \pm 10.49	59.65 \pm 9.86	61.47 \pm 11.74
Weight	81.80 \pm 14.31	84.38 \pm 14.99	76.29 \pm 10.97
Height	171.53 \pm 8.44	175.02 \pm 7.67	164.33 \pm 4.49
BMI	28.16 \pm 4.43	28.04 \pm 4.52	28.42 \pm 4.28

III. METHODS

A. Experimental setup

Data of 161 patients was analyzed within this research. Table I shows the distribution of male and female patients. Table II and Table III show the distribution of the numerical and categorical variables which consist of the patients' age, height, weight, BMI, lipid profile, glycated hemoglobin levels with respective glucose regulation and diabetes classes, history of heart, diabetes and other chronic illnesses, habitual behaviors (smoking, alcohol consumption, physical activity), medications intake (calcium channel blockers, BETA blockers, antiarrhythmics, AKE/ARB inhibitors, diuretics, statins, antiaggregation medication and anticoagulants).

Table IV presents the blood glucose i.e. glycated hemoglobin HbA1c and the lipid profile which is made up of four continuous variables: Total Cholesterol (Chol), LDL Cholesterol (LDL-C), HDL Cholesterol (HDL-C) and Triglycerides (TG). These are the main variables that will be subject to analysis for possible correlations between the lipid profile and glucose regulation.

B. Correlation methods

The research is based on the following correlation methods compliant to the research goal:

- *Pearson correlation*

Pearson's correlation captures only linear correlations between two continuous variables.

- *Spearman rank correlation*

In order to capture non-linear correlations as well, Spearman rank correlation was included. Spearman's coefficient captures all relationships, linear and non-linear.

IV. RESULTS

As HbA1c is the marker for glucose regulation diagnosis, we can further indirectly reevaluate the correlation between the cholesterol attributes and the glucose regulation by assessing the Pearson and Spearman coefficients. The results of conducting the Pearson and Spearman rank correlation methods are presented in Table VI.

TABLE III
CATEGORICAL STATISTICS OF THE PROPRIETARY DATASET (INNOVATION DOOEL)

Feature with classes	Distribution(%)		
	All	Male	Female
Known high blood pressure in the past			
Controlled with medication	60.6	55.0	72.5
No high blood pressure in the past	35.0	40.4	23.5
Uncontrolled	4.4	4.6	3.9
Known diabetes in the past			
No diabetes in the past	62.3	65.7	54.9
Regulated with medication	21.4	21.3	21.6
Regulated with insulin	13.2	11.1	17.6
Regulated with diet	3.1	1.9	5.9
Diabetes mellitus in the family			
No	82.9	84.3	80.0
Yes	17.1	15.7	20.0
Known heart disease in the past			
No heart disease	80.6	79.8	82.4
CAD (Coronary artery disease)	19.4	20.2	17.6
Known other chronic disease			
No chronic disease	93.8	95.5	90.0
Hyper / Hypothyreosis	2.5	0.9	6.0
Chronic lung disease	1.9	0.9	4.0
Chronic kidney disease	1.9	2.7	0.1
Alcohol consumption			
No	94.3	91.7	100.0
Recommended quantity	5.7	8.3	0.0
Smoking			
Current smoker	45.9	56.0	24.0
Nonsmoker	45.9	34.9	70.0
Former smoker	8.2	9.2	6.0
Daily physical activity			
Occasionally (about 3 hours/week)	38.4	44.0	26.0
Regularly (more than 3 hours/week)	32.7	41.3	14.0
None (less than 3 hours/week)	28.9	14.7	60.0
BETA Blockers			
No	67.7	64.5	74.5
Yes	32.3	35.5	25.5
Calcium Channel blockers			
No	98.8	99.1	98.0
Yes	1.2	0.9	2.0
Antiarrhythmics class 1-3			
No	99.4	100.0	98.0
Yes	0.6	0.0	2.0
AKE inhibitors/ARB			
Yes	83.2	83.6	82.4
No	16.8	16.4	17.6
Diuretics			
No	80.1	75.5	90.2
Yes	19.9	24.5	9.8
Antiaggregation medications			
Yes	90.7	92.7	86.3
No	9.3	7.3	13.7
Anticoagulants medications			
No	98.1	97.3	100.0
Yes	1.9	2.7	0.0
Statins			
Yes	90.1	92.7	84.3
No	9.9	7.3	15.7

This solidifies our findings about the lack of strong correlation between diabetes and cholesterol without splitting the data set. After the division, one coefficient and its p-value stand out the most and that is the HDL-C and HbA1c correlation for the female patients.

Table VII contains thresholds for the cholesterol variables for specific scenarios where we can decide if the patient has or does not have diabetes and the corresponding certainty of

TABLE IV
DISTRIBUTION OF CONTINUOUS CHOLESTEROL AND GLUCOSE VARIABLES OF THE PROPRIETARY DATASET (INNOVATION DOEL)

Abbreviatoin	Feature	Unit of measurement	mean \pm std		
			All	Men	Women
Chol	Total Cholesterol	mmol/L	5.21 \pm 1.23	5.16 \pm 1.30	5.33 \pm 1.07
TG	Triglycerides	mmol/L	1.96 \pm 1.23	1.98 \pm 1.25	1.91 \pm 1.20
LDL-C	Low-Density Lipoprotein	mmol/L	3.06 \pm 1.11	3.03 \pm 1.16	3.12 \pm 1.02
HDL-C	High-Density Lipoprotein	mmol/L	1.20 \pm 0.35	1.15 \pm 0.34	1.33 \pm 0.36
HbA1c	Glycated Hemoglobin	%	6.85 \pm 1.59	6.82 \pm 1.74	6.93 \pm 1.18

TABLE V
DISTRIBUTION OF GLUCOSE REGULATION CLASSES OF THE PROPRIETARY DATASET (INNOVATION DOEL)

ID	class	HbA1c(%)	Distribution(%)		
			All	Male	Female
W	Well regulation	≤ 6.4	52.8	56.4	45.1
B	Bad regulation	> 6.4	47.2	43.6	54.9

TABLE VI
PEARSON AND SPEARMAN CORRELATION COEFFICIENTS BETWEEN THE LIPID PROFILE AND HbA1c

feature1	feature2	Pearson c.	p-value	Spearman c.	p-value
All					
Chol	HbA1c	-0.264	0.002	-0.225	0.009
LDL-C	HbA1c	-0.297	< 0.001	-0.271	0.002
HDL-C	HbA1c	-0.280	0.001	-0.225	0.009
TG	HbA1c	0.178	0.03	0.116	0.18
Men					
Chol	HbA1c	-0.27	0.009	-0.23	0.027
LDL-C	HbA1c	-0.33	0.002	-0.31	0.003
HDL-C	HbA1c	-0.23	0.024	-0.16	0.112
TG	HbA1c	0.19	0.07	0.08	0.458
Women					
Chol	HbA1c	-0.24	0.14	-0.21	0.206
LDL-C	HbA1c	-0.17	0.31	-0.07	0.704
HDL-C	HbA1c	-0.52	0.001	-0.55	0.001
TG	HbA1c	0.16	0.328	0.2	0.226

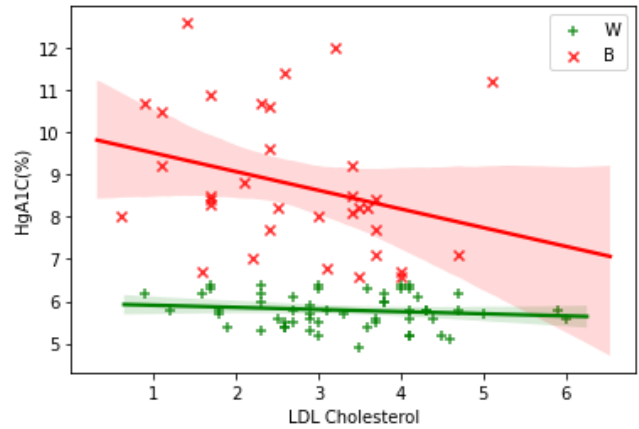


Fig. 1. Scatter Plot for LDL-C and HbA1c for male patients

TABLE VII
THRESHOLDS FOR SIGNIFICANT CORRELATIONS BETWEEN LIPID PROFILE AND GLUCOSE REGULATION IN DIFFERENT SCENARIOS

Feature	Class	Condition	Outcome	Certainty(%)
Men				
All	All	LDL-C ≥ 4.1	No Diabetes	89
High blood pressure	No	LDL-C ≥ 4	No Diabetes	99
BETA Blockers	Yes	LDL-C ≥ 3.8	No Diabetes	99
BETA Blockers	Yes	Chol ≥ 5.5	No Diabetes	85
High blood pressure	No	Chol ≥ 5.5	No Diabetes	90
Women				
All	All	HDL-C ≥ 1.6	No Diabetes	88
AKE /ARB inhibitors	Yes	HDL-C ≥ 1.5	No Diabetes	90
Diabetes in family	No	LDL-C ≥ 3.3	No Diabetes	82
Diabetes in family	No	Chol ≥ 5.5	No Diabetes	76

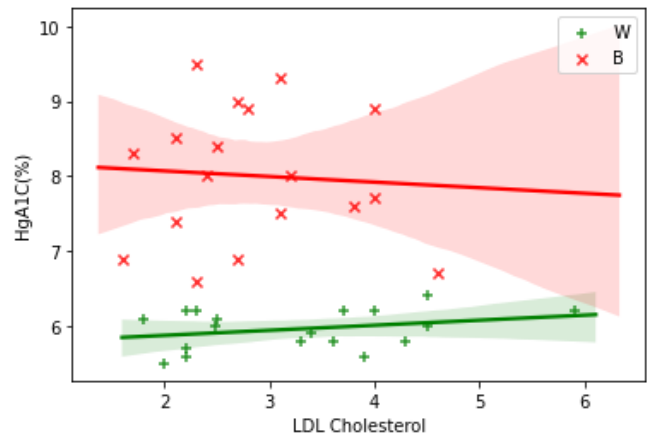


Fig. 2. Scatter Plot for LDL-C and HbA1c for female patients

the conclusion based on our data.

V. DISCUSSION

A. Pearson and Spearman rank coefficient

The Pearson and Spearman rank coefficients separated the HDL cholesterol in women as a strong stand-alone correlation with high statistical significance.

B. Thresholds for inferring

Since many factors could influence the glucose regulation classes alongside the lipid profile, we analyzed the correlations and possibilities for inference of positive or negative outcome for the glucose regulation classes (Well or Bad).

Table VII presents the outcomes of setting this threshold. The results (thresholds) were calculated from the percentage of the Well regulation and Bad regulation classes correspondingly and the scatter plots of HbA1c and the cholesterol variables were used to initially place the threshold and then move it around to find the best one.

Figure 1 and Figure 2 present the plots and express the differences in correlation between the male and female patients pointing the different gender behavior of lipid profile and glucose regulation.

C. Comparison to other research

Our findings overlapped in the correlation results regarding the Triglycerides which were positively correlated and HDL-C levels which were negatively correlated with glucose regulation and worse cases of diabetes.

There was a clash in the results when correlating Total Cholesterol and LDL-C levels, which we found to get lower with worse diabetes and glucose regulation classes, instead of elevated as the other papers revealed. This could be the result of some confounding factors in the data that we worked with.

VI. CONCLUSION

We have conducted a clinical research study on 161 patients in order to analyze if the glucose regulation is correlated to the lipid profile of the patient, knowing that the same autonomous nervous system is responsible for them.

Analyzing the correlations between the lipid profile and glucose regulation in patients led to different results when the analysis was done separately on men and women. Thus, better predictions and insights can be made dependent on gender.

The research results show that there are no strong stand-alone correlation when analyzing all data, but when the data was segmented in male and female records.

However, a strong negative linear ($r=-0.52$, $p=0.001$) and non-linear ($r=-0.55$, $p=0.001$) correlation was found for the HDL-C and glucose levels in female patients, while in men, statistically significant negative correlations with HbA1c were assessed for Chol ($r=-0.27$, $p=0.009$) and LDL-C ($r=-0.33$, $p=0.002$).

These findings motivate us to continue the research towards a deeper explanation about different correlations between lipid profile and glucose regulation level for men and women.

REFERENCES

- [1] Innovation Dooel, "Glyco project - measure ECG and glucose levels with a small, non-invasive, wearable monitor," 2019, project partially funded by Fund of Innovations and Technical Development, North Macedonia. [Online]. Available: <http://glyco.innovation.com.mk/>
- [2] S. Calanna, R. Scicali, A. Di Pino, F. Knop, S. Piro, A. Rabuazzo, and F. Purrello, "Lipid and liver abnormalities in haemoglobin a1c-defined prediabetes and type 2 diabetes," *Nutrition, Metabolism and Cardiovascular Diseases*, vol. 24, no. 6, pp. 670–676, 2014.
- [3] M. M. Prado, T. Carrizo, A. V. Abregú, and T. Meroño, "Non-hdl-cholesterol and c-reactive protein in children and adolescents with type 1 diabetes," *Journal of Pediatric Endocrinology and Metabolism*, vol. 30, no. 3, pp. 285–288, 2017.
- [4] S.-H. Kim, I.-A. Jung, Y. J. Jeon, W. K. Cho, K. S. Cho, S. H. Park, M. H. Jung, and B. K. Suh, "Serum lipid profiles and glycemic control in adolescents and young adults with type 1 diabetes mellitus," *Annals of pediatric endocrinology & metabolism*, vol. 19, no. 4, p. 191, 2014.
- [5] H. Drexel, S. Aczel, T. Marte, W. Benzer, P. Langer, W. Moll, and C. H. Saely, "Is atherosclerosis in diabetes and impaired fasting glucose driven by elevated ldl cholesterol or by decreased hdl cholesterol?" *Diabetes care*, vol. 28, no. 1, pp. 101–107, 2005.
- [6] S. REDDY, S. Meera, E. WILLIAM, and J. Kumar, "Correlation between glycemic control and lipid profile in type 2 diabetic patients: Hba1c as an indirect indicator of dyslipidemia," *Age*, vol. 53, pp. 10–50, 2014.
- [7] J.-B. Chang, N.-F. Chu, J.-T. Syu, A.-T. Hsieh, and Y.-R. Hung, "Advanced glycation end products (ages) in relation to atherosclerotic lipid profiles in middle-aged and elderly diabetic patients," *Lipids in health and disease*, vol. 10, no. 1, p. 228, 2011.
- [8] A. Hussain, I. Ali, M. Ijaz, and A. Rahim, "Correlation between hemoglobin a1c and serum lipid profile in afghani patients with type 2 diabetes: hemoglobin a1c prognosticates dyslipidemia," *Therapeutic advances in endocrinology and metabolism*, vol. 8, no. 4, pp. 51–57, 2017.
- [9] A. Ghari Arab, M. Zahedi, V. Kazemi Nejad, A. Sanagoo, and M. Azimi, "Correlation between hemoglobin a1c and serum lipid profile in type 2 diabetic patients referred to the diabetes clinic in gorgan, iran," *Journal of Clinical and Basic Research*, vol. 2, no. 1, pp. 26–31, 2018.
- [10] M. M. Eftekharian, J. Karimi, M. Safe, A. Sadeghian, S. Borzooei, and E. Siahpoushi, "Investigation of the correlation between some immune system and biochemical indicators in patients with type 2 diabetes," *Human antibodies*, vol. 24, no. 1-2, pp. 25–31, 2016.
- [11] Y.-C. Hwang, H.-Y. Ahn, S.-W. Park, and C.-Y. Park, "Apolipoprotein b and non-hdl cholesterol are more powerful predictors for incident type 2 diabetes than fasting glucose or glycated hemoglobin in subjects with normal glucose tolerance: a 3.3-year retrospective longitudinal study," *Acta diabetologica*, vol. 51, no. 6, pp. 941–946, 2014.
- [12] H. Khan, S. Sobki, and S. Khan, "Association between glycaemic control and serum lipids profile in type 2 diabetic patients: Hba 1c predicts dyslipidaemia," *Clinical and experimental medicine*, vol. 7, no. 1, pp. 24–29, 2007.
- [13] H. A. Khan, "Clinical significance of hba 1c as a marker of circulating lipids in male and female type 2 diabetic patients," *Acta diabetologica*, vol. 44, no. 4, pp. 193–200, 2007.
- [14] T. Grant, Y. Soriano, P. R. Marantz, I. Nelson, E. Williams, D. Ramirez, J. Burg, and C. Nordin, "Community-based screening for cardiovascular disease and diabetes using hba1c," *American journal of preventive medicine*, vol. 26, no. 4, pp. 271–275, 2004.
- [15] M. F. Lopes-Virella, H. J. Wohltmann, R. K. Mayfield, C. Loadholt, and J. A. Colwell, "Effect of metabolic control on lipid, lipoprotein, and apolipoprotein levels in 55 insulin-dependent diabetic patients: a longitudinal study," *Diabetes*, vol. 32, no. 1, pp. 20–25, 1983.
- [16] Y. Hu, W. Liu, R. Huang, and X. Zhang, "Postchallenge plasma glucose excursions, carotid intima-media thickness, and risk factors for atherosclerosis in chinese population with type 2 diabetes," *Atherosclerosis*, vol. 210, no. 1, pp. 302–306, 2010.

Using Educational Escape Room to Increase Students' Engagement in Learning Computer Science

Georgina Dimova
Center for innovations and digital
education DIG-ED
Skopje, North Macedonia
gina@dig-ed.org

Maja Videnovik
Center for innovations and digital
education DIG-ED
Skopje, North Macedonia
maja@dig-ed.org

Vladimir Trajkovik
"Ss. Cyril and Methodius" University -
Faculty of Computer Science and
Engineering
Skopje, North Macedonia
trvlado@finki.ukim.mk

Abstract— This paper presents the process of creating and implementing a large-scale experiment that involves students from more than 10 primary schools using the escape room style educational game in the classroom. The potential benefits and issues related to using technology in integrating this kind of games in education are explored and elaborated in the paper. An escape room is a game played by a team of people. They have to 'escape' from a room filled with challenges within a given time limit. In order to succeed, the players must solve the challenges using different hints and strategies. If these challenges incorporate education materials, students will have to master these materials, which will enhance their learning and increase their engagement.

Keywords— educational escape room, computer science education, game-based learning, gamification, collaborative learning

I. INTRODUCTION

Contemporary teaching is learner-centered. Teacher's role in guiding students in the learning environment is to enable them to progress within the learning process at their own pace. There is no "one approach for all" for learning-centered teaching. Teachers should enable each student to learn according to their previous knowledge, abilities, and skills. Students' learning should upgrade while they carry on to advantage from fostering, mentorship, and direction of their teachers.

Learning experiences should be structured to challenge students' thinking in the way students could construct new knowledge. According to Zaibon & Shiratuddin [1], learning is active process of acquiring and constructing knowledge through meaningful ways and interactions based on prior experience.

Student engagement has a critical role in student achievement within the student-centered learning process [2]. With governments interested in measuring student outcomes [3], and findings that student engagement can act as a proxy for quality [4], a clear understanding of what student engagement is, becomes essential. Students' engagement is a complex process that brings together diverse threads of research contributing to explanations of student success [5].

Different educational games can increase students' engagement in the learning process [6]. By using digital games, we can create new ways of learning in the classrooms. These interactive learning experiences can increase the engagement of students, and thus, achieving learning outcomes can be more quickly.

Digital games offer students opportunities to develop skills that are not focused just on learning facts. According to

Sung and Hwang [7], digital games enable the development of problem-solving, decision making, and strategic planning skills. Different challenges are facing the implementation of learning scenarios that use technological resources in and out of distance learning environments [8],[9].

Many games are used in an educational context, but most of them are not enjoyable for the students. On the other hand, it is challenging to match popular games to the curriculum in order to use them in the educational process [10]. Augmented reality is one of the technological tools that can be used in educational processes ranging from simple elementary school games to simulation tools [11]. In recent years, augmented reality is mostly associated with mobile platforms. Since its introduction, augmented reality managed to create a more active, productive, and meaningful learning process [12]. The merging of augmented reality with education allows students to be immersed in realistic experiences [13]. The mobile-based treasure hunt game using augmented reality can improve learning experiences by supporting active discovery and by balancing physical and digital interactions [14].

In addition to being a well-liked form of recreation, escape rooms have drawn the attention of educators due to their ability to foster teamwork, leadership, problem-solving, creative thinking, and communication in a way that is engaging for students. As a consequence, educational escape rooms are emerging as a new type of learning activity under the promise of enhancing students' learning through highly engaging experiences [15].

An escape room is a game played by a team of people where they have to 'escape' from a room filled with challenges within a given time limit [16]. In order to win ('escape'), the players must solve the challenges (puzzles) using hints, clues, and developing a strategy. If these challenges incorporate course materials within their puzzles, students will have to master these materials in order to succeed, which will enhance their learning and will increase their interest and engagement in learning.

Educators have now introduced the so-called escape-games into their teaching or training practices. During a limited time, a team of learners collaboratively solves puzzles related to educational content. For learners, the aim consists of "escaping" from a room. For educators, an escape-game contextualizes educational content into a meaningful and inspiring experience based on game-based and collaborative learning [17].

The objective of this paper is to present the process of creating and implementing large-scale experiment that involves students from 12 primary schools using simple treasure hunt [18] and escape room [19] style educational

game in the classroom. The potential benefits and issues related to integrating this kind of games in education using technology are explored and elaborated within the paper. The main aim of conducting the educational escape room was to provide an engaging activity beneficial for the students' learning.

Section 2 of this paper elaborates methodology and describes created educational escape room and materials. Section 3 discusses the obtained results and identified challenges. Section 4 concludes the paper.

II. METHODOLOGY

A. Designing an educational escape room

The educational escape room is an experience in which students had to solve a combination of puzzles in a limited amount of time in order to win in the game. The puzzles of the escape room were arranged as a treasure hunt - each puzzle unlocked the next one. Thus, students were required to solve the puzzles in a specific order. Arrangement of the puzzles in a sequence requires the whole team to engage in the puzzle at a specific time.

The designed educational escape room combined computer-based and physical puzzles, in order to create a highly engaging activity without compromising its educational value. Puzzles consisted of information from the course material, but also encouraged improving students' knowledge and skills through teamwork, communication, collaboration, and learning from each other. The escape room was hosted in the computer laboratory and designed to last a maximum of one hour.

The process of creation of an educational escape room was composed of several steps. First, the educational objectives

that should be incorporated in the game were defined, and the puzzles were created taking into account to have physical and computer-based puzzles. The number of puzzles was defined according to the previous students' experience in playing an escape room and considering its duration. The escape room consists of 6 puzzles, which should be solved. Afterward, puzzles, clues, and additional resources needed to conduct the escape room experience were created.

Puzzles were created very carefully, paying attention to their complexity. If a puzzle is too hard, the students will enter a state of frustration and give up; whereas, if a puzzle is too easy, the students will get bored and stop playing. Since it is essential to prevent students from getting stuck at one puzzle for too long (students can get bored, frustrated, or even angry), hints on demand when students get stuck were involved. First four hints on demand are free, and after that penalties are applied.

The overall theme of the educational escape room was finding and deactivating a device that threatens to turn all humans in robots. The activity starts by telling students a story that aliens are observing our planet, and they wanted to conquer it by turning all humans in robots. The story continues that in order to succeed in this, aliens have hidden devices in the classroom. To stop them, students must find the device and destroy it. To find the device, students have to solve several puzzles in sequence order, which will lead to the place where the device is hidden.

Table 1 summarizes all the treasures that were integrated into the educational escape room activity. Each treasure point consists of the puzzle, hint where to find the next puzzle which should be open with the solution of the previous one.

TABLE I. SUMMARY OF THE ESCAPE ROOM PUZZLES

Treasure	Puzzle	Solution	Hint
Piece of paper that includes binary representation of the letters	Students should translate given binary code into a word	Discovered word is the password to enter the next puzzle	Somewhere in the classroom, they have to find a QR code
QR code which should be scanned and opened with the password	A web site containing six questions concerning course material. Students should answer them	The first letter of each answer create the next puzzle's password	One of the computers is discreetly marked with red.
Marked computer which turns on with the password. Excel file should be opened which is in folder Escape room on Desktop	Students should discover numbers that are hidden behind corresponding cells in the Excel file	Numbers are the 4-digit number code for the next puzzle.	Classroom tables contain the next treasure – a box.
The box which can be unlocked by a 4-digit number code	The box contains VR glasses, and the puzzle is students to discover the device that does not belong in the group	A device (monitor) is a place where the next puzzle is put	The teacher is not so innocent in the game, is a hint for the next puzzle
Teacher's monitor	Students should find a hidden key and discover what that key can open	Locker	/
The locker that can be opened with a key and contains two pictures	The last puzzle is to find the difference between two pictures	Place where the hidden device is put	End of the game!

After an educational escape room was created, the puzzles were tested, and a simulation of the game to test it in a real setting was performed.

Additionally, instructions for teachers that will implement this educational escape room was created, consisting of the explanation of each puzzle, solution, and how the next puzzle should be discovered. Instructions for students are created, too. They consist of the story and how to play the game.

B. Method (experiment)

The educational escape room event was implemented during the initiative Computer Science Week, from 10th to 14th of February. In total, 12 teachers in 12 primary schools from North Macedonia have registered on the open call to implement this activity with the students. The distribution of the schools is given on Figure 1.

Most of the registered participants were Computer Science teachers. In total, 881 students have participated in the experiment. Students were from 6th till 9th grade. There was

no significant difference between the number of students in different grades. Pictures from the implementation of an escape room are given in Figure 2.

The main aim of this study was to evaluate the students' perceptions of the implemented educational escape room and to provide information whether it could increase students' engagement in learning Computer Science. Quantitative measures of students' achievement were not filed of interest.

Information about students' opinion about performed educational escape room were gathered from their feedback at the end of the activity. Additionally, an online survey was conducted with some of the students. The survey was adopted from a similar survey concerning the use of an educational escape room for teaching programming in a higher education setting (López-Pernas, et al., 2019). Participation in the survey was voluntary, and it was offered just to the students of two participating schools, which were around 20% of all participants.

Escape room

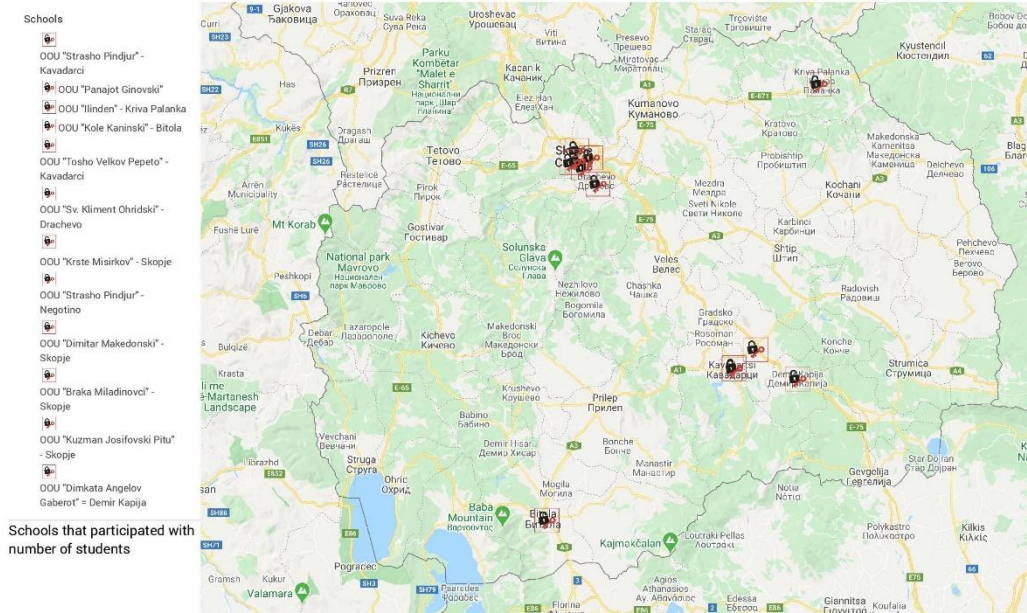


Fig. 1. Map of schools that have implemented educational escape room



Fig. 2. Educational escape room (permissions for taking students' photos are provided in advance)

The survey included some demographic questions, a list of statements with which students needed to agree or disagree using a 5-point Likert scale, and some opinion questions about further use of escape room. The questions aimed to assess the students' perceptions toward the use of the educational escape room as a learning activity, as a game, and the students' thoughts on the design of the escape room, as well as whether students preferred the escape room over a regular class.

III. RESULTS AND DISCUSSION

At the very beginning of the organization of the escape room in the educational context, massive interest among students was achieved. Students were very enthusiastic about participation in an educational escape room, and they show a huge interest in it. They were impatient when will the activity be performed and started making some arrangements among the team members on how to achieve the goal fastest.

Implementation of the educational escape room has led to a lot of excitement, fun, and satisfaction among the students. They were very competitive, trying to solve puzzles, to find the next treasure and to succeed. Students were enjoying during the activity and showed great interest in it.

After each finished escape room, a discussion with students was carried out. They were motivated to express their opinion about the activity, what they liked about it and what can be improved. Discussion about the puzzles and their level of difficulty was carried out. Students emphasized their strengths and weaknesses during the activity.

The most important fact that could be noticed during the activities in all participating schools was that students who had good teamwork and coordination among team members achieved the best results. Proper distribution of tasks among the students in the group had led to faster fulfillment of the task. Groups that on their initiative appointed a team leader who delegated tasks and managed the process were more successful. In that manner, the development of communication, collaboration, management, and leadership skills is one of the most important benefits of this kind of teaching strategy.

The time limit was beneficial in the implementation of the activity. Adding a timer created an urgency that drives student teams to engage with the content in a way that a traditional activity structure may not. They were running around the classroom, trying to organize themselves in order to finish the activity as soon as possible. Students had to work together to win or lose as a team. They took control of their learning and tried to manage the best that they can.

Developing critical thinking and problem-solving skills during solving puzzles and finding the next one, according to a hint, was another benefit from this kind of activity.

The combination of the computer-based and physical puzzles was a huge success because it made an activity more dynamic and excited for the students.

The initial guidance about the process was essential. Groups that did not understand the guidance well or did not listen carefully lose a significant amount of time in the beginning. For that reason, when organizing an educational escape room teacher should be sure that students understand the final goal of the escape room, the first puzzle that should be solved, what the solution of the puzzle means, and where

to find hints. They must understand that there is an option to ask for help if they get stuck during the activity.

Discussions with students lead to the conclusion that they like the activity very much, they were impatient to enter the escape room, to play it, and to reach the end of a game. Some of the students mentioned that they should be better organized next time and finish the game quicker. The interest and motivation for participation in the educational escape room were enormous. Students liked the puzzles; they were not too difficult to be solved, and they have stated that they learned new things during the activity. Overall, most students gave very positive comments that they thoroughly enjoyed the educational escape room experience and that they would like more hands-on activities like this one. They wished that other classes undertook similar initiatives and thanked the teacher for the experience.

It should be mentioned that students evaluated their work very well, successfully identifying their strengths and weaknesses during the game. Students were happy with the feedback received from their teacher about their activity during the game. They were even asking questions on how their activity could be improved, taking into account what they learn from some of their activities.

In addition, the evaluation survey was completed by a total of 95 students who volunteered to do so. Table 2 shows the results of the evaluation survey, including, for each question, the mean (M) and standard deviation (SD).

Of the 95 students of the sample, 35 were male (36.84%), and 60 were female (63.16%). It is interesting mentioning that there are no gender differences in answers to the survey's questions. These findings indicate that, although females seem to have reservations concerning games in general, the escape room attracted students of both genders equally. Educational escape rooms seem to be attractive and useful for both genders equally, which is a highly valuable insight into the research and educational communities.

Most students expressed a prior interest in games ($M = 4.82$, $SD = 0.51$), which was a good starting point in expecting students' interest in playing an escape room. The results of the survey show that students had a very positive overall opinion on the educational escape room ($M = 4.96$, $SD = 0.20$) and thought it was a fun experience ($M = 4.89$, $SD = 0.54$).

Most of the students were very competitive during participation in the escape room ($M = 4.74$, $SD = 0.61$), and they were impatient to open the next puzzle ($M = 4.81$, $SD = 0.53$). They have also noticed that teamwork and good task management are very important. They like working in teams ($M = 4.84$, $SD = 0.64$), too. The competition that was raised developed students' collaboration, interest, motivation, critical thinking, and problem-solving skills, increasing their engagement in the classroom activities at the same time.

A percentage of 98.95% of students stated they would recommend other students to participate in the escape room, and 100% claimed that they would like other classes to embrace similar activities. This excellent outcome obtained for student engagement confirms that educational escape rooms can be an excellent way to foster motivation and increase students' engagement in computer science.

TABLE II. RESULTS OF THE EDUCATIONAL ESCAPE ROOM SURVEY CONDUCTED AMONG STUDENTS

Question	Mean	SD
What is your general opinion on the Escape room? (1 Poor - 5 Very good)	4.96	0.20
Please state your level of agreement with the following statements (1 Strongly disagree, 5 Strongly agree)		
In general, I like to play games (video games, board games, etc.)	4.82	0.51
It was easy to reach the end of the Escape room	4.48	0.63
The Escape room was fun for me	4.89	0.54
The Escape room allowed me to improve my knowledge of the Computer Science	4.71	0.81
The Escape room was well organized	4.89	0.45
I was impatient to open the next puzzle in the Escape room	4.81	0.53
The Escape room encourage me for a competition	4.74	0.61
I like working in teams during Escape room	4.84	0.64
I like the Escape room more than a regular class	4.83	0.43
I can learn more with the Escape room than I would do during regular classes	4.56	0.81
	Yes	No
Would you recommend other students to participate in the Escape room	94 98.95%	1 1.05%
Would you like other classes to include activities like this?	95 100%	0 0.00%

Regarding learning effectiveness, students stated that the escape room helped them improve their knowledge of computer science ($M = 4.71$, $SD = 0.81$). These results were consistent with previous studies, which also found that educational escape rooms can improve students' knowledge on a specific topic. However, there is not much research concerning computer science in primary education.

Regarding the design of the escape room, students thought it was well organized ($M = 4.89$, $SD = 0.45$). Most of the students think that it was easy to solve the puzzles and to reach the end of the escape room ($M = 4.48$, $SD = 0.63$). When compared with the other classes of computer science, students declared that they prefer the escape room over them ($M = 4.83$, $SD = 0.43$).

Based on the previous discussion, it can be suggested that the appropriate use of educational escape rooms can have significant positive impacts on student engagement and learning in computer science. These findings provide evidence that educational escape rooms constitute a compelling way to increase student engagement.

Educational escape room has been very little used in the field of computer science, and there are just a few experiences reported in the literature, mainly in higher education. That is the reason why this first step in implementing an educational escape room is made just to see how students will accept it and whether it can improve their engagement in the classroom activities. Further work in this field could consider creating an educational escape room in order to guide their learning process or to evaluate students' achievements.

This experience can be applied to other subjects in different fields, such as mathematics, arts, or biology, too. For example, a biology teacher has already adapted created educational escape room to her classroom, adding puzzles connected to the biology curriculum. Students' feedback after the implementation of this escape room was very positive, enthusiastic, and motivated for new lessons conducted similarly.

Although the initial investment of time and effort on the teacher to design and create an educational escape room is, in principle, notably higher than that of other traditional classes, their significantly positive effect on student engagement as well as their ability to be reused in the following years makes it worthwhile.

IV. CONCLUSION

The designed educational escape room combined computer-based and physical puzzles, in order to create a highly engaging activity without compromising its educational value. Puzzles consisted of information from the course material, but also encouraged improving students' knowledge and skills through teamwork, communication, collaboration, and learning from each other. The escape room was hosted in the computer laboratory and designed to last a maximum of one hour. In total, 881 students from 12 different primary schools have participated in the experiment.

This paper is to present the process of creating and implementing a large-scale escape room in primary education. The main goal of conducting the escape room was to provide an engaging activity for the students.

The most important finding during the activities in all participating schools was that students who had good teamwork and coordination among team members achieved the best results. Proper distribution of tasks among the students in the group had led to faster fulfillment of the task.

This research throws up many questions in need of further investigation. For instance, a quantitative assessment of how educational escape rooms impact student academic performance on a specific topic is certainly needed in order further to understand the pedagogical utility of this novel teaching method.

REFERENCES

- [1] S. B. Zaibon and N. Shiratuddin, "Mobile game-based learning (mGBL) engineering model as a systematic development approach," *Proceedings of Global Learn*, pp. 1862-1871, 2010.
- [2] V. Trowler and P. Trowler, "Student engagement evidence summary," York, UK: Higher Education Academy, 2010.
- [3] N. Zepke and L. Leach, "Beyond hard outcomes: 'Soft' outcomes and engagement as student success," *Teaching in Higher Education*, 15, 661–673, 2010.
- [4] G. D. Kuh, "What student affairs professionals need to know about student engagement," *Journal of College Student Development*, 50, 683–706, 2009.
- [5] J. A. Fredricks, P. Blumenfeld, A. Paris, "School engagement: Potential of the concept, state of the evidence," *Review of Educational Research*, 74, 59–109, 2004.
- [6] V. Trajkovic, T. Malinovski, T. Vasileva-Stojanovska, M. Vasileva, "Traditional games in elementary school: Relationships of student's personality traits, motivation and experience with learning outcomes," *PLoS one*, 13(8), 2018.
- [7] H. Y. Sung and G. J. Hwang, "A collaborative game-based learning approach to improving students' learning performance in science courses," *Computers & education*, 63, 43-51, 2013.
- [8] G. Kimovski, V. Trajkovic, D. Davcev, "Resource manager for distance education systems," In *Proceedings IEEE International Conference on Advanced Learning Technologies*, pp. 387-390, August 2001.
- [9] G. Kimovski, V. Trajkovic, D. Davcev, "Negotiation-based multi-agent resource management in distance education," In *IASTED International Conference on Computers and Advanced Technology in Education including the IASTED International Symposium on Web-Based Education*, pp. 327-332, 2003.
- [10] M. Videnovik, L. Kionig, T. Vold, V. Trajkovic, "Testing framework for investigating learning outcome from quiz game: A study from Macedonia and Norway," In *17th International Conference on Information Technology Based Higher Education and Training (ITHET)*, pp. 1-5, 2018.
- [11] D. Bogatinov, P. Lameski, V. Trajkovic, K. M. Trendova, "Firearms training simulator based on low cost motion tracking sensor," *Multimedia tools and applications*, 76(1), 1403-1418, 2017.
- [12] S. Barma, S. Daniel, N. Bacon, M. A. Gingras, M. Fortin, "Observation and analysis of a classroom teaching and learning practice based on augmented reality and serious games on mobile platforms," *International Journal of Serious Games*, 69-88, 2015.
- [13] N. F. Saidin, N. D. A. Halim, N. Yahaya, "A review of research on augmented reality in education: advantages and applications," *International education studies*, 8(13), 1-8, 2015.
- [14] K. H. Ng, H. Huang, C. O'Malley, "Treasure codes: augmenting learning from physical museum exhibits through treasure hunting," *Personal and Ubiquitous Computing*, 22(4), 739-750, 2018
- [15] S. López-Pernas, A. Gordillo, E. Barra, J. Quemada, "Examining the use of an educational escape room for teaching programming in a higher education setting," *IEEE Access*, 7, 31723-31737, 2019.
- [16] M. Wiemker, E. Elumir, A. Clare, "Escape room games," *Game Based Learning*, 55, 2015.
- [17] E. Sanchez and M. Plumettaz-Sieber, "Teaching and learning with escape games from debriefing to institutionalization of knowledge," In *International Conference on Games and Learning Alliance* pp. 242-253. Springer, Cham, December 2018.
- [18] D. W. Kim and J. Yao, "A Treasure hunt model for inquiry-based learning in the development of a web-based learning support system," *J. UCS*, 16(14), 1853-1881, 2010.
- [19] S. Nicholson, "Creating engaging escape rooms for the classroom," *Childhood Education*, 94(1), 44-49, 2018.

Development of educational game for children with dyslexia

Aleksandra Sholdova
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia

Abstract—This paper will describe the development of an educational game designed for children from 4 to 10 years old to help them learn the letters, writing and reading if they are dyslexic. If people with dyslexia have a proper way of learning letters, writing and reading, then their condition can be corrected up to 80%. Also, if learning patterns for dyslexic children are applied to all children, it helps the children to develop their own logic and conclusions. In this paper will be explained how learning style is correlated with methods for correction of dyslexia, how educational games could be used as alternative learning methods and how several findings on dyslexia are used in the process of defining the functionalities and design of the educational game presented in this paper. If this game is added to the elementary school curriculum, this game could be an additional tool in the teaching process. This game will be helpful to teachers and parents to present the process of learning the alphabet, reading and writing to the children in an interesting way, where children will have the opportunity to learn and have fun at the same time. With the development of this educational game for children with dyslexia, I will contribute to increasing the availability of assistive technologies designed for Macedonian language.

Keywords—*educational game, dyslexia, assistive technology.*

I. INTRODUCTION

Dyslexia is a disorder that can affect reading (learning to read), writing (interpreting words, letters, and other symbols) and spelling of a person, but may also affect other areas such as working memory, sequencing, time management, orientation and much more [1, 2]. Worldwide, 5-10% of the population has dyslexia (or 1 in 10 people), but this number can reach as high as 17% depending on the language area [3]. Thus, in people from the English language area dyslexia is present at 15% and in the Slavic language area at 10% of the population [4]. People with dyslexia have difficulty to read fluently and they often read slowly and with errors. This can affect how they understand the text, but if someone else reads the text they have no problem with understanding the content [5]. When it comes to people with dyslexia, great attention should be paid to the learning style. In the beginning, it is good to find the learning style that suits them the most, so that later the teaching style can be adapted to their learning style. When children with dyslexia are given the right method of learning, they learn and progress constantly [6].

II. LEARNING STYLE

Learning style is a term that describes the factors that influence all aspects of learning. We all have different preferences for the way and style of learning. Multiple learning styles are defined, but the primary three types of learning styles are: visual learning, auditory learning, and kinesthetic learning [6, 7]. The multisensory learning style incorporates all of the above learning styles: visual, auditory and kinesthetic learning [8].

III. METHODS FOR CORRECTION OF DYSLEXIA

Multisensory learning style applies to all methods of correction of dyslexia.

A. Ron Davis Method

So according to Ron Davis's theory, people with dyslexia often think and understand with the help of images, that is, they think multidimensional and use all senses, which gives them the opportunity to see the world from a different perspective. The pictorial way of thinking provides them much more information than the verbal way. The pictorial way of thinking is 400 to 2000 times faster than the verbal way in which picture and speech take place at the same speed [9].

B. Orton-Gillingham method

The Orton-Gillingham method is also based on multisensory learning. Dr. Orton was the first doctor to make a miraculous and unexpected finding back in 1920. He found that children who were considered retarded because of reading difficulties often had average or above-average intelligence. He further predicted that by applying kinesthetic and tactile techniques, supplemented by visual and auditory techniques in learning letters and voices, the state of dyslexia could be corrected. Using these findings, talented teacher and pedagogue Anna Gillingham created a multisensory reading program, so in 1936 the first Orton-Gillingham Dyslexia Correction Program was introduced which included systematic and explicit study of sounds (phonemes), syllables, root of the word (morphemes) and spelling. This program is intended for English language, but also is in preparation of adaptation for Macedonian language [10].

C. AFS method

Also, the AFS method incorporates elements of multisensory learning, this method includes three basic help priorities: Attention, Function, and Symptoms. For people with dyslexia is more difficulty to have deliberate attention to letters and numbers, so most often dyslexia is considered with medical diagnoses such as attention deficit disorder, concentration deficiency and hyperactivity. Because people with dyslexia perceive differently and have a very fast thinking process, they need more time to recognize and learn symbols and letters. The AFS method determines which sensory perceptions function in a different way, because the dysfunction is not always synchronous at all sensory perceptions function. To establish a process that provides a gradual increase in attention and sensory perceptions, it is necessary to work with all the senses [11]. From this we can conclude that everywhere through dyslexia correction methods a multisensory learning style is found, it can improve memory and attention and increase the performance of cognitive and sensory abilities.

IV. USE OF EDUCATIONAL GAMES AS ALTERNATIVE LEARNING METHODS

Educational games are also increasingly used as alternative learning methods. Fun-guided learners are more easily motivated to continue the learning process by fulfilling meaningful activities and / or tasks that are defined in the context of the game. Before several years, the project “Grandma’s games” was conducted in five primary schools located in different areas in Macedonia, where in classroom environment were integrated traditional games. The results in this research showed positive change for the learning outcomes, increased collaboration, teamwork and increased level of interest and interactivity among children [12]. Also, in other research was shown that educational computer games with multi-sensor interfaces, could improve the quality of the learning experience and facilitate the acquisition of knowledge and understanding of the content through seamless integration between virtual and physical environments. [13].

V. FINDING USED IN THE PROCESS OF DEFINING THE FUNCTIONALITIES AND DESIGN OF THE EDUCATIONAL GAME

Therefore, in this educational game for children with dyslexia, several findings on dyslexia are considered in the process of defining the functionalities and design of this educational game [14].

- As all dyslexia correction methods are all adapted to a multi-sensory learning style, in the educational game for children with dyslexia is payed close attention to the game to have a multi-sensory playing environment.
- The choice of background and text colors plays a big role in reading ability when it comes to people with dyslexia. A specific combination of text and background colors should be selected to facilitate the reading process. So, for background colors it is the best to use pastel colors and the text to have less brightness and contrast than the background.
- The DyslexicFZF font, which is an adaptation of the DISLEXIE font, is created to allow writing in Macedonian and this font represents a type of assistive technology for people with dyslexia [14].

VI. THE EDUCATIONAL GAME DESCRIPTION

Because this educational game is designed for children from 4 to 10 years old, guided by their wishes and interests, we all know that children love cartoons. Therefore, in this educational game will be used illustrations and animations that will be presented as in a children's cartoon. To achieve authenticity, the design of the illustrations is specifically tailored to fit a story. Storytelling is an important element in creating a game where players are constantly involved and active. The educational game for children with dyslexia represents the story of the parrot Ricky. This story is about animals, because children love animals, and some keep them as pets and care for them.

So, in this educational game we need to take care of the parrot that is very hungry. The moment when the player approaches the game is late autumn. The parrot Rickey is worried that he will not have enough food to survive the winter. At the last minute, he asks his friends for help if they can gather food together to suffice for the winter until spring comes. Ricky is not just a usual parrot, he is a talking parrot, so he has a very specific taste for food, his favorite foods are

being the letters that allow him to speak. In addition to talking, he wants to learn the process of read, and his friends can help him by collecting food from all possible letters in the alphabet. So, by playing the game, the players collect food for the parrot Ricky, and they unconsciously learn the letters.

VII. TECHNOLOGIES USED IN THE GAME DEVELOPMENT PROCESS

The technologies used in the development of this educational game for children with dyslexia are:

- Scratch - a block-based visual programming language where the entire educational game is programmed and
- Adobe Illustrator - vector graphics software that was used to create digital illustrations for the educational game.

According to the elementary school curriculum, in the subject of Informatics, the basics of programming are taught through the Scratch - working environment [15]. If students are already familiar with the game for children with dyslexia, they will be able to easily recognize the game and associate it with the Scratch working environment. Projects created in Scratch that are published are open source projects and anyone can view, modify and download them. If children in fifth or sixth grade show an interest in learning computer science and programming, with the help of teachers, they can modify the game and make changes as they wish. With that in mind, this educational game could have a dual application. Students in the early grades can learn the alphabet by playing the educational game, and students in the older grades can learn the basics of programming by modifying the same educational game.

VIII. GAME ANALYSIS AND DESIGN

In this section will be shown the game analysis and design. Thus, figure 1 shows a use case diagram where you can see the actions that trigger an event. When a player accesses the game's website, two options for starting and leaving the game are displayed. If the player chooses to exit the game all activities are terminated and the game is closed. And if the player chooses to start the game, in the background this action accesses the database and retrieves information from the database to create backgrounds, to create the character through which the player can control the game, create the necessary levels and create objects. The player can choose one of the offered game levels. When a player's character is in the activated level, there are two options for closing the level: when the player chooses the option to exit the level, or when he has accumulated a total of 20 points to open the next level.

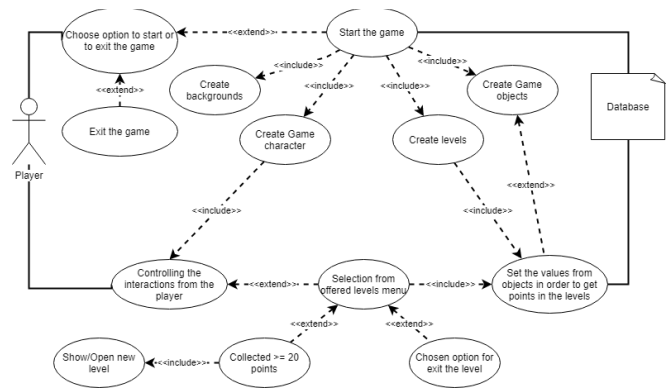


Fig. 1. Use Case diagram - User access to the game [14]

Figure 2 shows a class diagram, showing how the classes are linked, what data is stored in them, and what functions are used.

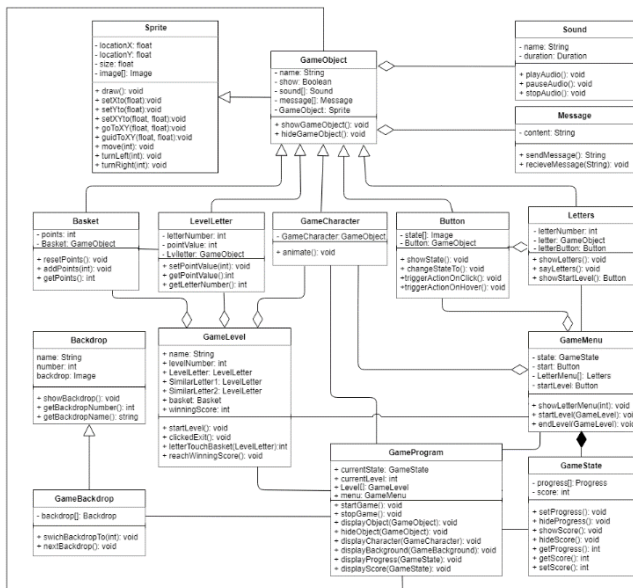


Fig. 2. Class diagram for the online-game the parrot Ricky [14]

Figure 3 shows the sequential diagram. From this diagram we can see how certain interactions the player undertakes, affect the activation of certain functions of objects, and how the activated functions facilitate the data exchange across objects in the sequential order in which these interactions take place.

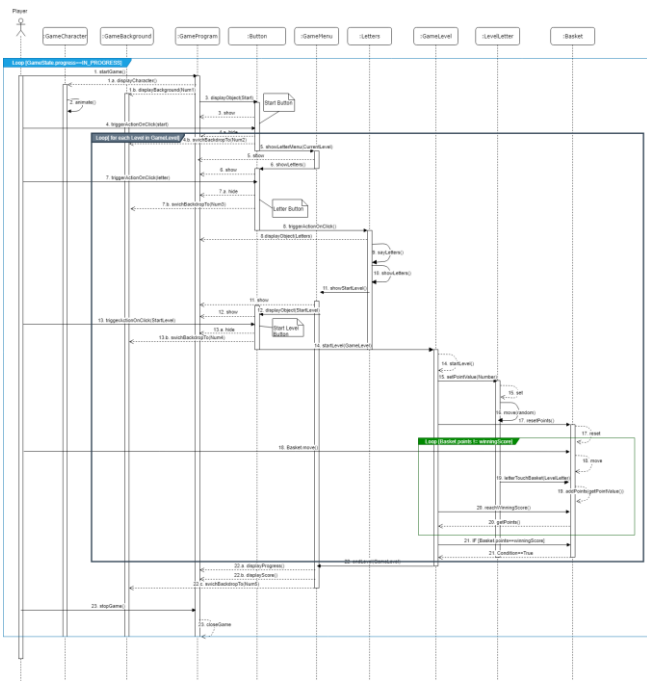


Fig. 3. Sequence diagram for the online-game the parrot Ricky [14]

IX. INTRODUCING THE GAME

When a player accesses the web site with the game, a screen shown in Figure 4 is displayed. To zoom in and out of the full screen, the player needs to click on the four arrow buttons pointing to the edges. Then by clicking on the green flag the game starts, it can also be stopped at any time by clicking on the red octagon. When the player has activated the

game, the parrot Ricky begins to guide the players through the game.

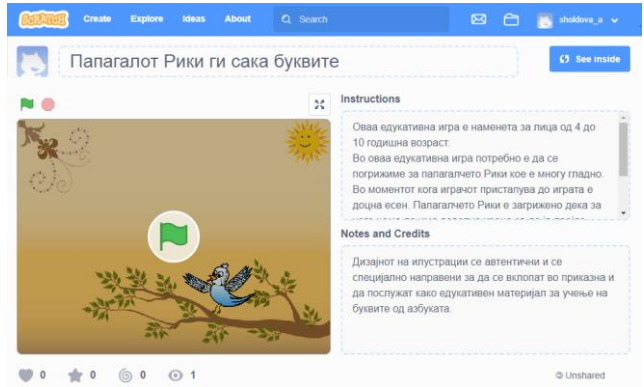


Fig. 4. Player accesses the web site with the game - the parrot Ricky [14]

Thus, the player first needs to press the "Start" button (Figure 5 on the left) in order to be displayed the first letter in the letter list from the alphabet (Figure 5 on the right).

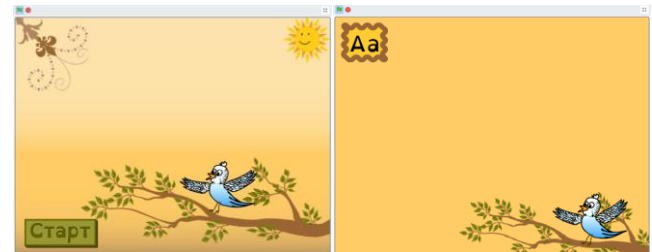


Fig. 5. Starting the game (left) and showing the first letter (right) [14]

When the player clicks on the letter button a new display opens, shown in Figure 6, where the parrot teaches the player the selected letter and needs to remember it well because it will be needed in the game level, then the parrot provides guidance on how to make it easier.

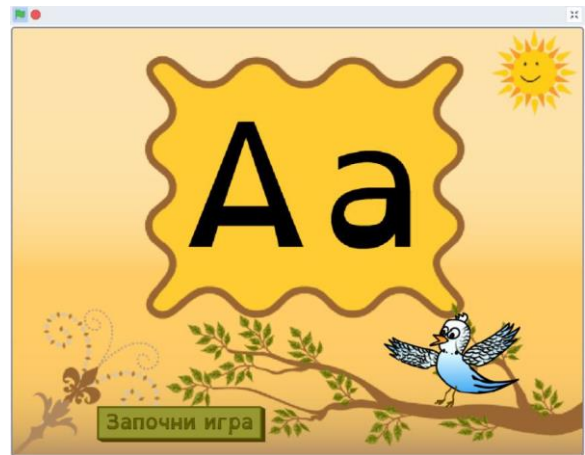


Fig. 6. Information about selected letter [14]

When the player clicks on the "Begin game" button a level for the selected letter is activated, shown in Figure 7. The player using the keyboard keys "Arrow Left" and "Arrow Right" moves the sack with food for birds where the player has to collect symbols from the selected letter. For every correct letter, the parrot says, "You're collecting a lot of letters, keep going." And for every incorrect letter, it says, "Play carefully, I'll be hungry." To complete a level, it is necessary to collect 20 symbols from the selected letter.

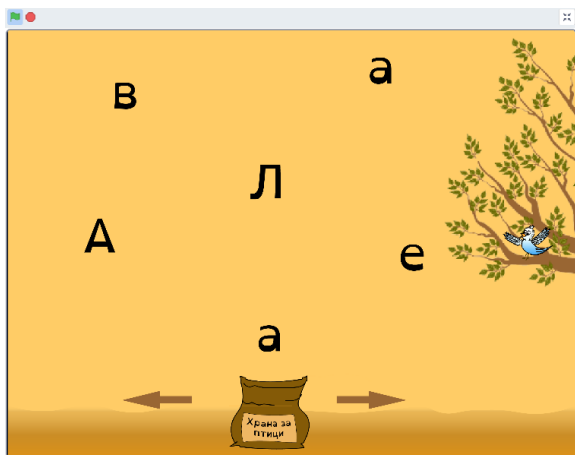


Fig. 7. Level in the game for the selected letter [14]

Upon completing the level, a new scene is shown, shown in Figure 8, where the parrot says "Bravo, you got the star of the letter A" and thus the player is rewarded for his hard work and successfully completed level.

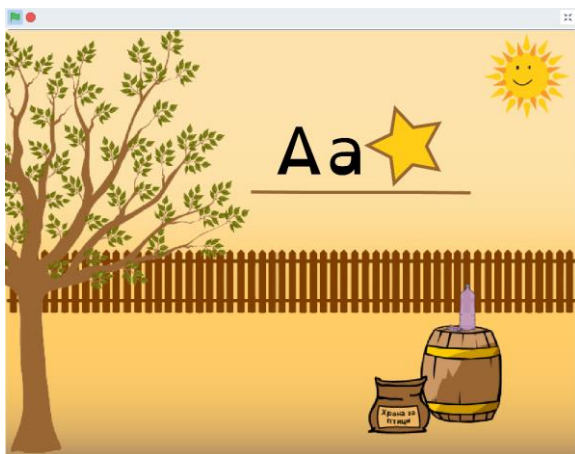


Fig. 8. Displaying won star for first letter [14]

After completing one level, the scripts are repeated for each next level, where the letter list is re-displayed and one letter from the alphabet is added to each subsequent level. When all letters are successfully passed the player can return to play all the letters if desired. If the player wants to stop the game, he can do so by clicking on the red octagon or turning off the browser by clicking the "X" button [14].

X. CONCLUSION

Playing the game presented in this paper is expected to make it easier for people with dyslexia to learn the letters and later the reading and writing processes. Also, through the process of playing the game, all students from elementary school, in the lower grades will be able to learn the alphabet very easily, while the students in the upper grades, in the field of informatics can rewrite the game because it is open source. Students can make their own versions of the game, and this may even apply to teachers, where they will be able to adapt it to their teaching styles and curriculum needs. Following the development of the game, it is planned to conduct a research and performance analysis in collaboration with the "Einstein" Dyslexia Association. If it is proven to be effective, it may find application in primary schools in our country, where all

students would benefit, as mentioned before, the way that dyslexic students learn is much simpler for all students.

From this we can conclude that the proposed solution in this paper could contribute to increasing the number of assistive technologies available in Macedonian language for people with dyslexia and that this solution is of great value to society and can have further multipurpose use.

ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my dear professor and mentor Vladimir Trajkovik Ph.D. for sharing his expertise and knowledge with me and his full support in the preparation and implementation of this project.

REFERENCES

- [1] "Definition of Dyslexia" LEXICO powered by Oxford. <https://www.lexico.com/en/definition/dyslexia> (accessed Nov. 12, 2019).
- [2] "How Can Assistive Technology Support Dyslexia?" [dyslexic.com. https://www.dyslexic.com/blog/how-can-assistive-technology-support-dyslexia/](https://www.dyslexic.com/blog/how-can-assistive-technology-support-dyslexia/) (accessed Nov. 12, 2019).
- [3] "Dyslexia Facts and Statistics" Austin Learning Solutions. <https://www.austinlearningsolutions.com/blog/38-dyslexia-facts-and-statistics.html> (accessed Nov. 12, 2019).
- [4] "Teachers need more education to work with students with dyslexia, (in Macedonian)" [disleksija.org.mk. http://disleksija.org.mk/educacija/](http://disleksija.org.mk/educacija/) (accessed Nov. 12, 2019).
- [5] "Dyslexia," [mayoclinic.org. https://www.mayoclinic.org/diseases-conditions/dyslexia/symptoms-causes/syc-20353552](https://www.mayoclinic.org/diseases-conditions/dyslexia/symptoms-causes/syc-20353552) (accessed Nov. 23, 2019).
- [6] "How Dyslexics Learn: Teaching to the Dyslexic Strengths," [homeschoolingwithdyslexia.com. https://homeschoolingwithdyslexia.com/dyslexics-learn-teaching-dyslexic-strengths/](https://homeschoolingwithdyslexia.com/dyslexics-learn-teaching-dyslexic-strengths/) (accessed Nov. 23, 2019).
- [7] "Learning Styles: Understanding How You Learn," [BeatingDyslexia.com. https://www.beatingdyslexia.com/learning-styles.html](https://www.beatingdyslexia.com/learning-styles.html) (accessed Nov. 23, 2019).
- [8] "How to Design a Multisensory Lesson," [child1st.com. https://child1st.com/blogs/resources/113530247-how-to-design-a-multisensory-lesson](https://child1st.com/blogs/resources/113530247-how-to-design-a-multisensory-lesson) (accessed Nov. 23, 2019).
- [9] "Ron Davis Method, (in Macedonian)" [www.dyslexia-info.com. https://www.dyslexia-info.com/4264-2/](https://www.dyslexia-info.com/4264-2/) (accessed Nov. 26, 2019).
- [10] "Orton-Gillingham method, (in Macedonian)" [www.dyslexia-info.com. https://www.dyslexia-info.com/orton-metod/](https://www.dyslexia-info.com/orton-metod/) (accessed Nov. 26, 2019).
- [11] "AFS Method, (in Macedonian)" [www.dyslexia-info.com. https://www.dyslexia-info.com/%d0%b0%d1%84%d1%81-%d0%bc%d0%b5%d1%82%d0%be%d0%b4/](https://www.dyslexia-info.com/%d0%b0%d1%84%d1%81-%d0%bc%d0%b5%d1%82%d0%be%d0%b4/) (accessed Nov. 26, 2019).
- [12] Trajkovik V, Malinovski T, VasilevaStojanovska T, Vasileva M (2018) Traditional games in elementary school: Relationships of student's personality traits, motivation and experience with learning outcomes. *PLoS ONE* 13 (8): e0202172. <https://doi.org/10.1371/journal.pone.0202172>
- [13] Covaci, A., Ghinea, G., Lin, C. et al. Multisensory games-based learning - lessons learnt from olfactory enhancement of a digital board game. *Multimed Tools Appl* 77, 21245–21263 (2018). <https://doi.org/10.1007/s11042-017-5459-2>
- [14] A. Sholdova, "Development of educational game for children with dyslexia, (in Macedonian)" B.S. thesis Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje. 2020. [Online]. Available: http://https://www.researchgate.net/publication/339941543_Graduation_the_sis_-_Development_of_educational_game_for_children_with_dyslexia.
- [15] "Curriculum - Sixth Grade Informatics, (in Macedonian)" [www.bro.gov.mk https://www.bro.gov.mk/wp-content/uploads/2019/08/Nastavna_programa-Informatika-VI_odd.pdf](https://www.bro.gov.mk/wp-content/uploads/2019/08/Nastavna_programa-Informatika-VI_odd.pdf) (accessed Dec. 23, 2019).

The effects of flexible work in the IT industry

Mile Davitkovski
Crazy Labs, Tel Aviv Israel
Research & Development Center
Skopje, Macedonia
davitkovski.mile@gmail.com

Smilka Janeska Sarkanjac
Ss Cyril and Methodius University
Faculty of Computer Science and
Engineering
Skopje, Macedonia
smilka.janeska@finki.ukim.mk

Branislav Sarkanjac
Ss Cyril and Methodius University
Faculty of Philosophy
Skopje, Macedonia
sarkanjac@hotmail.com

Abstract—Flexible work can come in a variety of types, such as fully flexible strategy where employees are fully flexible in carrying out their tasks from any location in the world to strategies that allow workers to work flexibly over several days, weeks or months at a time. There are many types of flexible work strategies for companies, and the most common are hiring digital nomads, freelancers, using co-working space crowds and so on. This paper tries to compare and analyze the findings of Stack Overflow research from 2017 with the survey conducted as a part of this research, in Macedonia in 2019 regarding flexible work arrangements in the IT industry. The comparative analysis will represent the workers' perspective. In the second part of the paper we will consider several obstacles and traps of flexible work arrangements from the companies' perspectives, and offer principles to overcome them.

Keywords—Flexible-working, IT-industry, Productivity, Satisfaction, Profitability

I. INTRODUCTION

Flexible work can come in a variety of types, namely a fully flexible strategy where employees are fully flexible in carrying out their tasks from any location in the world to strategies that allow workers to work flexibly over several days at a time, week or month. There are many types of flexible work strategies for companies, and the most common are hiring digital nomads, freelancers, using co-working space crowds and so on.

As Hill and his associates state, "In the organizational perspective, the goal of flexibility is to enable the organization as a whole to adapt to rapidly changing demands placed on the organization from either internal or external forces. By contrast, the goal of workplace flexibility from the worker perspective is to enhance the ability of individuals to meet all of their personal, family, occupational, and community needs. It is assumed, however, that as a byproduct the organization will indirectly benefit with increased efficiency, effectiveness, and greater productivity." [1]

Flexible work is not only suitable for employers and employees, research has shown that it proved more effective. According to a Gallup report [2], employees in various industries that worked 60 to 80 percent of their time flexibly had the highest productivity rates. As companies and employees become increasingly aware that work is "fluid" and can take place anywhere and at any time, they realize that simply having a dedicated employee just to fill a position for a future activity, no longer justifies cost, nor is it effective in achieving goals.

Previous studies note two ways of dealing with the effect of ubiquitous information technologies, blurring the boundaries between personal life and work activities: keeping work and personal life domains separated or integrated, that is, segmenting or blending of domains [3].

The growing trend of flexible working in technology companies shows no signs of slowing down. Instead, it gains momentum. Given the benefits that businesses have in the long run, this should come as no surprise given the technological advances of innovative startups that have made hiring flexible employees, especially IT professionals, more effective and successful than ever. In addition, flexible work is no longer just a benefit to attracting the best candidates, but it is certainly turning into a business strategy for companies actively preparing for the Fourth Industrial Revolution.

This paper tries to compare and analyze the findings of Stack Overflow research from 2017 [4] with the survey conducted as a part of this research, in Macedonia in 2019 regarding flexible work arrangements in the IT industry. The comparative analysis will represent the workers' perspective [1]. In the second part of the paper we will consider several obstacles and traps of flexible work arrangements from the companies' perspectives, and offer principles to overcome them [5].

II. STACK OVERFLOW SURVEY

Founded in 2008, Stack Overflow is the largest online community of IT professionals that enjoys great trust among them. IT professionals use this platform to learn, share their knowledge and build their careers. More than 50 million professionals in the IT industry visit Stack Overflow every month to help solve coding problems, develop new skills, and find new challenging job opportunities.

Since 2011, each year Stack Overflow has been conducting different types of surveys of IT professionals approaching the community to find out about their favorite technologies, coding habits, work preferences, the way they learn, share their knowledge and advance in their career.

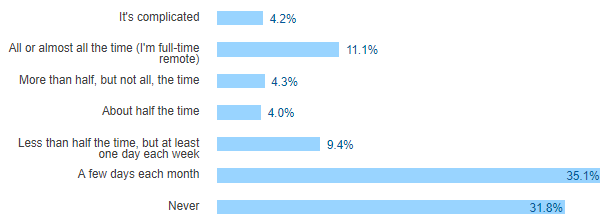
In 2017, Stack Overflow received the largest group of respondents in their history, numbered about 64,000 IT professionals. The research was done in January 2017.

Important information from Stack Overflow's research on "flexible work" with their interviewees is selected and highlighted below.

A. How often do developers work remotely?

64% of the respondents of Stack Overflow's survey answered that they work flexibly for at least 1 day a month, while 11.1% answered that their work is full flexible working

How Often Do Developers Work Remotely?



44,008 responses

time.

Fig. 1. How often do developers work remotely? (<https://insights.stackoverflow.com/survey/2017#remote-work>)

The next question of particular interest for this research is what developers value in compensation benefits, and the answers are presented in Fig. 2. 53.3% of the respondents said that flexible working hours are one of the 5 most important priorities in accepting a new job.

What Developers Value in Compensation/Benefits

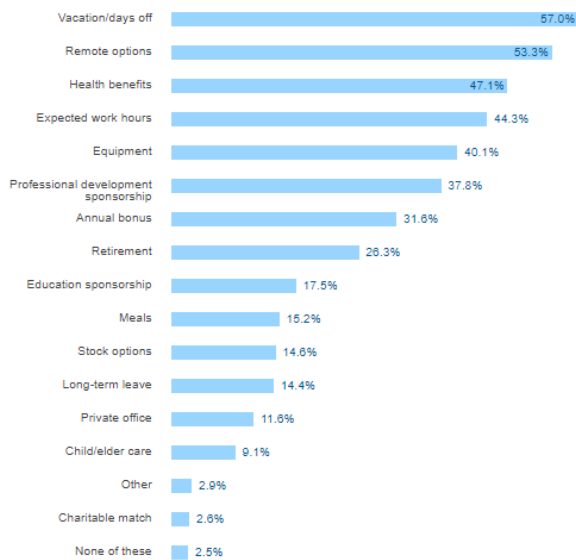
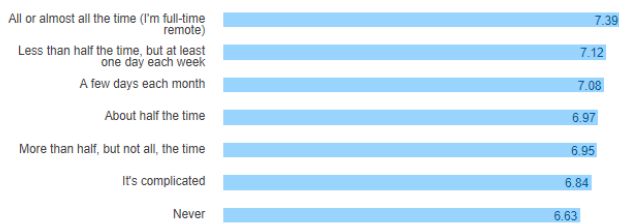


Fig. 2. What developers value in compensation benefits (<https://insights.stackoverflow.com/survey/2017#remote-work>)

Job Satisfaction Rating and Remote Work



Mean of 40,325 responses; satisfaction on a 0-10 scale

Fig. 3. Job satisfaction rating and remote work (<https://insights.stackoverflow.com/survey/2017#remote-work>)

As Fig. 3. indicates, there is a moderate correlation between flexible working time and satisfaction with the job and the company. The highest ratings for job satisfaction come directly from employees who have full flexible hours.

III. FLEXIBLE WORKING OF IT PROFESSIONALS IN MACEDONIA

In order to gain insight, we conducted an online survey on flexible working of IT professionals in Macedonia. The methodology was following:

- The questionnaire consisted of 14 questions;
- Each question contained from 1 to 30 options in the answer. If the respondent could not find satisfying answer in the options offered, each question had an additional field in which open answer can be added;
- The survey questionnaire was sent to 117 IT professionals from Macedonia through the social networks LinkedIn and Facebook;
- The time frame for filling in this questionnaire was from 13.02.2019 to 20.02.2019 (7 days).

A. Demographics

The demographic structure of the respondents is as follows:

- Regarding the age, between 18 – 23 years were 7.7%; between 23 – 30 years were 54.7%; between 30 – 50 years were 35.9%; between 50 - 64+ years were 0.9% and 0.9% did not want to share this information.
- Regarding the gender, male respondents were 59.8% and female were 39.3%. 0.9% did not want to share this information.
- Regarding the level of education, the structure is following: secondary education 8.5%; completed level of College/University studies without a degree 11.1%; Undergraduate studies 65.8; Master's Degree 12%; Doctoral studies and 0.9% and did not want to share this information / 1.7%.
- Regarding the type of IT professional, the structure of respondents was following: Full Stack Web Developer 16.2%; Front End Web Developer 8.5%; Back End Web Developer 10.3%; Desktop Application Developer 3.4%; Android Developer 0.9%; iOS Developer 0.9%; Embedded Developer 0%; Database Administrator 0%; System Administrator 3.4%; DevOps Specialist 0.9%; Machine learning specialist 0.9%; Quality Assurance Engineer 9.4%; Graphics programming 0.9%; Graphic Designer or Illustrator 5.1%; Educator or Academic 1.7%; IT Project Manager 1.7%; Product Manager 1.7%; Marketing or Sales Manager 4.3% ; C-suite Executive [CEO, COO or CTO] 2.6%; Did not want to share this information 15.4%; Data Analyst 0.9%; Game Designer 0.9%; App Developer 0.9%; Game level designer 0.9%; UI designer 0.9%; Senior Recruiter 0.9%; Customer support - Service management 0.9%; Integration Developer 0.9%; Database Developer 0.9%; Test Automation Engineer 0.9%; IT Law 0.9%; Java developer and Database administrator 0.9%; Animator 0.9% and Concept Artist 0.9%.

- Regarding the work status, the structure were following: Full time 82.1%; Freelance 5.1%; Self-employed 0%; Part time employee 0%; Full time/part time employee and freelancer 8.5%; Unemployed who is currently not looking for a job 0%; Unemployed who is currently actively seeking work 1.7%; did not want to share this information 2.6%.
- Work experience is distributed this way: less than 1 year 12.8%; 1-2 years 8.5%; 2-3 years 13.7%; 3-4 years 10.3%; 4-5 years 11.1%; 5-6 years 3.4%; 6-7 years 11.1%; 7-8 years 1.7%; 8-9 years 1.7%; 9-10 years 6%; 10-11 years 4.3%; and more than 11 years 12.8%.
- The company size distribution is as follows: less than 10 employees 12.8%; 10-19 employees 15.4%; 20-99 employees 37.6%; 100-499 employees 18.8%; 500-999 employees 6%; 1000-5000 employees 1.7%; more than 5000 employees 1.7%; does not know 1.7%; and did not want to share this information 4.3%.
- The company industry is following: Software - Development & Maintenance 50.4%; Internet and web services 6.8%; Finance, Banking and Insurance 3.4%; Media, advertising, publishing or entertainment 4.3%; Consulting 4.3%; Education 3.4%; Health services 0.9%; Telecommunications 0%; Retail or wholesale 1.7%; Civil Service (including Ministry of Defense) 1.7%; Computer hardware or consumer electronics 0%; Transportation, logistics or storage 0.9%; Automotive 0%; Aviation or Defense 0%; Gaming 11.1%; Industrial equipment/heavy machinery 0%; Energy production/distribution 1.7%; Pharmaceuticals and/or medical devices 0%; Non-profit/non-governmental organizations 0.9%; Construction 1.7%; Agriculture, forestry or fishing 0.9%; Mining or extraction of oil and gas 0.9%; Archive and data center 0.9%; Legal services / 0.9%; don't want to share this information 3.9%.
- The question whether the company is startup was answered like this: yes - 70.9%; no - 20.5%; don't want to share the information 8.5%.

B. Main results

The other part of the survey answers will be discussed in more details, since they serve as a base to analyze the effects of flexible work in the IT industry in Macedonia.

The answers on the question how satisfied are you from your current work are presented in the Figure 3.

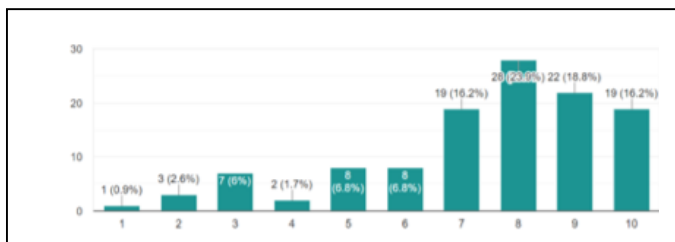


Fig. 4. How satisfied are you from your current work?

The answers on the question whether they look for a new job are as follows:

- I'm not actively seeking, but I'm open for new opportunities - 56.4%
- I'm not interested in a new job - 22.2%
- I'm actively seeking for a new job - 9.4%
- I don't want to share this information - 12%.

The next answer was what benefits and perks do you appreciate the most when you accept a new job offer. There were maximum of three possible answers on this question. The results indicate following:

- Free days - 44.4%
- Ability to work from home (flexible work) - 53%
- Health benefits - 35%
- Strictly planned and expected working hours - 29.9%
- Sponsorship for professional development - 51.3%
- Educational sponsorship - 28.2%
- Food and beverages during working hours - 11.1%
- Company shares / options - 12%
- Long term absence - 2.6%
- Private offices - 4.3%
- Weekly / monthly / yearly social work - 2.6%
- Sports packages - 14.5%
- None of the above - 9.4%.

The next question deepens the issue, asking how important is flexible work offer when you accept a new job. The results are shown at Figure 4.



Fig. 5. How important is flexible work offer when you accept a new job?

Next question is how often do you use flexible work for fulfilling your work duties. The answers are:

- Complicated to explain - 12%
- Fully flexible - 11.1%
- Half the time, but not always - 4.4%
- Half of the working time 0.9%
- Less than half, but at least one day during the working week 2.6%
- Few days in a month - 27.4%
- Never - 36.8%
- Very rarely, in time-bound, emergency situations. Annually maybe 5 times - 2.2%
- Once a month - 2.6%.

C. Comparative Analysis of Stack Overflow and Survey in Macedonia

1) Working status

The results obtained for the employment status of the survey questionnaire and the survey of Stack Overflow did have significant difference. In the Stack Overflow survey, 70.3% responded that they were employed full time, while 10.2% were freelancers. In the Macedonian survey questionnaire, 82.1% responded that they were employed full time, while 5.1%, exactly a half of the Stack Overflow's results were freelancers.

2) Are you satisfied with your work?

On a scale of 1-10, where 1 represents "not at all satisfied with their current job" while 10 represents "high level of satisfaction with their current job", in a survey by Stack Overflow, 22.2% rated their current job with a rating of 8 whereas, in the our survey questionnaire, 23.9% rated their current job with a score of 8 and an additional 18.8% rated their current job with a score of 9. Thus, we can conclude that respondents to the questionnaire have a high level of satisfaction with their current work.

3) Are you currently looking for a new job?

Nowadays, the constant fluctuation of IT professionals from company to company for various reasons is a huge problem for IT companies because of the impact on the productivity of the companies.

Although we have a high level of satisfaction with the respondents, in the survey by Stack Overflow, 62.1% responded that they are not actively looking for a job at the moment, but are open to new opportunities. 24.8% responded that they were not interested in new job opportunities.

In the survey questionnaire conducted in Macedonia, 56.4% responded that they are not actively looking for a job at the moment but are open to new opportunities. 22.2% responded that they were not interested in new opportunities.

From both surveys we note that this percentage is very high and companies must investigate why despite high satisfaction, many employees are interested in leaving the company on the first or better offer.

4) What do you most consider when accepting a job offer

In a survey of Stack Overflow, from a larger selection of benefits, respondents identified the following three as most important:

- Vacation - 57%
- Flexible work opportunity - 53.3%
- Health benefits - 47.1%

In Macedonia, the respondents selected the following benefits:

- Flexible working opportunity - 53%
- Sponsorship for professional development - 51.3%
- Holiday - 44.4%

Unlike the respondents of Stack Overflow, IT professionals in Macedonia are more interested in the

possibility of flexible working and sponsoring professional development. For respondents to Stack Overflow, sponsoring professional development has been replaced by health benefits.

The percentage of respondents that valued flexible work arrangements were almost the same in Macedonia and on the global level – 53% and 53.3%.

5) How often do you use flexible working hours to complete your activities

Among Stack Overflow respondents, 35.1% responded that they work flexibly for a few days a month, while 31.8% responded that they never use flexible working hours to complete their activities.

In our survey questionnaire in Macedonia, 36.8% said they never use flexible working time to complete their current tasks, while 27.4% responded that they use a few days a month for flexible work.

Although flexible work is one of the major benefits sought by Macedonian IT professionals, we can conclude that many companies do not have a flexible work program.

IV. DISCUSSION

Flexible work in a very short time has gone into the race for who will do it faster. While the media report on how much flexible work increases productivity and the skills needed for a worker to succeed in a flexible environment, the need for this change in companies' work strategies comes much faster than expected. It is worrying that some companies, focused on their profits, are still unaware of the need, as well as the risk of not implementing a flexible strategy for their employees. Despite all the research showing that workers are ready to leave the current company for a more flexible work environment, even if that means lower financial benefit, some companies are not even considering such a strategy.

We can safely conclude that the structure of the work is changing. However, not all companies are successful in change [6] [7], so they try to stay as close as possible to the most successful ones.

Also, not all countries have solid background in introducing and promoting flexible work during the last decades. Macedonia has very modest experience in this area. As to this broader systematic change some socioeconomic insights should be mentioned regarding the situation in Macedonia. The developed western European countries has long been practiced different and numerous models of flexible working time, just to mention few interesting ones: four day work week, staggered working time, regular part time (Turnusteilzeiten), work time based on trust (workers, employees allocate their time on mutual trust).

German company Bosch certainly has over 100 forms of employment – like BMW, Ikea and many others¹.

In Macedonia, working hours are extremely inflexible. And this is mainly due to the ineffective combination of a legacy of socialist classification of workers employment, and the neoliberal model of understanding the position of the employee that has ruled for nearly 30 years [8].

¹ See <https://www.zeit.de/karriere/2015-10/bosch-arbeitszeitgesetz-pausenzeit-christoph-kuebel/komplettansicht>.

The socialist classification consists of the two most common forms of employment - full-time, indefinite period employment, and fixed-term employment.

The neoliberal model has been grafted onto economy with high unemployment, run by new made employers who believed that he had all the options of the Manchester capitalism at their disposal, on the indolence of the worker who accepted exploitation because he had no choice. Macedonia is still a country of widely spread stance of de-ideologization among leading political parties in which distancing from Marxism and Socialism meant distancing from labor rights discourses and the role of workers unions.

After 30 years the percentage of black labor is still high, unemployment is reduced to about dull 20 percent. The influence of the ruling parties on economy and public life is also too high. In the Macedonian very small market with underdeveloped private sector, the state (the ruling party) is still the best employer.

It is difficult to discuss on culture of flexible working time in Macedonia. For illustration the state institution still confuse staggered working time, with flexible working time. Both the media and citizens are under the illusion that this is the only possible innovation of the flexible working time plan. The new EU Work-life Balance Directive is almost unknown in Macedonia

This, however, is not the case with the IT sector. The IT labor market is more dynamic, independent, and with much larger perspectives than other labor markets in Macedonia. Not only the region and Europe, but also the US and Canada are considered as part of the IT labor market for Macedonian IT experts. And exactly here lies the problem. The state and the market in Macedonia cannot contain young IT experts. Their incomes are pretty high, compared to other professions, but they are insufficient to keep young IT experts in Macedonia.

Something still could be done. In Robert Walters Employee Insights Survey, 37% of the professionals would accept a lower salary if that offered a better work-life-balance.² In Macedonia 53% of the respondents selected flexible working opportunity as benefit.

Still, in the Macedonian survey questionnaire, 82.1% responded that they were employed full time. It is clear that flexible working time is a serious option to be offered to young IT professionals if the state wants to stop the brain drain. Though, no national strategy can fight the need of young people for international experience (in Robert Walters Survey 39% of the respondents choose international experience as a reason for working abroad, and only 7% for better work-life-balance).

On a global level, although everyone is focused on the impact of artificial intelligence on the labor market, on the contrary, according to most studies, in the near future flexible work will have the greatest impact on the current labor market (the nature of the job), especially when hiring and retaining talented workers.

Yes, it should be mentioned that some companies have abandoned their flexible working strategy and have called

their employees back to the office, for example: IBM [9] and Yahoo [10]. However, a report released in 2018 on the state of the flexible labor market shows that 90 percent of flexible workers plan to continue working flexibly until the end of their careers. This is not surprising, given the high benefits of a flexible schedule, more time with family, a more comfortable working environment and avoiding specific office policies.

According to experience, the main issues that led to cancellation of flexible working arrangements in some companies are:

1. Co-locating employees in one place with all employees as a strategy for fostering innovation. Returning people physically back to the office is a "calculated risk", an effort to keep up with younger startups without corporate bureaucracy and able to better focus on product and re-design.

2. Flexible work is not suitable for every company and every employee.

3. Too much focus on technology and not enough on process.

The usual traps [11] in introducing flexible work arrangements are altered work-life dynamics, reduced fairness perceptions, and weakened organizational culture—and a balanced flexibility approach is required. The managers must become flex savvy to understand the options that exists in flexibility practices to align implementation with the workforce and organizational context. Also, implementing flexibility must be treated as a broader systematic organizational change empowering individuals and teams.

V. CONCLUSION

What can be concluded from the analysis of the current and previous situation in the labor market and the data obtained from the conducted surveys?

In order to make a successful flexible working policy, the following key principles must be established at the outset:

1. Communication. In a virtual environment, it can be difficult to explain complex ideas, especially if people are unable to ask questions and discuss in real time. Lack of face-to-face interaction limits social cues, which can lead to misunderstandings and conflict.

2. Coordination. Having processes is not enough. Managers must model and implement them until they are fully accepted. They should also evaluate the team members for how well they adhere to the protocol. Otherwise, they will return to their old habits. Working outside the established processes will shake the team's cohesion.

3. Culture. This principle is especially critical for virtual teams, but it is also important for individuals who work flexibly. Because these people rarely meet their peers face-to-face, they tend to focus on tasks and ignore the team. This may work for a while, but it must develop a culture to encourage engagement and maintain their performance in the long run.

4. Security practices and security risk management. The key to managing flexible workers' safety is to implement

² <https://www.robertwalters.de/content/dam/robert-walters/country/germany/files/reports/rw-eis-germany-2015.pdf>

applicable security measures right now, even if it costs a little money in advance

Implementing a successful flexible working policy is quite complex. This requires careful strategy and reliable execution. But when it's done well, the reward is high: increased productivity, happier employees, and cost savings (which can be invested in building a better business). With major changes in the workplace, such as growth of the millennial cohort and a blurred boundary between work and life, flexible work will become a key tool for recruiting and hiring employees.

Companies like Yahoo may try to reverse the trend, but it is better to rethink the issues that have led them to ban flexible work.

Countries like Macedonia should educate business managers and entrepreneurs about the flexible working arrangement models and the benefits for the companies and for the society.

During the preparation of this research for publication, a Coronavirus pandemic emerged. Many companies migrated their work online in a short period of time, under the pressure of extraordinary circumstances. As future research, it would be interesting to carry out what the results of the same survey would be after the end of the pandemic.

REFERENCES

- [1] Jeffrey Hill E, Grzywacz JG, Allen S, Blanchard VL, Matz-Costa C, Shulkin S, Pitt-Catsoupes M. Defining and conceptualizing workplace flexibility. *Community, Work and Family*. 2008 May 1;11(2):149-63.
- [2] Hickman A, Fredstrom T. How to Build Trust with Remote Employees. Gallup. com, Gallup, Inc. 2018 Feb;7.
- [3] Mellner C, Aronsson G, Kecklund G. Boundary management preferences, boundary control, and work-life balance among full-time employed professionals in knowledge-intensive, flexible work. *Nordic journal of working life studies*. 2014 Dec 1;4(4):7-23.
- [4] Stack Overflow (2018), *Developer Survey Results 2017*, URL: <https://insights.stackoverflow.com/survey/2017>
- [5] Kossek EE, Thompson RJ. Workplace flexibility: Integrating employer and employee perspectives to close the research-practice implementation gap. *The Oxford handbook of work and family*. 2016 May 17;255.
- [6] Blok MM, Groenesteijn L, Schelvis R, Vink P. New ways of working: does flexibility in time and location of work change work behavior and affect business outcomes?. *Work*. 2012 Jan 1;41(Supplement 1):2605-10.
- [7] Leonardi PM. When flexible routines meet flexible technologies: Affordance, constraint, and the imbrication of human and material agencies. *MIS quarterly*. 2011 Mar 1:147-67.
- [8] Le Grand J. Motivation, agency, and public policy: of knights and knaves, pawns and queens. Oxford University Press on Demand; 2003 Sep 18.
- [9] Jing Cao (2017), *IBM Touts Trump-Pleasing Hiring Plans While Firing Thousands*, URL: <https://www.bloomberg.com/news/articles/2017-01-23/ibm-touts-trump-pleasing-hiring-plans-while-firing-thousands>
- [10] Miller CC, Rampell C. Yahoo orders home workers back to the office. *The New York Times*. 2013 Feb 25:A1.
- [11] Kossek EE, Thompson RJ, Lautsch BA. Balanced workplace flexibility: Avoiding the traps. *California Management Review*. 2015 Aug;57(4):5-25.

Transition from the classroom to online educational environment: First impressions

Petre Lameski, Vladislav Bidikov, Kiril Kjiroski, Boro Jakimovski,
Eftim Zdravevski, Ivan Chorbev and Vladimir Trajkovik

Ss. Cyril and Methodius University in Skopje
Faculty of Computer Science and Engineering
Skopje, North Macedonia

{lameski, vladislav.bidikov, kiril.kjiroski, boro.jakimovski, eftim, ivan.chorbev, trvlado}@finki.ukim.mk

Abstract—With the world pandemic caused by the COVID-19 virus, the lectures at the Faculty of Computer Science and Engineering (FCSE), at the Ss. Cyril and Methodius University in Skopje, Republic of North Macedonia, were fully transferred online within the first week of the start of the imposed movement and public gathering limitations by the Government of North Macedonia. After two weeks of the start of the classes, we performed a survey for both students and faculty teaching staff to gather the initial sentiment for the online classes and to identify opinions, suggestions for best practices and possible obstacles. In this paper, we present the technological solutions used to give the lectures and the results of the initial surveys.

Index Terms—distance learning, moodle, e-learning, online education

I. INTRODUCTION

Due to the COVID-19 pandemic, the education process in North Macedonia has been rapidly transferred from the blackboard and smart classrooms to the online tools. Within the first month of the proclamation of the emergency state from the government, almost all educational institutions have transferred to the available online conferencing platforms such as Zoom [1], Moodle [2] and Microsoft Teams [3].

Given the need for rapid transfer to these technologies, the process's full digitization was done almost without an in-depth analysis of the technology acceptance. Some of the schools and faculties opted to stop the classes for the semester. They decided not to continue online due to the lack of technology acceptance by their teachers and students. The technology acceptance of online learning and education in general, has been discussed many times in the literature [4]. The reported survey suggests that faculties are more eager to accept online education if they perceive the usefulness of such technologies, regardless of the ease of use. The study shows that usefulness seems to be a major indicator of the acceptance of online learning. Authors in [5], find that teachers that teach more traditional courses are less likely to easily migrate to online learning and provide online courses and curricula. Be that as it may, since 2011, many universities have offered undergraduate, professional, and masters online degrees. The quality of experience and quality of service measurements are significant for online classes, primarily when classes are intended for students [6]. Among the first companies that offered fully online courses and degrees with the possibility of certification

were Udacity [7] and Coursera [8], which quickly became trendy tools for online learning. The number of faculties that offer online educational opportunities increases every day. The COVID-19 crisis further sped up the process of improving and increasing the volume of online education.

The technical faculties, especially FCSE, were at an advantage because online classes and conference meetings are not a new thing in the ICT sector. The Faculty has already participated in several projects for online education such as The ViCES Tempus project [9], which included Teleconferencing equipment and classroom in which online lectures were organized since 2011. The initial experiences were very optimistic and the quality of experience for students was on a very high level [10]. Furthermore, the courses in the Faculty were already hosted on the Moodle platform. The same platform was also used for organizing tests and exams within the Faculty laboratories and sometimes online.

FCSE at the Ss. Cyril and Methodius University in Skopje, the largest technological Faculty in North Macedonia, started the online lectures within one week from the government proclamation. In this paper, we present the technologies used by the Faculty staff and students and the initial student and professor feedback and satisfaction from the online lectures.

In the following sections, we describe the used technologies and present selected results of the performed survey for both students and faculty teaching staff.

II. TECHNOLOGY STACK

A. Hardware and software resources

Faculty of computer sciences and engineering has been a crucial part of the Macedonia e-Education Project. In the course of this project, a Container Data Center was installed on the Faculty premises. Employees at the Faculty Computer Center (FCC) took an integral part in the installation of all computing, storage and networking resources. That enabled us to plan, test and implement various virtualization and cloud systems. As a result of this effort, many national platforms are today hosted at Faculty premises, and we strive to test and improve on our previous experiences. Taking a portion of the installed equipment, FCC implemented a highly available and fault-tolerant virtualized environment for implementing the e-Education platforms. This platform's goal is to enable teachers

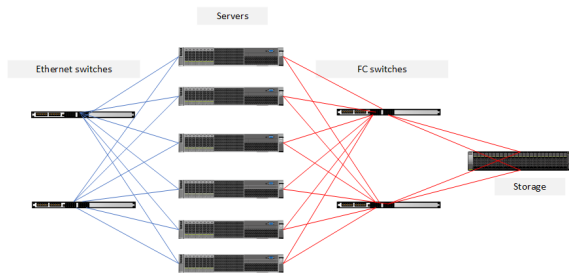


Fig. 1. Hardware connections for used equipment.

and students to take active participation in our efforts to provide an uninterrupted continuation of the education process in these moments of social distancing. Following hardware resources were (either partially or fully) used in this effort:

- six Huawei RH5885H V3 servers
- one Huawei OceanStor 5500 V3 FC storage system with redundant controllers
- two SNS2248 Fiber Channel switches
- two Huawei Cloud Engine 6851 10 Gbit ethernet switches in a stacked configuration

Fig. 1 represents hardware interconnections. All of the servers have the following configuration:

- Two Intel(R) Xeon(R) CPU E7-8880 v3 @ 2.30GHz
- 256 GB DDR3 RAM @ 2.13 GHz
- dual 2600 Series 16Gb Fiber Channel HBA
- dual Intel 82599EB 10 Gbit network adapter

Storage and servers are interconnected using SNS2248 FC switches, where each server and storage controller are connected to each of the switches for redundancy. Additionally, servers are connected to the rest of the network using stacked CE6851 switches, also in a redundant configuration. Thus, there is no single point of failure, and FCC put in place various monitoring mechanisms to enable prompt and accurate reaction in case of hardware failures.

For server virtualization, we used VMware ESXi, version 6.5.0, and vCenter Server Standard, version 6.5.0, for the centralized management and operation. VMware cluster is working in High Availability mode, with 10% of CPU and 10% of memory dedicated to maintaining this HA environment. Two 10 Gbit ethernet adapters per server are used for networking, both connected to the same virtual switch, thus providing redundant and load balancing network connection. Network is implemented through VMkernel ports, using two such ports. One for management and virtual machine connection purposes, and the other for vMotion.

Storage is organized in the following fashion:

- Performance Tier, comprising of ten 600 GB 15k rpm SAS HDD disks, organized in RAID 5
- High-Performance Tier, comprising of eight 900 GB eMLC SSD disks, organized in RAID 5
- Total usable capacity of 8 TB

Storage has redundant FC controllers, each equipped with two 16 Gbit FC adapters. Each of the controllers is connected

to both FC switches, thus providing redundant storage network connections. This configuration also enables load balancing between servers and storage, utilizing Round Robin for multi-path connections. This contributes to faster and more reliable communication with the storage. All virtual machines and corresponding files, such as disks and swap files reside on the storage, and can be migrated to another storage if necessary.

Between storage, servers and FC switches, there is no single point of failure, and monitoring has been implemented to alarm FCC staff in case of any hardware or software outage. FCC is monitoring computing, storage and networking elements using the following technologies:

- Monitoring of networking elements using SNMP monitoring and reporting tool LibreNMS
- Monitoring of hypervisors, virtual machines and storages using Veeam One Monitor
- Monitoring of virtual machines and services using Zabbix

The cluster network connectivity to the external environment was based on implementing VLANs through Virtual Machine Port Groups and using two 10 Gbit network interfaces as aggregated 20 Gbit network connectivity on virtual switches. This architecture enables implementing more than three hundred virtual machine network interfaces on more than fifty different VLANs using just these two physical interfaces. Three of these VLANs were used in our platform's implementation, as will be described in detail later.

In order to provide a better end-user experience for the users, additional work was done on the topic of network connectivity. FCSE is already well connected to the general internet based on two full Internet up-links as following:

- 1 Gbit upstream via GEANT project provided via the University network
- 500 Mbit upstream via a commercial ISP provided in cooperation with the Ministry of Education

Based on our knowledge of the way Internet service providers are connected between them and based on the Faculty project for national network connectivity and research IXP.mk we established the need to have appropriate network capacity for our future endeavors. It was decided to use the IXP.mk platform to allow for direct connections between the FCSE and several Macedonian Internet operators. We based our work on establishing a separate Autonomous number (AS) for the FCSE network and its' connection to the University network and commercial operators via a dedicated router based on the Bird routing software¹. In order to be able to scale the system and handle this expected additional load expected, it was decided to switch from a bare-metal hardware machine to a virtual machine running on the same infrastructure. This would make the router also highly available and eliminate the single point of failure which a bare metal machine was presenting.

On the software side, we already had both systems for courses and exams running on the Moodle LMS, so we only

¹<https://bird.network.cz/>

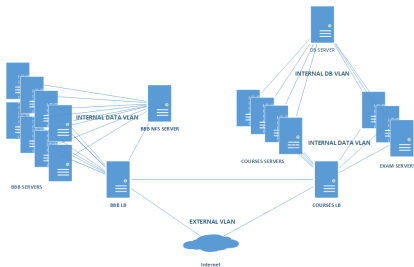


Fig. 2. Logical interpretation of server platform interconnections.

needed to add the part of the system, which will allow easy web conferencing for the teachers and students. Based on our previous experience, we were quite sure that we have the perfect platform for this based on the Big Blue Button² open source web conferencing system. The Big Blue Button web conferencing system was the perfect match since it already has all the required tools to aid in the teaching process. As expected, based on our size and number of students, we also saw the need to scale this component.

Scaling for the LMS part of the systems was generally an easier task. We only needed to be sure that we use the appropriate number of end nodes and use the already established Load Balancer server we have for most of our Faculty services. The scaling of the web conferencing component was a bit more challenging. It was resolved by implementing an open-source project which we had to customize based on our needs heavily.

At the end this is the final platform setup which we use now:

- Course learning platform using three front end nodes
- Course exam platform using four front end nodes
- Database server for course learning and course exam platforms
- Load balancer virtual machine serving both course learning and exam clusters to end-users
- Big Blue Button web conference platform comprising eight web conference front end nodes, one NFS storage and special BBB load balancing machine.

Previously mentioned VLANs are used for interconnection in the following way:

- External VLAN for load balancers serving content from course and web conference platforms to all users
- Internal VLAN for exchanging data between load balancers and application servers
- Another internal VLAN for database access between the database server and application servers

Logic interconnections and appropriate VLANs between web conference servers, recorded video server, course servers, exam servers, database server and load balance servers are depicted on Fig. 2. All lines drawn between the servers symbolize data connections between them and do not represent any existing physical logical connections.

²<https://bigbluebutton.org/>

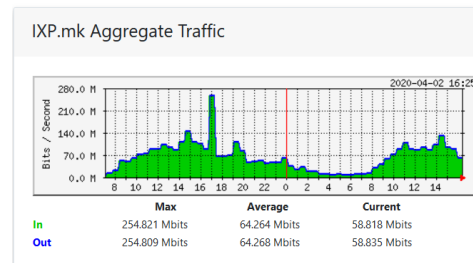


Fig. 3. Traffic over IXP.mk platform on 2020-04-02.

To fully utilise the internet capacity and provide better end-user experience for the web-conference platform, we decided that the Course learning and exam systems will use the University connectivity and the BBB platform will use the commercial and IXP.mk connectivity. This would allow audio and video traffic to travel the shortest path thanks to the IXP.mk platform, thus providing better quality, and in the long run, it will offload connectivity needs for the service providers where the teachers and students will connect.

Our effort on using the IXP.mk platform was quite well accepted with all Service Providers in Macedonia. In a matter of several days, we got connected with: Macedonian Telekom, Telekabel, Telesmart, Interspace, all combined with the presence of the University network. Also, the IXP.mk platform allowed for most of the traffic to flow over this high-capacity platform of loading this traffic from Commercial operators' internet links. This benefit is clearly shown in the following Fig.3, where total traffic is shown flowing on 2020-04-02, where we saw the maximum amount of traffic which we observed.

The presented platform gave the expected results, and we never saw any possible resource bottleneck based on the network's capacity or the available virtual resources.

The biggest challenge was providing a secure way to execute the partial exams that were coming and required that we use the exam platform in a form where students access it from their home computers. For this task, the only viable solution was already known to us as part of the Moodle LMS, but we had to test the system's sustainability and stability. Moodle already has support for a particular component known as Safe Exam Browser³, which allows for more precise control of exams done with the Moodle Quiz activity. We just needed to test how this component will work on the clustered exam system and how much more resources it will use since the Safe Exam Browser required an additional access module to be active on the Moodle LMS. After the initial testing, we saw that the system was with good quality and it was presented to the teaching staff in order to be used if required for exams.

The current development of Safe Exam Browser is still not completely in line with both supported platforms Windows and Mac OS X. However, based on its more frequent usage in this time of global pandemic, we see that development is gaining

³<https://safeexambrowser.org/>

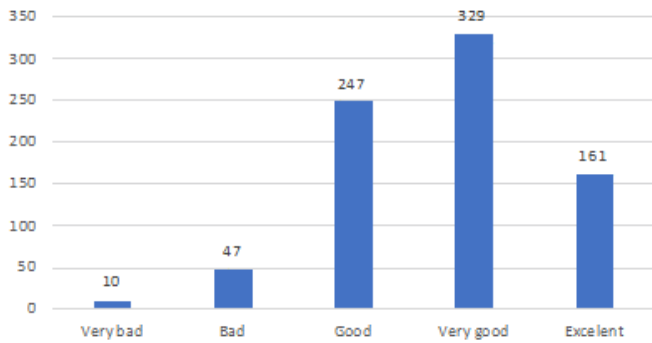


Fig. 4. General satisfaction of online education by Students

momentum, which means that we will probably have the technical capability to further enhance its usage into the FCSE LMS systems. The faculty computer center already committed some valuable feedback to both Big Blue Button and Safe exam browser Moodle LMS integration projects based on their own experience in using them in a very clustered environment.

III. EVALUATION PROTOCOL

The Faculty of Computer Science and Engineering has over 4000 active students and over 70 teaching staff that use the online system. For this purpose, we used our online system (Moodle) for surveying both the teaching staff and the students. The survey was anonymous. The students and teachers were asked a different set of questions, some of which included performance measurements and their home computer setup details. The results of the survey are presented and discussed in Section IV.

IV. RESULTS AND DISCUSSION

Much analysis can be performed from the filled-in surveys by the students and teachers. For the purpose of this publication, we have decided to use only the initial satisfaction from online lectures and compare it to the observed satisfaction from traditional lectures performed in classrooms. We focused our initial analysis on two questions: “How do you evaluate the general experience with the online lectures?”, and “Compare the traditional lectures with the online lectures?”. For the first question, the students had to choose between 5 options from best to worst: Excellent, Very good, Good, Bad, Very bad. For the second question, the students chose between 5 options too: Online much better, Online better, Equally good, Traditional better, Traditional much better.

On Fig 4 we can see the results of the general satisfaction from online lectures by students. We can observe that large majority of students are very satisfied with online classes. Only 10 students evaluated the online classes as very bad which scales to less than 2% of the students. More interesting observation can be made by seeing the results of Fig 5. It can be observed that the satisfaction from online classes generally grows as the students are more experienced with First and Second-year students being less satisfied than Third and Fourth-year students.

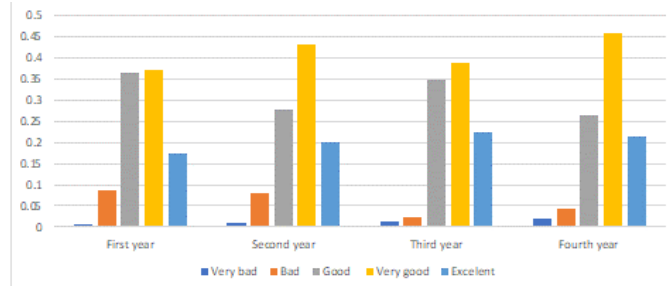


Fig. 5. General satisfaction of online education by Students

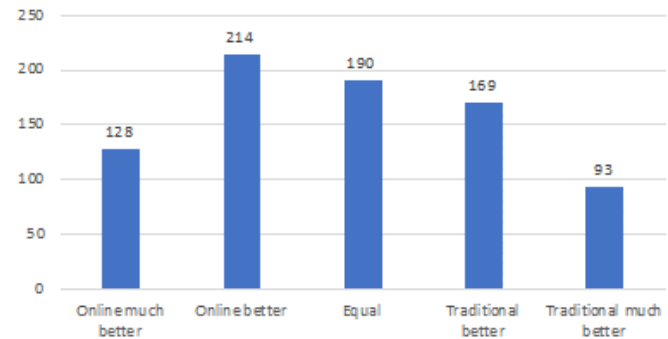


Fig. 6. General comparison of Online VS Traditional classes

A similar observation can be made when we compare the student satisfaction from online and offline classes. As can be observed in 6, the majority of the students evaluate the online classes to be equal or better than traditional offline classes. Furthermore, when we compare the distribution of answers per year (by normalizing for the different number of responses per year), we can see that similar to the observation about the general satisfaction, higher year students tend to like online education more than the lower year students. In general, the majority of the students from each year, prefer online education. However, the number of those who think that traditional is better or much better lowers down as the study year increases. When asked why they prefer online education in the survey, students explained that they like online classes more because they do not need to commute to the faculty premises and they can follow the classes comfortably from

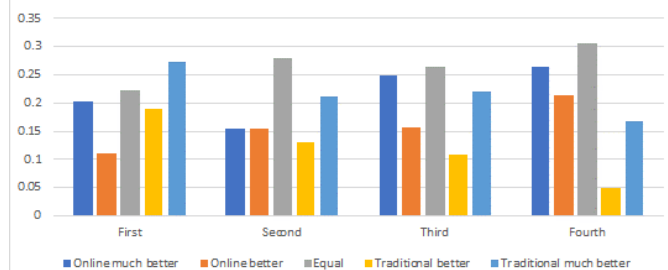


Fig. 7. Normalized per year distribution for comparison of Online VS Traditional classes

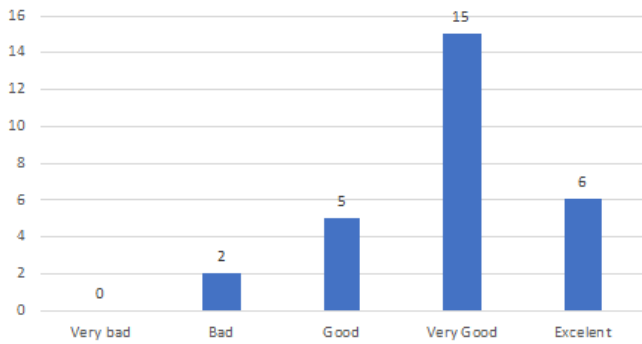


Fig. 8. General satisfaction of teachers from online classes

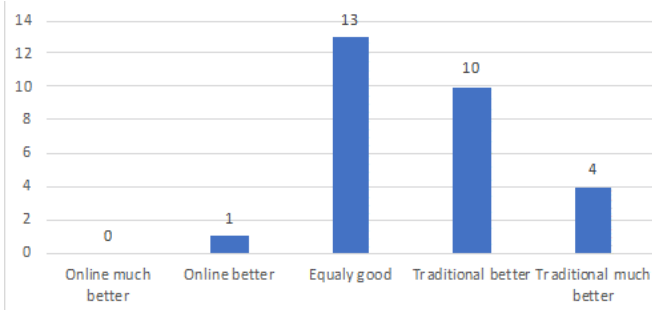


Fig. 9. Comparison of online VS traditional classes satisfaction by teachers

their homes. Another reason is that most of the classes are recorded and they can watch them later when they feel like it. On the other hand, students complain to lack of classroom context, lack of interaction and bigger dissatisfaction if the online classes are live and not recorded.

We did a similar survey for the teachers and the results were significantly different from the results from the students. For example, if we observe the general satisfaction from the online lectures by the teachers in Fig 8, we can see that the teachers are similarly satisfied with the online lectures. However, if we see Fig 7, we can observe that by a large margin, teachers believe that traditional classes are equal or better than online classes. Based on the followup answers, the main reason for the lower satisfaction from online classes is the lack of live interaction on the classes and the classroom atmosphere, which cannot be emulated by the online classes environments and tools. In a live classroom environment, teachers can get both audio and visual feedback from the students and estimate the degree of understanding. This is lacking in our online classes, where students mostly chose to connect only for listening and almost never turned on their camera.

V. CONCLUSION

Thanks to the rapid response of the teaching, administrative and IT staff FCSE has successfully implemented online learning and fully restarted the classes on 2020-03-17, with several classes being experimentally held in the previous period. This was done only seven days after the faculty premises were closed and the on-premises classes were banned. Based on

the initial survey, both students and teachers are generally satisfied with the online classes. However, there seems to be a different initial perception of the online courses from teachers and students. Furthermore, older students seem to prefer online lectures rather than traditional ones, while most teachers prefer traditional lectures or consider them the same. Additional surveys are planned after the end of the semester. Additional research is planned after the results from the exams to evaluate online education further and compare it with the traditional one. Based on the initial survey, we can conclude that FCSE has successfully migrated towards online education, especially for regular classes and auditory exercises. There are still many challenges to be overcome, such as limitations of students' equipment, organizing exams, and laboratory exercises, some of which are very difficult to implement in an online environment.

ACKNOWLEDGMENTS

The authors acknowledge the contribution of the teaching and administrative staff at the Faculty of Computer Science and Engineering without whose devotion, the online classes would not have been possible. The authors acknowledge the contribution of the students who filled in the surveys and gave their much-appreciated feedback.

The work presented in this paper is financed by the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia.

REFERENCES

- [1] "Zoom," <https://zoom.us/>, accessed: 2020-04-20.
- [2] "Moodle," <http://moodle.org>, accessed: 2020-04-20.
- [3] "Microsoft teams," <https://www.microsoft.com/en/microsoft-365/microsoft-teams/group-chat-software>, accessed: 2020-04-20.
- [4] S. G. Gibson, M. L. Harris, and S. M. Colaric, "Technology acceptance in an academic context: Faculty acceptance of online education," *Journal of Education for Business*, vol. 83, no. 6, pp. 355–359, 2008.
- [5] C. Stewart, C. Bachman, and R. Johnson, "Predictors of faculty acceptance of online education," *MERLOT Journal of Online Learning and Teaching*, vol. 6, no. 3, pp. 597–616, 2010.
- [6] T. Malinovski, V. Trajkovic, and T. Vasileva-Stojanovska, "Impact of different quality of service mechanisms on students' quality of experience in videoconferencing learning environment," *Turkish Online Journal of Distance Education*, vol. 19, no. 3, pp. 24–37, 2018.
- [7] www.udacity.com, accessed: 2020-04-20.
- [8] www.coursera.org, accessed: 2020-04-20.
- [9] T. Project, "Video conferencing educational services," *144650-TEMPUS-2008-IT-JPGR*, 2009/2012.
- [10] E. Caporali and V. Trajkovic, "Vices: Video conferencing educational services main project outcomes," *ViCES*, pp. 1–282, 2012.

Design optimization of Rectifier Transformers

Rasim Salkoski

Faculty of Machine Intelligence and Robotics
University for Information Science and Technology
Ohrid, R. of Macedonia
rasim.salkoski@uist.edu.mk

Ivan Chorbev

Faculty of computer science and engineering
University of Ss Cyril and Methodius
Skopje, R. of Macedonia
ivan.chorbev@finki.ukim.mk

Abstract— Optimization refers to finding one or more feasible solutions, which correspond to extreme values of one or more objectives. The need for finding such optimal solutions in a problem comes mostly from the extreme purpose of either designing a solution for minimum possible cost of fabrication, or for maximum possible reliability, or others. Because of such extreme properties of optimal solutions, optimization methods are of great importance in practice, particularly in engineering design, scientific experiments and business decision-making. Rectifier transformers deserve extensive treatment in the field of research and production, due to the fact that the electric energy undergoes several transformations on its way from generators to the consumers i.e. rectifiers. In this paper, an effective application of the population based search Differential Evolution algorithm is proposed with the aim of minimizing the cost of the active part of wound core rectifier transformers. The constraints resulting from international specifications and customer needs are taken into account. The Objective Function that is optimized is a minimization dependent on multiple input variables. All constraints are normalized and modeled as inequalities.

Keywords— Optimization, Rectifier transformer, Design optimization methodology, Differential Evolution algorithm, Optimization methods, Wound core type rectifier transformer.

I. INTRODUCTION

When using any population based search algorithm in general and DE in particular to optimize a function, an acceptable trade-off between convergence rate and robustness must generally be determined. Convergence rate implies a fast convergence although it may be to a local optimum. On the other hand, robustness guarantees a high probability of obtaining the global optimum. Because of the software design approach and the ease of making multiple iterations of the same design layout, it is easy to optimize the rectifier transformer using a minimal set of expensive materials. The difficulty in resolving the optimum balance between the cost of rectifier transformer and its performance is becoming even more complicated nowadays, as the main materials to produce (copper or aluminum for windings and steel for magnetic circuit) are stock exchange commodities and their prices vary daily. One area of great importance that can benefit from the effectiveness of such algorithms is Rectifier equipment systems.

The work in this paper introduces the use of an evolutionary algorithm, titled Differential Evolution (DE) in conjunction with the penalty function approach to minimize the rectifier transformer cost while meeting international

standards and customer needs. A simple additive penalty function approach is used in order to convert the constrained problem into an unconstrained problem.

The method is applied to the design of a rectifier transformer and the results are compared with a heuristic transformer design optimization methodology, resulting in cost savings.

II. RELATED WORK

In this paper, a single-objective differential evolution algorithm, which combines several features (penalty function) of previous evolutionary algorithms (EAs) in a unique manner, is proposed for the design of a Rectifier transformers.

Several applications have been recently proposed in the scientific literature. In [17], [20], [21] the DE algorithm has been applied by combining two of the various possible implementing strategies for this evolutionary approach. In particular, the DE/1/best/bin version [17] is used until the cost function has reached a predefined value; successively, the DE/1/rand/bin strategy [17] is applied. It has been found that the DE/1/best/bin strategy is quite able to rapidly locate the “attraction basin” of a minimum, but, since it uses the best individual of the population to perform the mutation, it can sometimes be trapped in a local minimum. This drawback is overcoming by switching, after a predefined threshold, to the DE/1/rand/bin strategy, which is able to explore more efficiently the search space, without modifying the previous best solution if it is inside the correct attraction basin. Concerning the choice of the control parameters, F has been chosen in the range [0.5,1.0], whereas good reconstructions have been obtained with $CR = 0.8$. In [16], the DE method has been used to suppress the sideband radiation patterns in time modulated linear array antennas. The DE algorithm has been found to be a very effective tool in optimizing the static excitation amplitudes and the “switch-on” time intervals of each element. In this application, the DE algorithm has been applied to optimize 32 variables. Moreover, the authors of [16] have found the DE method to be “more powerful” than the standard GA for the present application.

III. THE DIFFERENTIAL EVOLUTION (DE) ALGORITHM

Differential Evolution (DE) algorithm was introduced by Ken Price and Rainer Storn [5], [6] as a population-based stochastic method for global optimization problems over continuous domains. Unlike simple GA that uses binary coding for representing problem parameters, Differential Evolution (DE) uses real coding of floating point numbers. Among

the DE's advantages are its simple structure, ease of use, speed and robustness. The way in which the DE is applied to these Rectifier transformers problems is schematized in Fig. 1.

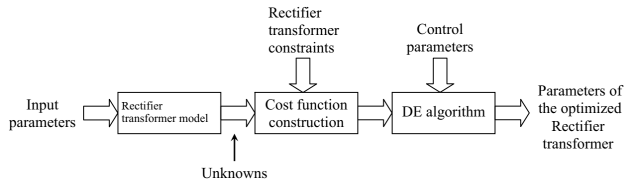


Fig.1 Schematic representation of the application of the DE algorithm to a Rectifier transformer

In all, just three factors control evolution under DE, the population size, NP ; the weight applied to the random differential, F ; and the crossover constant, CR . A notation the various DE-variants is defined by DE/x/y/z where x denotes the base vector, y denotes the number of difference vectors used, and z representing the crossover method. Price and Storn [5] gave the working principle of DE with single strategy [6]. They suggested ten different strategies for DE. The following are the ten different working strategies: 1. DE/best/1/exp, 2. DE/rand/1/exp, 3. DE/rand-to-best/1/exp, 4. DE/best/2/exp, 5. DE/rand/2/exp, 6. DE/best/1/bin, 7. DE/rand/1/bin, 8. DE/rand-to-best/1/bin, 9. DE/best/2/bin, 10. DE/rand/2/bin. However, strategy-7 (DE/rand/1/bin) appears to be the most successful and the most widely used strategy.

IV. RECTIFIER TRANSFORMER

A rectifier transformer(RT) is a transformer which includes diodes or thyristors in the same tank. Voltage regulation may also be included. Rectifier transformers are used for industrial processes which require a significant direct current (DC) supply. Typical processes would include DC traction, electrolysis, smelting operations, large variable speed drive trains, etc. The application for which the transformer is used, will drive the design considerations including: bridge type connection of the thyristors for higher voltages, interphase connection for low voltage - high current applications, number of pulses (6, 12 and higher with phase shifting), and eddy current and harmonic issues. Voltage regulation is achieved with no-load or on-load tap changers on the high voltage side. Fine levels of voltage regulation can be achieved using saturable reactors on the secondary side. Regulation units may be built in or separate.

The twelve pulses AC to DC converter are also popularly known as three-phase twelve pulse rectifier. As the number of pulses per cycle is increased, the output DC waveform gets improved. So, with twelve pulses per cycle, the quality of output voltage waveform would definitely be improved with low ripple content

One can actually increase the number of secondary windings to reduce the Total Harmonic Distortion(THD) of the Input Supply (Caused due to Rectification process), but this would increase the cost and number of pulses required in

the rectifiers. They are combined with a diode or thyristors rectifier. The comparison of different multi pulse converter has been shown in the Fig.2.

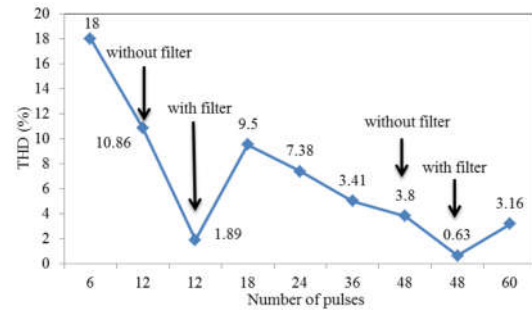


Fig.2 Performance comparison of multi-pulse converters

Regulating and rectifier transformer combinations that are applied to primary aluminum production (smelters) are commonly known as 'rectiformers'. A typical aluminum potline is built as a 60-pulse system with five parallel 12-pulse rectiformers, each with different phase-shift windings; a 60-pulse system can be achieved by the following phase shift angles: -12° , -6° , 0° , $+6^\circ$ and $+12^\circ$. As mentioned, one of the characteristics of rectiformers for aluminum plants is a very large regulating voltage range, from 0 Volts up to potentially 2,000 Volts (DC), depending on how many pots are connected in series.

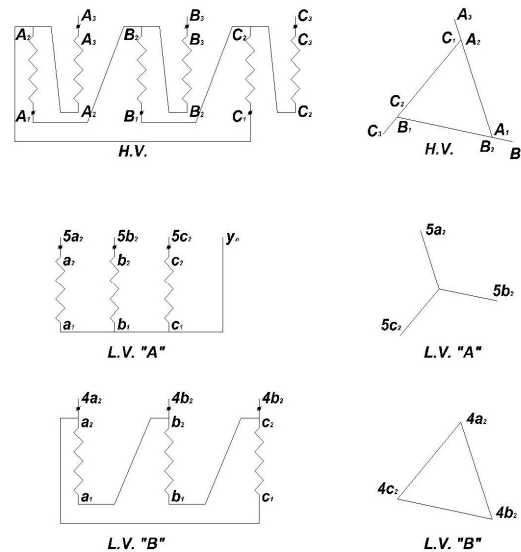


Fig.3 Vector relationship Dd -15° Dy $+15^\circ$ of a rectifier transformer under consideration(1470 kVA, 11000/(690/690) Volts)

V. MATHEMATICAL MODELLING AND OPTIMIZATION OF RECTIFIER TRANSFORMERS

A mathematical description of a global constrained minimization problem requires us to apply an appropriate model which has limited number of parameters (design

variables). In the mathematical notation consider the following optimization problem:

$$\text{Min } f(\mathbf{x}) \quad (1)$$

$$\text{s.t. } h(\mathbf{x}) = \begin{bmatrix} h_1(\mathbf{x}) \\ \vdots \\ h_M(\mathbf{x}) \end{bmatrix} = \mathbf{0} \text{ and } g(\mathbf{x}) = \begin{bmatrix} g_1(\mathbf{x}) \\ \vdots \\ g_L(\mathbf{x}) \end{bmatrix} \leq \mathbf{0} \quad (2)$$

$$\mathbf{x}^{LB} \leq \mathbf{x} \leq \mathbf{x}^{UB}$$

where $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$ is the vector of unknown quantities, $h(\mathbf{x})$ and $g(\mathbf{x})$ are the restriction constraints, which can be represented mathematically as equations and/or inequations and \mathbf{x}^{LB} and \mathbf{x}^{UB} are the lower and upper bound of the decision parameters, respectively. In order to find the global optimum design of a rectifier transformer, DE in conjunction with the penalty function approach technique is used. The goal of the proposed optimization method is to find a set of integer variables linked to a set of continuous variables that minimize the objective function (active part cost) and meet the restrictions imposed on the rectifier transformer. Under these definitions, a DE algorithm in conjunction with the penalty function approach is focused on the minimization of the cost of the rectifier transformer:

$$\min_x \sum_{j=1}^2 c_j \cdot f_j(\mathbf{x}) \quad (3)$$

where c_1 is the winding unit cost (€/kg), f_1 is the winding weight (kg), c_2 is the magnetic material unit cost (€/kg), f_2 is the magnetic material weight (kg), and \mathbf{x} is the vector of the five design variables, namely the width winding (a), the diameter of core leg (D), the winding height (b), the current density of winding (g) and the magnetic flux density (B). The minimization of the cost of the rectifier transformer is subject to the constraints:

$$S - S_N \leq 0; P_{CU} - P_{CUN} \leq 0; P_{FE} - P_{FEN} \leq 0; U_K - U_{KN} \leq 0 \quad (4)$$

where: S is designed rectifier transformer rating (kVA), S_N is rectifier transformer nominal rating (kVA), P_{FE} is designed no-load losses (W), P_{CU} is designed load losses (W), U_K is designed short-circuit impedance of a rectifier transformer (%), P_{FEN} is guaranteed no-load losses (W), P_{CUN} is guaranteed load losses (W) and U_{KN} is guaranteed short-circuit impedance (%). Accordingly, the objective function for the model is:

$$\text{Min } f(\mathbf{x}) = (41.7 \cdot x_5 + 248.5 \cdot x_3 + 3.2) \cdot 10^3 \cdot x_2^2 + 1.98 \cdot x_2^3 + (69.8 \cdot x_2 + 144.6 \cdot x_3 + 1.38) \cdot 10^4 \cdot x_3 \cdot x_5 \quad (5)$$

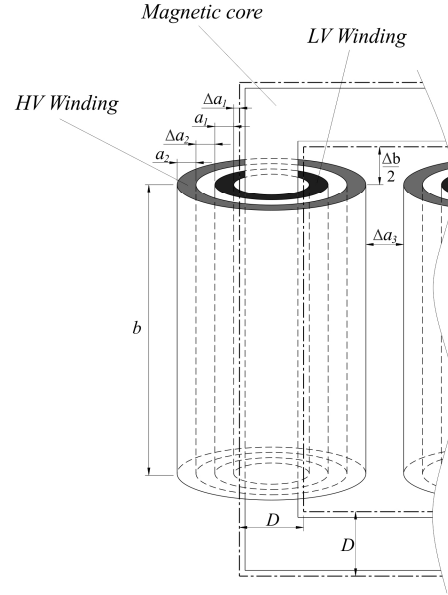


Fig.4 Active part of a rectifier transformer – main dimensions

The constraints of the analyzed mathematical model are entered as follows: Constraint 6 match to a rectifier transformer nominal rating, Constraint 7 match to guaranteed load losses, Constraint 8 match to guaranteed no-load losses and Constraint 9 guaranteed short-circuit impedance. Constants in front of decision variables have been taken from the Fig.4 and reference [9].

$$437.6 \cdot x_1 \cdot x_2^2 \cdot x_3 \cdot x_4 \cdot x_5 \cdot 10^3 - 1470 \leq 0 \quad (6)$$

$$(3.88 \cdot x_2 + 8.92 \cdot x_3 + 7.68 \cdot 10^{-2}) \cdot x_3 \cdot x_4^2 \cdot x_5 \cdot 10^{-7} - 16000 \leq 0 \quad (7)$$

$$(-0.48 \cdot x_1^2 + 1.60 \cdot x_1 - 0.06) \cdot ((41.7 \cdot x_5 + 246.5 \cdot x_3 + 3.20 \cdot 10^3) \cdot 10^3 \cdot x_2^2 + 1.93 \cdot x_2^3) \cdot 0.6 - 2400 \leq 0 \quad (8)$$

$$(85.0 \cdot x_2 + 188 \cdot x_3 + 322.8 \cdot x_3 + 17880.0 \cdot x_3^2 + 1.8) \cdot 10^{-4} \cdot 319.6 \cdot 0.019 \cdot x_3 \cdot x_4 / x_1 \cdot x_2^2 - 7.5 \leq 0 \quad (9)$$

These values are multiplied by a penalty co-efficient, which is then added to the objective function to continue the process of optimization. This process is often termed as a penalty function approach.

TABLE I THE OPTIMAL VALUE OF DECISION VARIABLES

Parameter	Value
X1	1.670110
X2	0.246260
X3	0.030270
X4	2.901650
X5	0.620400

TABLE II COMPARATIVE RESULTS OF TWO METHODOLOGIES

	B	g	D	a	b	<i>Cost of Active part</i>
DE Algorithm	1.67	2.90	246	30	620	6680
Lagrange with New.Rap.[10]	1.65	2.98	242	32	625	7410

The parameters X_1, X_2, X_3, X_4, X_5 match respectively to the magnetic flux density (B), the diameter of core leg (D), the width of secondary winding (a), the current density of secondary winding (g) and the core window height (b).

VI. CONCLUSION

In this study, DE with penalty function approach, an improved version of GA, is applied to designing of rectifier transformers. The rectifier technologies employed in industrial applications are commonly known as double star (DSS) or double bridge (DB). DSS systems use an interphase transformer and are predominately applied as 6- or 12-pulse units where high currents are required with very low nominal voltages. DB systems are applied as 6-, 12-, 24-, 48- or 60-pulse systems, as required to suit the harmonic mitigation and process stability requirements. A higher number of pulse groups can be applied but tend to be less commercially attractive

Our approach based DE with penalty function is integrates in a single unique algorithm and was tested on different devices which belong to power objects with non-rotating parts. The use of the DE computer program is applied to the analyzed mathematical model. Compared with the second methodology in the same table, the cost of materials for the active part of the reviewed object are lower by approximately 11 %.

REFERENCES

- [1] A. Vasan, "Optimization Using Differential Evolution." Water Resources Re-search Report. Book 22. Department of Civil and Environmental Engineering, The University of Western Ontario, Publication Date 7-2008.
- [2] A. Zamuda, J. Brest, B. Bošković, V. Žumer, "Differential Evolution with Self-adaptation and Local Search for Constrained Multi-Objective Optimization," IEEE Congress on Evolutionary Computation (CEC), pp. 195-202 (2009)
- [3] DE Homepage, <http://www.icsi.berkeley.edu/~storn/code.html>
- [4] Onwubolu, G. C., and Babu, B. V.: New Optimization Techniques in Engineering, Springer-Verlag, Germany (2004).
- [5] P.V. Kenneth., S.M. Rainer.: Differential evolution - A simple evolution strategy for fast optimization. Dr. Dobb's Journal, 22, 18-24 and 78. (1997).
- [6] P.V. Kenneth., S.M. Rainer., L.A. Jouni.: Differential evolution: A practical approach to global optimization. Springer-Verlag, Berlin, Heidelberg (2005).
- [7] B.V. Babu, M. Mathew, L. Jehan.: Differential Evolution for Multi-Objective Optimization. Chemical Engineering Department B.I.T.S. Pilani, India (2005).
- [8] E.I. Amoiralis, P.S. Georgilakis, M.A. Tsili.: Design optimization of distribution transformers based on mixed integer programming methodology. Technical University of Athens, Greece (2008).
- [9] R. Salkoski.: Selection of an optimal variant of 3-phase transformers with round and rectangular section of the magnetic core from aspect of minimum production costs. Master Thesis, Electrotechnical University in Skopje (2000).
- [10] Mezura and Montes.: E. Laboratorio NI Avanzada, Rébsamen 80, Centro, Xalapa, Veracruz 91090, Mexico, Velazquez-Reyes, J., Coello Coello, C.A.: Modified Differential Evolution for Constrained Optimization, pp 25 – 32, Conference Publications, Evolutionary Computation, CEC 2006 (2006).
- [11] U.K. Chakraborty (Ed.): Advances in Differential Evolution, Mathematics & Computer Science Department, University of Missouri, St. Louis, USA, Springer-Verlag Berlin Heidelberg (2008).
- [12] J. Rönkkönen, S., Kukkonen, K. V. Price.: Real-parameter optimization with differential evolution. Proc. IEEE Congr. Evolut. Comput., Sep. 2005, pp. 506–513, Edinburgh, Scotland (2005).
- [13] D. Zaharie.: Control of population diversity and adaptation in differential evolution algorithms. Proc. Mendel 9th Int. Conf. Soft Comput., R. Matousek and P. Osmera, Eds., Brno, Czech Republic, pp. 41–46, Brno, Czech Republic (2003)
- [14] U. K. Chakraborty, S. Das, A. Konar.: Differential evolution with local neighborhood. Proc. Congr. Evolut. Comput., pp. 2042-2049, Vancouver, BC, Canada (2006).
- [15] R. Salkoski, I. Chorbev.: Design optimization of distribution transformers based on Differential Evolution Algorithms, ICT Innovations2012 Web Proceedings ISSN 1857-7288, page 35-44. September 2012, Ohrid
- [16] Yang, S., Gan, Y.B., Qing, A.: Sideband suppression in time-modulated linear arrays by the differential evolution algorithm. IEEE Antennas Wireless Propagat. Lett. 1, 173-175.
- [17] M. Wolfram, A.K. Marten, D. Westermann: A comparative study of evolutionary algorithms for phase shifting transformer setting optimization, 2016 IEEE International Energy Conference (ENERGYCON), Leuven, Belgium, April 2016.
- [18] A. Lotfi, E. Rahimpour, "Optimum design of core blocks and analysing the fringing effect in shunt reactors with distributed gapped-core", ELSEVIER, Electric Power Systems Research 101(2013), pp. 63-70.
- [19] A. Lotfi, M. Faridi: "Design Optimization of Gapped-Core Shunt Reactors", IEEE Transactions on Magnetics, Vol.48, No. 4, April 2012.
- [20] S.K. Morya, H. Singh: "Reactive Power Optimization Using Differential Evolution Algorithm", IJETT-Volume 4 Issue 9, Sep. 2013.
- [21] F.S. Lobato, R. Gedraite, S. Neiro: "Solution of Flow Shop Problems using the Differential Evolution Algorithm", EngOpt 2012-3rd ICEO, Rio de Janeiro, Brazil, 01-05 July 2012.
- [22] Q. XU, L. Wang, B. HE, N. Wang: "Modified Opposition-Based Differential Evolution for Function Optimization", JCIS 7:5 (2011) 1582-1591.
- [23] M. Weber, F. Neri, V. Tirronen: "A study on scale factor in distributed differential evolution", Information Sciences 181(12) (2011) 2488-2511.
- [24] R. Salkoski: "Heuristic algorithm for multi-criteria optimization of power objects", PhD Thesis, April 4, 2018, University St. Cyril and Methodius Skopje, R.N. Macedonia

Design optimization of Earthing Transformers based on Differential Evolution Algorithms

Rasim Salkoski

Faculty of Machine Intelligence and Robotics
University for Information Science and Technology
Ohrid, R. of Macedonia
rasim.salkoski@uist.edu.mk

Ivan Chorbev

Faculty of computer science and engineering
University of Ss Cyril and Methodius
Skopje, R. of Macedonia
ivan.chorbev@finki.ukim.mk

Abstract— Differential Evolution (DE), a vector population based stochastic optimization method which is an improved version of Genetic Algorithm(GA) has been used for solving different problems with grounding of the isolated systems of distribution or interconnected networks. Earthing transformers deserve extensive treatment in the field of research and production, due to the fact that the electric energy undergoes several transformations on its way from generators to the consumers. In that regard, special interest is dedicated to the minimization of production and exploitation costs of the interconnected star earthing transformer. In this paper, an effective application of the combinatorial optimization algorithm based on Differential Evolution is proposed with the aim of minimizing the cost of the active part of wound core earthing transformers. The constraints resulting from international specifications and customer needs are taken into account. The Objective Function that is optimized is a minimization dependent on multiple input variables. All constraints are normalized and modeled as inequalities. Our approach provides very good results, which are highly competitive with those generated by the compared EAs in constrained evolutionary optimization.

Keywords— Optimization, Earthing transformer, Design optimization methodology, Differential Evolution algorithm, Optimization methods, Wound core type transformer.

I. INTRODUCTION

Optimization is a procedure of finding and comparing feasible solutions until no better solution can be found. Solutions are termed good or bad in terms of an objective, which is often the cost of fabrication, efficiency of a process, product reliability, or other. Classical optimization methods are in convenient to solve multi-objective optimization problems, as they could at best find one solution in one simulation run. However, Evolutionary algorithms (EAs) can find multiple optimal solutions in one single simulation run due to their population-based search approach.

The difficulty in resolving the optimum balance between the cost of earthing transformer and its performance is becoming even more complicated nowadays, as the main materials to produce (copper or aluminum for windings and steel for magnetic circuit) are stock exchange commodities and their prices vary daily.

The work in this paper introduces the use of an evolutionary algorithm, titled Differential Evolution (DE) in conjunction with the penalty function approach to minimize the earthing transformer cost while meeting international standards and customer needs. A simple additive penalty function approach is used in order to convert the constrained problem into an unconstrained problem. Due to this conversion, the solution falling outside the feasible region is penalized and the solving process is guided to fall into the feasible solution space after a few generations. The method is applied to the design of an earthing transformer and the results are compared with a heuristic transformer design optimization methodology, resulting in significant cost savings.

II. RELATED WORK

Price & Storn (1997) gave the working principle of DE with single strategy. Later on, they suggested ten different strategies of DE (Price & Storn, 2003). A strategy that works out to be the best for a given problem may not work well when applied for a different problem.

The key parameters of control in DE are: NP -the population size, CR -the crossover constant, and F -the weight applied to random differential (scaling factor). Babu et al. (2002) proposed a new concept called 'nested DE' to automate the choice of DE key parameters. As detailed above, the crucial idea behind DE is a scheme for generating trial parameter vectors. Basically, DE adds the weighted difference between two population vectors to a third vector. Price & Storn (2003) have given some simple rules for choosing key parameters of DE for any given application. Normally, NP should be about 5 to 10 times the dimension (number of parameters in a vector) of the problem. As for F , it lies in the range 0.4 to 1.0. Initially $F = 0.5$ can be tried then F and/or NP is increased if the population converges prematurely. A good first choice for CR is 0.1, but in general CR should be as large as possible. DE has been successfully applied in various fields. Some of the successful applications of DE include: digital filter design (Storn,1995) [17], optimal design of heat exchangers (Babu & Munawar, 2000; 2001) [8], B. V. Babu and M. Mathew Leenus Jehan in [7] have applied Differential Evolution with a Penalty Function Method and Weighting

Factor Method for finding a Pareto optimum set for the different problems. Mezura-Montes and Coello in [11] present a Differential-Evolution based approach to solve constrained optimization problems.

III. THE DIFFERENTIAL EVOLUTION (DE) ALGORITHM

The DE algorithm is a population based algorithm like genetic algorithms using the similar operators; crossover, mutation and selection. The main difference in constructing better solutions is that genetic algorithms rely on crossover while DE relies on mutation operation. This main operation is based on the differences of randomly sampled pairs of solutions in the population. The algorithm uses mutation operation as a search mechanism and selection operation to direct the search toward the prospective regions in the search space. The DE algorithm also uses a non-uniform crossover that can take child vector parameters from one parent more often than it does from others. By using the components of the existing population members to construct trial vectors, the recombination(crossover) operator efficiently shuffles information about successful combinations, enabling the search for a better solution space.

An optimization task consisting of D parameters can be represented by a D -dimensional vector. In DE, a population of NP solution vectors is randomly created at the start. This population is successfully improved by applying mutation, crossover and selection operators. The main steps of the DE algorithm are given below:

Initialization

Evaluation

Repeat

Mutation

Recombination

Evaluation

Selection

Until (*termination criteria are met*)

For each target vector a mutant vector is produced. The parent vector is mixed with the mutated vector to produce a trial vector. All solutions in the population have the same chance of being selected as parents without dependence of their fitness value. The child produced after the mutation and crossover operations is evaluated. Then, the performance of the child vector and its parent is compared and the better one is selected. If the parent is still better, it is retained in the population.

Fig. 1 shows DE's process in detail: the difference between two population members (1,2) is added to a third population member (3). The result (4) is subject to the crossover with the candidate for replacement (5) to obtain a proposal (6). The proposal is evaluated and replaces the candidate if it is found to be better.

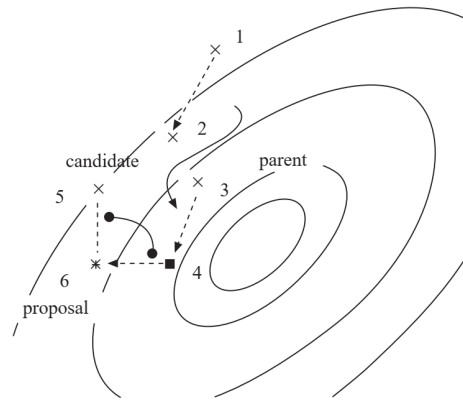


Fig.1 Obtaining a new proposal in DE

Price and Storn [5] gave the working principle of DE with single strategy [6]. They suggested ten different strategies for DE. The following are the ten different working strategies: 1. DE/best/1/exp, 2. DE/rand/1/exp, 3. DE/rand-to-best/1/exp, 4. DE/best/2/exp, 5. DE/rand/2/exp, 6. DE/best/1/bin, 7. DE/rand/1/bin, 8. DE/rand-to-best/1/bin, 9. DE/best/2/bin, 10. DE/rand/2/bin. However, strategy-7 (DE/rand/1/bin) appears to be the most successful and the most widely used strategy. In all, three factors control evolution under DE, the population size NP , the weight applied to the random differential F and the crossover constant CR .

IV. EARTHING TRANSFORMER

An Earthing or (Grounding) of power system is very important since the reliability, short circuit fault current withstand capability, over voltage and basic insulation levels, etc. depend on the characteristics of neutral grounding. The desirable quantities of earthing transformer are low zero sequence impedance and low losses (no load losses). Zero sequence impedance plays a significant role in the effectiveness of grounding, and the accurate prediction of the zero sequence impedance of earthing transformer is very important for power system designers, from a cost point of view as well as a safety point of view. The earthing transformer is usually of the wye delta or zig-zag connections, but in this paper we shall concentrate on the zig-zag connection, with the neutrals connected to earth. Fig.1, Fig.2 shows the Zig-Zag transformer connection with connection group ZNyn(d)5.

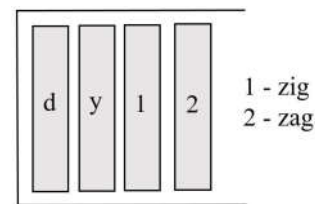


Fig.2 Earthing Transformer window – usual windings

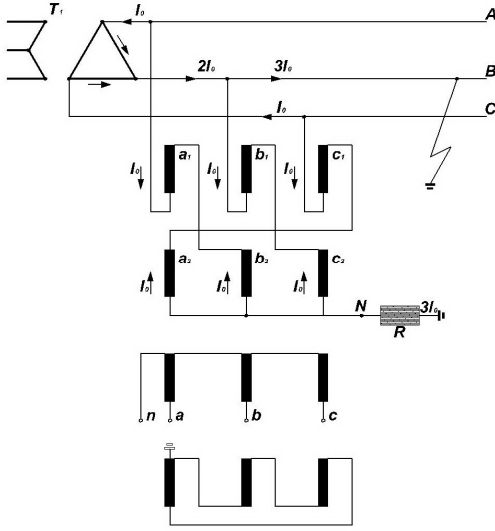


Fig.3 Earthing transformer with connection group ZNyn5, rated power on secondary winding 200 kVA(11000/400 Volts) and rated to carry 400 Amps continuous at Neutral

V. MATHEMATICAL MODELLING AND OPTIMIZATION OF EARTHING TRANSFORMERS

Constrained optimization problems, especially nonlinear optimization problems, where objective functions are minimized under given constraints, are very important and frequently appear in the real world. Let us consider nonlinear constrained optimization problems as follows:

$$\begin{aligned} & \text{minimize} && f(\mathbf{x}) \\ & \text{subject to} && g_j(\mathbf{x}) \leq 0, \quad j = 1, \dots, q \\ & && h_j(\mathbf{x}) = 0, \quad j = q + 1, \dots, m \\ & && l_i \leq x_i \leq u_i, \quad i = 1, \dots, n, \end{aligned} \quad (1)$$

where $\mathbf{x} = (x_1, x_2, \dots, x_n)$ is the n dimensional vector of unknown quantities, $f(\mathbf{x})$ is an objective function, $g_j(\mathbf{x}) \leq 0$ and $h_j(\mathbf{x}) = 0$ are q inequality constraints and $m - q$ equality constraints, respectively. Functions f , g_j and h_j are linear or nonlinear real-valued functions. Values u_i and l_i are the upper bound and the lower bound of x_i , respectively. In order to find the global optimum design of an earthing transformer, DE in conjunction with the penalty function approach technique is used. The goal of the proposed optimization method is to find a set of integer variables linked to a set of continuous variables that minimize the objective function (active part cost) and meet the restrictions imposed on the earthing transformer. Under these definitions, a DE algorithm in conjunction with the penalty function approach is followed on the minimization of the cost of the earthing transformer:

$$\min_{\mathbf{x}} \sum_{j=1}^2 c_j \cdot f_j(\mathbf{x}) \quad (2)$$

where c_1 is the winding unit cost (€/kg), f_1 is the winding weight (kg), c_2 is the magnetic material unit cost (€/kg), f_2 is the magnetic material weight (kg), and \mathbf{x} is the vector of the five design variables, namely the width winding (a), the diameter of core leg (D), the winding height (b), the current density of winding (g) and the magnetic flux density (B). The minimization of the cost of the earthing transformer is subject to the constraints:

$$S - S_N \leq 0; P_{CU} - P_{CUN} \leq 0; P_{FE} - P_{FEN} \leq 0; Z_0 - Z_{0N} \leq 0 \quad (3)$$

where: S is designed earthing transformer rating (kVA), S_N is earthing transformer nominal rating (kVA), P_{FE} is designed no-load losses (W), P_{CU} is designed load losses (W), Z_0 is designed zero sequence impedance of an earthing transformer (ohms/phase), P_{FEN} is guaranteed no-load losses (W), P_{CUN} is guaranteed load losses (W) and Z_{0N} is guaranteed zero sequence impedance (ohms/phase). Accordingly, the objective function for the model is:

$$\begin{aligned} f(x_2, x_3, x_5) = & (41.7 \cdot x_5 + 248.5 \cdot x_3 + 3.2) \cdot 10^3 \cdot x_2^2 + \\ & 1.98 \cdot x_2^3 + (69.8 \cdot x_2 + 144.6 \cdot x_3 + 1.38) \cdot 10^4 \cdot x_3 \cdot x_5 \end{aligned} \quad (4)$$

The constraints of the analyzed mathematical model are entered as follows: Constraint 5 match to an earthing transformer nominal rating, Constraint 6 match to guaranteed load losses, Constraint 7 match to guaranteed no-load losses and Constraint 8 guaranteed zero sequence impedance. Constants in front of decision variables have been taken from the reference [9].

$$402.4 \cdot x_1 \cdot x_2^2 \cdot x_3 \cdot x_4 \cdot x_5 \cdot 10^3 - 1682 \leq 0 \quad (5)$$

$$(4.34 \cdot x_2 + 7.71 \cdot x_3 + 6.23 \cdot 10^{-2}) \cdot x_5 \cdot x_4^2 \cdot x_5 \cdot 10^{-7} - 24500 \leq 0 \quad (6)$$

$$(-0.42 \cdot x_1^2 + 1.22 \cdot x_1 - 0.049) \cdot \quad (7)$$

$$((40.2 \cdot x_5 + 225.3 \cdot x_3 + 3.12 \cdot 10^3) \cdot 10^3 \cdot x_2^2 + 2.13 \cdot x_2^3) \cdot 0.57 - 1350 \leq 0$$

$$(72.3 \cdot x_2 + 172 \cdot x_2 \cdot x_3 + 294.3 \cdot x_3 + 18235.0 \cdot x_3^2 + 1.6) \cdot 10^{-4} \cdot 311.3 \cdot \quad (8)$$

$$0.016 \cdot x_3 \cdot x_4 / x_1 \cdot x_2^2 - 4.7 \leq 0$$

These values are multiplied by a penalty co-efficient, which is then added to the objective function to continue the process of optimization. This process is often termed as a penalty function approach.

TABLE I THE OPTIMAL VALUE OF DECISION VARIABLES

Parameter	Value
X1	1.510010
X2	0.231260
X3	0.056270
X4	2.801650
X5	0.420400

TABLE II COMPARATIVE RESULTS OF TWO METHODOLOGIES

	<i>B</i>	<i>g</i>	<i>D</i>	<i>a</i>	<i>b</i>	<i>Cost of Active part</i>
DE Algorithm	1.51	2.80	231	56	420	9170
Lagrange with New.Rap.[10]	1.50	2.88	226	58	435	9980

The parameters X_1 , X_2 , X_3 , X_4 , X_5 match respectively to the magnetic flux density (B), the diameter of core leg (D), the width of secondary winding (a), the current density of secondary winding (g) and the core window height (b).

VI. CONCLUSION

In this paper, DE with penalty function approach is applied to designing of earthing transformers.

The rating of earthing transformer is entirely different from that of a power transformer. Power transformers are designed to carry total load continuously, whilst grounding transformer carries no load, and supplies current only if one of the lines becomes grounded. Since it is almost working on no-load, dictates to have low iron losses. The kVA rating of a three phase earthing transformer is the product of normal line to neutral voltage (kV) and the neutral or ground amperes that the transformer is designed to carry current under fault conditions for a specified time. Most earthing transformers are designed to carry their ground current for a limited time only, such as 10seconds to 1 minute.

Moreover, this approach is easy to implement and its computational cost is relatively low. The use of the DE computer program is applied to the analyzed mathematical model. In the first methodology, the single objective DE optimization showed that single optimum could be obtained quickly, even when constraints in the penalty function method are complex. Compared with the second methodology in the same table, the cost of materials for the active part of the reviewed object are lower by approximately 8.8 %.

REFERENCES

- [1] A. Vasan, "Optimization Using Differential Evolution." Water Resources Re-search Report. Book 22. Department of Civil and Environmental Engineering, The University of Western Ontario, Publication Date 7-2008.
- [2] A. Zamuda, J. Brest, B. Bošković, V. Žumer, "Differential Evolution with Self-adaptation and Local Search for Constrained Multi-Objective Optimization," IEEE Congress on Evolutionary Computation (CEC), pp. 195-202 (2009)
- [3] DE Homepage, <http://www.icsi.berkeley.edu/~storn/code.html>
- [4] Onwubolu, G. C., and Babu, B. V.: New Optimization Techniques in Engineering, Springer-Verlag, Germany (2004).
- [5] P.V. Kenneth., S.M. Rainer.: Differential evolution - A simple evolution strategy for fast optimization. Dr. Dobb's Journal, 22, 18-24 and 78. (1997).
- [6] P.V. Kenneth., S.M. Rainer., L.A. Jouni.: Differential evolution: A practical approach to global optimization. Springer-Verlag, Berlin, Heidelberg (2005).
- [7] B.V. Babu, M. Mathew, L. Jehan.: Differential Evolution for Multi-Objective Optimization. Chemical Engineering Department B.I.T.S. Pilani, India (2005).
- [8] B.V. Babu, Munawar A. Shaik Differential Evolution Strategies for Optimal Design of Shell-and-Tube Heat Exchangers, July 2007 Chemical Engineering Science 62(14):3720-3739.
- [9] K. Zielinski, R. Laur: Constrained Single-Objective Optimization Using Differential Evolution, Evolutionary Computation, 2006. CEC 2006. IEEE Congress. January 2006.
- [10] R. Salkoski.: Selection of an optimal variant of 3-phase transformers with round and rectangular section of the magnetic core from aspect of minimum production costs. Master Thesis, Electrotechnical University in Skopje (2000).
- [11] Mezura and Montes.: E. Laboratorio NI Avanzada, Rébsamen 80, Centro, Xalapa, Veracruz 91090, Mexico, Velazquez-Reyes, J., Coello Coello, C.A.: Modified Differential Evolution for Constrained Optimization, pp 25 – 32, Conference Publications, Evolutionary Computation, CEC 2006.
- [12] U.K. Chakraborty (Ed.): Advances in Differential Evolution, Mathematics & Computer Science Department, University of Missouri, St. Louis, USA, Springer-Verlag Berlin Heidelberg (2008).
- [13] J. Rönkkönen, S., Kukkonen, K. V. Price.: Real-parameter optimization with differential evolution. Proc. IEEE Congr. Evolut. Comput., Sep. 2005, pp. 506–513, Edinburgh, Scotland (2005).
- [14] D. Zaharie.: Control of population diversity and adaptation in differential evolution algorithms. Proc. Mendel 9th Int. Conf. Soft Comput., R. Matousek and P. Osmera, Eds., Brno, Czech Republic, pp. 41–46, Brno, Czech Republic (2003)
- [15] U. K. Chakraborty, S. Das, A. Konar.: Differential evolution with local neighborhood. Proc. Congr. Evolut. Comput., pp. 2042-2049, Vancouver, BC, Canada (2006).
- [16] R. Salkoski, I. Chorbev.: Design optimization of distribution transformers based on Differential Evolution Algorithms, ICT Innovations2012 Web Proceedings ISSN 1857-7288, page 35-44. September 2012, Ohrid
- [17] R. Storn: Differential Evolution Design of an IIR-Filter with Requirements for Magnitude and Group Delay International Computer Science Institute, TR-95-026, June 1995.
- [18] A. Lotfi, E. Rahimpour, "Optimum design of core blocks and analysing the fringing effect in shunt reactors with distributed gapped-core", ELSEVIER, Electric Power Systems Research 101(2013), pp. 63-70.
- [19] K.R. Hameed: " Zig-Zag Grounding transformer modeling for zero sequence impedance calculation using finite element method", ISSN 1999-8716, Diyala Journal of Engineering Sciences, Vol.08, No 3, pp.63-87, September 2015.
- [20] A. Lotfi, M. Faridi: " Design Optimization of Gapped-Core Shunt Reactors", IEEE Transactions on Magnetics, Vol.,48, No. 4, April 2012.
- [21] S.K. Morya, H. Singh: " Reactive Power Optimization Using Differential Evolution Algorithm", IJETT-Volume 4 Issue 9, Sep. 2013.
- [22] F.S. Lobato, R. Gedraite, S. Neiro: " Solution of Flow Shop Problems using the Differential Evolution Algorithm", EngOpt 2012-3rd ICEO, Rio de Janeiro, Brazil, 01-05 July 2012.
- [23] C.Qiu, M. Liu, W. Gong: "Differential Evolution with Tournament-based Mutation Operators", IJCSI Vol.10 (2), No.1, March 2013.
- [24] Q. XU, L. Wang, B. HE, N. Wang: "Modified Opposition-Based Differential Evolution for Function Optimization", JCIS 7:5 (2011) 1582-1591.
- [25] M. Weber, F. Neri, V. Tirronen: " A study on scale factor in distributed differential evolution", Information Sciences 181(12) (2011) 2488-2511.
- [26] R. Salkoski: " Heuristic algorithm for multi-criteria optimization of power objects", PhD Thesis, April 4,2018, University St. Cyril and Methodius Skopje, R.N. Macedonia

Framework for Efficient Resource Planning in Pandemic Crisis

Nenad Petrovic
Faculty of Electronic Engineering
University of Nis
Nis, Serbia
nenad.petrovic@elfak.ni.ac.rs

Djordje Kocic
Faculty of Electronic Engineering,
University of Nis
Nis, Serbia
seriousdjoka@gmail.com

Abstract—Pandemics have dramatic consequences, both taking human lives and ruining economy leading towards crisis worldwide. It has also been the case with COVID-19 pandemic since the beginning of this year. In this paper, we present a framework aiming efficient resource planning during the pandemic crisis, making use of modelling, simulation, predictions based on deep learning, linear optimization and blockchain. As a case study, we target the current COVID-19 pandemic. According to the achieved results, the proposed framework has not only huge potential in cost reduction, but also enables the proactive approach to tackle the pandemic which can save many lives as well.

Keywords—blockchain, deep learning, coronavirus, COVID-19, linear optimization, modelling, simulation

I. INTRODUCTION

In the first days of this year, a new infectious flu-alike disease *COVID-19* was discovered in China, after several strange pneumonia cases in Wuhan's Seafood Wholesale Market [1]. Behind this disease is *SARS-Cov-2* virus, also referred to as *2019-nCoV* or popularly *coronavirus*. However, its distinctive feature compared to other influenza viruses is actually the fact that even asymptomatic people might be potential sources of infection, which first caused a dramatic outbreak in China [2]. Moreover, long and varying incubation periods (normally from 2 to 14 days, in extreme cases up to 27 days) [3] and high death rate among elderly and people with chronic diseases makes the situation even more difficult to control and handle [1, 2].

After the outbreak in China, in just few weeks, the first cases were reported in other continents around the world, as well. In Europe, the first cases were recorded during the last week of January [4]. Later, towards the middle of February, the number of infected people in northern Italy has started to rise dramatically making it one of the world's worst-affected countries quite soon. The new disease quickly turned from Chinese outbreak to worldwide pandemic, having disastrous consequences - huge number of lost lives, but also catastrophic financial losses and stagnation of economy as well [2]. Towards the middle of March, COVID-19 hit most parts of Europe which led to different region-level government responses, such as limiting the citizens' movement, city lockdowns, closing country borders and social distancing [5]. As for now, there is no specific treatment for COVID-19 proven by clinical trials [6]. For all these reasons, the COVID-19 pandemic crisis was inevitable.

In pandemic, it is crucial to plan resources efficiently and timely, both human and material [7, 8, 9]. Moreover, the protection of health and the economy of a country are tightly

connected [10]. On the other side, it was shown that simulation approach has many benefits in pandemic situations, leading to reduction of pandemic's consequences, especially at provincial and local level [9]. Therefore, in this paper, a software framework utilizing simulation is proposed targeting the efficient resource planning in context of current COVID-19 crisis.

In [8], simulation was used to tackle the past bird influenza pandemic. Recently, a simulation-based approach to prediction of COVID-19 spread in Iran was presented [11]. However, the work presented in our paper builds upon the approach approved in [12], combining energy consumption prediction and linear optimization for optimal blockchain-based energy trading within smart grids.

II. BACKGROUND AND RELATED WORKS

A. Blockchain Technology

The initial purpose of Bitcoin, the pioneer blockchain technology that emerged in 2009, was enabling the transfer of financial assets worldwide without involving intermediaries and transfer costs. Blockchain refers to a data structure (also called *ledger*) which represents an append-only sequence of blocks that hold the information about the executed transactions [13]. Moreover, the same term is also used for a distributed system that stores copies of the previously mentioned data structure within the peer-to-peer network of interconnected nodes. In this network, each node contains alphanumeric address, while both anonymity and transaction record transparency are kept at the same time. Apart from that, each block also includes a cryptographic hash of the previous block and timestamp to ensure that no modification or deletion is possible, once they are recorded within the ledger. The blockchain-enabled transaction represents the transfer of value and ownership of digital tokens between sender and receiver that is appended to the distributed ledger [13, 14]. Token can represent either tangible or intangible goods/assets. On the other side, *smart contract* refers to a protocol which has purpose to digitally facilitate, verify, or enforce the execution of a particular contract [14]. In blockchain technology, it refers to a software code that defines and executes transactions on the targeted platform, while the performed transactions are trackable, irreversible and do not involve third part.

In this paper, we consider the usage of Ethereum¹ blockchain in synergy with smart contracts that are written using a high-level object-oriented language Solidity². The output of linear optimization process is leveraged to generate smart contracts for resource exchange between cities,

¹ <https://ethereum.org/>

² <https://solidity.readthedocs.io/en/v0.6.4/>

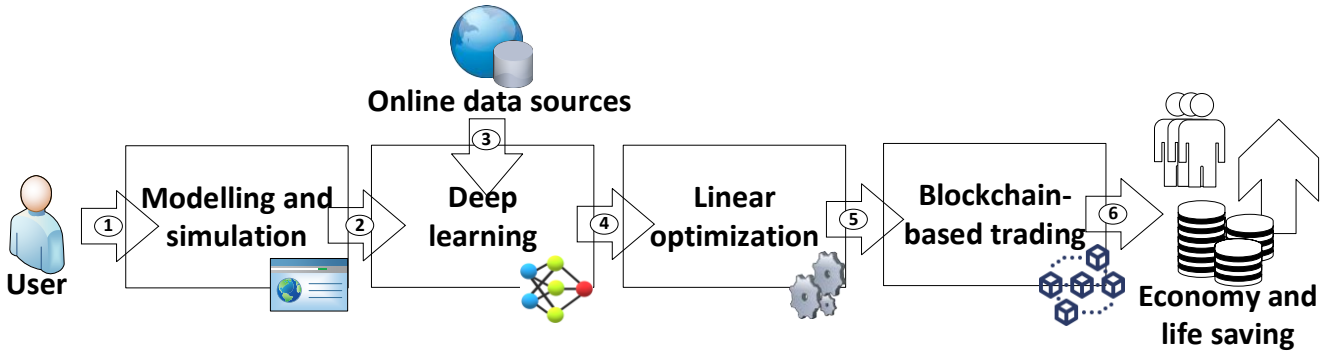


Fig. 1. Framework overview: 1-Model definition and simulation execution 2-City resource model 3-Aggregated and filtered public online data about new cases 4-Predictions based on past data 5-AMPL data file 6-Optimization output 7-Smart contract-based blockchain transactions

inspired by frameworks for energy trading relying on blockchain [12] and smart contract-driven music asset exchange approach from [15].

B. Deep Learning

Deep learning covers a family of various machine learning techniques and algorithms based on artificial neural networks. *Artificial Neural Network* (ANN) represents a collection of computational units interconnected via weighted links. These units are called *neurons* or *perceptrons*. Each of them has one or more weighted input connections, a transfer function that combines the received inputs and an output connection to forward the result to connected units [16]. Three types of layers are identified in neural networks [16, 17]: 1) *input layer* - corresponds to the input variables 2) *hidden layer* - placed between input and output layers 3) *output layer* - generates the output variables. A *deep neural network* (DNN) is an artificial neural network (ANN) with multiple layers between the input and output layers [17].

For example, in [18], Markov model was used to predict the spread of COVID-19 in Germany. On the other side, in [19], an approach based on neural network was used for country-wise COVID-19 risk predictions, showing promising results. In our paper, deep learning approach is used for prediction of new cases and deaths on regional- or city- level, necessary for resource demand estimation used crucial parameters in the proposed linear optimization models. The prediction model is trained on publicly available online data about COVID-19 cases and deaths from previous periods on daily basis. TensorFlow³ library for Python programming language was used for implementation of the prediction module. One of its advantages is the support for faster GPU-powered execution on CUDA-enabled hardware, which was used for energy consumption prediction in [15].

C. Linear Optimization

Linear optimization (also called *linear programming*) is a method which has a goal to achieve the best possible outcome (such as maximum profit or lowest possible cost), relying on a mathematical model, where the requirements are represented by linear relationships [20]. It refers to techniques for optimization of a linear *objective function*, subject to linear equality and inequality *constraints*. The vector of variables which are determined as a result of optimization process is called *decision variable*.

In this paper, linear programming is adopted to enable optimal resource planning and exchange between cities during COVID-19 pandemic crisis. The implementation is based on AMPL⁴, an algebraic modeling language for mathematical programming. It enables writing linear programs using the expressions similar to traditional algebraic notation. However, AMPL is not responsible for solving optimization problems, but it rather provides interface to other programs responsible for that. In this paper, CPLEX⁵ was used as a solver of optimization problems, while its implementation is based on simplex method [21].

III. IMPLEMENTATION OVERVIEW

A. Framework Architecture

In Fig. 1, an illustration of the proposed framework is shown.

First, user has to define models of cities and set parameters about the available resources for each city relying on a visual modelling environment tool run in web browser. A domain-specific notation is used for that purpose, while the modelling tool implementation is based on JavaScript, HTML and CSS on frontend and Node.js on backend.

Once the modelling is done, the simulation can be run within the tool. For that purpose, the city models are augmented with the data about expected resource demands based on the results of deep learning-based predictions. In order to enable making predictions, public data about number of new cases and deaths on daily basis from various online is collected, aggregated, filtered and prepared.

Furthermore, AMPL data file containing the city resource model augmented with predictions is transferred to linear optimization process with respect to pre-defined AMPL optimization models. Once the process of optimization is completed, its output which represents optimal resource exchange matrix is used for generation of Ethereum blockchain Solidity smart contracts. The aim of such contracts is to execute the proactive resource exchange transactions that would result saving both the city's economy and citizens' lives.

B. Modelling Notation

Domain-specific language is a programming language or notation specialized for solving problems from some domain of interest. In this paper, a domain-specific notation is used

³ <https://www.tensorflow.org/>

⁴ <https://ampl.com/>

⁵ <https://www.ibm.com/analytics/cplex-optimizer>

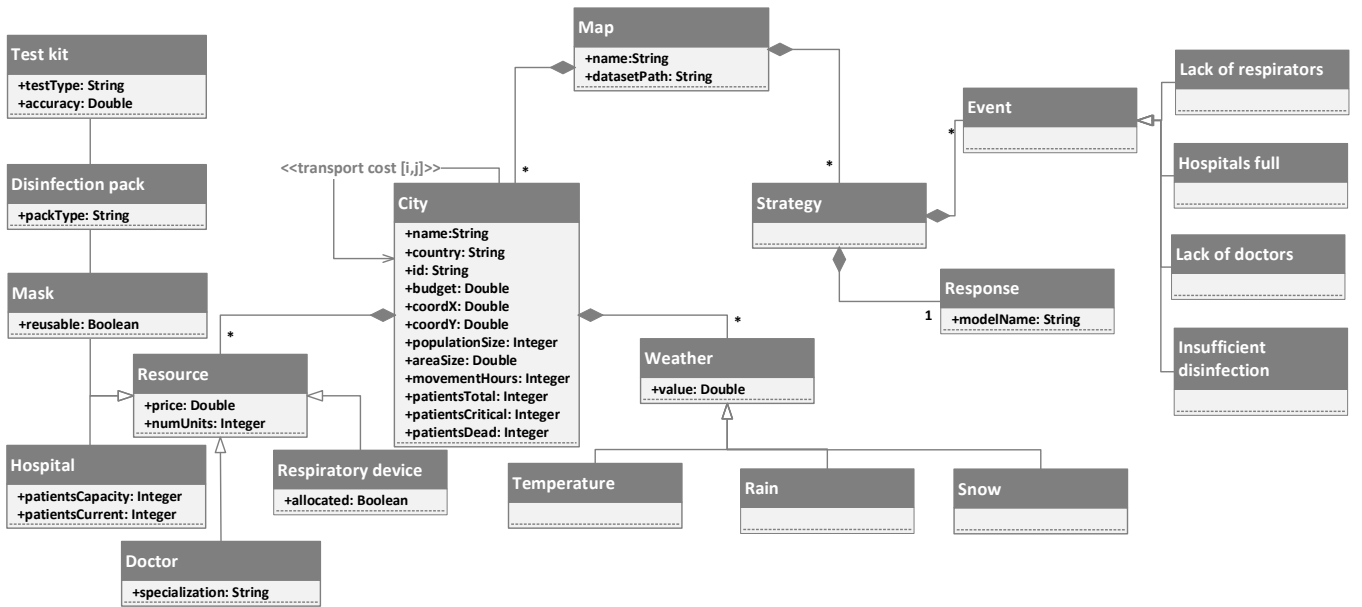


Fig. 2. City resource modelling notation metamodel represented as UML class diagram

for city resource modelling. In Fig. 2, a UML class diagram of the metamodel used for that purpose is shown.

The highest-level concept in modelling notation is *map*. It consists of a set of interconnected *cities* able to exchange resources during the pandemic crisis. The parameter describing the city connections is *transport cost*. In context of resource allocation, relevant parameters for cities are its area size, the total number of citizens, number of hours allowed for movement, total budget. Moreover, relevant parameters for pandemic situation are the total number of cases, number of critical cases and patients that died. There are several types of resources considered: *hospital buildings*, *doctors*, *disinfection products*, *respiratory devices*, *masks* and *test kits*. A common feature of all resources are their price and number of units available. Each hospital has a limited capacity of patients. At each moment, some of the places within the hospital are occupied. For COVID-19, respiratory equipment is crucial for saving lives in critical cases, therefore it is highlighted as one of the most important resources that are exchanged between the cities, apart from disinfection products and masks which can lead to reduction of infection spread [22]. Furthermore, it is possible to set the value of weather parameters, such as temperature, rain and

snow, which can affect the people movement and, indirectly, the disease spread.

The implementation of modelling and simulation environment is built upon [12] and [23]. The tool offers two views: (a) map view and (b) city view. Map view gives a global illustration of the affected cities, while city view is available for each city from the map by double-clicking on it and is used to set resource parameters. In Fig. 3, a screenshot of the described tool is shown.

C. Simulation Mechanisms

It is possible to check whether the number of respiratory devices is sufficient by checking if it is greater or equal to the number of critical patients:

$$n_critical_patients \leq n_respiratory_devices \quad (1)$$

Moreover, it can be checked whether the number of doctors available can handle the number of expected new cases based on estimation of how many patients each of them can treat:

$$\frac{n_expected_cases}{n_patients_daily_per_doctor} \leq n_doctors \quad (2)$$

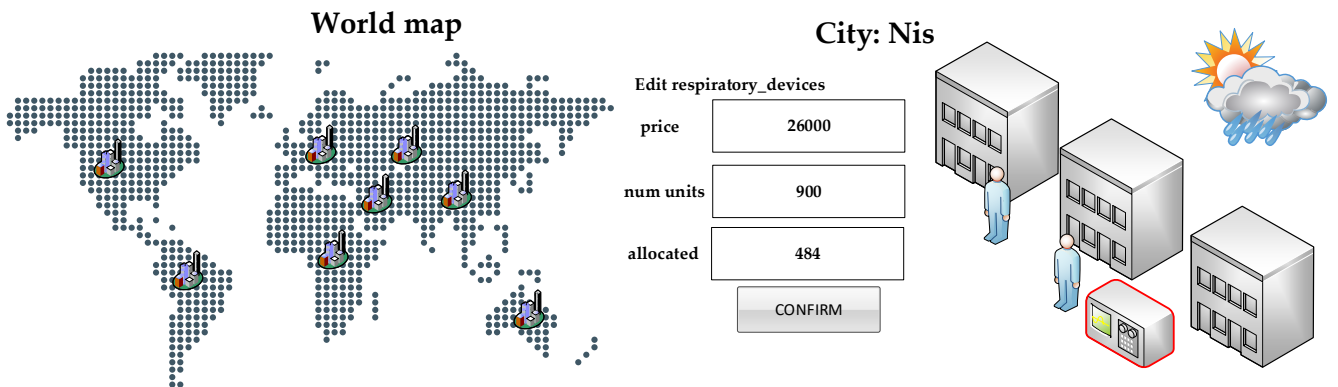


Fig. 3. A screenshot of modelling and simulation tool: (a) map view (b) city view

It is also checked if there are enough free places in city hospitals to allocate the patients:

$$n_expected_cases \leq \sum_{i=1}^{n_h} hospital[i].free_places \quad (3)$$

Furthermore, it is checked if the number of available test kits is above a given threshold that is 10 times the predicted number of cases:

$$0.1 \cdot n_expected_cases \leq n_test_kits \quad (4)$$

Finally, it is checked if the number of masks is at least 10 times greater than the city population size:

$$10 \cdot population_size \leq \sum_{j=1}^{n_p} pharmacy[j]n_masks \quad (5)$$

In case that one of the previous conditions is false, then the specific event is triggered in the simulation environment. After that, a user-defined resource exchange strategy based on optimization will be executed as response to that event:

$$if(event_i) \text{ then } optimization(response_i, modelName) \quad (6)$$

D. Deep Learning-Based Spread Prediction

The aggregated online data is split into two disjoint sets – the training set (70%) and test set (30%). It consists of total eight variables, as it is shown in Table I.

TABLE I. DISEASE SPREAD PREDICTION DATASET

<i>cid</i>	<i>Density</i> [p/km ²]	<i>m_age</i> [y]	<i>Day</i> [0..366]	<i>T</i> [°C]	<i>mh</i> [h]	<i>New</i> [p]	<i>D</i> [p]
------------	----------------------------------------	---------------------	------------------------	------------------	------------------	-------------------	-----------------

Cid is the identification number of the considered city. *Density* is number of persons per km² provides the information about population density within the area. *M_age* is the age that divides a population into two numerically equally sized sets, which might be of huge importance for prediction of death cases. *Day* is the ordinal number of a day in a year. *T* is the average daily temperature during the considered day which can have impact on the movement of citizens and disease spread, as proposed in [19]. Instead of temperature, other weather parameters might be leveraged, such as rainfall or snow. *Mh* represents the number of hours during the day when the citizens are allowed to go outside and may vary across countries and cities due to different government decisions. It is in range from 0 (total quarantine) to 24 hours. *New* refer to the number of new infected people identified within the area, as an average of 7 days which is approximately the average length of the incubation period [3]. *D* represents the number of people that died as a consequence of COVID-19 disease during the considered day. City id, density, age, day number, temperature and movement hours are treated as independent variables, while the number of new cases and deaths are dependent variables (predicted values). However, the information about population density, median age and movement hours are optional, but the predictions could also be made without them. The abbreviations of the measurement units have the following meaning in Table I: *p* - people, *y* - years, *p/km²* - people per square kilometer, *h* - hours.

The data from public domain dataset about the new cases and deaths on national level worldwide was used from [24], together with national-level detailed data about COVID-19 in Serbia from [25].

E. Optimization Model

Several similar optimization models for different resource exchange strategies are implemented within the tool (respirators, test kits, masks, medical personnel, disinfection, and patients exchange), but one of them will be presented.

Let us consider a network of n_C interconnected cities involved into resource trading, denoted as C . To each link between c travel cost $transport_cost[i,j]$ is assigned. Each $city[i]$ has predicted number of new cases $new_patients[i]$ and available test kits $tests[i]$. Each $city[i]$ has test kits of price $test_price[i]$. Furthermore, each $city[i]$ has amount of budget $budget[i]$ that is the maximum amount of money which can be spent.

To each connection between cities, a decision variable $x[i,j]$ is assigned to indicate the amount of test kits that will be sent from $city[j]$ to $city[i]$:

$$x[i,j] \geq 0 \text{ if trading between } city_i \text{ and } city_j \text{ will be performed,} \quad (7)$$

$$x[i,j] = 0 \text{ otherwise}$$

Additionally, there are several resource exchange constraints.

First, the overall amount of available test kits after exchange is always equal or greater than the number of predicted cases multiplied by 10 in each of the considered cities:

$$tests[i] + \sum_{j \in C} x[i,j] - x[j,i] \geq 10 \cdot new_patients[i], i \in C \quad (8)$$

Moreover, the total cost of test kit acquisition should not exceed the available budget of each city:

$$budget[i] \geq \sum_{j \in C} transport_cost[j,i] + x[i,j]test_price[j], i \in C \quad (9)$$

Finally, the objective function minimizes the overall sum of costs, considering both the travel cost and test kit:

$$minimize \sum_{i,j \in C} transport_cost[j,i] + x[i,j]test_price[j] \quad (10)$$

In Listing I, the AMPL code of the proposed optimization model for test kit exchange is given.

LISTING I. AMPL CODE FOR TEST KIT EXCHANGE MODEL

```

param nC;
set C:=1..nC;
param transport_cost {i in C, j in C};
param new_patients {i in C};
param tests {i in C};
param test_price {i in C};
param budget {i in C};

# variable declaration
var x {i in C, j in C} >= 0;

# objective function
minimize cost:
    sum{i in C, j in C} (transport_cost[j,i]+x[i,j]*test_price[j]);

subject to cover_estimated {i in C}:
    tests[i]+sum{j in C} (x[i,j]-x[j,i])>=10*new_patients[i];

subject to budget_threshold {i in C}:
    budget[i]>=sum{j in C} (transport_cost[j,i]+x[i,j]*test_price[j]);

```

F. Smart contract generation

The role of smart contract generation algorithm is to parametrize the smart contract templates for resource exchange. These contracts enable the execution of

transactions according to the optimal resource allocation that was obtained as output of corresponding optimization processes. In Listing II, the smart contract generation is given in a form of pseudocode.

First, the optimization problem for the resource allocation strategy is solved for the corresponding model using as a response to event that occurs in simulation environment. The resource exchange matrix is created as a result of optimization process. After that, the matrix is traversed. For each resource exchange between $city[j]$ to $city[i]$, the necessary parameters are retrieved and inserted to a smart contract template (similar to templates used in [12] and [15]): buyer and seller id, amount of resources that will be transferred, resource cost and transport cost. The amount of tokens is calculated as:

$$price = num_resources[i, j] \cdot resource_price[j] + transport_cost[j, i] \quad (11)$$

LISTING II. PSEUDOCODE OF ALGORITHM FOR SMART CONTRACT GENERATION LEVERAGING OPTIMIZATION PROCESS OUTPUT

Input: event in simulation environment
Output: A set of smart contracts
Steps:

1. Perform optimization according to model response.modelName
2. Store the resource exchange matrix obtained as result
3. For i=1 to matrix.n
4. For j=1 to matrix.m
5. If(matrix[i,j]>0) then
6. Retrieve city[i] and city[j] identifiers
7. Retrieve resource cost offered by city[j]
8. Retrieve transport cost from city[j] to city[i]
9. Price:=transport_cost[j,i]+matrix[i,j]*resource_price[j]
10. Parametrize_contract(city[i].id, city[j].id, price)
11. Else
12. do nothing
13. end for each
14. end

IV. EVALUATION AND RESULTS

A laptop equipped with Intel i7 7700-HQ quad-core CPU running at 2.80GHz, GTX1050 GPU with 2GB VRAM, 16GB of DDR4 RAM and 1TB HDD was used for emulation of the proposed implementation. In Table II, the achieved results in test kit exchange scenario for different number of cities involved are given, considering the execution time for different processing steps (in seconds) and overall cost reduction (as percentage).

According to the results, the optimization processing time increases with the model size, but it does not exceed one second in the executed experiments. On the other side, the reduction of resource exchange costs and smart contract generation time depend from the specific problem instance itself and vary from case to case.

The relative error of predictions was around 30%, which is worse than the energy consumption prediction using a similar approach in [12]. This can be explained by the fact that amount of data about new COVID-19 cases in Serbia and worldwide was not so detailed during the process of writing this paper. Moreover, the cost reduction in all cases was also lower, leading to the conclusion that more accurate prediction could also impose greater cost reduction. On the other side, the contract generation time is faster than [12], as it does not rely on semantic triple store which involves many queries for retrieval of the desired information.

TABLE II. EVALUATION RESULTS

Size [nC]	Prediction [s]	Optimization [s]	Contract gen. [s]	Cost red. [%]
4	0.31	0.019	1.31	27
6	0.33	0.028	3.34	41
8	0.49	0.081	1.12	33

V. CONCLUSION AND FUTURE WORK

The proposed approach seems promising, providing the ability to timely allocate resources in a pandemic crisis situation and respond proactively. However, the proposed framework is still a work in progress and more data about new COVID-19 cases in Serbia and other countries will be used in future for more accurate predictions, potentially leading to greater cost reduction when it comes to resource planning during pandemic.

ACKNOWLEDGMENT

This work has been supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia.

REFERENCES

- [1] "Cluster of pneumonia cases caused by a novel coronavirus, Wuhan, China" [online], European Centre for Disease Prevention and Control, Stockholm, pp. 1-10, 2020. <https://www.ecdc.europa.eu/sites/default/files/documents/Risk%20assessment%20-%20pneumonia%20Wuhan%20China%2017%20Jan%202020.pdf>
- [2] B. Cruz, M. Dias, "COVID-19: From outbreak to pandemic", GSJ Vol. 8 Issue 3, March 2020, pp. 2230-2238.
- [3] Coronavirus Incubation Period [online]. Available on: <https://www.worldometers.info/coronavirus/coronavirus-incubation-period/>
- [4] G. Spiteri et al., "First cases of coronavirus disease 2019 (COVID-19) in the WHO European Region, 24 January to 21 February 2020", Eurosurveillance, pp. 1-6, 2020. <https://doi.org/10.2807/1560-7917.ES.2020.25.9.2000178>
- [5] Considerations relating to social distancing measures in response to COVID-19 – second update [online]. Available: <https://www.ecdc.europa.eu/sites/default/files/documents/covid19-social-distancing-measuresg-guide-second-update.pdf>
- [6] Y. Song et al., "COVID-19 Treatment: Close to a Cure? – A Rapid Review of Pharmacotherapies for the Novel Coronavirus" [preprint], pp. 1-25, 2020. <https://doi.org/10.20944/preprints202003.0378.v1>
- [7] H. Eriksson et al., "A Cloud-Based Execution Environment for a Pandemic Simulator", AMIA Annual Symposium 2011, pp. 364–373, 2011
- [8] H. Deguchi et al., "Anti Pandemic Simulation by SOARS", 2006 SICE-ICASE International Joint Conference, pp. 4581-4586, 2006. <https://doi.org/10.1109/sice.2006.315130>
- [9] D. M. Prieto et al., "A systematic review to identify areas of enhancements of pandemic simulation models for operational use at provincial and local levels", BMC Public Health, 12(1), pp. 1-13, 2012. <https://doi.org/10.1186/1471-2458-12-251>
- [10] P. Vanini, "Protection of the Population and the Economy in a Pandemic" [preprint], pp. 1-21, 2020.
- [11] N. Ghaffarzadegan, H. Rahmandad, "Simulation-based Estimation of the Spread of COVID-19 in Iran", pp. 1-19, 2020. <https://doi.org/10.1101/2020.03.22.20040956>
- [12] N. Petrović, Đ. Kocić, "Data-driven Framework for Energy-Efficient Smart Cities", Serbian Journal of Electrical Engineering, Vol. 17, No. 1, Feb. 2020, pp. 41-63. <https://doi.org/10.2298/SJEE2001041P>
- [13] N. Balani and R. Hathi, Enterprise Blockchain: A Definitive Handbook, 2017.
- [14] A. Narayanan and J. Clark, "Bitcoin's academic pedigree", Communications of the ACM, 60(12), pp. 36–45, 2017.

- [15] N. Petrovic, "Adopting Semantic-Driven Blockchain Technology to Support Newcomers in Music Industry", CIIT 2019, Mavrovo, North Macedonia, pp. 2-7, 2019.
- [16] R. Lippmann, "An introduction to computing with neural nets", IEEE ASSP Magazine, vol.4(2), 1987, pp. 4-22.
- [17] Y. Bengio, "Learning Deep Architectures for AI", Foundations and Trends in Machine Learning 2(1), 2009, pp. 1-127.
- [18] J. R. Donsimoni, R. Glawion, B. Plachter, K. Walde, "Projecting the Spread of COVID19 for Germany" [preprint], pp. 1-26, 2020. <https://doi.org/10.1101/2020.03.26.20044214>
- [19] R. Pal, A. Sekh, S. Kar, D. K. Prasad, "Neural network based country wise riskprediction of COVID-19" [preprint], pp. 1-9, 2020.
- [20] G. B. Dantzig and M. N. Thapa, *Linear Programming 1 - Introduction*, Springer, 1997.
- [21] G. B. Dantzig, A. Orden and P. Wolfe, "The generalized simplex method for minimizing a linear form under linear inequality restraints", Pacific Journal of Mathematics, Volume 5, Number 2, 1955, pp. 183-195.
- [22] H. Fathizadeh et al., "Protection and disinfection policies against SARS-CoV-2 (COVID-19) ", Le Infezioni in Medicina n. 2, April 2020, pp. 185-191.
- [23] N. Petrović, "Approach to Dynamic Adaptivity Simulation in Fog Computing Scenarios", TELSIS 2019, pp. 58-61, October 2019. <https://doi.org/10.1109/TELSIS46999.2019.9002322>
- [24] Coronavirus Update (Live) [online]. Available on: <https://www.worldometers.info/coronavirus/>
- [25] Statistika COVID-19 u Srbiji [online]. Available on: <https://covid19.data.gov.rs/>

A Generalization of the Convolutional Codes

Dejan Spasov
 Faculty of Computer Science and Engineering
 Sts. Cyril and Methodius University
 Skopje, North Macedonia

Abstract—In this paper we propose a generalization of the convolutional codes. The proposed generalization of the convolutional codes offers the possibility to discover previously unknown convolutional codes. For example, convolutional codes with pseudo-random time-varying trellis diagram that may improve error-correcting capabilities of the convolutional codes. An important property of the proposed generalization of the convolutional codes is that the decoding complexity remains the same as the decoding complexity of the ordinary convolutional codes. We propose encoding and decoding schemes for the generalized convolutional codes.

Keywords—error-correcting codes, convolutional codes, turbo codes, LTE, 5G

I. INTRODUCTION

Channel coding plays an important part in modern communication systems. In [1], Shannon gave probabilistic proof that we can communicate with an arbitrary small probability of error as long as the communication rate is below the channel capacity. Error-correcting codes provide constructive solution to the Shannon's theorem. Many error-correcting codes were developed to provide low-power and reliable communication over unreliable channels.

Let F be a finite alphabet of $|F|$ letters and let F^n be the set of all strings of length n over F . In general, an *error-correcting code* C is a subset of F^n of M elements. Elements of the code $c_i \in C$ are called *codewords*.

Let $d(x, y)$ denote the *Hamming distance* between two strings $x, y \in F^n$. The *Hamming distance* $d(x, y)$ is the number of positions in which x and y differ. Let d denote the *minimum distance* d of the code C defined as

$$d = \min\{d(c_i, c_j) | c_i, c_j \in C, i \neq j\} \quad (1)$$

If the minimum distance d is known, we say that C is an (n, M, d) code. The larger the minimum distance d of a code C is, the better the error-correcting capability the code C is.

The code C is linear if its codewords form k -dimensional linear subspace in F^n . We will write $[n, k, d]$ to denote that the code C is linear. For linear codes there exist k basis vectors that are kept as rows in a *generator matrix* G . For each linear code C there is a generator matrix G of type $G = [I \ A]$, where I is the identity matrix. We may say that the generator matrix $G = [I \ A]$ is in *standard form*. It is well-known that for linear codes there exist additional matrix, known as the *parity check matrix* H , defined as

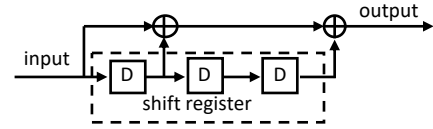


Fig. 1. A convolutional encoder

$$Hc_i^T = 0, \forall c_i \in C \quad (2)$$

Let $G = [I \ A]$ be the generator matrix of a code C , then $H = [-A^T \ I]$ is the parity check matrix of the code C .

A k -letter message may be encoded in a n -letter codeword c . The codeword c may be sent over a noisy channel. The noisy channel may alter few bits of the codeword. A receiver may receive a n -bit string x obtained from the codeword c in which some bits have been altered. The process of finding the nearest, in terms of Hamming distance $d(x, y)$, codeword $\hat{c} \in C$ to the string x , is known as *decoding*

$$\hat{c} = \operatorname{argmin}\{d(x, c_j) | c_j \in C\} \quad (3)$$

Some codes have efficient (polynomial) procedure to find the nearest codeword $\hat{c} \in C$ (in terms of Hamming distance) to the received string x . Codes with polynomial decoding procedures can be used for transmission of digital information over a noisy channel.

In general, error-correcting codes may be divided in two groups: block codes and convolutional codes [2], [3]. Block codes offer greater error-correcting capability, but their decoding algorithms are not very efficient. Convolutional codes are a class of error-correcting codes with polynomial encoding and decoding procedures. They are used in numerous applications to achieve reliable data transfer and reliable data storage. For example, convolutional codes are used in digital video storage, satellite communications, GSM networks, numerous standards: GPRS, EDGE, LTE, 3G, and so on. In this paper we propose a framework that generalizes the concept of convolutional codes. In Section II we briefly introduce certain aspects of convolutional codes that are important for presenting our idea. In Section III we present the new class of convolutional codes. We consider these codes to be generalization of the convolutional codes. We propose

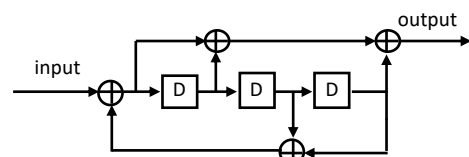


Fig. 2. A recursive convolutional encoder

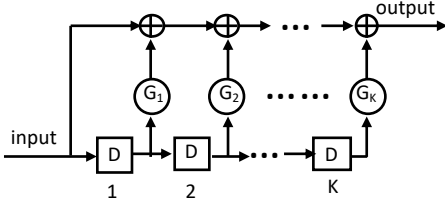


Fig. 3. A framework for a convolutional encoder with K -bit shift register

encoding and decoding mechanisms. In section IV we discuss possible applications of the new class of convolutional codes.

II. CONVOLUTIONAL CODES

Convolutional codes are one of the oldest known classes of error-correcting codes. They were introduced by P. Elias in 1955 [2]. Fig. 1 shows a circuitry that may be used for producing convolutional codes. This circuitry is known as *convolutional encoder*. The convolutional encoder is built with a shift register of length 3 and combinatorial elements, i.e. XOR gates. The shift register consists of 3 memory cells (denoted with D) that introduce one-bit delay. In one clock cycle, one information bit enters the shift register from the left side. Data bits that are already in the register are shifted to the right. Let a semi-infinite stream of information bits $x = (x_1, x_2, \dots, x_i, \dots)$ is provided to the input of the convolutional encoder. In this sequence an information bit x_i is considered older than the information bit x_{i+1} . Let the convolutional encoder be configured to output a semi-infinite sequence of *parity bits* $y = (y_1, y_2, \dots, y_i, \dots)$. The principle of operation of the convolutional encoder is simple: at time $t = 0$ the content of the shift register is 0; a parity bit y_i is computed as

$$y_i = x_i \oplus x_{i-1} \oplus x_{i-3}. \quad (4)$$

where, the symbol \oplus denotes the binary operation XOR.

Convolutional encoders may be divided in two major categories: *recursive* convolutional encoders and *non-recursive* convolutional encoders. The encoder on fig. 1 is an example of a non-recursive encoders. Fig. 2 shows an example of a recursive convolutional encoder. A recursive convolutional encoder is obtained from a non-recursive encoder with two branches of combinatorial elements that may output simultaneously two parity bits. One of the two

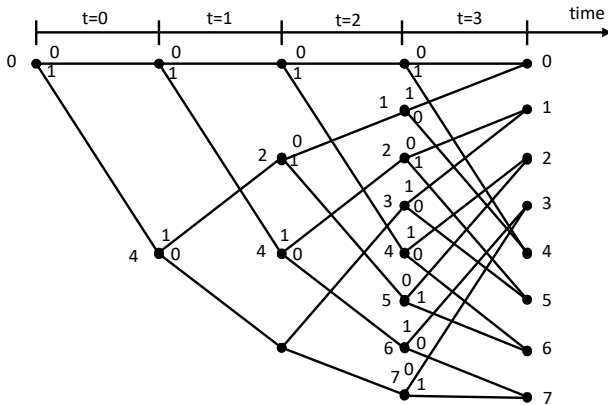


Fig. 4. Trellis diagram of a convolutional encoder

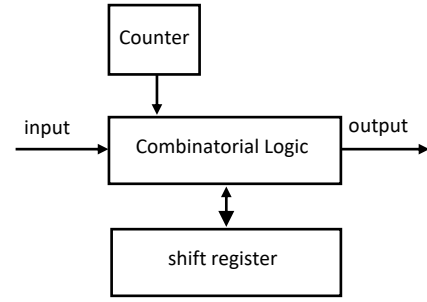


Fig. 5. Proposed convolutional encoder

branches is fed back to the input of the shift register. This way, in a recursive convolutional encoder outputted parity bit is influenced by all older information bits.

Fig. 3 shows a framework for a convolutional encoder with K -bit shift register. The K -bit shift register consists of K memory cells (denoted with D) that introduce one-bit delay. We can obtain a particular encoder using the framework by specifying $G_i = 1$ or $G_i = 0$ for all elements G_i , $i = 1, 2, \dots, K$. Here $G_i = 1$ denotes short circuit and $G_i = 0$ denotes high impedance. An encoder is completely defined by specifying all elements G_i . Particular values of the elements G_i may be represented with a *generator vector* $G = [G_1, G_2, \dots, G_K]$. Sometimes, the generator vector G may be given as octal number. For example, the generator vector for the encoder on fig. 1 is $G = 5$. Hence, outputted bits of a convolutional encoder with shift register of length K are computed as

$$y_i = x_i \oplus G_1 \cdot x_{i-1} \oplus \dots \oplus G_K x_{i-K} \quad (5)$$

The length K of the shift register is known as *constrained length* of the convolutional code. In general, larger shift register provides larger number of input bits that influence outputted parity bits. It is believed that larger constrained length K may provide a convolutional code that may recover larger number of errors. On the other side, larger constrained length K makes the decoding of the convolutional code more complex and demanding in terms of time, space, and hardware.

A convolutional encoder with K shift registers can be considered a finite state machine with 2^K states. A state in the finite state machine is identified with the content of the shift register. Transition from one state to another produces a parity

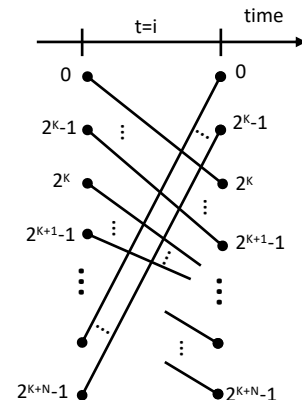


Fig. 6. Trellis diagram of the proposed convolutional encoder

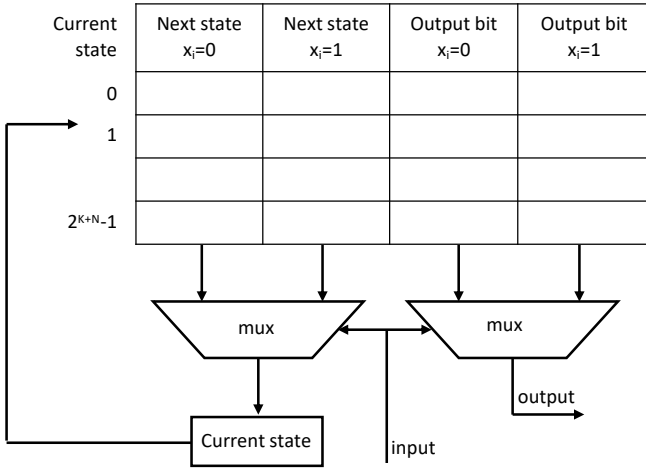


Fig. 7. Implementation of the proposed convolutional encoder

bit. Given a finite state machine, progression of the machine through the states with respect to the time is known as *trellis diagram* of the machine. Trellis diagram is labelled M -partite graph, in which every path represents a valid codeword. Fig. 4 shows the trellis diagram that corresponds to the convolutional encoder given on fig. 1. Vertices are labeled with decimal numbers from 0 to 7. Vertices represent all states of the shift register. Edge labels represent parity bit that will be outputted for that transition.

Encoding process of both the block codes and the convolutional codes may be described with trellis diagram. Trellis diagram of error-correcting codes give hint about the decoding. A decoder may be configured to guess the most likely path through the trellis. The block codes, in general, have exponential number of states; thus, the decoding process is much complex. The number of states of the trellis diagram of a convolutional code is determined by the size of the shift register. Hence, convolutional codes have better decoding complexity; although, convolutional codes have weaker error correcting capability. In general, all convolutional codes can be decoded with the Viterbi algorithm [5], [6]. Memory requirements of the Viterbi decoder are proportional with the constrained length K . Another algorithm that may be used for decoding is the BCJR algorithm [7]. Memory requirements of the BCJR decoder are proportional with the product of the length of the sequence and the number of states of the shift register.

Popular convolutional codes are the turbo codes. Turbo codes were the first error-correcting codes that demonstrated reliable communications near the channel capacity with practically feasible hardware [8]. Due to their excellent error-correcting capability, they are part of many modern communication technologies, like LTE [9]. An LTE turbo encoder is a systematic encoder made of two 8-state recursive convolutional encoders. Recursive encoder used in turbo codes is given on fig. 2. Generator polynomials of the LTE turbo codes are fixed and specified in a standardized specification [9].

III. GENERALIZED CONVOLUTIONAL CODES

From fig. 3 it can be observed that the framework with K -bit shift register can specify 2^K convolutional encoders. Thus, only 2^K non-recursive encoders can be built. Additionally, this framework specifies that only 2^{K+1} recursive convolutional encoders can be built. However, the number of

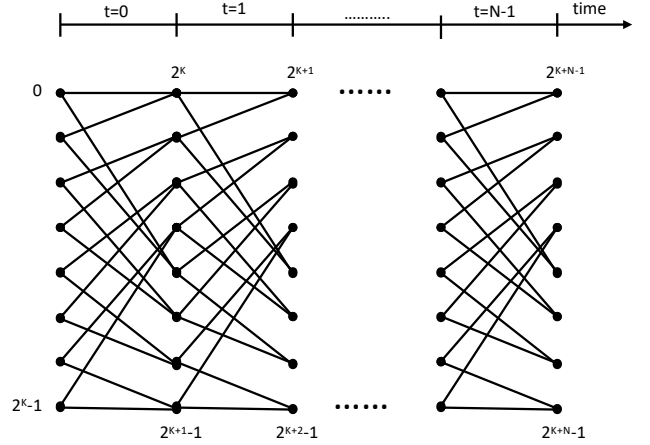


Fig. 8. Decoding of the proposed convolutional encoder

possible convolutional encoders is much larger than 2^K . In this paper, our goal is to provide a framework for a convolutional encoder that may specify all possible convolutional encoders. Principle of operation of the proposed framework is given on fig. 5.

The framework on fig. 5 generalizes the concept of convolutional encoder. The convolutional encoder consists of a K -bit shift register, a combinatorial logic, and a N -bit shift register. Principle of operation of the shift register and the combinatorial logic is similar to the principle of operation of the regular convolutional encoder (fig. 1 and fig. 2). The combinatorial logic convolutes the bits stored in the shift register and the bit provided to the input in order to produce an output bit. A counter is added to the framework. The counter is incremented with each input bit. In a simplest embodiment, the counter may be 1-bit counter. The purpose of the counter is to increase the number of states of the encoder. A state may be identified with the content of the counter and the content of the shift register. Thus, the proposed encoder has 2^{K+N} states. Each time the counter is incremented, the combinatorial logic changes the formula for computing the output bit. Hence, the parameters of the encoder appear to vary over time.

The proposed encoder may be considered as finite state machine with 2^{K+N} states. The trellis diagram of the finite state machine is given on fig. 6. It can be observed that the trellis diagram has larger number of states compared to the trellis of a regular convolutional encoder (fig. 4). An important feature of the trellis is that the 2^{K+N} states may be divided into 2^N disjoint sets, wherein each set comprises of 2^K states. Transitions are possible from one set of states to another set. This property makes the decoding of the proposed convolutional codes is with the same complexity as the regular convolutional codes.

Hardware implementation of the proposed convolutional encoder consists of a SRAM memory operated as table, a register to hold the current state and multiplexers (fig. 7). The table is indexed with the number of states of the finite state machine. The content of the current-state register is provided as index to the read ports of the table. The output of the table is provided as input to the multiplexers. An input bit is provided as selection control to the multiplexers. This way, the appropriate fields of the outputted entry are selected.

The proposed convolutional codes can be decoded with the Viterbi algorithm and with the BCJR algorithm. Decoding complexity of the proposed codes remains the same as decoding complexity of the regular convolutional codes. Fig. 8 shows another view of the trellis diagram of the proposed codes. The diagram shows only possible transitions from one set of states to another set of states at one point of time. It can be observed that at any time, the number of active states is the same as with the regular convolutional codes. Therefore, the decoding complexity remains the same.

IV. CONCLUSION

We have proposed a framework for generating a new class of convolutional codes. We consider the new class of codes as generalization of the concept of a convolutional code. The proposed framework offers possibility to specify convolutional codes that cannot be specified with the previously known methods for specifying convolutional codes. One important feature of the proposed convolutional codes is that the decoding complexity remains the same as the decoding complexity of the regular convolutional codes, although the number of states increases. The proposed encoder can be efficiently implemented in hardware using SRAM memory cells operated as table. The SRAM table may be populated with the possible transitions at the run-time. Therefore, the parameters of the convolutional codes need not

to be specified at the design stage of the communication system. Our future goal is to search over these codes and to find a convolutional code with improved error correcting capability.

REFERENCES

- [1] C. E. Shannon, "The mathematical theory of communication," Bell System Technical Journal, vol. 27, pp. 379–423, July 1948.
- [2] F. J. MacWilliams, N.J.A. Sloane, "The Theory of Error-Correcting Codes," North Holland, Amsterdam, 1977.
- [3] S. Lin, D. J. Costello, "Error Control Coding," Pearson, 2005.
- [4] Elias, P. "Coding for noisy Channels." *IRE Convention Record*, Part IV, pp. 37-46 (1955).
- [5] Viterbi, A. J. "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm." *IEEE Transactions on Information Theory* vol. 13, no. 2, pp. 260–269, 1967.
- [6] Forney, G. D. "The Viterbi Algorithm." *IEEE Transactions on Information Theory* vol. 61, no. 3, pp. 268–278, 1973.
- [7] Bahl, L., Cocke, J., Jelinek, F., Raviv, J. "Optimal Decoding of Linear Codes for minimizing symbol error rate." *IEEE Transactions on Information Theory*, vol. 20, no. 2, pp.284-287, 1974.
- [8] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo-codes," in Proc. ICC, Geneva, Switzerland, May 1993.
- [9] ETSI Technical Specification, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding," 3GPP TS 36.212 version 8.8.0 Release 8, Jan. 2011.

A meshfree formulation for the simulation of mould filling processes in casting

Felix R. Saucedo-Zendejo
Research Center on Applied Mathematics
Autonomous University of Coahuila
Saltillo, Mexico
fesaucedoz@uadec.edu.mx

Abstract—The numerical simulation of the mould filling step in casting is crucial since it provides a complete set of detailed information. However, the mesh based methods usually used to model this processes have some drawbacks when dealing with this kind of problems. In order to overcome some of these limitations in this work the application of an innovative meshfree computational tool based on the Finite Pointset Method is proposed. The main features and details of its implementation are presented and finally, the accuracy and robustness of this scheme is assessed through the numerical simulations of three-dimensional test problems in mould filling processes.

Index Terms—casting, meshless method, Finite Pointset Method, FPM

I. INTRODUCTION

Many industrial applications require the production of aluminum alloy and metal matrix components with precise shapes, dimensions and surface fineness, which could be very complex in general. Shape casting processes as lost foam casting, die casting, stir casting and sand casting are among the most widely used procedures to manufacture aluminum alloy products. Such processes begin with a step in which a mould's cavity is filled with molten material followed by a cooling and solidification step. The features in the manufactured components and the number of defects are the result of the coupling of the physical phenomena involved in these steps. Therefore, in order to get homogeneous casting components with the desired characteristics and an acceptable low amount of defects, it is necessary to know and control all the physical phenomena involved in the selected process and to conduct a proper design procedure [1].

Currently the foundry industry is interested in the best possible performance at the lowest cost. The required performance can be achieved only when the desired microstructure and an acceptable low number of defects are obtained in the end product. For a specific alloy, this can be achieved only by knowing and understanding the selected casting process so that it can be optimized. Moreover, nowadays the industrial casting product development is shifting from traditional heuristic and experience-based trial-and-error design procedures to a deep scientific proof-of-concept design procedure.

Numerical simulation is commonly used to analyze and improve different casting processes since it provides a large amount of information that cannot be obtained through other

techniques. It is the most technologically efficient, cost effective and powerful technology for the evaluation, analysis and prediction of the quality in the end products and the number of defects since it models the entire casting process and shows all the details and the dynamic behavior of the casting system in real working conditions. Therefore, the root causes of the quality in the end products and the casting defects are pinpointed and the possible solution routes to avoid them and to improve the overall quality could be determined, evaluated and analyzed with this tool [2].

Mesh-based methods such as the Finite Element Method (FEM), Finite Difference Method (FDM), Finite Volume Method (FVM), and more recently, meshless methods as Smoothed-particle hydrodynamics (SPH) have already been used to analyse mould filling processes [3]–[6]. The advantages of meshless methods over mesh-based methods are that they use a set of nodes to discretize the problem domain and its boundaries without requiring any information about the relationship between nodes such that they do not form an element mesh which lets to model deformations and discontinuities in the domain without the mesh-based methods drawbacks. Moreover, they have the flexibility to add or remove nodes wherever and whenever needed and it lets to easily develop adaptive schemes [7].

The so called Finite Pointset Method (FPM), member of the family of generalized finite difference methods (GFDM), is a meshfree method that has proven to be far superior to traditional mesh-based and some other meshfree methods to treat fluid dynamics problems involving complex geometries and heat transfer problems [8]–[12]. It has many advantages over other methods since it is able to naturally and easily incorporate any kind of boundary conditions without requiring any special treatment or stabilization and it is really simple to implement. Therefore, in this work we propose the application of this novel meshfree formulation for the numerical simulation of mould filling processes. The current paper is organized in the following manner: Section II shortly describes the governing equations and the numerical procedure used to solve them. Section III describes the basic ideas behind the meshfree formulation for the mould filling problem followed by the numerical test presented in Section IV with its corresponding results. Finally some conclusions are given in last section.

II. GOVERNING EQUATIONS AND NUMERICAL PROCEDURE

The governing equations for mould filling processes are the incompressible Navier-Stokes equations which in Lagrangian form read

$$\frac{D\mathbf{v}}{Dt} = -\frac{1}{\rho}\nabla p + \nu\nabla^2\mathbf{v} + \mathbf{f}, \quad (1)$$

$$\nabla \cdot \mathbf{v} = 0, \quad (2)$$

where ρ is the density, p is the pressure, ν is the kinematic viscosity, \mathbf{f} is the acceleration vector due to the body forces and \mathbf{v} is the velocity. This system of equations is completed with the following initial and boundary conditions

$$\mathbf{v}|_{t=0} = \mathbf{v}_0, \quad (3)$$

$$\mathbf{v}|_{\partial\Omega_1} = \mathbf{0}, \quad (4)$$

$$\mathbf{v} \cdot \mathbf{n}|_{\partial\Omega_2} = 0, \quad (5)$$

$$\left. \frac{\partial(\mathbf{v} \cdot \mathbf{t}_i)}{\partial \mathbf{n}} \right|_{\partial\Omega_2} = 0, \quad (6)$$

$$(\boldsymbol{\tau} - \mathbf{I}p)\mathbf{n}|_{\partial\Omega_3} = \sigma\kappa\mathbf{n}, \quad (7)$$

$$\mathbf{t}_i^T \boldsymbol{\tau} \mathbf{n}|_{\partial\Omega_3} = 0, \quad (8)$$

where $\partial\Omega_1$ is a solid wall boundary with no-slip condition, $\partial\Omega_2$ is a solid wall boundary with free-slip condition, $\partial\Omega_3$ is a free surface, \mathbf{v}_0 is the initial velocity over the entire domain Ω , $\boldsymbol{\tau}$ is the viscous stress tensor, σ is the surface tension, κ is the free surface boundary curvature, \mathbf{n} is an outward orthonormal vector and \mathbf{t}_i is a tangential unitary vector to the boundary, for $i = 1, 2$.

In order to solve the system of equations (1) and (2) with the boundary and initial conditions (3) - (8) in a natural and simple way, a semi-implicit Chorin-Uzawa's projection formulation of first order of accuracy will be used [13] which consists of the following steps [12]:

- 1) Explicitly update the nodes positions through

$$\mathbf{r}^{n+1} = \mathbf{r}^n + \Delta t \mathbf{v}^n. \quad (9)$$

- 2) Implicitly solve for the intermediate velocities

$$\mathbf{v}^* - \Delta t \nu \nabla^2 \mathbf{v}^* = \mathbf{v}^n + \Delta t \mathbf{f}^{n+1}, \quad (10)$$

with the boundary and initial conditions (3) - (8).

- 3) Implicitly solve for the artificial pressure

$$\nabla \varphi = \frac{\rho}{\Delta t} \nabla \cdot \mathbf{v}^*, \quad (11)$$

which must satisfy the following boundary conditions

$$\left. \frac{\partial \varphi}{\partial \mathbf{n}} \right|_{\partial\Omega_1, \partial\Omega_2} = 0, \quad (12)$$

$$\varphi|_{\partial\Omega_3} = 0. \quad (13)$$

- 4) Correct/Update the velocity field

$$\mathbf{v}^{n+1} = \mathbf{v}^* - \frac{\Delta t}{\rho} \nabla \varphi \quad (14)$$

- 5) Correct/Update the pressure field

$$p^{n+1} = \varphi - \rho \nu \nabla \cdot \mathbf{v}^*. \quad (15)$$

where \mathbf{r}^n and \mathbf{v}^n are initially given and they denote the nodes positions and its velocities at time t^n , respectively.

III. THE FINITE POINTSET METHOD (FPM)

In this section we describe the main ideas of the FPM method proposed by [14]. The FPM is a member of the family of the GFDM and it is based on the WLSM. Following [8]:

Let Ω be a given domain with boundary $\partial\Omega$ and suppose that the set of points $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$ are distributed with corresponding function values $f(\mathbf{r}_1), f(\mathbf{r}_2), \dots, f(\mathbf{r}_N)$. The problem is to find an approximate value of f at some arbitrary location $f(\mathbf{r})$ using its discrete values at particles positions inside a neighbourhood of \mathbf{r} . To define the set of nodes and the neighbourhood of \mathbf{r} , a weight function $w(\mathbf{r} - \mathbf{r}_i)$ is introduced

$$w_i = w(\mathbf{r} - \mathbf{r}_i) = \begin{cases} e^{-\gamma \|\mathbf{r} - \mathbf{r}_i\|^2 / h^2}, & \text{if } \frac{\|\mathbf{r} - \mathbf{r}_i\|}{h} \leq 1 \\ 0 & \text{else} \end{cases} \quad (16)$$

where h is the smoothing length, γ is a positive constant whose value is considered to be 6.5, and \mathbf{r}_i is the position of the i -th point inside the neighbourhood. A Taylor's series expansion of $f(\mathbf{r}_i)$ around \mathbf{r} reads

$$f(\mathbf{r}_i) = f(\mathbf{r}) + \sum_{k=1}^3 f_k(r_{k_i} - r_k) + \frac{1}{2} \sum_{k,l=1}^3 f_{kl}(r_{k_i} - r_k)(r_{l_i} - r_l) + \epsilon_i, \quad (17)$$

where ϵ_i is the truncation error of the Taylor's series expansion, r_{k_i} and r_k represent the k -th components of the position vectors \mathbf{r}_i and \mathbf{r} , respectively. f_k and f_{kl} ($f_{kl} = f_{kl}$) represent the set of first and second spatial derivatives at node position \mathbf{r} . The values of f_k and f_{kl} can be computed minimizing the error ϵ_i for the n_p Taylor's series expansion of $f(\mathbf{r}_i)$ corresponding to the n_p nodes inside the neighbourhood of \mathbf{r} . This system of equations can be written in matrix form as $\mathbf{e} = M\mathbf{a} - \mathbf{b}$, where $\mathbf{e} = [e_1, e_2, e_3, \dots, e_{n_p}]^t$, $\mathbf{a} = [f, f_1, f_2, f_3, f_{11}, f_{12}, f_{13}, f_{22}, f_{23}, f_{33}]^t$, $\mathbf{b} = [f(\mathbf{r}_1), f(\mathbf{r}_2), \dots, f(\mathbf{r}_{n_p})]^t$, $M = [\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_{n_p}]^t$, $\mathbf{s}_i = [1, \Delta r_{1i}, \Delta r_{2i}, \Delta r_{3i}, \Delta r_{11i}, \Delta r_{12i}, \Delta r_{13i}, \Delta r_{22i}, \Delta r_{23i}, \Delta r_{33i}]^t$, $\Delta r_{ki} = r_{k_i} - r_k$, $\Delta r_{kl_i} = (r_{k_i} - r_k)(r_{l_i} - r_l)$ and $\Delta r_{kk_i} = 0.5(r_{k_i} - r_k)(r_{k_i} - r_k)$, for $k, l = 1, 2, 3$ and $k \neq l$. The unknown vector \mathbf{a} is obtained through WLSM by minimizing the quadratic form

$$J = \sum_{i=1}^{n_p} w_i \epsilon_i^2, \quad (18)$$

which reads $(M^t W M)\mathbf{a} = (M^t W)\mathbf{b}$, where $W = \text{diag}(w_1, w_2, \dots, w_{n_p})$. Therefore, $\mathbf{a} = (M^t W M)^{-1} (M^t W)\mathbf{b}$. In this way we automatically get the values of f and its derivatives at points \mathbf{r} .

A. FPM formulation for the semi-implicit Chorin-Uzawa's scheme

Poisson equations as (11) and coupled vector boundary value problems as (10) have been already studied in [12], [15]. Following such works we present, for completeness, the corresponding FPM formulation to solve the equations (9 -

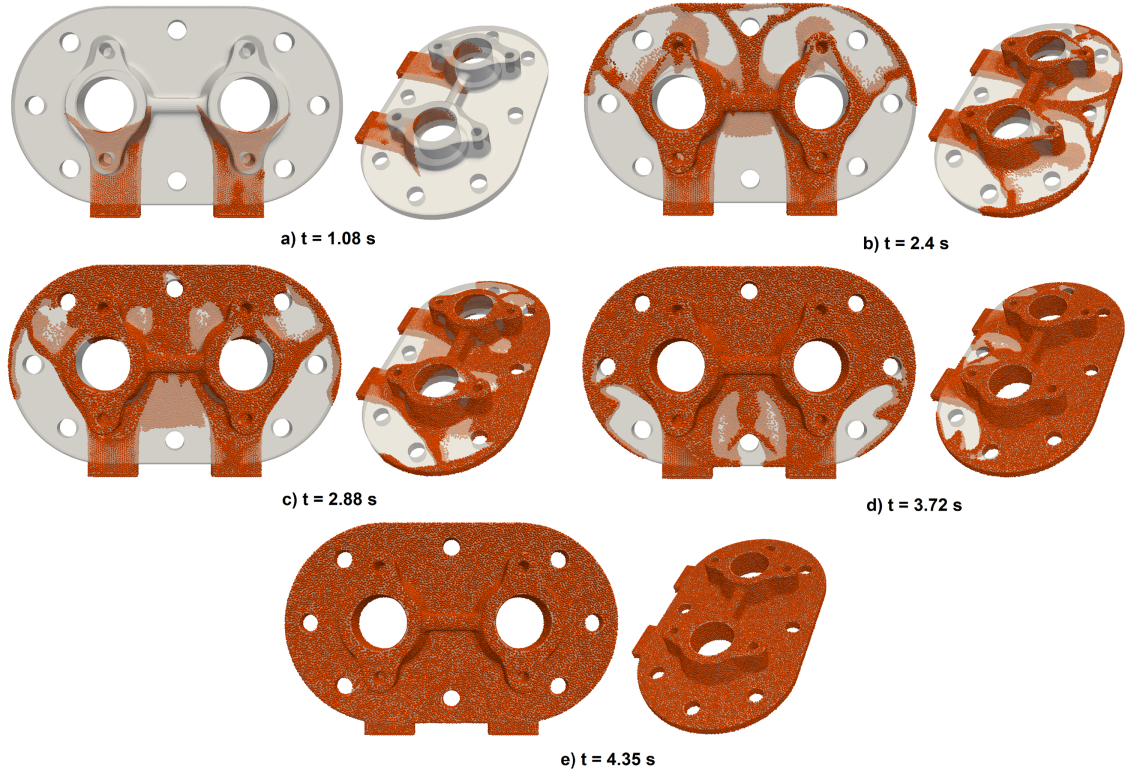


Fig. 1. Numerical filling patterns at different time steps.

15). Equation (11) is an elliptic partial differential equation, which can be written in the following general form

$$Af + \mathbf{B} \cdot \nabla f + C\nabla^2 f = D \quad (19)$$

and the boundary conditions take the general form

$$Ef + G\nabla f \cdot \mathbf{n} = H. \quad (20)$$

In the FPM representation of the above problem, equation (19) must be taken together with the system of n_p Taylor's series expansion of $f(\mathbf{r}_i)$ around \mathbf{r} . In this case, the matrices we need to compute by each particle in Ω take the following form: $\mathbf{b} = [f(\mathbf{r}_1), f(\mathbf{r}_2), \dots, f(\mathbf{r}_{n_p}), D]^t$, $M = [s_1, s_2, \dots, s_{n_p}, s_E]^t$ and $W = \text{diag}(w_1, w_2, \dots, w_{n_p}, 1)$, where $s_E = [A, B_1, B_2, B_3, C, 0, 0, C, 0, C]^t$ and $\mathbf{B} = [B_1, B_2, B_3]^t$. If $\mathbf{r}_i \in \partial\Omega$, additionally we have to add the general boundary condition (20). Therefore, in this case, the matrices we need to compute by each particle in $\partial\Omega$ take the following form: $\mathbf{b} = [f(\mathbf{r}_1), f(\mathbf{r}_2), \dots, f(\mathbf{r}_{n_p}), D, F]^t$, $M = [s_1, s_2, \dots, s_{n_p}, s_E, s_B]^t$ and $W = \text{diag}(w_1, w_2, \dots, w_{n_p}, 1, 1)$, where $s_B = [E, n_1, n_2, n_3, 0, 0, 0, 0, 0, 0]^t$.

If we define $\mathbf{q} = [q_1, q_2, \dots, q_{10}]^t$ as the first row of $(M^t W M)^{-1}$ and the terms in the moving least squares

solution $\mathbf{a} = (M^t W M)^{-1} (M^t W) \mathbf{b}$ are worked out, we can see that the following linear equations arises

$$\begin{aligned} f(\mathbf{r}_j) - \sum_{i=1}^{n(j)} w_i (q_1 + q_2 \Delta r_{1i} + q_3 \Delta r_{2i} + q_4 \Delta r_{3i} + q_5 \Delta r_{11i} \\ + q_6 \Delta r_{12i} + q_7 \Delta r_{13i} + q_8 \Delta r_{22i} + q_9 \Delta r_{23i} + q_{10} \Delta r_{33i}) f(\mathbf{r}_i) \\ = [Aq_1 + B_1 q_2 + B_2 q_3 + B_3 q_4 + (q_5 + q_8 + q_{10}) C] D \\ + (n_1 q_2 + n_2 q_3 + n_3 q_4) E, \end{aligned}$$

where $f(\mathbf{r}_j)$ denotes the unknown function value at particle j and $n(j)$ the number of j -th particle neighbours. Since equation (III-A) is valid for $j = 1, 2, \dots, N$, this can be arranged in a full sparse system of linear equations $L\mathbf{F} = \mathbf{P}$ which can be solved by iterative methods. In the same way, coupled vector boundary value problems as (10) are treated similarly. For further information on this kind of problems, we refer to [12]. Therefore, all kind of problems as (9 - 15) can be solved with this formulation, just adding appropriate entries in the corresponding systems of equations [12], [15].

IV. NUMERICAL EXAMPLE

In this section the suitability and feasibility of this FPM formulation in order to simulate 3D mould filling processes in metal casting will be evaluated considering the filling of a pump cover. In this example, the problem domain was discretized with approximately 180000 points with a

mean spacing of 0.0015 m. The inlet velocity was taken as $\mathbf{v} = [-0.1, 0, 0]^t$ m/s. The pressure in all particles as well as the atmospheric pressure were considered as zero. The surface tension forces and the gravitational acceleration vector were neglected. The density and viscosity of the fluid were considered as $\rho = 6964 \text{ kg/m}^3$ and $\mu = 0.0143 \text{ Pa}\cdot\text{s}$, respectively. These parameters corresponds to the physical parameters of the molten cast iron. Finally, the smoothing length used in the simulation was $h = 0.0045 \text{ m}$, a slip boundary condition was used at solid walls and the time step was chosen as $\Delta t = 0.004 \text{ s}$.

Two perspective views of the filling patterns at different time steps are depicted in Figure 1. There, the picture on the left shows the view exactly from the top whilst the second one shows the view from the top at some angle to the right. As it is shown in this figure, the leading material is divided in four liquid fronts when it impacts the two annular central sections of the die. Two central jets partially merge forming a single liquid front which is split again when it impacts the central cylindrical obstacle. The emerging jets move backwards and starts filling the rear part of the mould. Splashing droplets and liquid fragmentations are visible in this part. The remaining two jets flow around the curved outsides of the die until they collide with the fronts coming from the rear and central parts of the mould. In the two annular central sections of the mould, the liquid flow up into the upper extensions. At this point the rear part of the mould is substantially filled and the fluid flow is towards the front part of the mould. Afterwards, almost all the mould cavity is filled and the biggest voids are principally behind two of the cylindrical obstacles near the inflow jets. They are uniformly filled until the filling process finishes.

These pictures show the robustness of this FPM formulation for the simulation of complex 3D mould filling processes since the splashing into droplets, the fluid fragmentation into jets and the fronts collisions observed in this example are well reproduced and predicted by this approach.

V. CONCLUSIONS

Based on the numerical performance shown in the numerical example we can conclude that the current approach is suitable and feasible for the simulation of 3D mould filling processes. It is stable and it has enough accuracy in order to capture the splashing into droplets, the fluid fragmentation into jets and fronts collisions which are observed in this kind of processes. Since this formulation is a truly meshfree method it could be used for the study and analysis of complex problems involving high deformations and domain fragmentations with a great computational advantage since it does not need to compute any numerical quadrature and it does not need remeshing approaches. Further, it is able to naturally and easily handle any kind of boundary conditions without requiring any special treatment or stabilization and it is really simple to implement. Therefore, it could be a promising numerical tool for the simulation of industrial processes involving complex flows and other phenomena described by elliptic partial differential

equations as heat transfer. Consequently, it depicts a rich source of research opportunities.

REFERENCES

- [1] T. V. R. Rao, *Metal casting: Principles and practice*, 1st ed. New Age International, 2007.
- [2] A. Kermanpur, S. Mahmoudi, and A. Hajipour, "Numerical simulation of metal flow and solidification in the multi-cavity casting moulds of automotive components," *J. Mater. Process. Technol.*, vol. 206, no. 1-3, pp. 62–68, 2008.
- [3] R. W. Lewis and K. Ravindran, "Finite element simulation of metal casting," *Int. J. Numer. Meth. Engng.*, vol. 47, no. 1-3, pp. 29–59, 2000.
- [4] A. Kimatsuka, I. Ohnaka, J. D. Zhu, and T. Ohmichi, "Mold filling simulation with consideration of gas escape through sand mold," *Int. J. Cast Metal Res.*, vol. 15, no. 3, pp. 149–152, 2003.
- [5] H. Bašić, I. Demirdžić, and S. Muzaferija, "Finite volume method for simulation of extrusion processes," *Int. J. for Numer. Meth. Engng.*, vol. 62, no. 4, pp. 475–494, 2005.
- [6] P. W. Cleary, G. Savage, J. Ha, and M. Prakash, "Flow analysis and validation of numerical modelling for a thin walled high pressure die casting using SPH," *Comput. Part. Mech.*, vol. 1, no. 3, pp. 229–243, 2014.
- [7] T. Belytschko, Y. Krongauz, D. Organ, M. Fleming, and P. Krysl, "Meshless methods: An overview and recent developments," *Comput. Methods Appl. Mech. Engng.*, vol. 139, no. 1-4, pp. 3–47, 1996.
- [8] S. Tiwari and J. Kuhnert, "A meshfree method for incompressible fluid flows with incorporated surface tension," *Revue Europeenne des Elements*, vol. 11, no. 7-8, pp. 965–987, 2002.
- [9] —, "Particle Method for Simulation of Free Surface Flows," in *Hyperbolic Problems: Theory, Numerics, Applications*, Y. Hou and E. Tadmor, Eds. Springer Berlin Heidelberg, 2003, pp. 889–898.
- [10] —, "Modeling of two-phase flows with surface tension by finite pointset method (FPM)," *J. Comput. Appl. Math.*, vol. 203, no. 2, pp. 376–386, 2007.
- [11] A. Jefferies, J. Kuhnert, L. Aschenbrenner, and U. Giffhorn, "Finite Pointset Method for the Simulation of a Vehicle Travelling Through a Body of Water," in *Meshfree Methods for Partial Differential Equations VII*, ser. Lecture Notes in Computational Science and Engineering, M. Griebel and M. A. Schweitzer, Eds. Springer Berlin Heidelberg, 2015, vol. 100, pp. 205–221.
- [12] F. R. Saucedo-Zendejo and E. O. Reséndiz-Flores, "A new approach for the numerical simulation of free surface incompressible flows using a meshfree method," *Comput. Method. Appl. M.*, vol. 324, pp. 619–639, 2017.
- [13] A. Prohl, "Projection and quasi-compressibility methods for solving the incompressible Navier-Stokes equations," ser. *Advances in Numerical Mathematics*. Vieweg Teubner Verlag, 1997.
- [14] J. Kuhnert, "General Smoothed Particle Hydrodynamics," Ph.D. dissertation, Technische Universität Kaiserslautern, 1999.
- [15] S. Tiwari and J. Kuhnert, "Grid free method for solving the poisson equation," *Berichte des Fraunhofer ITWM*, vol. 25, 2001.

SHORT PAPERS

Smart City: Public Parking Dashboard

1st Ivan Klandev

*Back-end Engineer
IT Labs*

Skopje, Republic of North Macedonia
klandev.ivan@students.finki.ukim.mk

2nd Marta Tolevska

*Faculty of Computer Science and Eng.
St. Cyril and Methodius University*

Skopje, Republic of North Macedonia
marta.tolevska@students.finki.ukim.mk

3rd Dimitar Trajanov

*Faculty of Computer Science and Eng.
St. Cyril and Methodius University*

Skopje, Republic of North Macedonia
dimitar.trajanov@finki.ukim.mk

Abstract—It is widely known that finding an available parking slot during peak hour is a significant issue in cities. Drivers are spending large amount of time in search of a free spot where they can properly park their car. The purpose of providing these visualizations is to have a visualized preview of the data collected from different parking events, in order to spot improper parked vehicles, find causes for jammed areas, find places where there is a need for building a parking garage and etc. Considering the popularity and evolution of Internet of Things and their usage in the development of smart cities, these visualizations are providing crucial information. That information will help distinct areas that have a need of implementing some system regarding smart city, from another areas that are not that crowded and there is no need of any changes. All these benefits will also lead to reduction of expenses, pollution and traffic jams, while the income will rise.

Index Terms—smart city, public parking, visualization, dashboard

I. INTRODUCTION

The interest in the concept of smart cities has been increasing constantly. After the expansion of Internet of Things, the idea of creating a smart city seems to be much more realizable. As a result, consistent efforts are being made in the field of IoT in order to enlarge the productivity and reliability of urban infrastructure. Nowadays, parking is becoming one of the major problems for drivers in the cities, and it tends to become even harder as a result of the continuous increase of car users. Because of this, parking is limited in major cities including universities and major attractions all around the world [1]–[4]. For instance, finding parking space during major events such as during game day or graduation day is very challenging. Although parking lot occupancy could be significantly reduced by an increase of parking fee [5], that is not a durable solution and does not fix the problem. On the other hand, this situation can be seen as a motive for smart cities to take actions in order to reinforce the efficiency of the parking resources therefore leading to decrease in traffic congestion, air pollution and road accidents as well [6].

Considering the rise of urban population, the land reduction and traffic jams increase, it is beyond clear that the problems associated with parking impose significant societal costs, both economically and ecologically [7]. Furthermore, Manville and Shoup [8] surveyed the percentage of total parking areas in the central business district of different cities. Averagely, parking coverage takes 31% of land use in big cities, like San

Francisco, and even more, 81% in Los Angeles and 76% in Melbourne, while at the lower end we find New York (18%), London (16%), and Tokyo (7%). Such a super high parking coverage density in Los Angeles can be a constraint on urban redevelopment [9]. Another alarming fact is that 100% of Parisian drivers ever abandoned their trips because of annoying and endless parking searches, while other drivers park their cars on unauthorized areas (Association for European Transport 2006). Concerning the phenomenon that common parking service could not satisfy the increasing demand of the private vehicle owners [10], we created a dashboard that provides analysis and visualizations of data regarding occupation, income and duration of various parking events. Using an existing parking system as a data source, a software was created in order to address the issue and help more efficiently and easily understand the meaning of collected data, as well as discover models and correlations between that data. The current works on smart parking, as a component/part of the concept of smart cities, are complex and multidisciplinary [9]. It is known for a fact that when implementing a smart parking system, a large amount of time is needed in order to inspect all the areas and their parking lots occupancy. Consequently, we created this dashboard which helps easily get the necessary information, by having a simple but very straightforward and visualized preview. As an example, if an area is spotted that has less parking slots occupied and at the same time has less traffic, then that will lead to a conclusion that in that area there is no need for implementation of smart parking systems or building new parking garage. In contrary, if a location that has a lot of traffic and no available parking slots is identified, then that will be an indicator of the need to take proper actions. Moreover, another benefit of having this type of textual data visualized is that it could be seen as a cause for further research like, why in some areas there are more parking events made compared to other areas? Is that a consequence of the larger parking fee in that area or is it a result only by its location? Actually, these are the issues that the city council should be aware of and if resolved properly they will result in increase of public transportation's use rate and cities' revenues, while reducing traffic jams, consumption and pollution. On the other hand, without this information there is a great possibility that an inadequate decision could be made, about where and what techniques and technologies to deploy, which later can lead to a large capital loss.

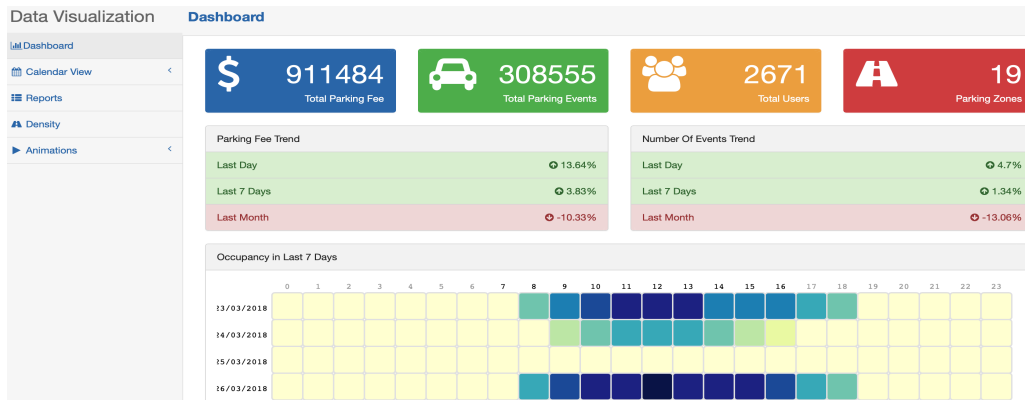


Fig. 1: Dashboard with summary view of data from parking events

II. ANALYTICS AND VISUALIZATIONS

This paper analyzes and visualizes parking events from an existing parking system in order to understand the meaning of the data, to identify patterns and correlations that are not easily recognizable when in a textual presentation. The data used in our application is obtained from PARX Ltd., a company that offers parking systems to municipalities and parking operators around the world. PARX Ltd. system called - EasyPark Mobile System, has been used in many different countries among which is Bermuda. This system collects the data by locating the person when launching a new parking event. If that person is using the mobile application, the system, if allowed, gets the exact coordinates where the event started. Having in mind that all of the parking zones are geotagged in the system, if some user does not allow the application to send his exact location, then the coordinates of the parking zone are obtained. The data set consists of parking events starting from 01.01.2014 to 31.03.2018. Each parking event consists of the following attributes:

- STARTDATE - the date the parking event started
- STARTTIME - time when the parking event started
- ENDDATE - the date when the parking event is over
- ENDTIME - time when the parking event is over
- LONGITUDE - longitude where the parking event occurred
- LATITUDE - latitude where the parking event occurred
- DURATION - duration of the parking event in minutes

- CLIENT - identifier of the client who parked
- DAYTYPE - day of the week when the parking event occurred
- PRICE - parking event fee in dollars
- ZONEID - identifier of the parking area where the parking event occurred
- MONTH - month of the year when the parking event occurred

The application for visualizing parking events is implemented with the Django framework. The Django framework enables easy server-side implementation of the Python programming language in a web application. Bootstrap CSS is used for the design of the application and the D3.js library is used to display graphics, which enables easy drawing of graphics in JavaScript. In the following chapters, the visualizations will be explained and presented in detail.

A. Summary View

”Summary view” is the introductory page of the application. This page shows summary of parking information, where at the top of the page are summarized data and trends grouped into four groups, Figure 1.

- Earnings - Total earnings from all parking events in US dollars
- Parking events - The total number of parking events
- Users - Total number of different users who parked in the parking zones

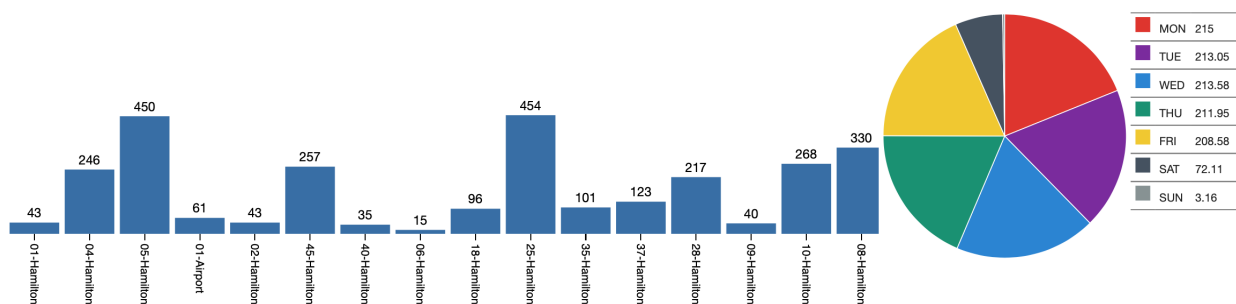


Fig. 2: Pie chart representing average parking event duration per day of the week and per parking area

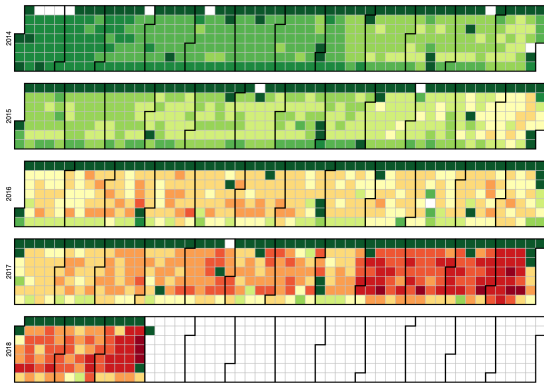


Fig. 3: Calendar view of total number of parking events per day, summed for all parking zones

- Parking Zones - Total number of parking zones where one can park

This page also shows the earnings trend and the number of parking events, Figure 1. The percentage of increase or decrease is shown for each trend. Each of the trends displays the following information:

- Previous day - This shows the upward or downward trend in earnings or the number of parking events compared to the previous day and the day before.
- Previous 7 days - This shows the upward or downward trend in earnings or the number of parking events compared to the last 7 days and the 7 days before them.
- Previous month - This shows the upward or downward trend in earnings or the number of parking events compared to the last month and the month before it.

The Summary page also shows the time distribution of parking events in the last 7 days, Figure 1. The y-axis shows the dates of the last 7 days while the x-axis represents the hour (24 hour format) that the parking event started. The visualization shows the total number of parking events that started at a given hour and on a given date represented by density cubes where a larger number of parking events means a darker color displayed.

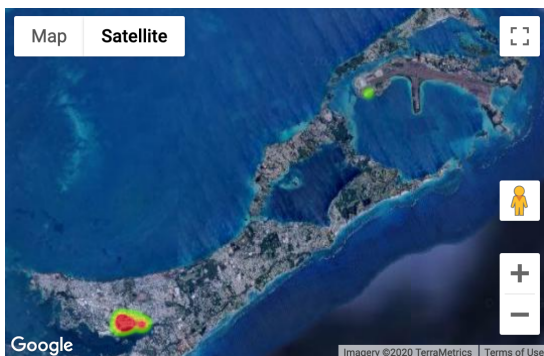


Fig. 5: Geographical distribution of parking events during workday

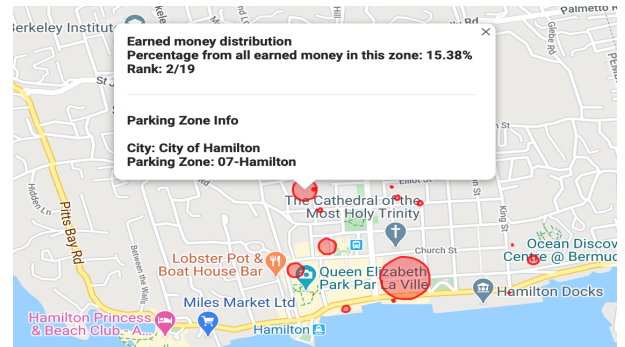


Fig. 4: Distribution of earnings from parking events by parking area

B. Calendar View

The "Calendar view" visualizes parking events and earnings. Figure 3 shows the total number of parking events per day summed for all parking zones, so that the greater the number of parking events the more the color moves from green to yellow which represents some intermediate level and ultimately to red.

C. Distribution of Parking Events by Parking Area

This visualization shows the total number of parking events in each zone. Each zone is represented on a geo-map with a proportional circle where the center of the circle is the geo location of the parking zone, Figure 4. The radius of the circle depends on the total number of parking events in the zone, that is, the greater the number of parking events, the greater the radius and thus the larger circle on the geo map.

Clicking on one of the zones will display an info window showing parking zone information:

- Percentage of all parking events - What is the percentage of parking events in this zone out of the total number of parking events in all parking zones?
- Rank - Ranking compared to other zones in terms of percentage of parking events
- Parking zone information - Parking zone information such as name, number and city



Fig. 6: Geographical distribution of parking events during weekend

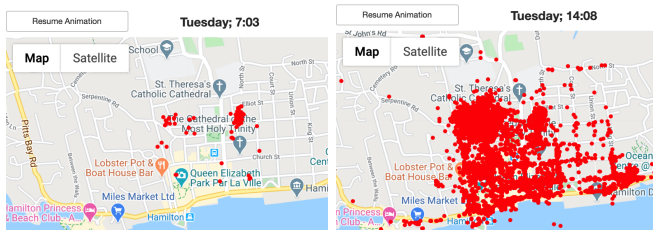


Fig. 7: Point map of parking events at chosen time and day

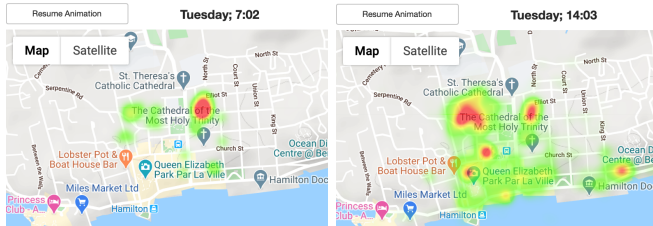


Fig. 8: Heat map of parking events at chosen time and day

D. Average Parking Duration Per Day of The Week and Per Parking Area

This visualization shows the average duration of a parking event in minutes per day of the week for each parking zone, Figure 2. On the right, the legend shows the average duration of a parking event for each day of the week. This average duration is calculated based on all parking events grouped by day from all parking zones. By clicking on the appropriate day, the graph changes and displays the average duration of a parking event for that day and for each zone respectively.

E. Geographical Distribution of Parking Event

In this part, on a map is shown the density of parking events for each day of the week. Because there is an identical distribution pattern for the days from Monday to Saturday, in this paper are shown the map of density of parking events for Friday and Sunday, Figure 5 and Figure 6. The geo location of parking events is obtained from the mobile apps for Android and iOS platforms. When the user launches the parking event through the mobile app, besides the parking data (parking zone, registration, user etc.) if available, data is also sent on the current location of the user (mobile device).

III. ANIMATIONS

The animations depict parking events. Each parking event is displayed on a geo map depending on the start date and time of the event and stays on the map until the hour when the parking event is over. Two types of animations are visualized:

- Density animation of parking events. As shown in Figure 8, the animation applies to 28/03/2018 and there can be a different density of parking events at different times of the day.
- Detailed map showing each parking event. As shown in Figure 7. So, the animation applies to 28/03/2018 and a different distribution of individual parking events can be observed at different times of the day.

The constant rise of everyday car users and lack of available parking slots, results in having many improper parked vehicles and consistent traffic jams. Not only does this lead to large amount of driver's time wasted, but also increases the air pollution. When implementing a smart parking system as a component of a smart city, it is known for a fact that a large amount of time is needed for doing a proper research by inspecting all the areas and their parking lots occupancy. In order to help speed up that process, by having an easier way of pattern recognition and spotting repetitive problems, we provide a dashboard with simple and straightforward visualized preview of data. Another benefit of having different types of data visualized, like daily income, number of parking events per hour and area, as well as trends in earnings or number of parking events, is to show different companies from economical and ecological spheres how they could be affected and what they can do to improve the situation.

The next step that we are planning to do is to incorporate the collected data in a much bigger system which will be used for smart parking prediction and navigation to nearby available space in real time, according to data from previous similar events. The visualizations will help easily spot and distinct areas where there is a need of implementing this system, from areas that there is not.

ACKNOWLEDGEMENT

We would like to thank Pelagus IT, the company where Ivan had worked while writing this paper, and PARX Ltd. for their assistance, informative discussion and for providing us with the parking data for visualization, from the EasyPark Mobile System used in markets around the globe for the past decade.

REFERENCES

- [1] G. Yan, W. Yang, D. B. Rawat, and S. Olariu, "Smartparking: A secure and intelligent parking system," *IEEE Intelligent Transportation Systems Magazine*, vol. 3, no. 1, pp. 18–30, 2011.
- [2] H. Zhao, L. Lu, C. Song, and Y. Wu, "Ipark: Location-aware-based intelligent parking guidance over infrastructureless vanets," *International Journal of Distributed Sensor Networks*, vol. 8, no. 12, p. 280515, 2012.
- [3] M. Boltze and J. Puzicha, "Effectiveness of the parking guidance system in frankfurt am main," *Parking Trend International*, pp. 27–30, 1995.
- [4] R. Grodi, D. B. Rawat, and F. Rios-Gutierrez, "Smart parking: Parking occupancy monitoring and visualization system for smart cities," in *SouthEastCon 2016*. IEEE, 2016, pp. 1–5.
- [5] D. Teodorović and P. Lučić, "Intelligent parking systems," *European Journal of Operational Research*, vol. 175, no. 3, pp. 1666–1681, 2006.
- [6] A. Khanna and R. Anand, "Iot based smart parking system," in *2016 International Conference on Internet of Things and Applications (IOTA)*. IEEE, 2016, pp. 266–270.
- [7] S. Mathur, T. Jin, N. Kasturirangan, J. Chandrasekaran, W. Xue, M. Gruteser, and W. Trappe, "Parknet: drive-by sensing of road-side parking statistics," in *Proceedings of the 8th international conference on Mobile systems, applications, and services*, 2010, pp. 123–136.
- [8] M. Manville and D. Shoup, "Parking, people, and cities," *Journal of urban planning and development*, vol. 131, no. 4, pp. 233–245, 2005.
- [9] T. Lin, H. Rivano, and F. Le Mouél, "A survey of smart parking solutions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 12, pp. 3229–3253, 2017.
- [10] F. Zhou and Q. Li, "Parking guidance system based on zigbee and geomagnetic sensor technology," in *2014 13th International Symposium on Distributed Computing and Applications to Business, Engineering and Science*. IEEE, 2014, pp. 268–271.

Home security system based on drone automation - IoT approach

Darko Kostadinov
School of Computer Science and
Information Technology
University American College Skopje
Skopje, Macedonia
kostadinov.darko91@gmail.com

Veno Pachovski
School of Computer Science and
Information Technology
University American College Skopje
Skopje, Macedonia
pachovski@uacs.edu.mk

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

Abstract—Drones have already established a significant role in many industries. As technology progresses, in the next few years, drones will be considered useful tools for a variety of consumers. The paper aims to show the benefits and advantages of using a drone as a dynamic security camera (strengthened by using image recognition software) for home security purposes. Two applications were created as proof of concept to present the idea. This approach can be beneficial not only for home security, but also for other industries such as: cinematography, agriculture, energy, civil engineering and constructions, safety, security in general, governmental use, oil refining and other environmental sciences.

Keywords— Home security, IoT, drone, image recognition, web sockets, tensorflow.js, AWS, Node.js

I. INTRODUCTION

Internet is one of the most important things ever created which affected humanity and the modern lifestyle. Internet of things, or IoT, is a system of connected computing devices that have the ability to transfer data over a network without the need for human interaction [1]. They are smart devices that are using built-in sensors or other embedded systems to collect and send data that they have acquired from the environment around them [2]. In the last couple of years, the hype around the Internet of things has been constantly rising and will continue to do so. The number of connected devices installed worldwide is increasing and that trend won't stop any time soon. The total installed base of IoT connected devices is projected to amount to 75.44 billion worldwide by 2025, which is a fivefold increase in ten years [3]. The increasing number of connected devices helped the global market value for IoT to rise continuously. A report done by Fortune Business Insights, shows that the overall market for IoT, which was valued at \$190 billion in 2018, is expected to reach \$1,111.3 billion by 2026 at a yearly improvement rate of 24.7%. The banking and financial industries are expected to occupy the biggest part of that sum [4].

Drones have already been used for various purposes. However, the growth and variety of their usage will continue to rise in the following years; there will be increase in productivity, safety, wellbeing, and even new job openings for the drone industry. Some of other advantages are reduced manual labor and lower risk (since there will be no need for locating labor in dangerous areas), as well as getting a more extensive field of view that drones provide. Nowadays, drones are used not only for military purposes. They also made a big leap in the consumer market; forecast for the end of 2020 predicts \$100 billion market value for drones which is generated by the increased demand from the commercial and civil government sectors. Although the defense sector will

have the biggest share of the market, with a projected \$70 billion by the end of 2020, consumer market is expected to be the runner up with \$17 billion in market share, while the commercial/civil usage projection is evaluated to be \$13 billion [5].

Home security plays an important role in our lives and in people's wellbeing in general. Modern era allows us to have UAV patrolling the house and letting us know of any inconsistencies and intruders. The idea of this paper is to give a contribution in that direction. The approach presented further on, relies on two custom applications that communicate with each other using web sockets. The drone app is communicating with the drone hardware and depending on the video stream and the data provided by the drone, makes decisions such as autopilot, auto-land, object detection, face detection, etc. The video frames are analyzed using tensorflow.js, which is an open-source library for building, training, or running machine learning models entirely in the browser [6]. The library is providing coordinates and sizes of the recognized objects in the frame and a given command is executed. Also, the drone application is sending notifications to the second application which serves as a mobile app, so that the user/landlord can preview flights or be notified if needed. The data provided by the drone is saved on Amazon Web Service (AWS), in particular the DynamoDb service and the Simple Storage Service (S3). DynamoDb is a managed NoSQL service that has strong consistency and stores the data as key-value pairs. Also, it doesn't require complex manual setup [7]. The drone application uses S3 to store the video data from the whole flight. Amazon S3 serves as a storage that is designed for large-capacity at a lower price and can be easily provisioned to other geographical regions [8].

II. THE ARCHITECTURE OF THE MODEL

A. System Overview

The location of the research is a backyard with different objects added to create continuous changing scenery, and thus enriching the viewport of the image recognition software during the automated flights of the drone. The hardware used in the research is: Dell XPS 15 laptop serving the drone application hosted on localhost with a node/express.js application and also a mobile application which is presenting the data fetched from AWS DynamoDB and AWS S3 bucket. The actual drone hardware is a DJI Tello which is providing a

wireless network that can be accessed using three ports: video stream port, control commands port and read commands port.

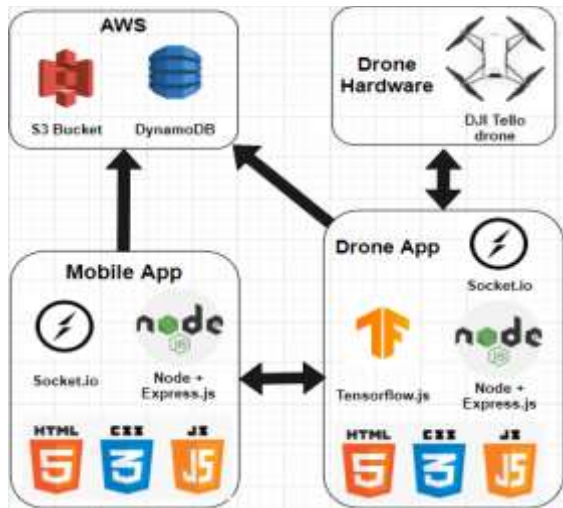


Fig. 1. Overview of software architecture

The drone SDK is providing various text commands that can be used to control the drone [9].

Drone Application

When the application is bootstrapped for the first time, the autopilot path is initialized with all the necessary variables. Then, the models from tensorflow.js are fetched including faceapi.js which is built as a wrapper from tensorflow.js to help with the face recognition feature in order to differentiate if the person is known or unknown. The recognized faces are first fetched from DynamoDB and later from an S3 bucket.

Next, a web socket connection is created with the drone; one for the video feed and another for read commands and control communication. The drone is flying on a predefined path around the house on autopilot, but when it detects an unknown person, it sends a notification to the user.

The landing is also dependent on image recognition. If the landing pad sign is in sight and the flight time is near the end, the drone starts adjusting for better accuracy before landing, so it can land on the same spot where it took off. The flight, with all the executed commands and detected faces during flight are saved on DynamoDB while the video of the flight is saved on the S3 bucket.

B. Mobile application

This application is used by the user/landlord and it has the features for previewing history of flights and familiar faces. Moreover, it has a feature for adding a new face which is directly updated and used instantaneously by the drone application.

III. RESULTS AND ANALYSES

A. Object recognition

Object recognition is one of the most important features of this application. Therefore, it requires a systematic approach. So far, there are three classes of 'familiar' objects, in which the stop sign is a class by itself. At present, there is only one image with that classification, but there could be much more, for reasons of security or obscurity.

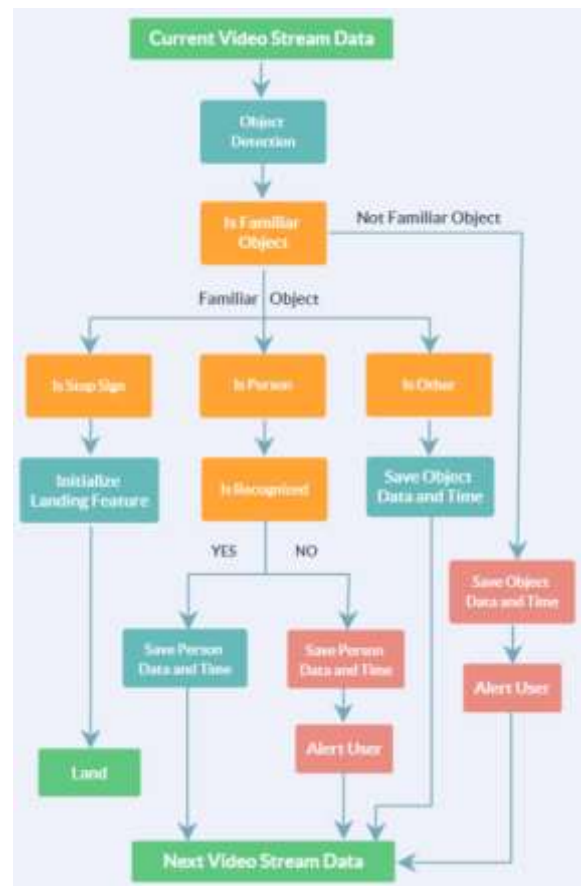


Fig. 3. Workflow of the image recognition

For the purposes of this research, the drone application is tested using dogs as objects. The results are based on the analysis of 100 screenshots taken from the drone camera - both moving and non-moving objects and 4 different dogs, all of them also from different breed. The research was done using the coco-ssd model which is capable of detecting 90 types of objects. It managed to recognize 73% of the objects provided as type dog. Each of the predictions has an appropriate confidence score, presented in the table below.

TABLE I. OBJECT RECOGNITION

Confidence Score	Dog
<0.6	4%
0.6-0.7	8%
0.7-0.8	15%
0.8-0.9	34%
0.9>	39%



Fig. 2. DOG images – an example of recognition

The efficiency of the system depends on the object current rotation or the drone perspective which is constantly changing, so in the next frame the recognition would get the correct object type. When tested in a video of 1 minute, while the object was in the viewport of the camera and constantly moving, it was correctly recognized in more than 90% of the time.

B. Face Recognition

In order to determine if a person is known or unknown, the faceapi.js library is used, built also with tensorflow.js which is based on the familiar faces fetched from AWS DynamoDB and S3 bucket. It waits until an object of type person is in the viewport and then activates the face recognition feature which tries to compare familiar faces with the one in the viewport. This was tested by having a person in front of the drone. It was successful in recognizing a person at distances of 1 and 3 meters, but it failed at 5 meters.

All of the data provided above is generated with the built-in camera of the drone which has a resolution of 720p. If equipped with a better camera, the results can be significantly improved; with a clearer image, it will be easier for tensorflow.js to predict objects.

C. Landing analyses

Since the drone does not have any GPS hardware, the landing feature is based on image recognition in order to land on the same point it took off. The data provided by tensorflow.js for each of the prediction boxes and its dimensions, are used to determine if any adjustment is needed (up, down, left, or right). Then the drone moves forward until it reaches the distance that is needed to position in front of the sign for landing, and, if everything is ok, it executes the landing.



Fig. 4. Drone in front of the landing image

The landing feature of the drone application has a result of 98% successful landings and in only 2%, where the drone missed the landing pad. The following heat map shows the points of (each of the) 50 landings done on landing pad with dimensions 70x70 centimeters:

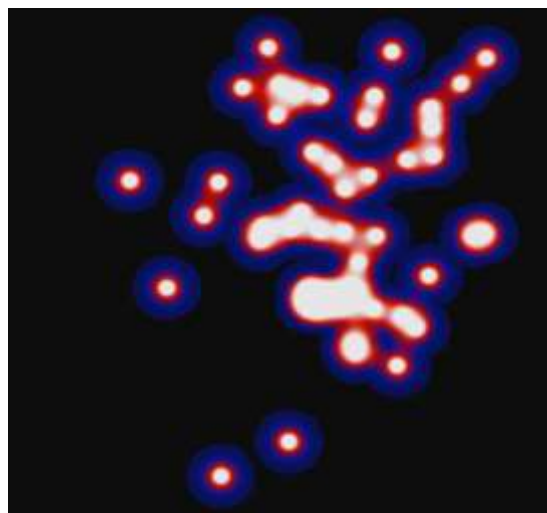


Fig. 5. Heatmap containing each of the landing points

With the current implementation of the drone application software and taking into consideration that the DJI Tello drone is a small one, when even light weather conditions can easily affect it, it still managed to execute more than 90% of the landing on a dimension of less than 50x50 centimeters

D. Aws Rekognition

AWS Rekognition is a visual analysis service from Amazon that uses deep neural network models, so it can detect objects, faces, and text in your images [10]. Since all of the results presented in Table I were done by using tensorflow.js, and in order to make a comparison between open source and proprietary software, the images were put to the test but this time using AWS Rekognition. It produced much better results and recognized 94% of dog type. The recognition results are presented in the table below.

TABLE II. AWS REKOGNITION

Confidence Score	Dog
<0.6	2%
0.6-0.7	3%
0.7-0.8	5%
0.8-0.9	24%
0.9>	66%

Compared with the confidence score (Table I), this test shows much higher confidence.

Again, as previously mentioned, with a better camera which has better image stabilization, the results would have been significantly improved (with a clearer image, either tensorflow.js or AWS Rekognition will be more precise at predicting objects).

Due to the drone camera capabilities, all of the image data tested was done during daylight. For night testing, there are more advanced drone cameras with night sight mode. where these methods can be applied and consequently, see what kind of results will they provide.

IV. ADDITIONAL FEATURES AND IMPROVEMENTS

The drone application also has other features such as voice recognition and speak functionalities, by using the built-in objects from the browser: SpeechRecognition and SpeechSynthesisUtterance. Currently, the drone has the

feature of understanding English sentences and executing drone commands by a given user command, and also replying back in English for each of the commands. This could be helpful for industries where a user needs to exercise a voice control on the drone (by a set of predefined voice commands). Finally, text generation feature could be included; to generate a full report about the outcome of the patrol - flight path registered and/or images discovered.

In order to make the recognition more accurate, an implementation of AWS Rekognition can be added as a feature to the drone application, so that the user can have a more detailed analysis.

Next step in improving this home security system would be to add a wireless charging functionality. A charging coil could be added to the drone and a wireless charging pad to the landing location, so that the whole system would become autonomous.

V. CONCLUSION

Research presented in this paper was based on two custom web applications connected to AWS and a DJI Tello drone. The data that was provided by the drone video stream was analyzed by image recognition software on different video frames during the drone flights. Moreover, there have been numerous tests for the landing feature of the drone application, which used only image recognition to decide/detect the point of landing. Recorded videos and images were also run through AWS rekognition in order to compare open source and a proprietary image recognition software. Finally, possible improvements and ideas for further investigations were suggested.

REFERENCES

- [1] Margaret Rouse , “Internet of things”, 2016, [Online] Available: <https://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT> , [Accessed: 06- Mar- 2020]
- [2] Jen Clark , “What is the Internet of Things”, 2016, [Online] Available: <https://www.ibm.com/blogs/internet-of-things/what-is-the-iot/> , [Accessed: 06- Mar- 2020]
- [3] Statista Research Department, “Internet of Things (IoT) connected devices installed base worldwide from 2015 to 2025”, 2016, [Online] Available: <https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/> , [Accessed: 06- Mar- 2020]
- [4] fortunebusinessinsights.com , “Internet of Things (IoT) Market”, 2019 [Online] Available: <https://www.fortunebusinessinsights.com/industry-reports/internet-of-things-iot-market-100307> , [Accessed: 06- Mar- 2020]
- [5] goldmansachs.com , “Drones reporting for work”, [Online] Available: <https://www.goldmansachs.com/insights/technology-driving-innovation/drones/> , [Accessed: 06- Mar- 2020]
- [6] Josh Gordon and Sara Robinson , “Introducing TensorFlow.js: Machine Learning in Javascript”, 2018 [Online] Available: <https://medium.com/tensorflow/introducing-tensorflow-js-machine-learning-in-javascript-bf3eab376db> , [Accessed: 06- Mar- 2020]
- [7] Chandan Patra , “Amazon DynamoDB: 10 Things You Should Know”, 2019, [Online] Available: <https://cloudacademy.com/blog/amazon-dynamodb-ten-things/> , [Accessed: 06- Mar- 2020]
- [8] Hemant Sharma , “Deep Dive into Amazon Simple Storage Service”, 2019, [Online] Available: <https://www.edureka.co/blog/s3-aws-amazon-simple-storage-service/> , [Accessed: 06- Mar- 2020]
- [9] ryzerobotics.com , “Tello SDK”, [Online] Available: https://dl-cdn.ryzerobotics.com/downloads/tello/20180910/Tello%20SDK%20Documentation%20EN_1.3.pdf , [Accessed: 06- Mar- 2020]
- [10] amazon.com , “Amazon Rekognition FAQs”, [Online] Available: <https://aws.amazon.com/rekognition/faqs/> , [Accessed: 06- Mar- 2020]

Exploratory data analysis and statistical inference for students` results on Discrete Mathematics and Probability and Statistics at Faculty of computer science and engineering

Lenche Jovova
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
lenche.jovova@students.finki.ukim.mk

Abstract— During the past decade Faculty of Computer Science and Engineering (FCSE) has become the most attractive faculty in North Macedonia. The interest for this faculty arises every year and this motivated me to investigate the student's results in Probability and Statistics and Discrete Mathematics and to check for correlation with the results from the Final state exam. Hypothesis testing and exploratory data analysis were made, and the results are presented.

Keywords— *hypothesis testing, exploratory data analysis, statistical inference, FCSE*

I. INTRODUCTION

Exploratory data analysis (EDA) is a powerful framework when it comes to having a big picture of which properties describe the process or the population of interest. It can never be the whole story, but nothing can serve as the foundation stone, as the first step [1]. When data analysis is done, EDA is often the first step taken after data is cleaned. On the other hand, to know how reasonable our model or our beliefs for the data are, hypothesis testing come into play. Our beliefs for the process or the population can be justified or not depending on the sample that describe the population. To draw conclusions for our process or population of interest, Statistical inference can be used.

In this paper, EDA and Statistical inference are used for drawing conclusions for the student's results for Probability and Statistics and Discrete Mathematics. In Section II the used datasets are described, Section III includes the analysis of the results from Probability and Statistics and Section IV analyzes the correlation between the results from Discrete Mathematics and Probability and Statistics.

II. DATASET DESCRIPTION

For the course Discrete Mathematics data is available from the last three academic years (2016/2017, 2017/2018, 2018/2019) and for Probability and Statistics data is available

from the last four academic years (2015/2016, 2016/2017, 2017/2018, 2018/2019).

The datasets for Discrete Mathematics and Probability and Statistics include the following attributes:

- **Id number** – Identification number of the student, ordinal type
- **Name and Surname** – First name and Last name of the student, string type
- **Program**–the program the student follows, categorical variable
- **Professor** – the professor that teaches the course, string type
- **Total Points**– Total points the student has for the subject, continues data type
- **Grade** – Final grade of the student

Additionally, for Probability and Statistics despite the above attributes, the following are also used:

- **ExercisePart1** – points gained on the first partial exam on exercises
- **ExercisePart2** – points gained on the second partial exam on exercises
- **TheoryPart1** – points gained on the first partial exam on theory
- **TheoryPart2** - points gained on the second partial exam on theory

III. ANALYSIS OF THE RESULTS FROM PROBABILITY AND STATISTICS

The focus of the analysis for this subject is to compare the results for the students that follow the course regularly

according to their year of enrollment in the faculty and the students that have previously followed the course but have not passed it. Also, the percentage that pass the subject from the regular students will be analyzed and the rank of the students according the results for Probability and Statistic will be compared with the rank of the students according their results from the Final state exam.

A. Regular vs. Non regular students

According to the data from the first and second partial exam for exercises, the regular and the students that once followed the subject seem to have different distributions.



Fig. 1. Probability density function for the points of the first partial exam

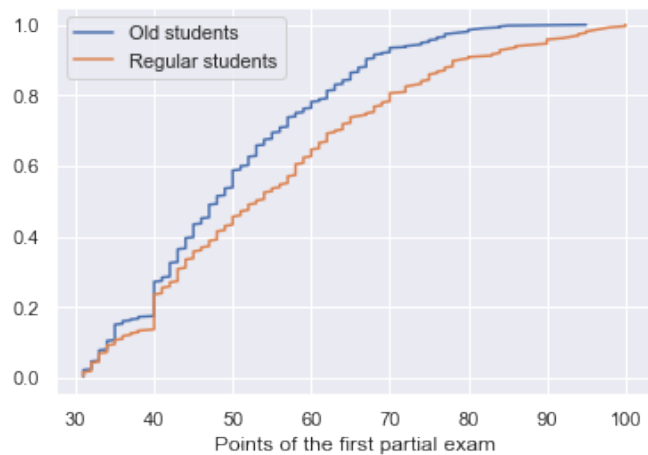


Fig. 2. Cumulative density function for the points of the first partial exam

From the graphs themselves is obvious that the regular students have slightly better performance on the exam. The distributions seem to have concession from normality, and we will perform Kolmogorov-Smirnov test, a non – parametric statistical test which tests the null hypothesis that the two groups come from the same population based on their cumulative density functions [2]. The test computes the distance between the cumulative density functions and based on that calculates Kolmogorov-Smirnov Statistic.

H_0 : The two groups come from the same population

H_a : The two groups come from different populations

Tested groups: Old students that followed the course previously and failed and students that follow the subject regularly.

Statistical Test: Kolmogorov-Smirnov

Level of Significance: 0.05

When the test was performed on the two groups, a p-value of 0.0002 was obtained, which is below the level of significance of 0.05, hence I reject the null hypothesis in favor of the alternative.

The data from the second partial exam seem to slightly have some normality, as shown in the figure below:

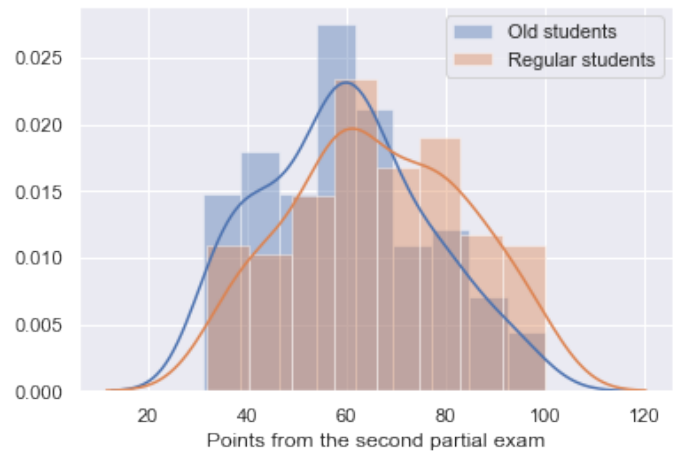


Fig. 3. Probability density function for the points of the second partial exam

For this data t-test for the mean of the two groups was performed. The two-sample t-test is parametric test for testing the means of two groups. The result of the test is a test statistic based on the difference of means and the pooled variance of the two groups [3].

H_0 : The two groups have the same mean

H_a : The two groups have different means

Tested groups: Old students that followed the course previously and failed and students that follow the subject regularly.

Statistical Test: t-test for means

Level of Significance: 0.05

When the test was performed on the two groups, a p-value of 0.0001 was obtained, which is below the level of significance. The null hypothesis is rejected in favor of the alternative. Additionally, the test statistic has a value of – 3.82 which suggest that the regularly enrolled students for the course have higher mean than the students that previously failed the subject.

B. Analyze of the frequency of students that passed the subject throughout the years

For the available data I calculated contingency table in order to analyze the number of students that passed the exam

successfully (either on partial exams or in any exam session) versus the number of students that did not pass any of the exams in the respective academic year. The analysis is done only for the students that regularly follow the course, i.e. students that follow the course for the first time. The table I bellow shows the number of the students that passed and did not pass the exams for Probability and Statistics.

TABLE I. CONTINGENCY TABLE FOR THE NUMBER OF STUDENTS THAT PASSED VERSUS THE NUMBER OF STUDENTS THAT FAILED

Academic Year	2015/16	2016/17	2017/18	2018/19	All
Passed					
No	114	122	156	175	567
Yes	69	54	72	82	277
All	183	176	228	257	844

Chi2 test for the contingency table will be performed. The null hypothesis that will be tested is the ratio of the passed versus failed students is not changing throughout the year [4].

H_0 : The ratio of the passed/failed students is not changing throughout the years

H_a : The ratio of the passed/failed students is changing throughout the years

Tested groups: Students that regularly follow the course through four academic years

Statistical Test: Chi2 test for contingency table

Level of Significance: 0.05

When the test was performed, a p-value of 0.95 was obtained which suggests that we should accept the null hypothesis. Given the data from the last 4 years, the percentage of the students that have passed the exam has not changed.

C. Comparing the rank according the total points on Probability and Statistics and the rank from the total points on the Final state exam

The Final state exam is a national exam consisting of 4 subjects that the students must pass in order to make an application for Faculty enrollment. The students that want to enroll for FCSE must have chosen Mathematics in the Final state exam. The last three numbers (or the last two numbers for the program KNIA on FCSE) from the Id number that the student get depend on the Rank the student have according the points from the Final state exam, i.e. the student that had the highest points on the Final state exam gets Id number xxx001. We will compare if the first students of the Final state exam are the first students on the exams of Probability and Statistics as well. For this purpose, we will use Kendall Tau statistic for ordinal data [5]. The Kendall Tau statistic is based on the difference between the number of concordant and the number of discordant pairs in the two vectors. The Tau test is a non-

parametric test for statistical dependence based on the Kendall Tau statistic.

The tested hypothesis and the groups, as well as the test and level of significance are the following:

H_0 : The two groups are positively dependent

H_a : The two groups are independent

Tested groups: Rank if the students according to the Final state exam and results from the exams from Probability and Statistics from the last 4 years

Statistical Test: Tau

Level of Significance: 0.05

The test resulted with very low p values and more important, tau correlations very close to zero, for every year. These results not only lead to rejecting the null hypothesis, but also lead to the conclusion the results from the Final state exam maybe are not relevant.

IV. CORRELATION BETWEEN DISCRETE MATHEMATICS AND PROBABILITY AND STATISTICS

I wanted to check if there is any correlation between the results on Discrete Mathematics and the results on Probability and Statistics for the same students. On the scatter plot on Figure 4 are shown the results from both subjects. There can be noticed a lot of outliers, students that have gained very high points in Discrete Mathematics, have very low points in Probability and Statistics.

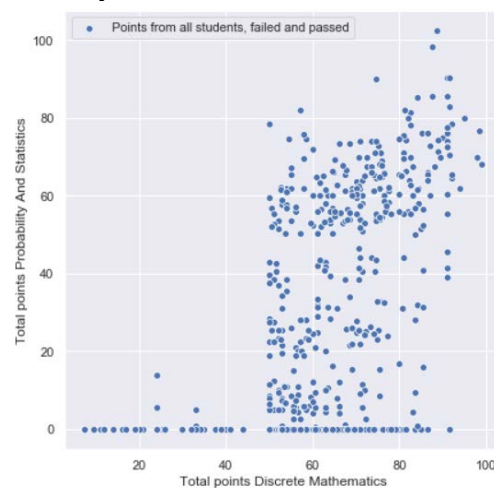


Fig. 4. Discrete Mathematics versus Probability and Statistics for all students

A moderate linear relationship is present and can be noticed especially in that part of the scatter plot where are the students that have gained more than 30 points in Probability and Statistics. On the scatter plot in Figure 5 is shown only that part of the plot.

V. CONCLUSION

From the obtained results can be concluded that the students who follow Probability and Statistics for the first time have slightly better results than the students which have once followed the same subject. In the second group there are also high scores and further investigation need to be done to check whether the previous time they followed Probability and Statistics, they had focus on another subject, e.g. Algorithms and Data structures, which is in the same semester as Probability and Statistics and is also extensive course.

The ratio of the passed/failed students that is not changing through the years according to the results, suggests that there is a constant percentage of students that would pass Probability and statistics on their first following of the course.

The lack of correlation with the results from the Final state exam can be a sign that this condition for enrolling on the faculty is not the best choice, but to validate this, in future more courses will be included in the investigation.

VI. FUTURE WORK

The goal was to check differences between regular and students that once followed the course, investigate relationship between the mentioned subjects and check for correlation with the Final state exam. My future work will include more detailed and extensive analysis as well as more subjects and more academic years. The obtained results can be then used to draw more precise conclusions that could help improving the faculty's strategy.

ACKNOWLEDGMENT

Thanks to Marija Mihova, PhD for the provided anonymized datasets.

REFERENCES

- [1] John W. Tukey, " Exploratory Data Analysis. Addison-Wesley", 1977,pp.1-3
- [2] Jean Dickinson Gibbons,Subhabrata Chakraborti, "Non parametric statistical inference", Fourth Edition , Revised and Expandedn , pp.239-244
- [3] Konstantin M.Zuev , "Statistical inference", unpublished
- [4] https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient
- [5] Peter D.Hoff, "A first course in Bayesian statistical methods", Fourth Edition , Revised and Expandedn , pp.149-154

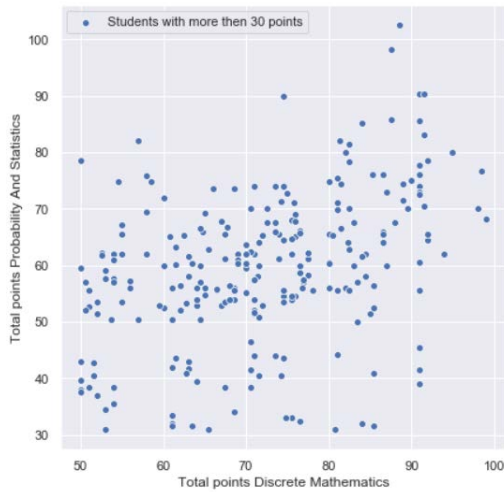


Fig. 5. Discrete Mathematics versus Probability and Statistics for students that have more than 30 points in Discrete Mathematics

A linear model will be made for the two variables Total Points in Discrete Mathematics and Total Points in Probability and Statistics for the students that obtained more than 30 points Probability and Statistics. The Pearson correlation coefficient for the two variables is 0.51 which confirms that there can be a moderate linear relationship. A linear regression model with ordinary least square method was performed [5]. The results from the model showed that the Total Points in Discrete Mathematics is significant variable and for every 1 point in Discrete Mathematics the total points for Probability and Statistics will increase for 0.43 The regression line is given in the figure 6.

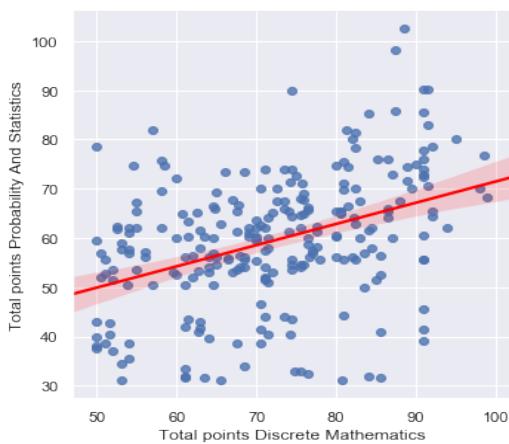


Fig. 6. Linear regression line for the model

The model leads us to the conclusion that based on the results in Discrete Mathematics, we could know what to expect in Probability and Statistics for the students that will pass the exam.

Quality of Online Teaching in Higher Education – the Case of South East European University (SEEU), North Macedonia

Veronika Kareva
Faculty of Languages, Cultures and Communication
South East European University
Tetovo, North Macedonia
v.kareva@seeu.edu.mk

Daniela Kirovska-Simjanoska
Language Center
South East European University
Tetovo, North Macedonia
d.kiroska@seeu.edu.mk

Abstract- Online teaching has become obligatory for all higher education institutions in North Macedonia, like in many other countries in the world, during the pandemic crisis. Various platforms and numerous opportunities have been on disposal to university professors depending on their ability to utilize them, their creativity and willingness. The aim of this paper is to present the results of the first stage of a research project about the effectiveness of the online teaching in higher education. Its purpose is to provide answers to the following questions: Which instruments can be used for measuring the effectiveness of the online teaching? To what extent are university professors knowledgeable about the online activities that are the most engaging for students and bring about the best results? How can they gain more information about the quality of their online instruction? The research will examine the following main hypothesis: the characteristics of good quality online teaching resemble closely those of the traditional classroom. Data about online teaching activities of one hundred and ten (110) professors at the South East European University (SEEU) are being collected and analyzed. Results from a student evaluation survey of one hundred and sixteen (116) undergraduate courses at SEEU are examined in order to be compared to the results of the evaluation of the same courses taught online in the next phase of the project. Recommendations based on the conclusions are expected to be useful to policy makers, university management and the professors in order to organize and deliver good quality online instruction.

Keywords—quality teaching, educational platforms, higher education

I. INTRODUCTION

The application of technology to the learning and teaching process or the so called digital or online learning started to boost in the early 1990s when word processors were introduced in schools. The commercialization and wider utilization of the Internet created many new opportunities and development of different categories of digital technology that could assist learning: they ranged from simpler ones, such as interactive boards and projectors, to more sophisticated interactive technologies that enabled application of modern pedagogical approaches. Online quizzes, flipped classroom,

virtual reality, massive open online courses etc. are some of the most recent trends in digital learning (Carrier, 2017). Nowadays, technology prevails in all spheres of life and because of that, students of present time are used to learning with the help of technology. The so-called digital natives are surrounded by technology and learn best by application and interaction with peers through multimedia (Liton, 2014).

Responding to the newly created circumstances due to the COVID 19 pandemic crisis, the Ministry of Education of the Republic of North Macedonia (RNM) has adopted a regulation in a very short time, according to which online education and distance learning has become a must for all higher education institutions during that period. South East European University (SEEU) started with the online instruction even before the governmental decision. From its foundation, it has utilized different learning management systems (LMSs). In the past five years, Google Classroom (GC) use has been obligatory for all teaching staff with mechanisms for control of its application. With the shift to entirely online instruction, every member of the academic staff has been obliged to meet his/her students in real class time, according to the regular schedule, and in addition to that to continue with the already established practice of posting materials and assignments on GC.

1.1 Institutional Quality Assurance (QA)

Since the establishment, as a result of the international orientation and the then lack of national strategy for QA, SEEU has focused on development of a QA system that is heavily student related. It has introduced a quality structure and methods of external review which actively add insight into strategic and operational planning and improvement, but has also focused on internal procedures for enhancing learning and teaching, including and considering students and their feedback. Internally, there are different instruments for getting student feedback and ensuring student-centered learning and teaching. The staff is encouraged constantly to use modern teaching methodologies such as flipped classroom and digital technologies in order to be in line with the international trends. Student evaluation of academic courses is one of the instruments for QA. It is conducted

online at the end of every semester. At least one course per professor is evaluated through an anonymous survey utilizing GC.

The quality of learning and teaching is maintained and developed through an annual Teaching Observation scheme, yearly student evaluations, individual staff evaluation and professional development opportunities.

1. Observation of learning and teaching

One of the most significant factors in ensuring that we are continuously improving and developing what we are providing to students is the quality of our teaching. Therefore, the University has an annual observation scheme in order to:

- support the University's strategic aim of continuous improvement and development of learning and teaching
- provide evidence of quality assurance at Faculty and University level
- ensure that the students' learning experience is of the highest quality across each Faculty
- acknowledge excellent practice and facilitate the sharing of good practice across each Faculty and the University
- support continuous, individual staff development
- inform other relevant processes, specifically, the annual self-evaluation process and the allocation of staff bonuses
- ensure that learning and teaching is inclusive and addresses the University's commitment to equality of opportunity

2. Student Evaluation

In order to give students the opportunity to evaluate academic staff, courses and the effectiveness of the administrative services and facilities, students at SEEU complete an annual Evaluation Survey. This is implemented in a confidential and anonymous way so that we receive honest and constructive comments to help us improve.

3. Staff Evaluation

The University recognizes that staff are a key resource and aims to acknowledge achievement, support continuous development and manage individual and overall performance. Therefore, there is an annual process of evaluation for both academic and administrative staff which meets these aims and provides information for the processes of contract renewal and promotion. The procedure utilizes other evaluative mechanisms and evidence, for example, success data, teaching observation and student evaluation.

1.2 The Problem

During regular teaching periods, GC is used only as additional support to traditional classroom teaching, while when the University was closed for students, it became the only way for organizing the instruction. Such a rapid move to complete online operation did not allow any time for preparation and training of the academic staff for full and successful implementation of the teaching process with all its components, such as presentation of new material, interaction with students, practice, assessment and feedback. Among many challenges for the institution that aroused from this change, the issue of quality was the one that could be addressed with immediate action. The aim of this paper is to analyze the effectiveness of the online teaching at university level and answer some of the following questions: Which

instruments can be used for measuring the effectiveness of the online teaching at tertiary level? To what extent are university professors knowledgeable about the online activities that are the most engaging for students and bring about the best results? The main research hypothesis that the study in progress will investigate is that the characteristics of good quality online teaching resemble those of the traditional classroom. In other words, we intend to prove that no matter what the mode of instruction is, the characteristics of good teaching, as pointed out in the literature and perceived by students, are the same.

II. LITERATURE REVIEW

Researchers argue that the main role of any kind of instruction is to promote learning (Anderson, 2008; Smidt, Bunk, Kochem and McAndrew, 2017) and that the online teaching is not an exception to this. According to Shelton (2011), since its emergence, online education has been critiqued and compared to traditional teaching and these criticisms implied low quality. As a result, many different approaches for evaluating quality online education exist in the literature. The same author points out a model by Lee and Dziuban (2002, in Shelton, 2011), suggesting that "the overall success of online education greatly depends upon the quality evaluation strategies integrated with the program" (para V). Constructivism, as one of the most influential learning theories is considered to offer the basic principles underling the quality of online teaching (Koohang, Riley and Smith, 2005). According to the constructivists' view, knowledge is constructed based on learner's prior experience. These authors advocate a model for online teaching based on constructivism, which includes three categories: the design of learning activities, learning assessment and instructor's role. The design of learning activities includes collaboration, cooperation, multiple perspectives, real world examples, scaffolding, self-reflection, multiple representations of ideas, and social negotiation. The learning assessment consists of instructor assessment, collaborative assessment, and self-assessment while the instructor's role is the one of a coach, mentor, feedback provider and assessor. According to this model, good quality online courses have to integrate all the previously mentioned elements. Furthermore, researchers also claim that the experience of a distance learning student should be "as rich, both intellectually and affectively, as the experience of a student in a traditional classroom" (Bower, 2001, p. 3). In this regard, Ascough (2002, in Yang and Cornelious, 2005), points out that because of the fact that most instructors have been trained in traditional instruction, it would be challenging for them to adjust to the role of a facilitator instead of being the leading speaker in class. Smidt, Li, Bunk, Kochem, and McAndrew (2017) advocate a few main quality features that should be emphasized in every online course: clarity, availability, feedback and interaction. The student evaluation survey that will be used as an instrument for measuring the quality of online teaching at SEEU has been created based on the principles of constructivism, pointed out in the literature and the last mentioned quality characteristics.

III. METHODOICAL APPROACH

In order to test the main research hypothesis and examine if the characteristics of good quality online teaching resemble the ones in a traditional classroom, as well as to answer the research questions, the following actions are or will be undertaken:

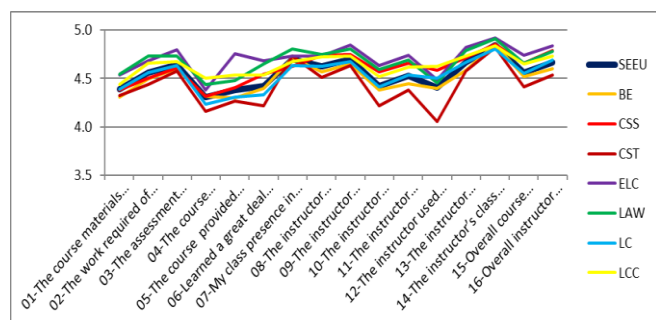
1. Analysis of the weekly reports of all GC activities of 110 professors at undergraduate level – completed already
2. Creation of a student evaluation survey based on the characteristics of good quality online teaching described in the literature - created (Appendix 1)
3. Distribution the survey to students of 116 undergraduate courses – to be completed at the end of the current semester
4. Comparison of the results from this student evaluation survey at individual, Faculty and University level, with the results from the student evaluation of the same professors in the previous semester – final stage
5. Bring conclusions and offer recommendations

IV. INITIAL FINDINGS

The analysis of the reports of all teaching activities by 110 professors at undergraduate level has shown that the main self-reported activities have been lecturing and posting materials on GC. It was very indicative that only few professors have entered in the reports description learning experiences that promote online interaction in terms of a paradigm shift to a student-centered classroom with students as active participants and not passive receivers of knowledge (Jacobs and Hayirsever, 2016). On the other hand, the results from student evaluation from the previous semester indicated that in the traditional classroom, professors did assign tasks to students and provided feedback. University average value on the item examining students' perceptions about this issue was 4.5 on a scale from 1 to 5. Students rated the traditional course content with 4.5 and the professors even higher, with 4.6.

Figure 1 below illustrates these values at Faculty vis-vis University level for the winter semester, academic 2019/20.

1. Figure1. University and Faculty average values from student evaluation survey, winter, 2019/20



V. CONCLUSIONS AND NEXT STEPS

The main conclusion from the analysis of the reports about online teaching activities is that professors are barely aware of what constitutes an effective and good quality online teaching pedagogy. They do not seem to understand that technology is only a tool and that it is the choice of activities that work well and not technology what drives effective online instruction. The opportunities for learning, created by selection of appropriate methodology in the online mode should be enhanced by technology, not dictated or subsumed by it. The elements of the new survey, selected according to what the literature has pointed out to be the characteristics of good quality online teaching, have shown that they equal those of the traditional classroom. This provides some evidence in favor of confirming the main hypothesis of the study: good quality teaching has the same characteristics; it is the mode of instruction (traditional or online) that differs. Additional evidence will be provided in the final stage of the project by comparing the results from the two surveys. It is expected to indicate whether according to students' perceptions, the same quality features apply to both modes of instruction. The final conclusions will serve for offering recommendations for actions and identifying areas for improvement and training at institutional level, but also national and wider since SEEU is considered to be a pioneer in establishing and cherishing a quality culture in the region. It is almost without any doubt that these findings will become even more relevant in the period that follows as distance education will be gaining on popularity and application after such a wide use all over the world.

REFERENCES

- [1] Anderson, T. (Ed.), The theory and practice of online learning. Edmonton, AB, Athabasca University Press, 2008.
- [2] Carrier, M., Damerow, R. & Bailey, K. (Eds.), Digital language learning and teaching: Research, theory, and practice. Rutledge New York, 2017.
- [3] Jacobs, D.B., Hayirsever, F., "Student Centered Learning. How does it work in practice?", British Journal of Education, Society & Behavioural Science. 18/3 pp. 1-15, 2016.
- [4] Koochang, A., Riley, L., & Smith, T., "Learning and Constructivism." Interdisciplinary Journal of E-Learning and Learning Objects vol. 5, pp. 91-109, 2009.
- [5] Liton, H., A., "Exploring Teachers' Attitude towards ICT integration into ESP and EFL Classroom" International Journal of Instructional Technology and Distance Learning 11/5, 2014 Retrieved https://www.academia.edu/10156678/Exploring_Teachers_Attitudetowards ICT_integration_into_ESP_and_EFL_Classroom, December, 27, 2019.
- [6] Shelton, K., "A Review of Paradigms for Evaluating the Quality of Online Education Programs." Online Journal of Distance Learning Administration, vol. 4, number 1, 2011. Retrieved from, <https://www.westga.edu/~distance/ojdla/spring141/shelton141.html>, April, 16, 2019.
- [7] Smidt, E., Li, R., Bunk, J., Kochen, T., & McAndrew, A., "The Meaning of Quality in an Online Course to Administrators, Faculty, and Students." Journal of Interactive Learning Research. 28/1 pp. 65-86, 2017.
- [8] Yang, Y., Cornelious L., "Preparing Instructors for Quality Online Instruction.", Semantic Scholar, 2005. Retrieved, <https://www.semanticscholar.org/paper/Preparing-Instructors-for-Quality-Online-Yang-Cornelious/1ad185beeb9c3a1cd8ac35e0af1699c207c34d8a>, April, 05, 2019

APPENDIX 1

Student evaluation survey

Instructions: Tick the answer that best explains your opinion: SA (strongly agree), A (agree), N (neutral), D (disagree); SD (strongly disagree)

Course/instructor:

1. The course materials posted online (topics, lectures, videos, etc.) were well organized and user friendly (easy to follow). SA A N D SD
2. The information and instructions by the instructor for following the course through the GC platform were clear.
3. The selection of materials and tasks on GC was relevant and appropriate for reaching the course objectives.
4. The instructor used the opportunities for online interaction with students (discussions, chats, assignments, etc.)
5. The assessment (mid-term exam, assignments, quizzes, projects etc.) was well organized and appropriately reflected the syllabus.
6. The instructor provided regular feedback on the given assignments.

Summary statements

(1) Please respond to questions 7 and 8 using a 5 point scale, from excellent (5) to poor (1).

- | | | | | | |
|----------------------------------|---|---|---|---|---|
| 7. Overall course content rating | 5 | 4 | 3 | 2 | 1 |
| 8. Overall instructor rating | 5 | 4 | 3 | 2 | 1 |

Comments

Please provide any comment or suggestion related to the course.

Object detection and instance segmentation of fashion images

Sandra Treneska

*Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius, Skopje, Macedonia
sandra.treneska@students.finki.ukim.mk
Skopje, 2020*

Sonja Gievska

*Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius, Skopje, Macedonia
sonja.gievska@finki.ukim.mk
Skopje, 2020*

Abstract—Over the past few years, fashion brands have been rapidly implementing computer vision into the fashion industry. Our research objective was to analyse a number of methods suitable for object detection and segmentation of apparel in fashion images. Two types of models are proposed. The first, simpler, is a convolutional neural network used for object detection of clothing items on the Fashion-MNIST dataset and the second, more complex Mask R-CNN model is used for object detection and instance segmentation on the iMaterialist dataset. The performance of the first proposed model reached 93% accuracy. Furthermore, the results from the Mask R-CNN model are visualized.

Index Terms—object detection, instance segmentation, semantic segmentation, computer vision, fashion images

I. INTRODUCTION

More than 25% of the entire revenue in E-Commerce is attributed to apparels and accessories. A major problem they face is categorizing these apparels from just the images. This poses an interesting computer vision problem.

Analyses of fashion images are popular research topics in recent years because of their huge potential in the industry. Detection of clothing items from a single image can have huge commercial and cultural impact on society. Many researches in this field have recently progressed from recognition-based clothing retrieval tasks to understanding-based tasks. That means that the models can not only recognize the attributes of fashion images but can also understand the meaning of the combination of those attributes.

Object detection and segmentation have a wide spectrum of computer vision applications for fashion, including online shopping, personalized recommendation and virtual try-on. Many fashion brands are already using machine learning techniques to predict and design what will be the next fashion trend [1] or for visual search [2].

However, real-world application remains a challenge, because of deformations, occlusions and discrepancies between consumer and commercial clothing images. Also, problems may occur due to wide variations in appearance, style, brand and layering of clothing items. At the same time, very subtle differences can exist that cannot be easily distinguished, for

example images of the same product can often look different under different conditions. Therefore, these tasks are being extensively studied in computer vision community by many research groups in both academia and industry.

Instance segmentation is challenging because it requires the correct detection of all objects in an image while also precisely segmenting each instance. It combines traditional object detection and semantic segmentation. The goal of object detection is to classify individual objects and localize each using a bounding box. The goal of semantic segmentation is to classify each pixel into a fixed set of categories.

II. RELATED WORK

The DeepFashion2 paper [3] is a benchmark for detection, pose estimation, segmentation and re-identification of clothing images. They try to fill the gap of a previous paper, Deep Fashion and address its issues including single clothing item per image, sparse landmarks, and no per-pixel masks. They also propose a baseline, Match R-CNN, which builds upon Mask R-CNN and solves the issues in an end-to-end manner.

ModaNet [4] provides a dataset of street images fully annotated with masks (polygons) of a single person. ModaNet aims to provide help for evaluating the progress of the latest computer vision techniques that rely on large data for fashion understanding. The rich annotation of the dataset allows to measure the performance of algorithms for object detection, semantic segmentation and polygon prediction of images in detail.

FashionAI [5] presents a hierarchical dataset for fashion understanding. They realize that fine-grained attribute recognition is critical, but it was missing from existing datasets. FashionAI addressed this by building a well-structured hierarchical knowledge and precise annotations of fashion apparel.

III. MODELS

A. Datasets

Fashion-MNIST [6] is a dataset of article images consisting of a training set of 60,000 examples and a test set of 10,000

examples. Each example is a 28x28 gray-scale image, associated with a label from 10 classes. The label descriptions are the following: T-shirt/top, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, Ankle boot.

iMaterialist [7] dataset is provided by FGVC6 workshop at the Conference on Computer Vision and Pattern Recognition (CVPR) 2019. It contains a total of 50,000 clothing images from daily-life, celebrity events, and online shopping. Their taxonomy contains 46 apparel objects (27 main apparel items and 19 apparel parts), and 92 related fine-grained attributes.

B. CNN model

In this part, a CNN model is trained on the Fashion-MNIST dataset for the purpose of classifying fashion items in images.

For the preprocessing of the images, the pixels are normalized and then every image is reshaped as a numpy array of pixels. Resizing the images is not necessary since all the images are already the same size. Next, the data is split into train, validation and test sets and we have 48,000, 10,000 and 12,000 images in each set respectively.

Three CNN models of different complexity are created. The first one has one convolutional layer with 32 filters which results in 173,738 trainable parameters. The second model has two convolutional layers with 32 and 64 filters, resulting in 515,146 trainable parameters. Finally, the third model has three convolutional layers with 32, 64 and 128 filters resulting in 1,421,194 parameters. All the models have dropout layers which help the models to not overfit. At the end, all models have two dense layers, one with ReLU and one with softmax activation function. For training, sparse categorical cross entropy is used and an Adam optimizer.

Kaggle notebooks are used for the training, so the computations are done faster, on a GPU.

C. Mask R-CNN model

In this part, a Mask R-CNN model [8] with COCO pre-trained weights [9] is trained on the iMaterialist dataset for the purpose of object detection and instance segmentation of fashion images.

Mask R-CNN is a recent advanced framework developed by FAIR (Facebook AI Research) for object instance segmentation. It detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance. It stands for Mask Regional Convolutional Neural Network and it extends Faster R-CNN. Additionally, Mask R-CNN is easy to generalize to any task and can also be used for key-point detection.

The model can be roughly divided into 2 parts — a region proposal network (RPN) and binary mask classifier. The first step sets bounding boxes that could possibly contain an object of relevance, this is called ROI (Region of Interest) Align. These boxes are then refined using a regression model. In the second step instance segmentation is applied to each box. The instance segmentation model is trained like a binary classifier, meaning 1 represents the presence of an object and 0

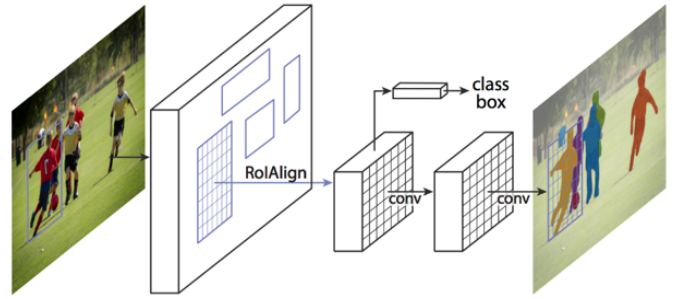


Fig. 1. Diagram of Mask R-CNN

represents the background. The architecture of Mask R-CNN is shown on Fig 1.

For the configuration of the model, the steps per epoch and validation steps are lowered so the training can finish in less than a day. Also the image size is set to 512x512 and all images are resized to those dimensions. All the other hyperparameters were left as default.

The images have masks containing the pixels where the fashion items are. One image can have multiple masks, meaning that this a multi-label problem. Some of the masks are visualized, shown on Fig 2.



Fig. 2. Masks

Again, Kaggle notebooks were used for the purpose of training the model faster on a GPU. The training was done on 36,156 images that contain 264,949 segments (masks) in total, while the validation set had 9,039 images and 66,264 segments.

IV. RESULTS

Accuracy and loss were measured for all the three CNN models. The results can be seen below on Fig 3 and Fig 4 for 10 and 50 epochs respectively.

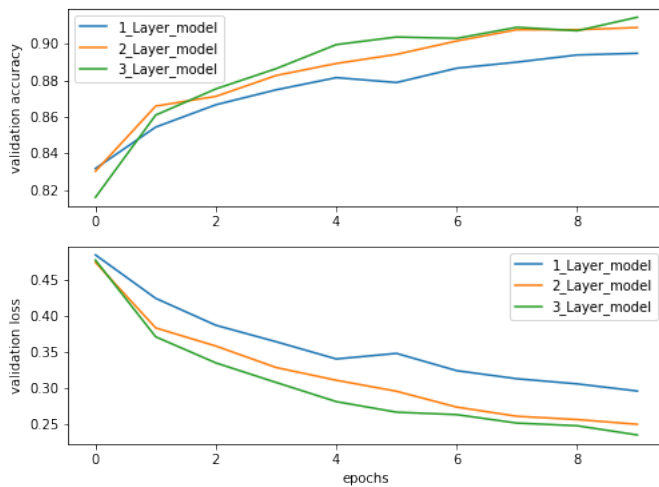


Fig. 3. 10 Epochs accuracy/loss graph

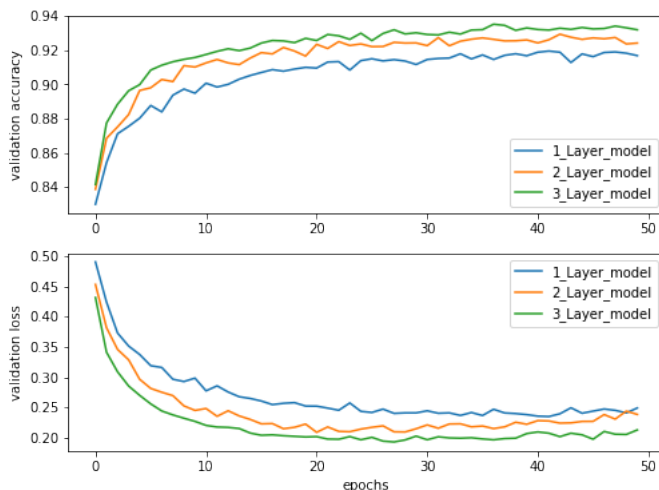


Fig. 4. 50 Epochs accuracy/loss graph

After 10 epochs we can say that there isn't much difference in the performance of the second and third model. But, in the long run we can see that the third model outperforms the others and after 50 epochs it has 93% accuracy. That was expected, since the third model has over a million parameters and naturally it needs more time to train.

The training accuracies for the first, second and third model respectively are 0.916, 0.928, 0.934. The F1-scores, micro and macro, are the same as the accuracies.

Since iMaterialist is a competition, the real classes of the test images weren't provided, so evaluation metrics couldn't be used for the Mask R-CNN model. But, it was still possible to visualize the predictions that the model made for the test images. Below, on Fig 5 are the testing images visualized.

For each image the model predicts the class, a bounding box, a mask and a confidence factor of every predicted fashion item. The model performed the best on images where the person is facing straight to the camera, and it was less accurate

on sideways pictures or images where the clothes are captured from a different angle.

V. CONCLUSION AND FUTURE WORK

The fashion industry has lately attracted a lot of attention with its huge economic potential and practical value. There are many researches and competitions that analyze how computer vision can be integrated in the fashion industry.

In this paper the focus was on object detection and instance segmentation of fashion items in images. By using convolutional neural networks and Mask R-CNN for these tasks we were able to produce meaningful results.

Future work could further improve the model's performance. Both models can be improved with more training time and computational resources as well as parameters tuning and data augmentation techniques. Additionally, classifying fine-grained attributes for the fashion items in iMaterialist dataset could be implemented.

REFERENCES

- [1] Stitch Fix <https://algorithms-tour.stitchfix.com/>
- [2] Pinterest visual search <https://newsroom.pinterest.com/en/post/introducing-the-next-wave-of-visual-search-and-shopping>
- [3] Ge, Yuying, et al. "Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
- [4] S. Zheng, F. Yang, M. H. Kiapour, and R. Piramuthu. Modanet: A large-scale street fashion dataset with polygon annotations. In ACM Multimedia, 2018.
- [5] Zou, Xingxing, et al. "FashionAI: A Hierarchical Dataset for Fashion Understanding." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019.
- [6] Fashion-MNIST dataset <https://github.com/zalandoresearch/fashion-mnist>
- [7] iMaterialist dataset <https://sites.google.com/view/fgvc6/competitions/imat-product-2019>
- [8] He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.
- [9] COCO <http://cocodataset.org/>



Fig. 5. Predictions from the Mask R-CNN model

A novel platform for sharing and renting clothing to reduce environmental pollution

Ana Todorovska, Evgenija Krajchevska, Dimitar Trajanov, Sasho Gramatikov
"Ss. Cyril and Methodius" University in Skopje
Faculty of Computer Science and Engineering
"Rugjer Boshkovikj" 16, 1000 Skopje, Republic of Macedonia

Abstract—In today's rapidly developing, technology driven world, the environment is being polluted in numerous ways, significantly more than in the past. Considering the fact that the fashion industry is the second largest polluter worldwide, it is evident that changes are needed in the way clothing is viewed and used.

In this paper, we present a possible solution to this problem, which is built using the constantly advancing concept of shared economy. The platform we present is intended to create an online space that is used to share clothes that are not considered useful by their owners. Furthermore, the rental of expensive clothing for special occasions is provided.

Adhering to the importance of the business aspect of any novelty being introduced, a market research was also conducted in our country. This paper includes these results, together with the conclusions drawn and the possible improvements suggested.

Index Terms—Sharing economy platform, Sharing economy, Fashion, Pollution

I. INTRODUCTION

Inexpensive fashionable clothing is widely available today, which is the impetus to change in the way items of clothing are viewed, making them disposable.

This means that garment consumption has been made more democratic, however not without negative consequences. The environment is being polluted in the process of growing water-intensive cotton, by the release of dyes into local water sources and by the millions of tons of textile waste in landfills [1].

Considering the threat posed by pollution globally, efforts to develop a sustainable solution are made perpetually.

Research on how to increase corporate social responsibility is currently being done [2], in hopes of introducing change into the production of clothing.

Other solutions can possibly be found by looking into the idea of clothing reusability. The concept of sharing is not new, which is one of the reasons sharing economy is successful [3]. Taking into consideration the changes brought by sharing economy, it can be used as an integral part of such a solution. In this paper we present a sharing economy platform that supports clothing reusability, together with market research to assess the possibility of its successful implementation in our country.

The structure of our paper is as follows. Section II details the market research that was conducted. Section III describes the business and technical aspect of our solution. Finally, section IV presents our conclusions.

II. MARKET RESEARCH

In this paper, we present a possible improvement regarding the pollution caused by the fashion industry.

Following the definition of the idea of a software solution that uses the concept of shared economy, an analysis of the global market was conducted. The analysis showed that several globally successful businesses have evolved from a similar idea. The most prominent are "Rent the Runway" in USA, "Girl Meets Dress" in UK, "Chic-by-choice" in EU and "Glam Corner" in Australia.

Considering these findings, the next logical step was to conduct a local research in our country.

A. General research

The initial research consisted of a survey which had a purpose of evaluating the consumption of digital content and inclination to share and/or rent clothing online. 360 people were reached by this survey in a period of four months at the end of 2019. The results were carefully analysed.

Firstly, the demographics of the sample was studied. The vast majority of the population are within the age range of 18 to 40 years old. Moreover, 73.5% of them are female. Taking into consideration the nature of the solution being presented, these demographics were expected.

Secondly, the consumption of digital content was examined. This was done by asking questions about the daily internet usage as well as the usage of some prominent e-commerce platforms. More than 79% of the population are active internet users and about 21% of them have used at least one e-commerce platform, as shown in Fig.1.

This statistics is a key indicator that the reach of e-commerce is growing in our country, even though rather slowly when compared to the USA, for example.

The following questions were all problem specific, and were carefully analyzed using appropriate statistic tools.

A result of high importance was the confirmation of the presence of the problem in our country, with around 94% of the population stating that they possess clothing which is unused, as can be seen in Fig.2.

Moreover, when asked about their current solutions to this problem, 63.8% answered that their unused clothing is thrown or given away, as shown in Fig.3.

More detailed question was also asked about the recipients

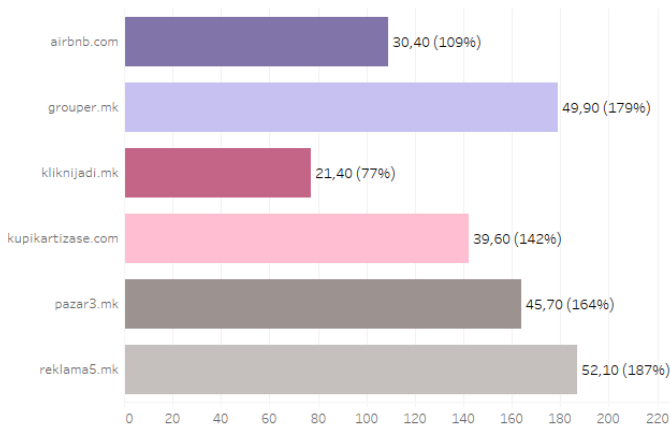


Fig. 1. Initial research results about the usage of e-commerce platforms

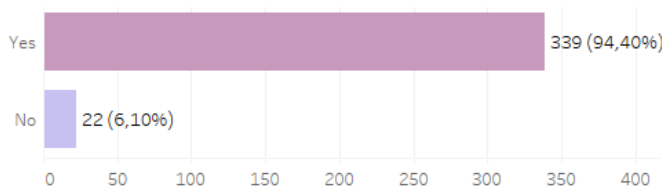


Fig. 2. Initial research results about the presence of unused clothing

of the gifted clothing, which in 87.9% were relatives and acquaintances.

Finally, the inclination to share and/or rent clothing was evaluated, by asking about the willingness to use an online platform enabling that. The results show that around 80% of the population is willing to use this kind of an online platform, as can be seen in Fig.4.

Following the individual analysis of each question, a Pearson's coefficient correlation matrix was computed, so that analysis of the dependency between the given answers could be performed. The interpretation of the correlation coefficient was done considering the notion of its dependency on context and purpose [4] [5].

Seeing that the correlation coefficient between the answers about the feasibility of the platform and the willingness to use it is around 0.7, we can state that there is a high correlation between them. Other significant dependencies could not be

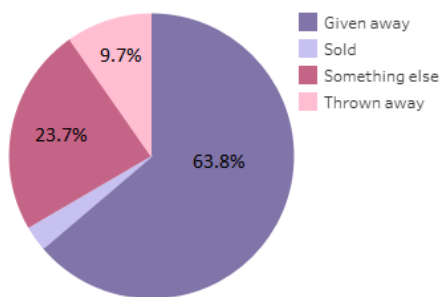


Fig. 3. Initial research results about current solutions for unused clothing

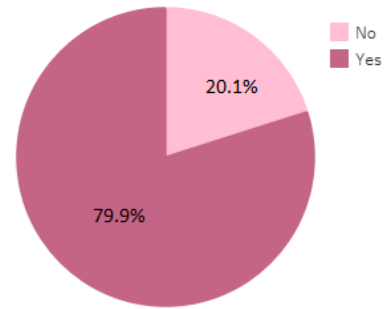


Fig. 4. Initial research results about the willingness to use the platform

concluded.

The main conclusions from the initial research are that there is a large population that could benefit greatly from our solution and that the inclination to share and/or rent clothing is rather satisfactory.

In addition, once some less frequent responses were studied in detail, the concept we initially proposed was altered to include sharing, swapping and renting of clothing. This changes, together with the need for more detailed statistics brought about further research.

B. Further research

Further research was done by conducting a survey which focused on presenting the idea in more detail, so that the results could be later used as guidance in the development of the software solution. The accent was put on sharing and swapping of unused clothing, as well as renting of more expensive special occasions garments. 317 people were reached by this survey in approximately two weeks.

The demographics of the sample was similar to the one in the initial research, consisting of mostly female individuals under the age of 40.

Additionally, the questions about the consumption of digital content were examined and showed similar results, when compared to those from the initial research.

All the other questions were concerning the specific problem. The results were studied, but the responses were more scattered than in the initial research, which made the analysis of the overall data more difficult.

The willingness to shop at a second hand store was of high interest, when thinking about the sharing aspect of the solution. The results shown in Fig.5. make the unwillingness to do so more than evident, with about 74% of the population stating that they do not shop at such stores.

Another question of importance was the one about the willingness to share clothing with friends and family, again adhering to the sharing concept. The percentage of the population that swap or give clothing to relatives and friends is 54.6%.

The following questions were designed to cover the renting aspect of the solution. Around half of the population stated that they would rent an outfit for some special occasion, however only around 30% stated the same when asked about

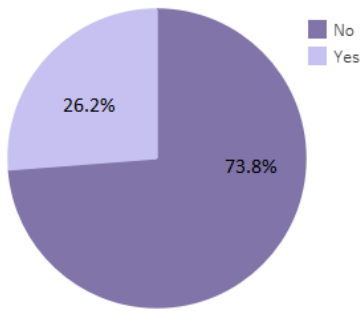


Fig. 5. Further research results about shopping at second hand stores

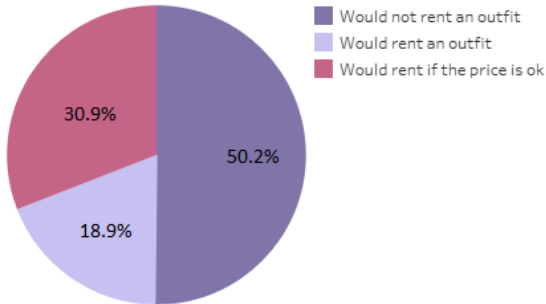


Fig. 6. Further research results about the willingness to rent an outfit for a special occasion such as graduation, a formal event etc.

their wedding dress or tuxedo. The results are shown in Fig.6.

Finally, the overall willingness to use an online platform to share and swap, or to rent clothing was examined. This was done using separate questions, so that more detailed results could be obtained. The results given in Fig.7. show that around 54% of the population is willing to use an online platform for sharing and swapping of garments and slightly less, around 48% are willing to use such a platform for renting.

These results are significantly worse than the ones obtained during the initial research. Dependency between the answers to the questions was examined for this survey as well, showing significant dependency in some cases. The first notable correlation is between the willingness to rent clothing in general, and to do that using an online platform. This

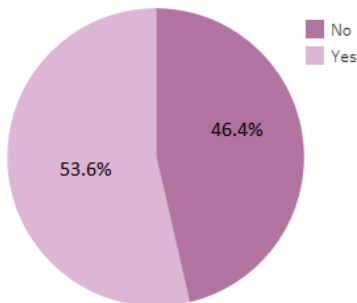


Fig. 7. Further research results about the willingness to use the suggested online platform for sharing

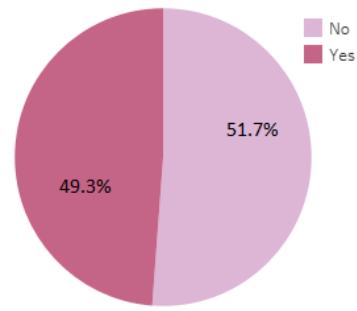


Fig. 8. Further research results about the willingness to use the suggested online platform for renting

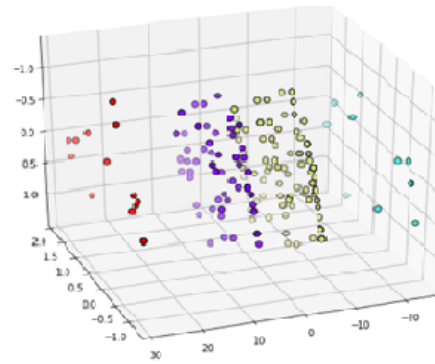


Fig. 9. Customer segmentation

shows that people who are inclined towards renting more expensive clothing, have no problem of doing that online.

Another correlation can be seen between the willingness to rent luxury clothing, and to rent clothing for weddings, including wedding dresses and tuxedos.

Furthermore, there is an important correlation between the willingness to rent and the willingness to share or swap garments. This shows that people who would swap or share clothing are more likely to embrace the idea of renting clothing as well.

The data from this survey was further used to define four customer segments. This was performed by clustering of the answers given by the respondents, which can be seen in Fig.8. and Fig.9.

The first customer segment is made of people who are not open to the idea of renting garments. Their opinions about swapping and sharing of clothing are mixed.

The second segment of customers is made of respondents who are willing to rent formal or luxury clothing, even for their weddings.

Most of the sampled population is found in the following two segments, made of people who are willing to rent clothing, only if such a rental is a good bargain. One group of these respondents are not open to renting clothing for their wedding day, while the other group are embracing that idea as well.

The fact that only about half of the population are open to renting and/or sharing and swapping of garments is of utmost interest, when the realization of this solution is considered.

The main conclusion from this extended research is that there is a possibility for success of this sharing economy platform. However, that can not be easily accomplished at the moment. Extensive marketing and advertising needs to be done first, so that the benefits of such a solution are well understood by the general public.

At that moment, this platform can possibly achieve the success of the prominent businesses that do this worldwide.

III. SOFTWARE SOLUTION

In this section we present the business and technical aspect of our solution.

A. Business aspect

The importance of the business aspect in creating any software solution should never be underestimated, as it is a major contributor towards the success of any application.

Every business can be elevated with the Business model canvas [6] which uses nine building blocks to describe and analyze the model. The strengths and weaknesses of our idea are presented using this model.

Value proposition is the combination of products and services a solution provides to its customers [6]. It is based on solving some problem or satisfying a need. The value proposed with our solution comes from the fulfillment of the need of constantly buying new clothing as well as disposing of unused clothing.

Concerning the market that is targeted by this solution, at the moment, only a hypothesis can be made about who would be our customers.

However, the distribution channels that reach the customers can be clearly stated. Our solution has a combination of virtual and physical channels, both equally important. The platform itself is the virtual channel that allows customers to make reservations, order, or just browse the items. However, the process of delivering and maintaining the products is done by the physical channel.

When sharing economy is discussed, reviews and ratings are one of the key aspects that straighten the relationship with the customers, which is of great importance. This concept is included into our solution.

An essential aspect of the business model is the revenue acquisition. Taking into account that the solution is designed to be free of charge for customers, the revenue acquired is highly dependent of our partnerships with fashion brands and local stores, which would pay for the usage of our services. Additional revenue could be acquired through advertisement. At last, it is of paramount importance that the online activities provided are well-developed, so that customers are guaranteed enjoyable experience and safety of information.

B. Technical aspect

The solution we created has three integral parts, where the API application is the middle-ware between the database

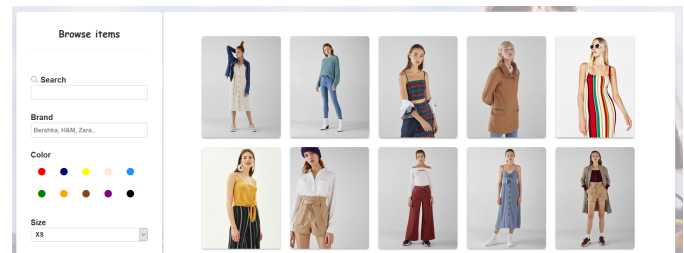


Fig. 10. Search component

system and the client application. It has several main functionalities, which we present through the client application. It is important to state that these functionalities are only available to users that are registered.

The search component which can be seen in Fig.9, provides the ability to search items of clothing. Various filters are included to enable easier use and therefore greater customer satisfaction. A component was created to view the details for each garment, which are provided by the customer that has posted it for sharing, or by the company that rents it. Therefore, a component was also created to enable creating posts by the customers. Another important functionality that was enabled is the ability to add and view reviews.

There is still a lot to be done in order to completely cover the desired functionalities of our idea, and we are currently working on such improvements to our solution.

IV. CONCLUSION

The purpose of our solution is to reduce the pollution of the environment that is caused by the fashion industry, by providing a sharing economy platform that enables sharing and renting of clothing online.

Research was conducted in order to assess the possibility for a successful realization of this solution in our country. The results show willingness to use our platform by at least half of the population.

Consequently, we are optimistic about the future possibilities of our solution, which is why we are working on developing more features of our software solution, while continuously conducting more research locally.

REFERENCES

- [1] Rachel Bick, Erika Halsey and Christine C.Ekenga. The global environment injustice of fast fashion. BioMed Central Journal 2018
- [2] Anika Kozlowski, Michal Bardecki and Cory Searcy. Environmental Impacts in the Fashion Industry. Greenleaf publishing 2012
- [3] Will Sutherland and Mohammad Hossein Jarrahi. The Sharing Economy and Digital Platforms: A Review and Research Agenda. International Journal of Information Management July 2018
- [4] Andrzej Buda, Andrzej Jarynovski. Life-time Of Correlation And Its Application (volume 1). Wydawnictwo Niezależne 2010, pages 5-21.
- [5] Jacob Cohen. Statistical power analysis for the behavioral sciences, 2nd edition. Lawrence Erlbaum Associates 1988.
- [6] Alexander Osterwalder, Yves Pigneur. Business Model Generation. John Wiley & Sons, Inc. 2010
- [7] Kenneth C. Laudon, Carol Guercio Traver. E-commerce 2017, business, technology, society. Pearson, 2017.

Educational robots in preschool education

Kristina Todorovska
Master student of ICT Education at
Faculty of Computer Science and Engineering
Skopje, North Macedonia
kikitodorovska@gmail.com

Abstract--Along with robot technology development, researchers and educators have employed robots to support the education. This paper gives an overview of using robots in education, describes the types of robots used in preschool education and the experience of utilizing robots in preschooler's everyday life - how successfully they learn, collaborate, share information, and master the basic concepts of programming as direct users of educational robots. Crucial importance is attached to robots that use tangible elements, objects that are visually accessible and offer to the young students an opportunity to manipulate.

Keywords: education, robots, communication, technology

I. INTRODUCTION

It is widely accepted that overcoming STEM competencies in preschool lead to faster development of cognitive abilities and the ability to solve problems that appear in children [14]. In this process, the educational robots used in the learning process is of great help. It gains wide popularity. Robots that take wide swing in education allow children to gain practical experience in accepting the constant changes in the living environment and adapting to them, as well as using the acquired knowledge in real-time situations.

II. RESEARCH REVIEW

In recent years, robots used in the educational process to learn and master new content have become very popular tools, as the authors Causo and Chen [8] have pointed out in their paper. In particular, there is a growing interest in using tangible robots that are defined as physical manipulative that can directly affect and stimulate the digital environment in which we live. Author Bers [2] points out that tangible, manipulative robots are of particular interest in early childhood learning because they correspond to traditional learning toys created by Maria Montessori and Friedrich Frobel, designed to lead to gaining knowledge and learning mathematical concepts.

The authors in [17] suggest that tangible educational robots allow a child to enter digital information by manipulating with physical objects, instead of using a screen, keyboard, or mouse (computer components). Belpaeme, Kennedy, Ramachandran [17] also provide a theoretical basis suggesting that the use of these types of robots enables interaction with the physical world, which is the basis for teaching young children. Tangible robots enable the introduction of children into the world of

Ana Madevska Bogdanova
Faculty of Computer Science and Engineering
Skopje, North Macedonia
ana.madevska.bogdanova@finki.ukim.mk

STEM, while leading to the learning of basic concepts, such as sequencing, abstract thinking, orientation, decomposition that allow solving the problems that children face.

Authors Randelli, Nardi, Venanzi in [5] emphasize the importance of manipulative robots and include them as the main supporters of learning and creative expression of young children.

It also enables the development of teamwork and collaboration, and children learn basic concepts of behavior and thinking, how to tell a story from start to the end, how to sort numbers starting from the smallest, sort objects by size, shape, and color. An organized environment abounding with new tools such as educational robots fully influences children's socio-emotional development - learning sides, directions, sharing resources and toys, says author Bers [2].

Weinberg and Yu, in their study, stated that robotics creates a unique learning environment by providing physical embodiment of computation; students receive strong, visceral feedback from physically experiencing their work [6]. In robotics classes, as the authors stated, students explore, make hypotheses about how things work, and conduct experiments to validate their beliefs and assumptions. Robots are useful aids for teaching mathematics and physics; they can be used in classrooms for explaining difficult concepts, as they capture the imagination of many younger people [9]. Furthermore, the plug-and-play characteristic of educational robots, like LEGO Mindstorm RCX, makes it easier to learn complex engineering subjects without having prerequisite knowledge. Another study examined the effectiveness of a LEGO robotics course on students' understanding of gear functions and mechanical advantages. The authors of this study concluded that robotic sessions improve students' understanding of gear function in relation to direction of turning, relative speed, and number of revolutions [7]. Martin in [4] applied the "Programmable Brick", a new educational technology that was an extension of

LEGO is suitable for introducing technology to students. The Programmable Brick combined the functionality of the desktop computer and the interface to the LEGO motors and sensors into a single brick. Martin found that the Programmable Bricks expand design and learning possibilities and children effectively learn technology when they are engaged in design, construction, and debugging activities.

III. THE IMPORTANCE OF TANGIBLE EXERCISES DURING THE PRESCHOOL PERIOD

During the preschool period children learn how to understand the functionality of objects and appearances they encounter constantly, how to remember them without seeing them. In early childhood, children struggle with abstract appearances, so they often rely on physical appearances, objects, and impressions that help them formulate, ask and answer, and understand how the world in which they live works. Although mastering cognitive concepts may seem daunting, children can easily overcome it because they have a special learning tool that helps with the abovementioned things, and that is the play.

Typically, the play involves manipulating with physical objects that surround and encourage learning and allow children to explore and understand abstract concepts in the educational domain. As children play, they enter the physical world using their five senses. Hence, special importance is attached to traditional toys that allow the development of their small hands and fine motor skills, by modeling structures that they see in their daily lives or design and build completely new ideas that arise from themselves.

All this supports the curricula and the study of the provided contents in the form of puzzles, counting cubes, sorting by shape and colour, stacking puzzles and many others.

IV. TANGIBLE ROBOTS AS A TOOL IN THE LEARNING PROCESS

Encouraged by traditional learning methods, created by Montessori and Froebel, we attach special importance to tangible types of robots, which enable support for the learning process and implementation of curricula using technology on one hand, without exposure to screens, on the other.

Technologies that offer tangible elements encourage socialization and interaction and help children develop self-confidence and their potential. The integration of tangible parts of technology into the learning process based on solving problems, stimulates children's initiative, motivation, perseverance and curiosity as the main drivers of their development [16].

Tangible robots are well-suited to introduce children to the basics of programming in order to develop their skills. With the help of tangible robots and the wooden blocks and cubes they offer, children have the opportunity to connect them by themselves, to create, and in that way to understand the basics of programming [20]. Tangible robots are designed to allow children to engage in the digital and 3D worlds. These robots have components in the form of blocks, directional cubes, movements, built-in sensors, etc. By using them and playing with them and their sensors, young children can become engineers from an early age.

Tangible robots allow interaction through motion sensors and gestures. Using these robots is actually accepting the view

that "Gesture is an innate skill that is adopted in communication", while inserting various sensors that lead to new and interesting interactions between humans and robots, the authors in [5] suggest.

The robots that offer the use of tangible elements and their use and creation of basic programs, enable development of fine motor skills of children as well as development of skills for organization and responsibility. Tangible robots are easily portable and encourage user mobility and development of rough motor skills, which is also crucial for the preschool period and early childhood development.

Through the daily experience of using tangible educational robots with preschool children, we became aware of the fact that learning through play using tangible objects stimulates imaginative thinking and reveals new opportunities for expression and research. The use of hands and manipulative objects changes the way a child learns, acknowledges information, and connects them to everyday situations he or she is constantly confronted with.

The following photographs (Figure 1, Figure 2, Figure 3, Figure 4) are taken in the kindergarten where the first author works. They show the daily activities with the interactive robot Cubetto with tangible blocks. This type of robot encourages the development of fine motor skills, teaches the basics of programming, development of perseverance and curiosity, and the desire to explore.



Figure 1. Sorting of tangible blocks



Figure 2. Presentation of components of tangible robot Cubetto



Figure 3. Perform a step-by-step task with sequencing



Figure 4. Perform multiple steps at once with tangible blocks

CONCLUDING REMARKS

Using educational robots in early childhood, which leads to learning and developing computer programming skills in preschool institutions, is a new challenge and innovation for education, especially when it comes to young children. The use of technology in preschool institutions should meet the needs of preschool children and all this should be done through play, in a fun and interesting way and easily acceptable to them. With this, children strengthen their cognitive skills and connect them with new areas they face in their daily lives.

Using robots requires a variety of thinking and problem-solving skills through the creation of various programs. In the initial stage, the most important thing is to motivate children to use robots for new challenges, to encourage the formation of new concepts and solutions, to complement and strengthen the competencies of educators as moderators of the teaching process, and also to introduce the parents as motivators and supporters of child development, through training and coaching in order to understand their importance.

Every child deserves an environment rich in innovative methods and tools that enable and encourage their overall development.

REFERENCES

- [1] A. Khanlari, "Teachers' Perceptions of using robotics in primary/elementary schools in Newfoundland and Labrador", Memorial University of Newfoundland, June 2014.
- [2] A. Strawhacker, M. Bers, "I want my robot to look for food": Comparing Kindergarten's programming comprehension using tangible, graphic, and hybrid user interfaces, August 2014.
- [3] "Beginning computer programming for kids", an introductory guide, 2016.
- [4] F. Martin, "Kids learning engineering science using LEGO and the programmable brick," Paper presented at the annual meeting of the American Educational Research Association, New York, NY, 1996.
- [5] G. Randelli, M. Venanzi and D. Nardi, "Evaluating tangible paradigms for ground robot teleoperation", IEEE Conference, September 2011.
- [6] J.B. Weinberg, and X. Yu, "Robotics in education: Low cost platforms for teaching integrated systems," IEEE Robotics & Automation Magazine, vol. 10, no. 2, pp. 4-6, 2003.
- [7] J. M. Chambers, M. Carbonaro, and H. Murray, "Developing conceptual understanding of mechanical advantage through the use of Lego robotic technology," Australasian Journal of Educational Technology, vol. 24, no. 4, pp. 387-401, 2008.
- [8] L. P. E. Toh, A. Causo, Pei-Wen Tzuo, I-Ming Chen and S. H. Yeo, "A Review on the Use of Robots in Education and Young Children", Educational Technology & Society 19(2):148-163, January 2016.
- [9] M. Cooper, D. Keating, W. Harwin, and K. Dautenhahn, "Robots in the classroom: Tools for accessible education," in Assistive Technology on the Threshold of the New Millennium, C. Buhler and H. Knops, Eds. Amsterdam: IOS Press, 1999, pp. 448-452.
- [10] M. U. Bers, "Blocks to robots: Learning with technology in the early childhood classroom", New York, NY: Teachers College Press, 2008.
- [11] M. U. Bers, "Designing digital experiences for positive youth development", From playpen to playground. Cary, NC: Oxford, 2012.
- [12] M. U. Bers, "Learning how to program robots in early childhood", Eliot Pearson Department of Child Development, Tufts University, chapter 8, pp.133- 145.
- [13] M. U. Bers, "The Tangible robotics program: Applied computational thinking for young children", Early Childhood Research and Practice, 2010.
- [14] N. DeJarnette, "Implementing STEAM in the Early Childhood Classroom", European Journal of STEM Education, 3(3), 18, 2018.
- [15] R. Isnaini, C. Budiyanto, "The Influence of Educational Robotics to Computational Thinking Skill in Early Childhood Education", Conference: The 1st International Conference on Computer Science and Engineering Technology At: Kudus, Indonesia, November 2018.
- [16] S. Schiffer, A. Ferrein, "Early Robotics Introduction at Kindergarten Age", Multimodal Technologies Interact, September 2018.
- [17] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati and F. Tanaka, "Social robots for education", Science Robotics Vol. 3, Issue 21, August 2018.
- [18] T. Sapounidis, Stavros N. Demetriadis, "Educational Robots Driven by Tangible Programming Languages", Conference Paper, pp.205-214, March 2017.
- [19] University "St. Cyril and Methodius" - Faculty of Philosophy - Institute of Pedagogy, Ajdinski, G. (2015), International Journal of Education, Research and Training, Skopje, MAR-SAZH
- [20] Yasaman S. Sefidgar, P. Agarwal, and M. Cakmak, "Situating Tangible Robot Programming", Computer Science & Engineering, University of Washington 185 Stevens Way, Seattle.

Security Situation in Republic of Macedonia Using Semantic Algorithms for Open Data

1st Zivka Jovevska
FON University
Republic of Macedonia
zivkajaneva@gmail.com

2nd Daniel Jovevski
Ministry of information society and
Administration
Republic of Macedonia
daniejovevski92@gmail.com

3st Leonid Djinevski
FON University
Republic of Macedonia
l.djinevski@fon.edu.mk

Abstract—Semantic algorithms are a web of information related to such a way that can easily be processed from the machines globally. It is an efficient display of information on the World Wide Web network or can be presented as a globally connected database. This paper aims to investigate the security situation in the Republic of Macedonia using Web Crawler and Semantic Algorithms for its research. The data was taken from the news page of the Ministry of Interior website. Some result-graphs of the research we conducted on the data obtained from the Ministry of Interior will be shown. The expected results of the paper are to get a clear picture of life in Macedonia, in individual cities and whether the security is improving or deteriorating.

Keywords: Internet, Ministry of Interior, Security, Web Crawler

I. INTRODUCTION

Many of the technologies and ideas developed during the creation of the semantic network are refined and live in various applications. What was not possible in 2001 is now possible: You can easily create applications that use data across the network. The difference is that you had to sign in for each API separately, which, beside the tedious manual integration, gives the one who hosts the API great control over how to access their data.

Because we hear and read about crimes daily over the news and on internet, we decided to make a research and find out what was the crime situation in Republic of Macedonia in 2019.

Using web crawling we will get data and based on the results we will get a clear picture about the security situation. Our motive for this research is to find out whether the security situation in our country is improving, or it is getting worse.

The research will be based on three parameters: Theft crime cases, Homicide cases and Drug trafficking cases. The results will be shown in the graphs on a monthly basis. In the end, we will make a conclusion based on the results.

II. SEMANTIC WEB

In 2001, Tim Berners-Lee, inventor of the World Wide Web, published an article in Scientific American. Burners-Lee, along with two other researchers, Ora Lasila and James Chandler, who want to give the world an overview of the revolutionary new changes they've seen coming. Since its introduction just a decade ago, the network has become the world's fastest document sharing tool. Now, the authors promise that the network will evolve to include not only

documents but any kind of data that could be imagined. They called this new website the Semantic Web.

The effort to build the Semantic Network consists of four phases. The first phase, which is from 2001 to 2005, is the golden age of semantic network activities. Between 2001 and 2005, the W3C issued new standards that represent the underlying technologies of the semantic future.

The most important of these is the Resource Description Framework (RDF).

Berner-Lee's article begins the second phase of the development of the Semantic Network, where the focus has shifted from standard settings and example building to the creation and popularization of large RDF databases. Perhaps the most successful of these databases was DBpedia, a huge repository of RDF triples extracted from Wikipedia articles. The third phase of Semantic Network development involves adapting W3C¹ standards to suit real web developer practices and preferences. In 2008, JSON begins its rapid rise in popularity.

W3C is still working on the Semantic Network under the title "Data Activity", which can be called the fourth phase of the Semantic Network project. But that says the latest project "data activity" is a study of what the W3C must do to improve the standardization process. Even the W3C now seems to admit that few of its semantic web standards are widely accepted and that simpler standards would be more successful.

Semantic web technologies can be viewed as layers, each layer relying on and depending on the functionality of the layers beneath it. Although the semantic network is often presented as a separate entity, it is an extension and improvement of an existing website, not a replacement.

Semantic algorithms are a network of related information in such a way that it can be easily processed by machines globally. It is an effective representation of information on the WWW² network or can be presented as a globally connected database.

The semantic network is an abstract representation of WWW data, based on RDF standards and other standards to be defined. This is developed by the World Wide Web Consortium (W3C), with contributions from academic researchers and industry partners. Data can be defined and linked in such a way that there is more efficient detection, automation, integration, and reuse in different applications. The semantic network is a sequel to the Internet (WWW) that allows people to share content beyond the boundaries of applications and websites.

¹ World Wide Web Consortium

² World Wide Web

If HTML and Web pages together make online documentation look like one big book, the semantic network makes all the data in the world look like a global and huge database.

The great promise of the Semantic Network is that it can be understood not only by humans but also by machines. Web pages will be important for software programs - they would have semantics - allowing programs to communicate with the network in the same way as people. Programs can exchange data over the Semantic Network without being explicitly designed to talk to each other.

XML bits are a way of expressing metadata for a website. We are all familiar with metadata in the context of a data system: When we look at a file on our computers, we can see when it was created, when it was last updated, and by whom it was originally created. Similarly, Semantic Web sites will be able to tell our browser who is the author of the site and perhaps even where he went to school, or where that person is currently employed. Theoretically, this information will enable Semantic Web browsers to answer questions through a large collection of web pages. In his article for Scientific American, Berner-Lee and his co-authors explained that you can, for example, use the Semantic Network to look for a person you met at a conference whose name you only partially remember.

Indeed, in many cases, the <meta> HTML tag is abused in an attempt to improve the visibility of their websites in search results. Search engines have once experimented with using keywords, delivered via the <meta> tag, to index results, but soon discovered that unscrupulous website authors include tags that are not related to the actual content of their website. As a result, search engines are starting to ignore the <meta> tag in favor of using complex algorithms to analyze the actual content of a website.

III. WEB CRAWLING

Web Crawler also known as a web spider or web robot is a program or automated script that can browse any site in a methodological, automated way. This process is called web crawling or networking. Many legitimate sites, especially in search engines, use crawling as a tool of providing up-to-date analytics data.

There are several libraries for downloading and parsing in C#. Some of them are:

- HtmlAgilityPack- HTML parser that builds read/write DOM and supports plain XPATH or XSLT
- Abot- is a free C # web crawler that is fast and flexible. Refers to the low level of HTTP requests, deployment, link parsing. Only event registration is required to process data from websites.
- Wangkanai.Detection.Crawler – is a reference to ASP.NET Core, which is used to crawl data from web pages.
- AbotX Web Crawler - a C# web crawler that simplifies advanced crawling features. It is an upgrade to the Abot crawler, which offers many extensions.
- Spidey – is a library designed for crawling and parse the web content.

- InfinityCrawler – a simple but powerful library for web content crawling.
- Aspose.HTML for .NET is a cross-platform library that allows you to perform a wide range of HTML tasks directly within your .NET applications. Aspose.HTML supports parsing HTML5, CSS3, SVG and HTML to construct the DOM.
- DotnetSpider.Core is a standard .NET crawling library similar to Web Magic and Scrapy.
- RestSharp is a large library, free HTTP client that works with all kinds of .NET technologies. It can be used to build robust applications that will facilitate the interface to public APIs and fast access.

In my data download research, we use the RestSharp library with C# language. It is one of the many ways you can create a web service or web application in .NET. RestSharp is a comprehensive open-source library that works with all types of .NET technologies. It can be used to build robust applications by simplifying the interface to public APIs and allowing quick and easy access to data without the hassle of sending a large number of HTTP requests. RestSharp offers enormous advantages and saves time with a simple, clean interface, making it one of the best and most used tools today.

With its simple API and powerful library, Rest Architecture is a tool for developers who want to build detailed programs and applications. The RESTful architecture provides resource-oriented information access for creating web applications. It also offers common tasks such as generating URIs, load parsing, and authentication as configuration options, ensuring that developers no longer have to worry about low-level tasks such as networking.

We made 2 crawlers. The code is written in a C# console application where the RestSharp and dotNetRDF libraries are linked.

The first crawler download 3 .csv files from www.meteoblue.com. From this site, each file is downloaded by clicking on a radio button and then “Download as CSV” button. Each radio button displays data about a particular wind direction. My crawler downloads files by entering the city take today's date for “date to” and takes 1 day to get the “date from”, which we will use for variables to form a link from where the data for the given day is downloaded. However, if we want data for an extended period, it can only be inserted into the code at a specific date and subtracted from the appropriate number of days to retrieve from when we want it and by activating the console application the data is downloaded to the CSV file locally. For downloading and writing files locally, we use the following code:

```

var dateTo = DateTime.Now;
var dateFrom = dateTo.AddDays(-1).ToString("yyyy-MM-dd");
var d2 = dateTo.ToString("yyyy-MM-dd");
var dateRange = "?daterange=" + dateFrom + HttpUtility.UrlEncode(" to ")
+ d2 + "&params=" + HttpUtility.UrlEncode("32;10 " +
"m above gnd;31;10 m above gnd");
var city = "shtip_north-macedonia_785482";
var client1 = new RestClient(
"https://www.meteoblue.com/en/weather/archive/windrose/" +
city + dateRange +
"&polarunit=hour&degree_resolution=22.5&value_resolution=5"
+"&windspeedunit=KILOMETER_PER_HOUR&submit_csv");
var request1 = new RestRequest(Method.GET);
var name = "historyExport" + d2;
var path = Environment.GetFolderPath(Environment.SpecialFolder.Desktop) +
"/MeteoDownloads/";
client1.DownloadData(request1).SaveAs(path+ name+"-1.csv");

```

Fig. 1. Download first file from meteoblue about Wind(10m above ground)

The code about the other 2 downloads is similar, the only parameter for wind is different. The downloaded results is for Shtip, if we want the other city we need to change value for variable "city".

The second courier is downloading news from the Ministry of Interior website on which we will conduct my security research in the Republic of Macedonia.

We first call the Ministry of Interior website, the news section, and in response, we get their homepage. We use this answer to send another call where we send the login information called cookies, and the generated data we already received from the first call for __VIEWSTATE and __EVENTVALIDATION.

```

try
{
var client = new RestClient(url);
var initialRequest = new RestRequest(Method.POST);

client.UserAgent = "Mozilla/5.0 (Windows NT 10.0; Win64; x64) +
" AppleWebKit/537.36 (KHTML, like Gecko) Chrome/" +
"78.0.3904.108 Safari/537.36";
initialRequest.AddHeader("cache-control", "no-cache");
initialRequest.AddHeader("Connection", "keep-alive");
initialRequest.AddHeader("Accept-Encoding", "gzip, deflate");
initialRequest.AddHeader("Host", "mvr.gov.mk");
initialRequest.AddHeader("Cache-Control", "no-cache");
initialRequest.AddHeader("Accept", "text/html,application/" +
"xhtml+xml,application/xml;q=0.9,image/webp,*;q=0.8");
initialRequest.AddHeader("Content-Type",
"application/x-www-form-urlencoded");

IRestResponse initialResponse = client.Execute(initialRequest);

```

Fig. 2. Part of the code about the crawling

The web site form is made in ASP.NET which works with the doPostBack() function and an application model that enables a particular page to validate and process its own data.

When we have already received the data from their site, it is then parsed in RDF format.

We use the dotNetRDF library to parse in RDF format. This library is written in C# and is designed to get a simple yet powerful RDF data API. As such it has many classes for executing different tasks for reading and writing RDF data and more. Creating triples of data in the form of subject, predicate, and object.

An RDF document can be considered to form a graph, so we represent it as a set of RDF triples as graphs. The assignment of values to locations is done using the

appropriate triple in which the object represents the location which has a proper identifier, the predicate or link represents the name of the location and the object represents the value of the location data. Other triplets for other locations are appropriately defined. All library graphs are IGraph interface implementations and generally derive from the abstract BaseGraph class that implements some of the basic interface methods, allowing specific implementations to focus on specifications as storage/security thread persistence. Implementation of IGraph is the representation of memory in the RDF document. The most commonly used application of IGraph is the Graph class. The library operates first on the level of triplets, graphs and Triple Stores and provides very limited interface support and no direct OWL support. The triples can be added to the graph using the Assert (..) method. The method takes a triple or list of triples.

IV. RESULTS FROM THE RESEARCH

Having already written the data in the FDF format, we did some research on them with SPARQL queries.

If we enter a word in the text field and search for it, then we check to see if the data contains the appropriate term. Since the word may be singular or plural so that we have no problems in the search we cut the last letter and extract the relevant data. If we enter a sentence or several words, then we remove the words consisting of 1 or 2 letters and search the content for the remaining words.

We first did research into which part of Macedonia had the most thefts during 2019.

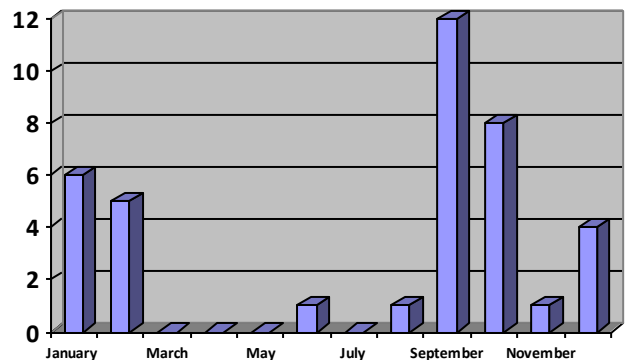


Fig. 3. Results from the first search about "theft"

From the results, we can conclude that the most thefts has been reported in September.

The second survey is in which part of Macedonia has the highest number of homicides in 2019.

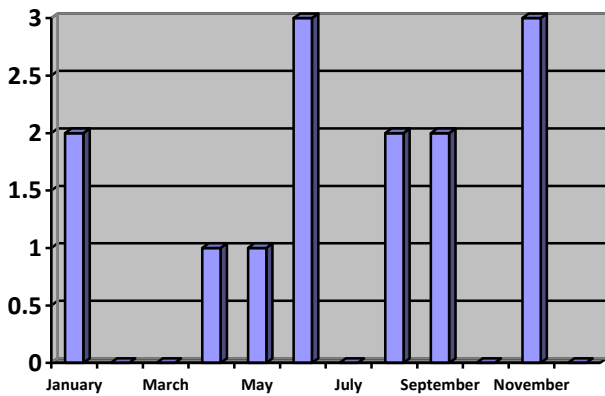


Fig. 4. Results from the second search about “killings”

We can see that the most homicides were reported in June and November. And the third research concerns where there was drug trafficking in 2019.

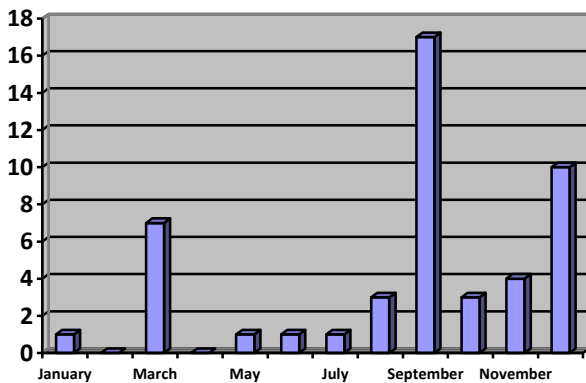


Fig. 5. Results from the third search about “drug trafficking”

Figure 5 has shown us that most reported cases about “drug trafficking” were reported in September.

V. CONCLUSION

Using semantic web and web crawling we can get all the data from any public news that are available and based on the collected data we can make research about any specific field. With this method we can get all the data for longer period which then is parsed in RDF format, from where using queries we filter the news with the parameters we need. We used this method to gather the needed data for our research and based of the data we conduct our investigation.

We can conclude that there were 38 thefts in all of Macedonia. Most thefts were reported in Skopje, a total of 10. No thefts were reported in Kriva Palanka. By the time of execution, most were in September, a total of 12, while none in April and May. There was a total of 14 homicides in Macedonia, mostly in Skopje, 5. Most of the cities do not mention the word homicide.

3 murders were committed on June, 2 in Skopje and 1 in Strumica.

In February and March, there was no mention of murder.

“Drug trafficking” is mentioned 49 times in total. It is often mentioned about the city of Skopje, 14 times in total, and never in Kriva Palanka.

It is mentioned 17 times in September, while in February and April it is never mentioned.

According to these data, we can conclude that Kriva Palanka is the safest city to live in, and Skopje is the riskiest city to live in. We can conclude that security in Macedonia is not improving at all, as drug trafficking is often mentioned in December, and there have been several thefts. The investigation is not 100% valid because, in the context of the extract, it has published a story which states, for example, “murder”, but refers to finding a murderer for an old crime, or for “drug trafficking” news that there was raid in certain cities but no drug-trafficking was found and committed in that month.

REFERENCES

- [1] Introduction to Semantic Algorithms <https://twobithistory.org/2018/05/27/semantic-web.html> May, 2018
- [2] Berners-Lee, Tim, James Hendler, and Ora Lassila. “The Semantic Web.” Scientific American, May 2001
- [3] Introduction to Semantic Algorithms. https://en.wikipedia.org/wiki/Semantic_Web#Applications
- [4] Considered nuggets for data mining. <https://nugetmusthaves.com/Tag/crawler>
- [5] Downloaded data on the nuggets <https://www.nuget.org/>.
- [6] Crawler for downloading data from open pages. <https://nugetmusthaves.com/Package/HtmlAgilityPack>
- [7] Ministry of Interior of the Republic of Macedonia. Published news. <https://mvr.gov.mk/vesti> 2019
- [8] Data Writing Tool in .RDF format. <https://www.nuget.org/packages/dotNetRDF>
- [9] Description of the dotNetRdf frame. <https://github.com/dotnetrdf/dotnetrdf/wiki/UserGuide-Library-Overview> June, 2017
- [10] RDF Schema Syntax. <https://www.w3.org/2001/sw/RDFCore/Schema/200212/> November, 2002
- [11] Meteoblue wind directions. https://www.meteoblue.com/en/weather/archive/windrose/shtip_north-macedonia_785482

Weekly Analysis of Moodle Log Data in RStudio for Prediction

Neslihan Ademi
Computer Engineering Department
International Balkan University
Skopje, North Macedonia
neslihan@ibu.edu.mk

Suzana Loshkovska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
suzana.loshkovska@finki.ukim.edu.mk

Abstract— This paper presents the results of the study to find in how many weeks the prediction/ classification can be done through a learning management system. The results are obtained by analyzing the effects of learners' online presence and activities in Moodle on their grades at the early stages. Analyses are done by partitioning log data of a course in three years by time in terms of weeks. For this purpose we used RStudio and developed script to automatize the analysis. We found that starting from the third week of the lecture period online presence of the students becomes stable and classification can be done starting from that time and accordingly different material and assessment methods can be offered to the students in LMS.

Index Terms—Adaptive Learning Systems, Moodle logs, Learning Analytics, Data Mining, Online Presence

I. INTRODUCTION

Learning management systems (LMSs) such as Moodle are commonly used in the modern education settings as they offer many opportunities to the students and to the tutors. On the other side they open a door for Educational Data Mining (EDM). Huge amount of data stored in LMS databases and log files give the possibility of gaining knowledge about the students following a course. The focus of this paper is the knowledge which can be achieved through the log files by analyzing weekly presence of the students.

Log data collected through LMSs provide descriptive overview of users' online behavior. According to [1] observing behavior provides insights about how people interact with existing systems and services and reveals surprises.

Observational Log Studies contain two common ways to partition log data; by time and by user. Partitioning by time is interesting because log data often contains significant temporal features, such as periodicities (including daily, weekly, and yearly patterns) and spikes in behavior during important events. It is often possible to get an up-to-the-minute picture of how people are behaving with a system from log data by comparing past and current behavior. It is also interesting to partition log data by user characteristics [1].

Recently there are many studies in the literature about log analysis in e-learning environments [2]–[6]. In [2], authors profile the students' learning approaches through the Moodle logs. Some of them are for only visualization purposes [3], some are for the prediction like [4] and [5].

Data mining and Learning Analytics can be useful to explore, visualize, and analyze in order to understand students' learning behavior and apply the gained knowledge for the adaptation of the learning contents.

The e-learning data mining process consists of four steps like in the general data mining process. Collect data, Pre-process the data, Apply data mining, Interpret, evaluate and deploy the results. [7]

In a previous study [8] we examined the relationship between variables which are obtained from Moodle logs and students' grade. We found positive correlations which show the linear relationship between all Moodle activities and the course grade of the students. In [9], we applied three Machine Learning algorithms; namely Decision Tree(DT), Bayesian Network(BN) and Support Vector Machine(SVM) to make prediction of the grades based on online activities of the students. In [10], we found the possibility of early predictions of grade through Moodle log data considering 7 weeks out of 14 weeks of lecturing time and we suggested early predictions to be used for preventing drop outs.

The purpose of this study is to find the relation between the online presence and the grades in an early stage. The research questions which we want to find answer for are: "1. How many weeks would be enough for an adaptive system to categorize the learners according to their online activities?" and "2. Which online activities can be taken into consideration in that time?"

This paper presents the possibility of making predictions in earlier weeks of studies by making weekly analysis of online activities. The second section defines the used methodology for the study, while the third section gives results and discussion; finally the last section is the conclusion.

II. METHODOLOGY

In this section, the used methodology is explained. For the study, log files are taken from Moodle which is installed and used at the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, North Macedonia. Recorded log files of Moodle LMS are downloaded in the form of .csv and they contain all online activities of the students for a one semester bachelor's degree course of User Interfaces. The course is in the third year of the studies. For the analysis to find the trends in online presence three academic years in a row are taken: 2016-2017, 2017-2018 and 2018-2019, where the teaching process was designed as blended learning. Moodle was used to support classroom teaching to distribute course material,

lectures, homework, laboratory exercises and to provide discussion through the forums within the course.

The standard retrieved fields in the log files are: Time, User full name, Affected user, Event context, Component, Event name, Description, Origin, IP address. Log files were composed of approximately 150K rows for each. Number of regular students was 255 for the first year, 205 for the second year and 260 for the third year.

A. Data Pre-processing

Applied pre-processing steps to the data were, data cleaning, transformation and integration. We used sqldf package which allows complex database queries in Rstudio. All pre-processing steps in details are listed below:

1. Clean the log files from the events performed by instructors and administrators as the focus is the students' actions.
2. Remove log data produced by the system by filtering the data where component field is system.
3. Remove duplicate records.
4. Extract userID from the description text and generate new columns for userID to be used instead of user full name to provide anonymity.
5. Transform the file into a data frame with the name of "events.csv" which will contain data attributes given in Table I.
6. Integrate the new file (events.csv) to the scores file (scores.csv) which gives the final grades of the students.

TABLE I. ATTRIBUTES AFTER DATA TRANSFORMATION

Name	Description
UserID	ID number of the student
Visits	Total number of visits by the student
Quizzes	Number of quizzes taken by the student
Assignments	Number of submitted assignments by the student
ForumCreated	Number of forum creations by the student
ForumView	Number of forum views by the student
CourseView	Number of course views by the student
FileSubmission	Number of file submissions by the student
GradeView	Number of grade views by the student

B. Data Exploration and Statistical Analysis

Data exploration and statistical analysis consist of following steps; visualizing online presence of the students for each week, correlations of total visits, quizzes, assignments, forum creations, forum views, file submissions, and grade views with the course grade.

Correlations are found by using Pearson correlation test. Equation 1 gives the formula for Pearson correlation coefficient. The value of r is always between -1 and $+1$. $r = -1$ or $+1$ indicates a perfect linear relationship, where sign indicates the direction and $r = 0$ indicates no linear relationship.

$$r = \frac{\sum (x - \bar{X})(y - \bar{Y})}{\sqrt{[\sum (x - \bar{X})^2][\sum (y - \bar{Y})^2]}} \quad (1)$$

We have used Pearson correlation coefficient to discover whether there is a linear relationship between the attributes given in Table I and the grade.

III. RESULTS AND DISCUSSION

A. Frequencies of student activities

Upon the analysis of the log data; we noticed that, for the first year in the first week of the semester only 143 students visited the course page on Moodle. This means 44% of the students did not have any online presence. In the second week online presence is increasing to 91%, and in the third week to 96%. Fig.1 gives the online presence in the course in terms of number of active students on Moodle for the first seven weeks of the semester for the academic year 2016-2017

Fig.2 gives the online presence of the students for the academic year 2017-2018. As it can be calculated from the figure, in the first week online presence is 68%, in the second week 90% and in the third week 99%.

Fig.3 shows the same trend as in Fig.1 and 2; starting from the third week online presence is regular.

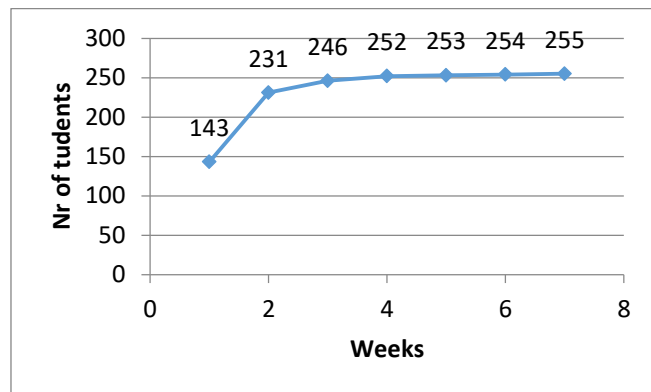


Fig. 1. Number of active students on Moodle per week in 2016-2017

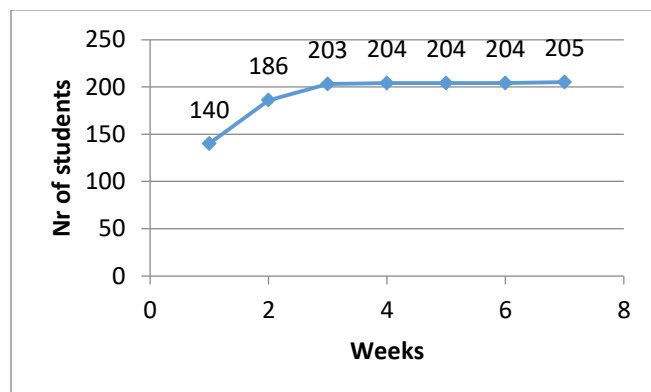


Fig. 2. Number of active students on Moodle per week in 2017-2018

TABLE II. CORRELATIONS OF TOTAL VISITS AND COURSE VIEWS WITH THE GRADE BY WEEKS (2016-2017)

	1 week	2 weeks	3 weeks	4 weeks	5 weeks	6 weeks	7 weeks
Total Visits	0.17	0.26	0.34	0.33	0.37	0.35	0.35
Course Views	0.13	0.25	0.31	0.30	0.29	0.27	0.28

TABLE III. CORRELATIONS OF TOTAL VISITS AND COURSE VIEWS WITH THE GRADE BY WEEKS (2017-2018)

	1 week	2 weeks	3 weeks	4 weeks	5 weeks	6 weeks	7 weeks
Total Visits	0.12	0.22	0.30	0.28	0.30	0.25	0.25
Course Views	0.11	0.21	0.26	0.21	0.22	0.21	0.21

TABLE IV. CORRELATIONS OF TOTAL VISITS AND COURSE VIEWS WITH THE GRADE BY WEEKS (2018-2019)

	1 week	2 weeks	3 weeks	4 weeks	5 weeks	6 weeks	7 weeks
Total Visits	0.01	0.11	0.22	0.26	0.27	0.26	0.27
Course Views	-0.04	0.05	0.16	0.2	0.23	0.24	0.24

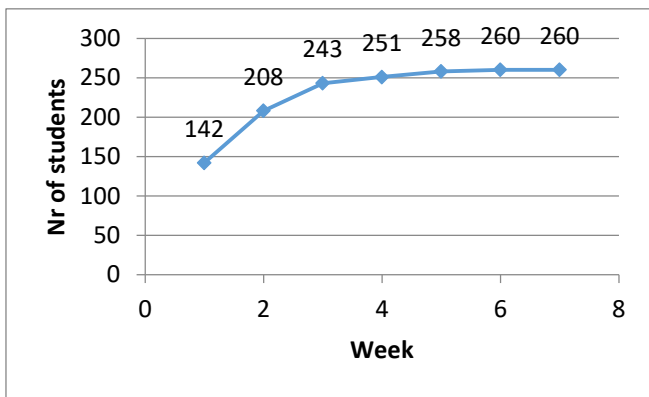


Fig. 3. Number of active students on Moodle per week in 2018-2019

Overall picture of three academic years show that starting from the third week of the lectures online presence becomes more stable.

B. Correlations

Table II, III and IV give the correlations of total visits and course views with grade by weeks for the academic years 2016-2017, 2017-2018 and 2018-2019 respectively. As it can be seen from these tables, after the 3rd week of the lectures students' online activities have similar correlations with their grade, they are following the same trend (also see Fig.4, 5 and 6).

In Table II, III and IV, only first seven weeks of the lectures are given as our focus is to see how much earlier the correlations start to be stable for early predictions. In the study, correlations of other online activities extracted from log files (given in Table I) with the grade are calculated but they are not shown in Table II, III and IV. Some of those correlations are not available in early stages such as grade views, forum views and forum creations or they are not significant with very low correlations such as quizzes, file submissions and assignments.

When the total course period of 14 weeks was taken into consideration as in our previous study [8], correlation coefficient of the total visits was 0.55 and correlation coefficient for the course views was 0.43 for the academic year 2016-2018. For the seventh week it is calculated here

for the same year as 0.35 for total visits and 0.28 for course views.

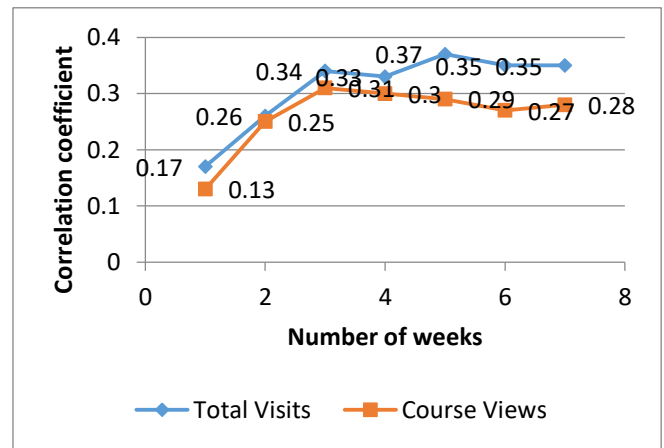


Fig. 4. Correlation coefficients of total visits and course views per week in 2016-2017

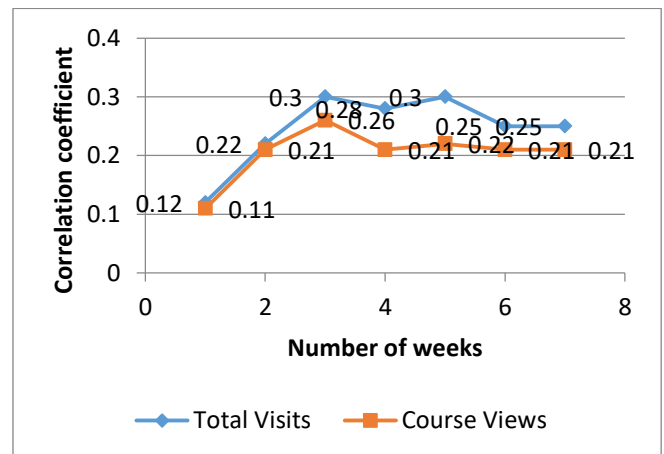


Fig. 5. Correlation coefficients of total visits and course views per week in 2017-2018

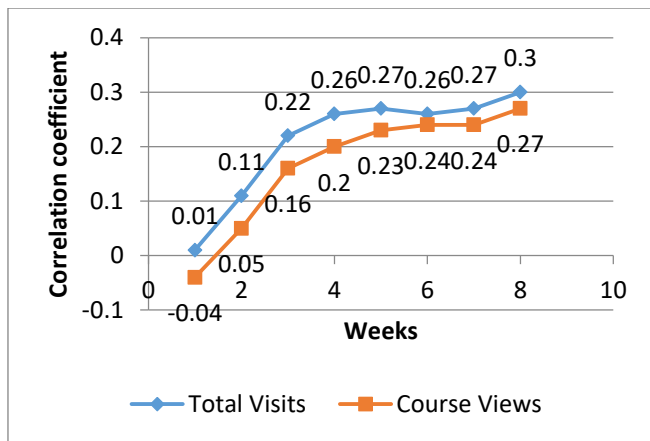


Fig. 6. Correlation coefficients of total visits and course views per week in 2018-2019

IV. CONCLUSION

In this paper we used some learning analytics techniques such as pre-processing and statistical analysis; to discover the effect of students' online presence on their grades for the early stages of the learning period. This knowledge will be a base of our future data mining process. We also developed an R script to automatize these analyses for the future use in our more detailed examinations and predictions.

Present study showed the relationship between variables which are obtained from Moodle logs and students' grade for the first half of the learning period. The correlations for the attributes such as forum views, forum creations, file submissions, assignments and quizzes are not as significant as for the complete period of the learning process. Some of those correlations are not available in early stages such as grade views, forum views and forum creations or they are not significant with very low correlations such as quizzes, file submissions and assignments. We found correlations between total visits and course views with grade are higher and these correlations tend to be stable starting from the third week of the lectures. So for the future adaptive learning system architecture total visits and course views can be taken as two of the criteria in terms of online presence for classification purpose in the early weeks, starting from the third week. And after the second half of the course period the other attributes such as forum views, file submissions, quizzes and assignments can be included.

Future direction of our research will be analyzing also preferences of the students in terms of learning materials and assessment methods for applying data mining techniques to provide adaptivity in learning management systems.

REFERENCES

- [1] S. Dumais, R. Jeffries, D. M. Russell, D. Tang, and J. Teevan, "Understanding User Behavior Through Log Data and Analysis," in *Ways of Knowing in HCI*, New York, NY: Springer New York, 2014, pp. 349–372.
- [2] G. Akçapınar, "Profiling Students' Approaches to Learning through Moodle Logs," *Proceedings of Multidisciplinary Academic Conference on Education, Teaching and E-learning in Prague 2015, Czech Republic (MAC-ETeL 2015)*, no. December, p. 7, 2015.
- [3] A. Konstantinidis and C. Grafton, "Using Excel Macros to Analyse Moodle Logs," *UK Research.Moodle.Net*, no. September, pp. 4–6, 2013.
- [4] Á. Figueira and Álvaro, "Mining Moodle Logs for Grade Prediction," in *Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality - TEEM 2017*, 2017, pp. 1–8.
- [5] T. Käser, N. R. Hallinen, and D. L. Schwartz, "Modeling exploration strategies to predict student performance within a learning environment and beyond," in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK '17*, 2017.
- [6] M. Cocca and S. Weibelzahl, "Log file analysis for disengagement detection in e-Learning environments," *User Modeling and User-Adapted Interaction*, 2009.
- [7] C. Romero, S. Ventura, and E. García, "Data mining in course management systems: Moodle case study and tutorial," *Computers and Education*, vol. 51, no. 1, pp. 368–384, 2008.
- [8] N. Ademi and S. Loshkovska, "Exploratory Analysis of Student Activities and Success Based On Moodle Log Data," in *16th International Conference on Informatics and Information Technologies*, 2019.
- [9] N. Ademi, S. Loshkovska, and S. Kalajdziski, "Prediction of Student Success Through Analysis of Moodle Logs: Case Study," in *11th International Conference, ICT Innovations 2019, Ohrid, North Macedonia, October 17–19, 2019, Proceedings*, S. Gievska and G. Madzarov, Eds. Springer, Cham, 2019, pp. 27–40.
- [10] N. Ademi and S. Loshkovska, "Early Detection of Drop Outs in E-Learning Systems," *Academic Perspective Procedia*, vol. 2, no. 3, pp. 1008–1015, Nov. 2019.

Artificial Intelligence: Simulating Human Emotion and Surpassing Human Intelligence

Filemon Jankuloski

*School of Computer Science and Information Technology
University American College Skopje
Skopje, Macedonia
filjankuloski@gmail.com*

Adrijan Božinovski

*School of Computer Science and Information Technology
University American College Skopje
Skopje, Macedonia
bozinovski@uacs.edu.mk*

Veno Pachovski

*School of Computer Science and Information Technology
University American College Skopje
Skopje, Macedonia
pachovski@uacs.edu.mk*

Abstract—In this paper, we explore the potentials of artificial intelligence and the benefits which can be brought about through its advancements. The purpose of the paper is to discuss how closely AI devices are able to mimic Human Intelligence and if there is a possibility that machines will be able to surpass this intelligence. In order to achieve this, first, we focus on the history of AI and many of its accomplishments over a period of 70 years. Not only do we take a look at the first instance in which an individual questions the difference in Machine and Human Intelligence, but we also look at how AI's foundation was built by the head of Artificial Intelligence, John McCarthy, and his four colleagues. Next, we discuss different AI types classified by two different aspects: capability and functionality. By defining the classifications of AI, we are then able to pinpoint how far humanity has come in creating machines which can mimic Human Intelligence. We analyze Kismet, the very first robot to simulate human emotions, and Sophia, the current pinnacle of emotional Artificial Intelligence machinery and ascertain which category of AI they fall under. Finally, the paper concludes by discussing the future of AI advancements and the possible outcomes that come with reaching Superior Artificial Intelligence, the most powerful, and yet challenging AI machinery.

Keywords—*Artificial Intelligence, Machine Intelligence, Human Intelligence*

I. INTRODUCTION

Artificial intelligence has become increasingly prevalent in our current society. We can find artificial intelligence being used in many aspects of our daily lives, such as through the use of Amazon Alexa, the self-driving Tesla vehicle, and AlphaZero, the neural network based chess playing program.

However, there is no universally accepted definition of intelligence and, consequently, there is no universally accepted definition of Artificial Intelligence. We will, however, mention a definition of intelligence given in a textbook of Artificial Intelligence: “Intelligence is an ability of a matter in an available time to process relevant information” [1]. There is no clear and concise definition for Artificial Intelligence, since it is a subject which still has massive space for improvement and has only recently been making most of its significant advancements. Leading textbooks on AI define it as the study of “intelligent agents”, which can be represented by any device that perceives its environment and takes actions that maximizes its chances of achieving its goals [2]. Although this is the formal definition, most individuals think of Artificial Intelligence by comparing it to Human Intelligence. Essentially, human beings possess natural intelligence, AI devices possess machine intelligence, and one could say that the goal of artificial intelligence is to make it so that machine intelligence simulates human intelligence as closely as possible, to achieve some purpose for the benefit of humanity. However, this begs the question: just how closely can artificial intelligence simulate human intelligence and human emotion? Is it possible for a machine to perceive its own existence in the same way a human being can? Before we can reach an answer to this, it's imperative to dive into the history of artificial intelligence and how the subject was conceived.

II. HISTORY

The subject of artificial intelligence has been in existence for roughly 70 years, since 1950. Although he did not coin the term “artificial intelligence”, Alan Turing was the first

individual to suggest that human intelligence and machine intelligence are comparable, in his famous 1950 article called "Computing Machinery and Intelligence" [3]. In this article, he explained that if individuals were incapable of making the discernment between a machine and a human being in a teletype dialogue, then it would not be farfetched to say that a machine is capable of intelligence. The true birth of artificial intelligence, however, occurred in 1956 at a workshop in Dartmouth College, where the term "Artificial Intelligence" was coined by John McCarthy [4]. Interestingly enough, the true purpose of this coinage was so that individuals would be able to distinguish between Artificial Intelligence and Cybernetics; Cybernetics being the study of the control and the communication of machines. Originally, Dartmouth College was meant to hold a conference, but due to skepticism and a lack of interest, no more than five people consistently sat through the conference, including McCarthy himself. However, John McCarthy, Allen Newel, Marvin Minsky, Herbert Simon, and Arthur Samuel were the sole five people who built the foundation for Artificial Intelligence to thrive.

The five founders of AI and their students began creating the world's first AI based programs. For example, computers were learning chess strategies starting in 1954 and, by 1959, these computers had become better than the average human at playing chess [5]. Chess was not the only thing computers were able to learn at the time, as there were computers also solving word problems in algebra, proving logical theorems, and speaking English [6]. By the mid-1960s, Artificial Intelligence had become a massive success and was heavily funded by the Defense Advanced Research Projects Agency (DARPA) [6]. Unfortunately, advancements made in Artificial Intelligence were halted by two large fundamental problems: low memory capacities and incredibly slow processing speeds. As a result, funding was cut from Artificial Intelligence and interest in the subject gradually died off. This stretch of time in which AI struggled to acquire funding was known as the First AI Winter [7].

The First AI Winter ended with the introduction of "Expert Systems" in the 1970s, which were adopted, developed, and integrated by competitive companies globally [7]. The main focus on Artificial Intelligence was now to utilize the accumulated knowledge of experts in several different fields to create programs. Expert systems were able to answer questions and solve problems in several different fields. Due to the simplistic design of expert systems, companies would be able to design, create, and update programs with relative ease.

During the 1970s, a field in Artificial Intelligence emerged which was related to neural networks. Although the field first emerged in 1943, it had a renaissance in 1986 after the book of Rumelhart, McClelland, and the PDP group [8–9]. This field of research experienced rapid growth, whereas classical Artificial Intelligence, based on Expert Systems, eventually declined in popularity.

In the early 1990s, Artificial Intelligence finally made another big breakthrough in the form of "intelligent agents". Intelligent agents are used for news retrieval services, online shopping, and browsing the web in the form of personal digital or personal virtual assistants [7]. A modern day example of such an assistant would be the Amazon Alexa. Intelligent agents, however, were not the only breakthrough made in Artificial Intelligence. In 1986, a robot was controlled using speech commands [10]. In 1997, reigning world chess champion Gary Kasparov was defeated by IBM's Deep Blue, a chess playing computer program [11]. In the same year, speech recognition software developed by Dragon Systems would be implemented into Windows for the first time [11]. The speech recognition software would be a large stepping stone for developing the aforementioned virtual assistants such as Amazon Alexa.

III. ARTIFICIAL INTELLIGENCE CLASSIFICATIONS

History is incredibly important for understanding how the foundation of Artificial Intelligence came to be, but it is equally important to understand the distinct classifications of Artificial Intelligence. The latter can be divided according to two different properties: capability and functionality. As far as capabilities go, machines can possess Artificial Narrow Intelligence (ANI), Artificial General Intelligence (AGI), and Artificial Superintelligence (ASI), whereas, for functionality, machines can be one of four different types: reactive machines, limited memory, theory of mind, and self-aware [12].

All AI based technology, including the most advanced machines capable of self-learning, fall under the category of Artificial Narrow Intelligence. The purpose of Artificial Narrow Intelligence is to accomplish one task or a very basic set of tasks. These machines typically possess a prearranged set of functionalities and act autonomously. A good example of this would be IBM's Deep Blue, as its only task is to defeat the opposing player in a game of chess. These types of machines are only capable of doing what they are programmed to do and whatever is within their scope.

At a minimum, any machine which possesses Artificial General Intelligence is capable of accomplishing any task an intellectual human being can accomplish with the same efficiency. Unlike Artificial Narrow Intelligence, the Artificial General Intelligence systems will be able to independently build multiple competencies in several fields and form connections and generalizations across multiple domains, which means that these systems will be proficient in several different fields [13]. The broadness of the tasks that an Artificial General Intelligence system will be able to fulfill is far more extensive than that of Artificial Narrow Intelligence. These systems will also be able to rationalize and make important decisions much like an intellectual human being can. Currently, there is not a single system which can be classified as an Artificial General Intelligence system, but there are global research labs with an abundance of funding attempting to make the first Artificial General Intelligence system a reality.

Artificial Superintelligent systems are currently only hypothetical and far from becoming realized, since we have yet to even fully develop an Artificial General Intelligence system. If Artificial General Intelligence systems are meant to be equivalent to above average human intellect, then Artificial Superintelligence is meant to be far superior to even the most gifted of human minds. Artificial Superintelligent systems will excel human beings in all logical functions because of greater memory capabilities, faster data processing and analysis, and more efficient decision making [14]. The existence of such a system brings up ethical dilemmas, because the capabilities of these systems would enable them to replace human beings as the most intellectual creatures to exist.

Reactive machines are amongst the most basic of all the AI technologies. Much like their name implies, they are only capable of appropriately reacting to particular stimuli. Reactive machines cannot store memories or past experiences and use them to influence or optimize their decision making process. They are often programmed to accomplish a single task. A good example of this would yet again be IBM’s Deep Blue. Deep Blue’s sole task is to be able to defeat any opponent in a game of chess. When the opponent makes a move, Deep Blue analyzes the positions of all chess pieces on the board and reacts accordingly, making the optimal decision to accomplish its given task. Deep Blue does not possess any learned historical data in which to base its decisions, but it might make some predictions based on the current state of the chess board. The historical data are preprogrammed.

Limited memory machines are much like reactive machines, but a step further. Limited memory machines are, as their name implies, machines which are able use their short term memory to be able to better accomplish their assigned tasks. A good example of this would be autonomous driving, which is now possible with Elon Musk’s creation: the Tesla. The concept of using limited memory is incorporated through sensors in the car that are able to detect several things such as: when a pedestrian is crossing a walkway, poor road conditions, weather, traffic lights, lane detection, and more [15]. These sensors within Teslas play a large role in avoidance of autonomous driving accidents. Unlike the Deep Blue, Teslas use both preprogrammed knowledge as well as knowledge taken from observations from their sensors in order to fulfill their purpose optimally. This observational knowledge is what distinguishes limited memory machines from reactive machines.

The “Theory of Mind” refers to a psychological ability which all humans possess. The theory of mind is the understanding that people, creatures, and objects in the world can have thoughts and emotions that affect their behavior [16]. In the case of Artificial Intelligence, this would mean that a machine would need to understand that there are others who have thoughts and emotions which affect their behavior, and that the machine would need an adequate reaction depending on the type of behavior that a person is exhibiting. These types of machines are fully capable of social interactions, so much

so that it would simulate two human beings conversing rather than a machine and a human. Such machines have already been built to some extent and will be discussed subsequently.

Self-awareness is the final stage of Artificial Intelligence and is much like an extension of the theory of mind. Machines which possess self-awareness will not only have the ability to recognize and replicate human emotions, but they will also have the ability to think for themselves, have desires, and understand their own feelings [17]. Reaching this stage would be the pinnacle of Artificial Intelligence, because machines being sentient would mean that there is almost no differentiation to be made between the intelligence of a human being and the intelligence of a machine. In this current time, we have only reached the third stage of the functionalities of Artificial Intelligence. The machines which will be discussed subsequently are machines which do not possess self-awareness, but are still able to recognize and replicate human emotions and have social interactions with human beings.

IV. THE EVOLUTION OF SENTIENT MACHINERY

Table I. History of AI Devices

Name	Year of Creation	Artificial Intelligence Type	Main Characteristics
ELIZA	1964	Reactive	ELIZA runs scripts which give it the ability to process user inputs and give outputs which would result in seemingly normal discourse.
WABOT-1	1970	Reactive	Humanoid robot which possesses a limb control system and a vision system. It is able to communicate in Japanese and grip onto objects.
WABOT-2	1980	Reactive	This robot is the same as WABOT-1, but it uses its vision system to read musical scores, and uses its limb control system to play songs of average difficulty on the

			electric organ.
Crossbar Adaptive Array	1981	Reactive	The first AI system capable of using emotions to reach a desirable state.
Kismet	1997	Limited Memory	Kismet is capable of displaying human emotions and reacting to the emotions of others.
Sophia	2016	Limited Memory	Sophia is capable of displaying over 100 emotional expressions, partaking in small talk in predefined subjects, walking, and can also remember people it has previously met.

A. ELIZA

As aforementioned, Alan Turing suggested that machine intelligence and human intelligence are comparable if individuals can't make the differentiation between man and machine. The Turing test was attempted with the creation of ELIZA, a natural language processing machine. ELIZA ran a script named DOCTOR such as to mimic the responses of a psychotherapist in a psychiatric interview [18]. It gave the illusion of giving deceptively intelligent responses in a conversation by simply reflecting what an individual would write to it. For example, if someone were to write to it "I am feeling depressed?", ELIZA would ask "Why are you feeling depressed?" as a response. ELIZA's similarity to a human being is in fact so convincing, that when it was used for conversational therapy, many individuals reported that they had forgotten they were talking to a machine [18]. Despite being able to deceive many individuals, ELIZA's mimicry of human beings is a facade, as it is incapable of bringing context into conversations, and truly understanding what is being said to it on an emotional level.

B. WABOT

In 1970, four laboratories in Waseda University, Japan began developing the world's first humanoid robot, WABOT-1 [19]. This robot was capable of controlling its arms and legs, had a vision system, and could communicate with others in Japanese. WABOT-1 could also use its arms to

grip and carry objects, and was capable of measuring distances between itself and other objects using external sensors. In 1980, the same four laboratories came together to create an entirely separate project known as WABOT-2 [19]. The sole purpose of this robot was to be able to perform songs of average difficulty on the electric organ. WABOT-2, however, was also capable of using its vision system to read musical scores and could use its communication system to "accompany" others while it plays on the electric organ. Unlike WABOT-1, which is capable of a broader range of functions, WABOT-2 is a specialized robot, designed and created only to achieve one purpose. While neither robot is capable of displaying or reacting to emotions, they were the first robots to achieve a human-like physical appearance and capable of performing human-like functions.

C. Crossbar Adaptive Array

Crossbar Adaptive Array (CAA) was the first working AI system capable of effectively using emotions [20]. In it, emotions were defined as state evaluations. Examples of the values were desirable, undesirable and neutral states or situations. Emoticons (☺, 😊, ☹) were used to represent emotion values. The CAA was built around a crossbar memory which was able to compute both emotion of being in a situation and a behavior to meet that situation. Emotions were used in the learning system of CAA. The CAA was first built and tested at the Computer and Information Science Department of University of Massachusetts at Amherst.

D. Kismet

One of the first machines capable of recognizing and simulating emotion was a robot named Kismet, which was made in 1997 by Cynthia Breazeal at Massachusetts Institute of Technology [21]. The way in which Kismet interacts with human beings is meant to be infant-like in nature, which simulates a caretaker/infant relationship. The architecture of Kismet's system consisted of six distinct subsystems: the low level feature extraction system, the perception system, the attention system, the motivation system, the behavior system, and the motor system [22]. The motivation system is much like the perception system in that they both heavily influence which emotion Kismet will simulate. In the behavior system, all behaviors act as self-interested, separate entities which fight for priority and an arbitration system is necessary in order to determine which behavior will remain active and for how long, given that Kismet has several different motivations to tend to [22]. The motor system is what allows Kismet to express these behaviors. They are responsible for allowing Kismet to perform vocal acts, move different parts of its face and body, and also change the orientation in its face and eyes. Kismet does an exceptional job at simulating emotions and recognizing emotions with the help of its complex system. However, this does not mean that it is a sentient, self-aware machine, as all of its functionalities are fully preprogrammed and structural. If it were truly sentient, it would be conscientious of its own emotions and think or act on its own

terms rather than simply react to stimuli in a flowchart fashion.

E. Sophia

Kismet was created in 1997, but how far has society come in reaching Artificial General Intelligence and self-awareness in machines and have any improvements been made? On February 14, 2016, Sophia of Hanson Robotics was activated for the first time [23]. Sophia, thus far, is one of the only robots which closely resembles a normal sized human adult. It is able to imitate over a hundred human gestures and facial expressions. In January 2018, she was given the ability to walk and tread on terrains of any kind [24]. She also has long term memory capabilities as she is able to recognize individuals she has spoken to before. Also, unlike Kismet, she is able to orally simulate social interaction and make simple small talk in predefined subjects [25]. Sophia has been designed to constantly improve in her social skills through conversational analysis, and, as years pass by, she is likely to give quicker responses, make fewer errors in social interactions, and answer more complex questions with higher accuracy [26].

V. CONCLUSION

In a span of 70 years, Artificial Intelligence has shown promise and has made major advancements even in the face of adversity. Although society has yet to construct a machine with Artificial General Intelligence, robots like Kismet and Sophia show us that we are consistently taking steps in the right direction. Furthermore, the AlphaZero chess playing engine has surpassed human intelligence significantly, when it comes to chess – the best human chess players have a ranking around 2800, whereas AlphaZero’s ranking is around 3500.

At the current rate, within 100 years, it is possible that all machines will possess Artificial General Intelligence, and machines will be able to coexist and cooperate with human beings in order to efficiently complete important tasks such as construction work or other kinds of physical labor.

While there are certainly benefits that come with Artificial General Intelligence systems, there are also concerns which come with these machines. For example, one could say that the purpose of Artificial Intelligence is to outclass human beings in making efficient and optimal decisions. However, there is a possibility that if machine intelligence were ever to evolve to the level of Artificial Superintelligent Systems, human beings may be seen as obsolete by these systems and as a result, human beings could be in serious danger because these systems may choose to “erase” anything which they deem unnecessary. Another concern is the ethical dilemma that comes with anything at the level of Artificial General Intelligence or higher. If a machine is able to think and feel and is fully aware of its own existence, then do machines also deserve human rights and a fair opportunity at a happy life?

There are both benefits and consequences that may come with artificial, sentient life forms. While it is fascinating to think that society could create perfect life forms in the form of

machines, it is important as a society that we think of the worst case scenario and how we can take precautionary measures to avoid anything potentially threatening to humanity. Artificial Intelligence is a fascinating and exciting subject, but as we continue making advancements in the field, we should be absolutely certain that each step is taken with the utmost caution.

REFERENCES

- 1) Bozinovski, S. 1994, The Artificial Intelligence. (In Macedonian) Gocmar.
- 2) Russell, S. and Norvig, P., 2003. Artificial Intelligence: A Modern Approach. 2nd ed. Upper Saddle River, New Jersey: Prentice Hall, p.55.
- 3) Intelligence, A., AI, W. and Europe, C., 2020. History Of Artificial Intelligence. [online] Artificial Intelligence. Available at: <<https://www.coe.int/en/web/artificial-intelligence/history-of-ai>> [Accessed 3 March 2020].
- 4) McCarthy, J., 1996. Defending AI Research. New York: Cambridge University Press, p.73.
- 5) Samuel, A., 1959. Some Studies In Machine Learning Using The Game Of Checkers. Oxford [u.a.]: Pergamon Press, pp.210-229
- 6) McCorduck, P., 2004. Machines Who Think. 2nd ed. Natick, Massachusetts: A.K Peters, p.34.
- 7) DATAVERSITY. 2020. A Brief History Of Artificial Intelligence - DATAVERSITY. [online] Available at: <<https://www.dataversity.net/brief-history-artificial-intelligence/#>> [Accessed 4 March 2020].
- 8) McCulloch, W., Pitts W., 1943, A logical calculus of the idea immanent in nervous activity, Bulletin of Mathematical Biophysics 5: 115-133
- 9) Rumelhart, D., McClelland J., and the PDP Group, 1986, Parallel Distributed Processing. Explorations in the Microstructure of Cognition. MIT Press
- 10) Grujovski, G., Bozinovski S., 1989, Realization of a system for speech control of a mobile robot, (In Macedonian) Proceedings of the 6th Yugoslav symposium on Applied Robotics and Flexible Automation, Novi Sad, p. 227-235
- 11) Science in the News. 2020. The History Of Artificial Intelligence - Science In The News. [online] Available at: <<http://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/>> [Accessed 4 March 2020].
- 12) www.javatpoint.com. 2020. Types Of Artificial Intelligence - Javatpoint. [online] Available at: <<https://www.javatpoint.com/types-of-artificial-intelligence>> [Accessed 7 March 2020]
- 13) Joshi, N., 2020. 7 Types Of Artificial Intelligence. [online] Forbes. Available at: <<https://www.forbes.com/sites/cognitiveworld/2019/06/19/7-types-of-artificial-intelligence/#7e287ba6233e>> [Accessed 7 March 2020].
- 14) Koombea – Bringing apps to life. 2020. Understanding The 7 Types Of Artificial Intelligence. [online] Available at: <<https://www.koombea.com/blog/understanding-the-7-types-of-artificial-intelligence/>> [Accessed 7 March 2020].
- 15) The Tech Edvocate. 2020. The 4 Types Of Artificial Intelligence: What Educators Should Know - The Tech Edvocate. [online] Available at: <<https://www.thetechedvocate.org/the-4-types-of-artificial-intelligence-what-educators-should-know/>> [Accessed 7 March 2020].
- 16) Hintze, A., 2020. From Reactive Robots To Sentient Machines: The 4 Types Of AI. [online] livescience.com. Available at: <<https://www.livescience.com/56858-4-types-artificial-intelligence.html>> [Accessed 7 March 2020].
- 17) Reynoso, R., 2020. 4 Main Types Of Artificial Intelligence. [online] Learn.g2.com. Available at: <<https://learn.g2.com/types-of-artificial-intelligence>> [Accessed 7 March 2020].
- 18) Weizenbaum, J., 1993. Computer Power And Human Reason: From Judgement To Calculation. San Francisco: Penguin Books, pp.2, 3, 6, 182, 188, 189.
- 19) Humanoid.waseda.ac.jp. n.d. Humanoid History -WABOT-. [online] Available at: <http://www.humanoid.waseda.ac.jp/booklet/kato_2.html> [Accessed 4 May 2020].
- 20) Bozinovski S., 1988, Adaptive Industrial Robots, (in Macedonian) Final report of the project, Republic Assembly for Scientific Activities of Macedonia
- 21) Menzel, P. and D'Aluisio, F., 2000. Robo Sapiens: Evolution Of A New Species. Cambridge, Massachusetts: MIT Press, p.66
- 22) Ai.mit.edu. 2020. Sociable Machines - The Framework. [online] Available at: <<http://www.ai.mit.edu/projects/sociable/kismet.html>> [Accessed 8 March 2020].
- 23) Mallonee, L., 2020. Photographing A Robot Isn't Just Point And Shoot. [online] Wired. Available at: <<https://www.wired.com/story/photographing-a-robot/>> [Accessed 8 March 2020].
- 24) Video, T., 2020. Sophia The Robot Takes Her First Steps. [online] The Telegraph. Available at: <<https://www.telegraph.co.uk/technology/2018/01/08/sophia-robot-takes-first-steps/>> [Accessed 8 March 2020].
- 25) Hanson Robotics. 2020. News - Hanson Robotics. [online] Available at: <<https://www.hansonrobotics.com/news/>> [Accessed 9 March 2020].
- 26) Hanson Robotics. 2020. Sophia - Hanson Robotics. [online] Available at: <<https://www.hansonrobotics.com/sophia/>> [Accessed 9 March 2020].

A survey of covert channels: Benign and malicious usage, conditions for creation and countermeasures

Ema Stamenkovska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University, Skopje
e.stamenkovska92@gmail.com

Vesna Dimitrova
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University, Skopje
vesna.dimitrova@finki.ukim.mk

Abstract - Secure communication is of great importance to everyone, but how to secure a communication when a method that can hide its very existence exists? This can be achieved by a covert channel which can be used for a secret transfer of information. The following article gives a survey of the types of covert channels, the basic conditions for creating an undetectable covert channel as well as the countermeasures that can be taken for its limitation or elimination. An example of a covert channel attack (more specifically DNS tunneling) will be shown. The need for a data leak prevention system, a tool that helps safeguard the sensitive data will be punctuated. Some of the covert channels' uses for corrupt causes will be mentioned. In some cases, security can be actually improved by the use of covert channels.

Key words - covert channel, DNS attacks, DNS benchmark, data loss prevention

I. INTRODUCTION

Network information hiding is a discipline that deals with the concealment of network communications or their characteristics by encapsulating the capabilities that a particular module has behind an abstract interface. The secret communication channel created by the data hiding method is referred to as a network covert channel and such channels exist throughout the network protocol architecture. It is very important that network designers and security managers understand that while a serious threat even for single hosts exists, the potential for covert channels in computer networks is greatly increased. In computer networks, overt channels, such as network protocols, may be used as carriers for covert channels and their network systems may be effectively subverted in a wide variety of locations within the network [1,2,3,4]. Information hiding can be used for benign purposes, like network authentication, copyright protection etc. but the malicious ones are unfortunately more frequent. Some of them will be described in this survey.

Section II contains information about the classification of covert channels. In section III the basic conditions for creating an undetectable covert channel and the countermeasures that can be taken against covert channels will be viewed more closely. In section IV the main tool that helps safeguard the sensitive data, the data loss

prevention system is explained. Section V shows the vast usage of covert channel: for copyright protection, authentication and authorization, for corrupt causes etc. In section VI an example of DNS tunneling is described.

II. CLASSIFICATION OF COVERT CHANNELS

Mainly, there are two types of network covert channels that can be created [5]:

- Covert storage channel, where a process writes data to a storage location, such as hard drive, after which another process of lower clearance is able to read it. So, person A at a low security level is able to read data written by person B at a higher security level. Establishing this type of covert channel is achieved by manipulating the header fields of packets in protocols, like: TCP, IP, ICMP and HTTP. In IP communication that includes changing mostly the ID, source IP address and/or the option field. The option field is used less than the others because it usually gets removed by routers and firewalls [6]. In the transport layer the TCP Initial Sequence Number and the Acknowledgement (ACK) field are used a lot and in the application layer the Domain name system (DNS) ID field is an example of a field exploited for establishing a covert channel [7].
- Covert timing channel, where a process relays information to another by modulating its use of system's resources based on different timing. They actually convey information through arrival timing of packets, and that method is more secure than the storage one, but because of their unpredictable nature (jitter, packet loss and packet reordering events predominant in the Internet) it is possible that they will not be decoded accurately. Some network timing channels require synchronization between the encoder and decoder. Also, these channels are disturbing the traffic patterns and inter-packet intervals of the IP packets. Higher layered protocols may be bothered by this, so by

analyzing those inconsistencies, the timing channel could be detected at a higher layer [8].

III. CONDITIONS FOR CREATING A COVERT CHANNEL AND COUNTERMEASURES AGAINST ITS EXISTENCE

This chapter is divided into two sections, in which firstly are mentioned the basic conditions for creating an undetectable covert channel. Then, some countermeasures that can be taken against covert channels will be shown and explained.

A. Basic conditions for creating an undetectable covert channel

When referring to network information hiding, the warden is the network entity that wants to reveal the existence of the covert channel and later eliminate it [1].

If person A wants to transmit information to person B via a covert channel, the following two conditions needed for achieving covert communication must be met:

- The detecting ability of the warden is practically equal to random guessing
- The probability that B will make a mistake when recovering A's message is close to 0 [9]

Of course, in networking, covert channels are not restricted to unicast (one-to-one) channels. This means A could also send messages to B, C, D and E simultaneously if the channel allows multicast (one-to-many) communication [4].

B. Countermeasures that can be taken against covert channels

The stage of designing a system is crucial, because it can be exploited by the two causes of covert channels:

- Design oversights, that may be repaired if discovered early
- Weaknesses inherent in the system design, that can not be repaired without a system redesign [4]

The possible countermeasures that the warden can take when a covert channel is detected are the following:

- Remove the covert channel [4]

Blocking a connection is the best protection, but it can be implemented only partially by blocking unneeded traffic [7].

- Limiting the bandwidth, which works if the capacity of the channel can be closely estimated [4]
- Auditing the channel by a warden [4]

The shared resource matrix methodology is often used, where all shared resources that can be modified by someone are enumerated and then each one of them is examined whether it can be used to transfer information

covertly [10]. Auditing is more useful if the audit event is rare.

- Monitoring and documenting the covert channel, even if it has insignificant capacity [4]

Proxies can be used as a second layer of protection between an enterprise and the Internet. Incoming and outgoing connections finish at the proxy and the proxy server hides the user's IP address. All the hidden information stored in the lower layers is going to be lost. Most of the proxies also work in the application layer and check for abnormalities. They can filter the content, looking for known patterns [7].

The job of taking countermeasures should be entrusted to the best warden. There is a distinction between a passive warden, who can only spy on the channel, and an active warden who is able to modify the messages, but their context must remain the same [4]. This is a generalization because there are many subtypes of wardens. The best warden is a dynamic adaptive warden that with as fewer rules and effort as possible is adaptively limiting the capabilities of those who are using the covert channel and makes it difficult for them to deduce warden's strategy [1].

IV. SOLUTION FOR DATA LOSS PREVENTION

All modern enterprises and organizations in every industry have data and depend on data sharing. For preventing a data breach to occur, a data leak (or loss) prevention system is needed. That is a tool that helps safeguard the sensitive data. In case a violation occurs, the data loss prevention (DLP) solution must provide historic information for forensic analysis. There are different types of DLPs needed for achieving a secure work space in a multi-organization environment [11].

End-point based DLP includes data-in-use, because this data is "used" by the enterprise's employees on end-point devices. The DLP applies end-point security methods, so that data can travel safely in the case when there is no Internet connection. Sometimes employees use USBs for storing and passing company data. A USB memory device is developed that can erase the data if the USB is lost and the sensitive data can be passed only to the allowed servers and PCs [11].

Network based DLP takes care of data leaving the enterprise via a network. It includes the data-in-motion. A great concern is monitoring and controlling outbound Internet communication, where data can be lost through many channels, such as web mail and FTP transfer. [12].

Data-at-rest, static data stored on enterprise devices also should be protected by the DLP system. Most enterprises should apply control access rights to sensitive data not only located outside, but also inside the company. The security technology that achieves that is called Enterprise Rights Management (ERM) [11]. It manages usage

restrictions (printing, editing, viewing etc.). This can be accomplished with a server. [11, 13].

After discovering the three types of company data, the DLP solution should identify the sensitive data and enforce policies to that data. The existing DLP systems rely on hashing, keywords, regular expressions, fingerprinting etc., so they actually rely on humans. They can not completely recognize and classify the sensitive information by themselves. There is a need for this classification to be automated by efficient machine learning algorithms [13].

Even if the DLP systems manage to successfully face the challenges of performance and accuracy [13], as cloud computing becomes used more often, a problem arrives, where a company and the customers have to have trust in the cloud provider, because they share and confide the management of their information to a third party [11]. A combination of Public Key Infrastructure (PKI), Lightweight Directory Access Protocol (LDAP) and Single sign-on (SSO) can address most of the concerns about integrity, authenticity, confidentiality and availability of data and communications [14]. The cloud providers are having themselves audited by certification systems, because they realize that it is essential that they gain the companies' and the customers' trust [11, 14].

V. USAGE OF COVERT CHANNEL

This chapter is divided into two sections, in which examples of different uses of covert channel, both benign and corrupt will be mentioned and explained.

A. Usage of covert channel for copyright protection, authentication and authorization

The data hiding method can be used for copyright protection, authentication and authorization. This could be achieved by diverse applications such as copyright protection for digital media, watermarking, fingerprinting and steganography. The communication patterns could be detected if we use only encryption. So there is a need for technologies that can protect content even after it is decrypted [15, 16]. Such technologies are the following:

- In watermarking applications, the message contains information (such as owner ID and a digital time stamp) placed within the content where it is never removed during normal usage [15, 16].
- The fingerprint sensor reads the ridge pattern on the finger surface and converts the analogue reading in a serial number that uniquely identifies the owner [17]
- Steganography is covert channel's oldest form (in a way, covert channels equals network steganography) [7]. The main difference between steganographic and cryptographic methods is that the first ones hide information by making it difficult to notice, and the cryptographic ones do that by making it difficult for recognition [2].

Steganography hides a secret message within the host data set which does not noticeably change the data. A simple example for this is changing the least significant bits to embed information in an image file. Its presence can not be detected (the picture will look the same) when nobody suspects transmission of a hidden data [6, 15].

- Port-knocking allows authorised users to access open firewall ports and to all other users these ports appear to be closed [4]

B. Usage of covert channel for corrupt causes

The covert channels can be used for illicit manner, like changing and leakage of confidential data. Almost every protocol can be used as a covert channel by carrying another protocol with a technique called tunneling. Hackers use the hidden network to escape from firewalls and IDSs (intrusion detection systems). There are many examples of uses of covert channels for corrupt causes.

Computer viruses and worms can use covert channels to self-replicate and spread copies of themselves undetected and/or exchange information for distributed processing, like brute-force attacks [4]. Most Trojans and botnets use covert channel communication, mostly over HTTP [7]. Trojans may send an instruction to a server on a compromised system through a covert channel. If the IDS consider it to be ordinary, the Trojans can communicate to the hacker (the client component) and not be detected. "Bots" is derived from "robot", because it is a malware that can act similar to a human being, and can launch DoS attacks and use covert channels to gather passwords, confidential information, log keystrokes, analyze traffic etc [18]. It is interesting that for DoS attacks, a packet traceback technique that uses covert channels for filtering the attack traffic and isolating the attacker is developed [4].

It was mentioned that covert channels can be used for authentication, but they can also break anonymity. There is a covert-channel based technique that can identify locations of sensors with probe response attacks [4].

VI. AN EXAMPLE OF A COVERT CHANNEL ATTACK THAT EXPLOITS DNS TUNNELING

The Domain Name System (DNS) is a network service that translates host names into numerical IP addresses. For almost any network, a communication with a DNS server is needed.

Unfortunately, one of the most often used backdoor by hackers is the DNS. This is due to its wide availability and the fact that usually it is not monitored by common security measures, like firewalls, proxies and IDS because it is not intended for transfer of data [19]. IDC reports that 82% of companies worldwide have faced a DNS attack over the past year, which is based on a survey IDC conducted on behalf of DNS security vendor EfficientIP of 904 organizations across the world during the first half of 2019. The average costs associated with a DNS attack rose

by 49% compared to 2018 and most companies resolved the DNS attack after more than a few days [20]. In a research conducted by Palo Alto Networks Unit 42 in 2019 was found that up to 80% of malware uses DNS to establish command and control [21].

Hackers may achieve this attack successfully by establishing a covert communication channel. The channel is set between the device inside the network running a tunneling technique like Iodine, NSTX, DNSCat2 etc. and a server on the Internet and then they communicate back and forward through the DNS tunnel to control the compromised device.

There is a second more malicious approach done by dropping a malware. The malware encrypts a file usually containing sensitive information, dissects it and sends it in a form of multiple DNS queries to a server controlled by the attacker. This approach is shown at Figure 1 [22].

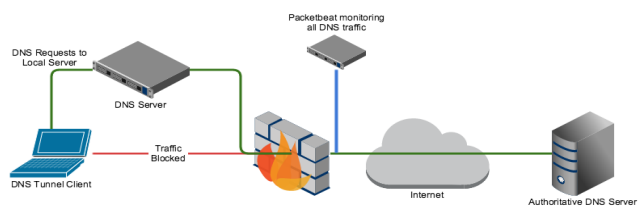


Figure 1. DNS approach done by dropping a malware

A common example of deploying robust secret communication is done by a simple steganography scheme through spam e-mails. This scheme can offer unidirectional asynchronous one-to-one or one-to-many covert channel facilities that are able to bypass firewalls and traffic analyzers [23]. DNS cache poisoning is often found in URLs sent via spam emails. By exploiting system vulnerabilities, attackers can inject malicious data into your DNS resolvers' cache. This is an attack technique often used to redirect victims to another remote server [22].

As a response toward the great DNS tunneling concern, researchers are tending to use Machine Learning Techniques (MLTs) to detect tunneling. These detection techniques can be grouped into two categories: payload analysis and traffic analysis [19]. Infoblox, for example, is a patented technology that uses machine learning and can even show exactly which devices or employees are trying to steal data [24].

It is also useful to notice that there are many free DNS projects that offer different Internet experience. The one that Google offers is the most known (with ip address 8.8.8.8). The newest DNS projects are made by: CleanBrowsing (185.228.168.9), Quad9 (9.9.9.9) and Cloudflare (1.1.1.1) [25].

We added the four addresses mentioned above to DNS Benchmark freeware, we ran the benchmark on the laptop and we got the results that can be seen at Figure 2 and Figure 3.

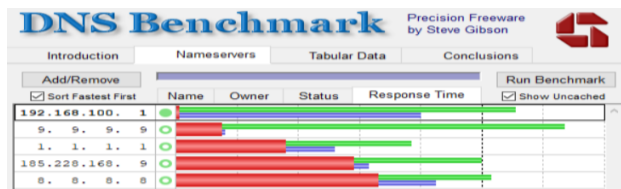


Figure 2. DNS Benchmark results 1

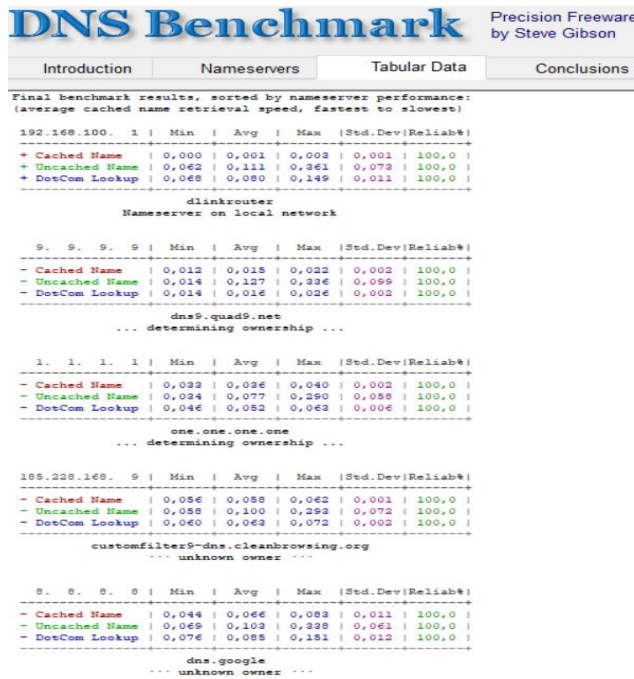


Figure 3. DNS Benchmark results 2

The first address is from the router and it has the best results related to latency. The Google's server had the worst results compared to the "new DNS players". CleanBrowsing is mostly used for content filtering and offers browsing the Internet without unwanted surprises [26]. Quad9 uses response policy zones to prevent tunnelling and phishing sites. When a user tries to go to such site, he encounters a walled garden and is warned of risky behaviour [21]. Information about personal id is not collected by the system [27]. Cloudflare is audited by KPMG to ensure that the customers' IP address and what they do is not kept by them. They delete the logs within 24 hours of their appearance. DNS has been largely unencrypted since its creation. Lately, there has been a push for using techniques for encrypting DNS over TLS and over HTTPS. Now, on the server side, Google DNS and Cloudflare's DNS support them. These techniques can increase latency, but it can be amortized over many queries. Chrome will probably adopt Google DNS and Firefox – Cloudflare, which could mean centralization of encrypted private information to them instead of unencrypted private information to many ISPs.

VII. CONCLUSION

Network covert channels can be used by attackers to help them steal information from compromised hosts. The DNS tunneling attack that was mentioned as an example is

simple to execute, and still can cause serious problems. It is a great challenge to maximize data leak protection system's performance and accuracy. On the other hand, the covert channels can also be used for delivering privacy information, like passwords, social security numbers, trade secrets etc. more securely [5, 11].

For some applications, the warden is an antagonist (for e.g. a censor in an oppressive regime). For others, it is trying to prevent the actions of an antagonist (for e.g. a malware) [1].

It is crucial for more to be done about better regulation.

REFERENCES

- [1] W. Mazurczyk, S. Wendzel, M. Chourib, J. Keller, "Countering adaptive network covert communication with dynamic wardens", *Future Generation Computer Systems* Volume 94, May 2019, pp. 712-725
- [2] W. Mazurczyk, S. Wendzel, S. Zander, A. Houmansadr, K. Szczypiorski, "Information hiding in communication networks: fundamentals, mechanisms, applications, and countermeasures", Wiley Online Library, 2016
- [3] T. G. Handel, M. T. Sandford II, "Hiding Data in the OSI Network Model", *International Workshop on Information Hiding IH 1996: Information Hiding*, pp. 23-38
- [4] S. Zander, G. Armitage, P. Branch, "A survey of covert channels and countermeasures in computer network protocols", *IEEE Communications Surveys & Tutorials*, July 2007
- [5] Y. Qian¹, T. Sun, J. Li, C. Fan, H. Song, "Design and analysis of the covert channel implemented by behaviors of network users", 2012 Sep, 31(9): 1611-24. DOI: 10.1007/s00299-012-1275-3. Epub 2012 May 20.
- [6] S. Z. Gober, B. Javed, N. A. Saqib, "Covert Channel Detection: A Survey Based Analysis", 9th International Conference on High Capacity Optical Networks and Enabling Technologies HONET 2012, DOI: 10.1109/HONET.2012.6421435
- [7] J. Selvi, "Covert Channels Over Social Networks", 2019
- [8] X. Luo, E. W. W. Chan, R. K. C. Chang, "TCP Covert Timing Channels: Design and Detection", *Conference Paper*, July 2008, DOI: 10.1109/DSN.2008.4630112 .
- [9] J. Wang, W. Tang, Q. Zhu, X. Li, H. Rao, S. Li, "Covert Communication with the Help of Relay and Channel Uncertainty", 2018
- [10] R. A. Kemmerer, "Shared Resource Matrix Methodology: An Approach to Identifying Storage and Timing Channels", *ACM Transactions on Computer Systems*, Vol. 1, No. 3, August 1983, pp. 256-277
- [11] T. Takebayashi, H. Tsuda, T. Hasebe, R. Masuoka, "Data Loss Prevention Technologies", *FUJITSU Sci. Tech. J.*, Vol. 46, No. 1, pp. 47-55, 2010
- [12] <https://www.cisco.com/c/en/us/products/collateral/security/web-security-appliance/solution-overview-c22-738537.html>
- [13] M. Hart, P. Manadhata, R. Johnson, "Text Classification for Data Loss Prevention", 2011
- [14] D. Zissis, D. Lekkas, "Addressing cloud computing security issues", *Future Generation Computer Systems* Volume 28, Issue 3, March 2012, Pages 583-592
- [15] M. M Amin, M. Salleh, S. Ibrahim, M.R.K Atmin, M.Z.I. Shamsuddin, "Information Hiding using Steganography", February 2003, DOI: 10.1109/NCTT.2003.1188294
- [16] I. J. Cox, M. L. Miller, T. Kalker, "Digital Watermarking and Steganography" Elsevier, 2007
- [17] D. Maltoni, "Handbook of Fingerprint Recognition", Springer, 2009
- [18] https://tools.cisco.com/security/center/resources/virus_differences
- [19] S. Yassine, J. Khalife, M. Chamoun, H. el Ghor, "A Survey of DNS Tunnelling Detection Techniques Using Machine Learning", Published in BDCSIntell 2018
- [20] <https://www.networkworld.com/article/3409719/worst-dns-attacks-and-how-to-mitigate-them.html>
- [21] T. Olzak, "DNS Tunnelling Identification and Defence", 2019
- [22] <https://securitytrails.com/blog/most-popular-types-dns-attacks>
- [23] A. Castiglione, A. De Santis, U. Fiore, F. Palmieri, "An asynchronous covert channel using spam", *Computers & Mathematics with Applications*, Volume 63, Issue 2, January 2012, pp. 437-447
- [24] <https://www.infoblox.com/solutions/service-providers/secure-dns-caching/>
- [25] <https://medium.com/@nykolas.z/dns-resolvers-performance-compared-cloudflare-x-google-x-quad9-x-opendns-149e803734e5>
- [26] <https://cleanbrowsing.org/>
- [27] <https://quad9.net/about/>

A Note on a Successful WEP Attack

Stefan Pavlov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University, Skopje
stefan.pavlov47@gmail.com

Vesna Dimitrova
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University, Skopje
vesna.dimitrova@finki.ukim.mk

Abstract

The wireless networks security threats are rising with the development of the wireless technology. The implementation of the security mechanisms in the networks is provided in a way that the mechanisms are incorporated through the wireless network security standards. With the time passing by, there are plenty of efforts in the creation of a multi-layer security system for every wireless network, which would represent the factor of surprise for the attacker, along with the risk of attacks and penetration.

However, unlike other types of data transmission, the wireless data transmission is transmitted in all directions and can be received from any device that is compatible with the wireless adapter. Due to this, it is very important that security mechanisms are implemented so that unauthorized access would be prevented from accessing your network or data.

In order to achieve the best of the security of our data, we need to work for a bigger and deeper implementation of the security measures in every domain of the wireless telecommunication systems. Whether it is about a computer, a mobile or any other sort of device connected to a wireless network, their vulnerability is almost the same, because each and one of them is exposed to a variety of threats and attacks.

That is why the goal of this paper is the research about the attack done on a wireless network.

Keywords—wireless networks, security, privacy, data, threats.

I. INTRODUCTION

The 21th Century is defined by human scientific breakthroughs. One of them is the global use of the Internet. Almost every educated person nowadays knows more or less about the internet. Thereby, a large percentage of the human population make use of the internet on a daily basis. Wireless internet access is thereby made affordable and accessible for everyone and that has boosted human communication and economy development into a new era. However, along with benefits, wireless internet access also possesses several risks and threats of security. The Internet has become a fertile land for criminals of all kinds to operate. Most of all time what they try to take is personal information, some of them extreme values and sensitivity, from naïve and unaware users. In that sense, a comprehensive study on how cybercriminal carry out their attacks and how to avoid and actually prevent such attacks if possible would be beneficial for the righteous netizens. [1]

Devices commonly used for wireless networking include portable computers, desktop computers, hand-held computers, personal digital assistants (PDAs), cellular phones, pen-based, computers, and pagers. Wireless networks work similar to wired networks however, wireless networks must convert information signals into a form suitable for transmission through the air medium.

Wireless networks serve many purposes. In some cases, they are used as cable replacements, while in other cases they are used to provide access to corporate data from remote locations.

Wireless infrastructure can be built for very little cost compared to traditional wired alternatives.

Wireless networks allow remote devices to connect without difficulty, independently these devices are a few feet or several kilometers away. This has made the use of this technology very popular, spreading rapidly. [2]

As we move on, in this section, we will face with two new important issues regarding wireless networks security. The first one of them, is Kali Linux, and the second one, is the process of virtualization.

Kali Linux is an enterprise-ready security auditing Linux distribution based on Debian GNU/Linux. Kali is aimed at security professionals and IT administrators, enabling them to conduct advanced penetration testing, forensic analysis, and security auditing.

The first Kali release happened in March 2013, and was based on Debian 7 “Wheezy”, Debian’s stable distribution at the time. In the first year of deployment, Kali team packed hundreds of pen-testing-related applications and built the infrastructure.

During the first two years following version 1.0, Kali released many incremental updates, expanding the range of available applications and improving hardware support, thanks to newer kernel releases. [3]

As mentioned above, the last part of the introduction is about the process of virtualization. Basically, virtualization is a fundamental part of cloud computing, especially in delivering Infrastructure as a Service (IaaS). Exploring different techniques and architectures of the virtualization helps us understand the basing knowledge of virtualization

and the server consolidation in the cloud with different architecture. In this paper we use virtualization as the tool that helps us access Kali without having it installed on a virtual machine. The virtualization tool that we are using here is VMWare. [4]

Then, in Section II, we describe the very basics of wireless networks, and we discuss the main concern of wireless networks – wireless networks security threats.

Afterwards, in Section III, the main focus is set on wireless networks attacks and their classification.

As we move on the next Section IV, the paper gives a brief overview of the latest wireless networks' security standards. Which can be noted as a big deal, due to the fact that, in order to attack, you need to know how to defend at first, and then, you need to know all about standards and policies.

Last but not least, in the last Section V, lies the heart of this paper. In the last section, there is a detailed and deep explanation given on how to conduct an attack on a network using WEP.

II. WIRELESS NETWORKS SECURITY THREATS

Wireless networks are very common, both for organizations and individuals. Many laptop computers have wireless cards pre-installed for the buyer. The ability to enter a network while mobile has great benefits. However, wireless networking has many security issues. Crackers have found wireless networking relatively easy to break into, and even use wireless technology to crack into non-wireless networks. Network administrators must be aware of these risks and stay up to date on any new risks that arise. Also, users of wireless equipment must be aware of these risks, so as to take personal protective measures. Due to this, wireless networks security threats can be classified as home and public.

A. Home wireless threats

The need to secure traditional wired internet connections was felt long before. However, there is a growing trend of shifting to wireless connection at home. This involves a process where the user connects a device to his DSL or cable modem that broadcasts the Internet connection through the air over a radio signal to his computer. If traditional wired connections are susceptible to security tribulations, there is a great risk of security breach that may arise when a user opens his Internet connection to the airwaves. An unsecured wireless network coupled with unsecured file sharing can be disastrous. There are, however, steps one can make to protect the wireless network. The following are some of the possible security steps:

- Make the wireless network invisible by disabling identifier broadcasting.
- Rename the wireless network and change the default name.
- Encrypt the network traffic.
- Change administrator's password from the default password. If the wireless network does not have a default password, create one and use it to protect the network.

- Use file sharing with caution. If the user does not need to share directories and files over his network, he could disable file sharing on his computers.
- Keep the access point software patched and up to date.
- Check internet provider's wireless security options as it may provide information about securing your home wireless network.
- Do not auto-connect to open Wi-Fi networks.
- Turn off the network during extended periods of non-use, etc.

B. Public wireless threats

The risk to users of wireless technology have increased exponentially as the service has become more popular. There were relatively few dangers when wireless technology was first introduced. Currently, however, there are a great number of security risks associated with wireless technology. Some issues are obvious, and some are not. At a corporate level, it is the responsibility of Information Technology (IT) department to keep up to date with the types of threats and appropriate counter measures to deploy. Security threats are growing in the wireless arena. Crackers have learned that there is much vulnerability in the current wireless protocols, encryption methods, and in the carelessness and ignorance that exists at the user and corporate IT level. Cracking methods have become much more sophisticated and innovative with wireless. Cracking has become much easier and more accessible with easy-to-use Windows-based and Linux-based tools being made available on the web at no charge. IT personnel should be somewhat familiar with what these tools can do and how to counteract the cracking that stems from them. Accessing the internet via a public wireless access point involves serious security threats. These threats are compounded by the inability to control the security setup of the wireless network. The following steps can be taken to protect oneself at public places:

- Be careful while dealing in an online environment if the network is not properly secured. Avoid online banking, shopping, entering credit card details, etc.
- Connect using a virtual private network (VPN) as it allows connecting securely. VPNs encrypt connections at the sending and receiving ends and keep out traffic that is not properly encrypted.
- Disable file sharing in public wireless spaces as it is more dangerous than it is on your home wireless network.
- Be aware of your surroundings while using a public wireless access point. If an internet connection is not essential, disable wireless networking altogether. [5]

III. WIRELESS NETWORKS ATTACKS CLASSIFICATION

This section contains seven sub chapters that address different attacks against popular wireless protocols and

systems. Common themes will emerge throughout it, but each wireless technology has its own unique quirks that make it useful to attackers in different ways, making understanding all of them important to overall security as rarely is just one wireless technology in use at home or office.

A) 802.11 Wireless – Infrastructure Attacks

The ubiquitous 802.11 wireless network is hard to be avoided without running into it. It has become an invaluable resource for both home and office for networking and Internet access. Wireless networking is also incredibly valuable to attackers as it gives the attacker the opportunity to access networks at a safe distance, almost as if they were connected to the wired network. These attacks focus on the infrastructure of these networks and the security implications of their use and how and how not to secure them. They may be ubiquitous, but that doesn't mean they are secure.

A) Wireless – Client Attacks

Wireless clients, those devices that talk to the rest of the wireless network, are mentioned here. Attackers, stymied by increasing amounts of security on the infrastructure side, are changing tactics and attacking client devices directly. At home or away, wireless clients and the information they contain and communicate are tempting targets for pranksters and thieves alike.

B) Bluetooth Attacks

Bluetooth is the subject in these types of attacks. This common protocol was meant to replace cable clutter but has become so much more. While it is meant for short range, any distance can be a comfort for an attacker. Modern devices carry a great deal of information, tempting for a new era of digital pick pockets. You could lose everything without losing anything.

C) RFID Attacks

RFID is a technology most people are not even aware of despite the billions of tags in use every day. As the subject of this type of attacks, RFID is looked at with an eye of how its perceived benefits can actually be their greatest vulnerability and how they can be thwarted by those with ill intentions. RFID is all around us and knowing how to identify it and how to protect it is a very important topic not often understood by many people.

D) Analog Wireless Devices

Even the most modern of wireless devices often at their heart are just radios. Often these new devices are using age-old radio techniques to allow their communication. This type of attack identifies these devices and understands the risks associated with their use and how vulnerabilities apparent over 100 years ago are still around to make life interesting.

E) Bad Encryption

A common solution to wireless security problems is to add encryption. The common problem though with wireless security is bad encryption. Poor design choices, hardware limitations, and cost can all turn a good security idea into a failure at record speed. This problem is being looked over with a number of real-world examples and shows how something that was supposed to protect communications can end up providing less security than advertised.

F) Cell Phones, PDAs, and other hybrid devices

It's impossible to escape them, but cell phones are everywhere. Today's modern smart phones and other hand-held gadgets are at their hard, computers in their own right and have their own unique security issues that need to be considered. In these types of attacks, the main focus is on new generation devices and how their small size, portability and communication capacity make them interesting and tempting targets for today and the future. [6]

IV. WIRELESS NETWORKS SECURITY STANDARDS

It is important to understand why standards are needed, and the role they play in the adoption and use of technology. The most important reason is for interoperability, enabling multiple vendors to supply equipment that can be integrated into complete systems, ranging from phones plugging into RJ-11 connectors, VCRs connecting to televisions, and in the case of wireless, mobile telephones and wireless modems communicating with wireless networks.

One of the benefits to customers is that an accepted standard reduces technology risk because there are multiple vendors available to choose from. Consumers can change vendors if another vendor offers better prices or features or if a vendor stops supplying equipment. One of the most used standards is known as IEEE 802.11b, which has ignited the WLAN industry.

To be successful, there are three questions a standard must address:

- 1) Is it useful,
- 2) Is it complete,
- 3) Is it practical,

In Figure 1 is presented how we can evaluate a standard. [7]

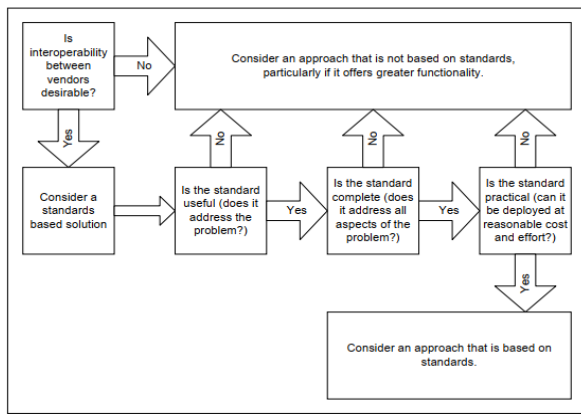


Fig. 1. Decision process for standards-based solutions

When standards do not properly address needs, the marketplace resists adoption. Because Mobile IP does not fully address all of the issues surrounding mobility, key vendors like Microsoft are not supporting the standard. This further erodes motivation for other vendors to support it, and the standard simply has not been able to achieve critical mass.

TABLE I. MERITS OF DIFFERENT STANDARDS

	Strengths	Weaknesses
TCP/IP Protocol Suite	Universally accepted. Excellent interoperability.	Not Optimized for wireless connections. No mobility support.
Mobile IP	Provides a mechanism for forwarding packets to a host operating in another network.	Does not address other mobility aspects such as loss of connections or link optimization. Not widely accepted.
IPv6	Increased address space. Mobility Support. Security Mechanisms.	No deployment today. Huge logistics issues for widespread deployment.
IPSec	Security Mechanisms.	No built-in mobility support.

In this Table 1, we can see the merits of the different standards being used in present day. [7]

V. WIRED EQUIVALENT PRIVACY

In the beginning it's believed that WEP offers impenetrable resistance to eavesdroppers/hackers. However, as wireless networks began to grow in popularity, many crypt analysts and researchers discovered flaws in the original WEP design. Many of the WEP flaws would have been caught in the early design phase if it's design and implementation specifications had been reviewed thoroughly. For most of the wireless networking users, WEP is the only choice available until new security mechanisms are added to the IEE 802.11 standard. But as people say "something is better than

nothing", even with its known weaknesses, WEP is still more effective than no security at all.

The design objectives of WEP as per section on 8.2.2. of the IEE 802.11 standard states the following:

- "It's reasonably strong: The security afforded by the algorithm relies on the difficulty of discovering the secret key through a brute force attack. This in turn is related to the length of the secret key and the frequency of changing keys. WEP allows for the changing of the key (K) and frequent changing of Initialization Vector (IV)."
- "It is self-synchronizing: WEP is self-synchronizing for each message. This property is critical for data-link-level encryption algorithm, where "best effort" delivery is assumed, and packet loss rates may be high."
- "It may be exportable: Every effort has been made to design the WEP system operation as to maximize the changes of approval, by the U.S. Department of Commerce, of export from the U.S. products containing a WEP implementation. However, due to legal and political climate toward cryptography at the time of publication, no guarantee can be made that any specific IEE 802.11 implementations that use WEP will be exportable from the USA."
- "It is optional: The implementation and use of WEP is an IEE 802.11 option."

From the above objectives, it's clear that WEP was not designed to provide a high military level security. The intention was to make it hard to break-in as opposed to impossible to break-in. [8]

VI. ATTEMPTING AN ATTACK ON WEP

In this section of the paper, we will discuss about how one of the flaws of Wired Equivalent Privacy protocol is being exploited by attackers. As we mentioned above, there are many different attacks that take place on a daily basis everywhere in the areas of IT. By this, we mean either physical or digital attacks occur reaching high numbers. In the example that follows, we will exploit the weaknesses of a wireless network, and the security protocol it uses, WEP.

One of the easiest and the most useful attacks for wireless networks are those that require the least time and effort. One of those types of attacks is a wireless network password cracking attack.

To conduct this type of attack we will be needing:

- Computer (either PC or laptop),
- Additional network card,
- VMWare application installed,
- Kali Linux OS installed on the VMWare virtual machine,
- Kali Linux OS wireless network tool exploiter – Aircrack,
- Network that provides us with Internet access,

- Network that we will be attacking,
- Txt. file that provides us with possible passwords.

The attack occurs as follows:

- 1) Once we have VMWare installed on our computer, we create a virtual machine, with the dedicated resources needed to run the OS we need for Aircrack – Kali Linux,
- 2) Then we power up the virtual machine with Kali Linux running on it,
- 3) When the machine is powered up, we then configure the set-up of the two network cards we provided our computer with. The first one, which is the integrated, provides us with Internet access. While the second one, which is the additional, provides us with the so-called spoofing that shows us which networks are active in the antenna coverage that we have,
- 4) Afterwards, we move on to launching Aircrack,
- 5) After Aircrack is launched, a terminal pops-up and we execute the command in the terminal window that provides us with the networks that are broadcasted in our antenna coverage,
- 6) Since we have all the networks running in our antenna coverage, we choose which network we want to penetrate, and then, we choose which user we want to de-authenticate, by doing this, we are executing two commands, one for network scan, other for user de-authentication selection,
- 7) When the de-authentication part is done, the user's device automatically tries to reconnect itself to the wireless network access point, and by sending the handshake, which is intercepted by us, it provides us with the information needed to crack the network's password,
- 8) Last, we need to execute the command that starts the brute force attack, by checking each and every one of the words put in the .txt file,
- 9) When the password is cracked, we get information details about the network cracking, and the most important part of the cracking – the network password.

CONCLUSION

The purpose of this paper is to overview the basics of wireless networking technologies, which includes:

- Wireless Network Security,
- Wireless Network Standards, and
- Wireless Network Attacks.

The analysis given in this paper gives a framework fit to describe everything you need to know when conducting an attack on a wireless network or defending one from an attacker. It is fundamental to know how to use wireless networks properly, or securely.

Besides the Wireless Networking Technologies, two more essential points are being monitored here. One is the Kali Linux OS and the other one is VMWare. These two are the key ingredients if you are broke and you cannot provide yourself with a physical machine capable of wireless network penetration, due to the fact that both, Kali Linux and VMWare are free.

To conclude, this paper could prove as a good reference to someone who is new to digital forensics. The topics that this paper provides are the fundamental key elements in getting to know wireless networks and security issues regarding them.

REFERENCES

- [1] Hoa Gia Bao Nguyen, "Wireless Network Security," Lahti University of Applied Sciences, Lahti, Finland, p. 2, 2018.
- [2] Jordi Salazar, "Wireless networks," Czech Technical University of Prague, Prague, Czechia, p. 6, 2017.
- [3] Raphael Hertzog, Jim O'Gorman and Mati Aharoni, "Kali Linux Revealed: Mastering the Penetration Testing Distribution," Offsec Press, Cornelious, NC, USA, pp.2-3, 2017.
- [4] Hyungro Lee, "Virtualization Basics: Understanding Techniques and Fundamentals," School of Informatics and Computing, Indiana University, Bloomington, IN, USA, p. 1, 2014.
- [5] Inyama H. Chibueze, Achi I. Ifenayi and Agwu O. Chukuemeka, "Threats and security measures on wireless local area networks," Department of Electronic and Computer Engineering, Nnamdi Azikwe University-Awka, Nigeria, p.2, 2014.
- [6] Brad Haines, "Seven Deadliest Wireless Technologies Attacks," Syngress, Elsevier, Burlington, MA, USA, pp. 14-16, 2010.
- [7] Peter Rysavy, "Networking standards and Wireless Networks," NetMotion Wireless, Seattle, WA, USA, pp.2-6, 2002.
- [8] Shivaputrapa Vibhuti, "IEE 802.11 WEP(Wired Equivalent Privacy) Concepts and Vulnerability," San Jose State University, CA, USA, pp.2-4, 2005.

Fog Necessity Over Cloud Computing for Healthcare Applications

Beyza Ali

University of Information Science and
Technology “St. Paul the Apostle”
Ohrid, North Macedonia
beyzaali34@gmail.com

Natasa Paunkoska Dimoska

University of Information Science and
Technology “St. Paul the Apostle”
Ohrid, North Macedonia
natasa.paunkoska@gmail.com

Ninoslav Marina

University of Information Science and
Technology “St. Paul the Apostle”
Ohrid, North Macedonia
ninoslav.marina@gmail.com

Abstract— The increased amount of data generated through the communication devices, hampered the manner of data storage locally in the computers. Innovative solution for dealing this problem are cloud computing systems. On the other side, smart devices, as a part of the rapid technology evolution, become an important part of our daily lives, to the extent that the quality of our lives depend on them. This led to a necessity for development of new technological architecture that will support the cloud systems towards overcoming their drawbacks appearing at the edge network. In that direction, fog computing is a novelty that solves the cloud issues by placing the computational analysis and storage closer to the network edge. In this paper, the effect of fog computing paradigm in healthcare applications is investigated with respect to the latency and network usage. A detailed analysis regarding both performances are proved, first, when only cloud environment is present and then when fog computing is added. The results prove that managing data closer to the network edge, or by applying fog, decreases latency and network usage by a considerable amount, than compared to transferring data to the cloud which increases the network traffic congestion.

Keywords— Internet of Things (IoT), fog computing, cloud computing, sensors, healthcare.

I. INTRODUCTION

In today's world whether it will be a smartphone, a smart car, or a juice dispenser we are bound to use the technologies and this have become an important part of our lives. Their massive usage creates a significant amount of data every day [1]. In old times the purpose of internet was to connect two or more end-users and their generated data was stored in the computers' memory locally. By a time as the quantity of generated data starts increasing, the computer's memory became limited and insufficient to handle this issue. This led to a shift in the technology, which is known as cloud computing [2].

With its emergence, cloud computing became the number one solution for storing and processing data. Since it offers a centralized computing model with a high computation power, efficient network management functions and storage capability has grown to crucial solution for companies that were looking for scaling their computational operations [1], [3]. Data centers being located close to the core network makes data transfer difficult in sense of producing significant amount of time for its realization, known as latency. Latency is an important drawback in cloud computing systems, especially, for applications and situations that are time sensitive, like, real-time applications.

Although cloud computing has its benefits, the extensive increase of devices connected to internet, known as Internet of Things (IoT), poses a challenge for the usage of cloud computing [1], [3]. By Cisco Systems [4] it is predicted that an estimated 50 billion “things” will be connected to the internet by 2020. But the cloud models are not designed in a

way that can meet the needs of IoT, which motivates the emergence of a new technological paradigm called fog computing. In this paper, we will focus on the advantages that comes with the new technology, emphasizing the two parameters improved latency and network usage that are evident at the end users. Both parameters will be analyzed and compared applying healthcare applications, when only cloud computing environment is considered and then when fog computing is added. The results show that the trend of massive data production cannot be served enough quality, without the use of fog computing.

The paper is organized as following: Section II gives comparison between the cloud and fog architecture. Section III gives overview of existing literature. Section III explains for the importance of fog in healthcare application and section IV shows the results. Section V concludes the paper.

II. CLOUD VS. FOG COMPUTING

Traditional cloud computing architectures move all the data from the network edge to the cloud. The data is stored in data centers. After being analyzed in cloud the required actions are taken and the required data is transferred back to the end-users. But the bigger the data is, the longer the transfer time becomes. Access time is one of the most important points to be considered, as the purpose of IoT is to enhance peoples' life quality, [5]. Table I gives a comparison of cloud computing versus fog computing discussing the advantages and disadvantages of both technologies with respect to the most important aspects. We see that both cloud and fog have high complexity level. On one hand, fog is more advantageous in the means of response time, mobility, distance to end-devices, communication mode, bandwidth costs, energy consumption, location awareness, and geo-distribution. On the other hand, cloud is more advantageous when it comes to computational and storage capability.

TABLE I. CLOUD COMPUTING VS. FOG COMPUTING

Operates on	Cloud cloud	Fog network edge
Complexity level	high	high
Response time	high	low
Mobility	limited	supported
Distance to end-devices	far	close
Communication mode	IP network	wireless
Bandwidth costs	high	low
Computation capability	strong	weak
Storage capability	strong	weak
Energy consumption	high	low
Location awareness	no	yes
Geo-distribution	centralized	decentralized

Fog computing was introduced to tackle the limitations of cloud computing, and its main goal is to increase scalability and reduce bandwidth. Cloud system is still the number one solution for the storage of big data but the needs has changed.

Although storing data is still important, it is not the only important issue. IoT generates data constantly and in many situations rapid analysis is required [6], [7].

Fig. 1 shows the hierarchical architecture of fog computing consisting of three layers. Cloud layer consists of multiple high performance servers and provides high storage capacity and powerful computing capabilities. Fog layer consists of fog nodes and is located at the proximity of the end-users.

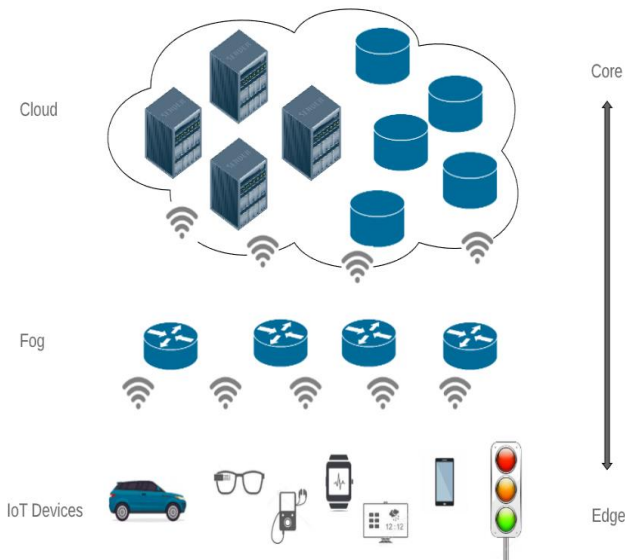


Fig. 1. Hierarchical architecture of fog computing.

Fog nodes are widely distributed and can be static or mobile. Terminal layer it is the layer closest to the physical environment. It consists of various IoT devices that are responsible for sensing and collecting the data. In this architecture the fog nodes are connected to the fog layer by wireless technology, and can also be interconnected. If they are interconnected, the intercommunication is made possible via wired or wireless network.

IoT's concept is incompetent when it comes to conceptual power, battery, storage and computing abilities. Therefore it needs the support of a more complex and strong concept, i.e. the fog computing layer [8], [9].

When a new technology arises it often takes time to come into a consensus about its definition. The most comprehensive yet simple definition is given by L.M. Vaquero et al. [10], as "a scenario where a huge number of heterogeneous (wireless and sometimes autonomous) ubiquitous and decentralized devices communicate and potentially cooperate among them and with the network to perform storage and processing tasks without the intervention of third-parties."

III. RELATED WORKS

L.M. Vaquero et al. [10], in their paper, have discussed fog computing on the domain of device ubiquity. They have examined two of the most important aspects of technology and technological devices, the size and battery lifespan, in the meaning of cost, portability and power consumption. Finally they have come to a conclusion that "the fog is nothing but the convergence of a set of technologies that have been developing and maturing in an independent manner for quite some time". Stojmenovic et al. [11] have addressed on a very

important issue, and that is the security. The main focus is on the security and privacy issues and they have discussed these issues by investigating a typical attack, known as the man-in-the-middle attack in detail. In a paper published by Cisco Systems [4], which is the worldwide leader in networking for internet, they have proposed a definition for the fog computing and discussed how the fog work. A comparison of cloud versus fog is done with respect to response time, application examples, how long IoT data is stored, and geographic coverage.

P. Hu et al. [3] have discussed the motivation behind the emergence of fog computing. While they have worked on the architecture of fog computing they have also analyzed the key technologies fog computing depends on. The authors have also discussed the applications of fog computing through real life scenarios.

M. Ahmadi et al. [8] have worked on the shift of technological paradigm and how it has shifted from cloud computing to fog computing. The authors have discussed both the advantages and the challenges of this new technological paradigm with a touch on reliability, security, management, availability, privacy, interoperability and applicability.

IV. THE IMPORTANCE OF FOG COMPUTING FOR HEALTHCARE APPLICATIONS

Healthcare is the most fundamental issue in human life. Unfortunately, as the time changes and the world evolves humankind is facing new diseases and chronic illnesses. This in turn creates the demand for resources. This high demand puts a lot pressure on the healthcare systems [6],[12].

The focus in healthcare has shifted from treating patients at hospital after an incident to delivering a high – quality healthcare to prevent serious incidents or illnesses. The first steps towards achieving this goal is monitoring the conditions of healthy people in order to keep them out of hospitals. But monitoring people individually in hospitals is impossible. Therefore it is safe to say that the solution is remote monitoring. Sensors will also provide accurate and precise measurements of peoples' health conditions, since they are capturing data continuously [13].

The system architecture of a cloud – based IoT healthcare application is mainly composed of three layers. The wearable sensors, i.e. devices that are connected to the patients' users' body. Once they are turned on they start to collect data about the patients' health. Smart phones provide an interface for the data to be sent to the cloud datacenter. Additionally, smart phones ask for user authentication. Cloud data center keeps all the data collected. The essential data is being extracted from the raw data and is simplified for further use. After the analysis the users' health conditions are being evaluated and if necessary the user or the doctor is notified.

The system architecture of fog – based healthcare applications, in addition to cloud – based applications, have and additional layer between the smart phones and the cloud datacenter, the fog layer, which has specialized networking devices called "fog nodes". These fog nodes are used for and capable of performing computational tasks, storing information and simplifying the massive amount of data since not every data collected need to be sent to the cloud. These data are eliminated in the fog layer, and access to the needed part of the massive data is being made easier. Fog nodes can be activated and/or deactivated as and when needed [14].

V. RESULTS

In order to give a better insight about the differences between cloud and fog computing we will discuss their system architectures in healthcare applications and analyze the results of several studies. We are going to compare the results of four different simulation experiments with a focus on network usage and latency. The simulation experiments that use only cloud is done using the CloudSim toolkit, which is a widely used library for the simulation of cloud – based environments. The simulations that use the fog layer are done using the iFogSim, which is an open source toolkit used to model and simulate the networks of edge computing, IoT and fog computing.

In [15] authors have proposed a “tri-tier architecture for context – and latency – sensitive health monitoring using cloud and fog computing”. The aforementioned tri-tiers are the sensors, fog computing, and cloud computing where the sensors work in conjunction with one another. The flow of information between these levels has to be managed efficiently, privately, and securely. The sensors tier, which consists of wearable devices, gather the data and send it to the fog tier. The fog computing tier aggregates the data that comes from the sensor tier and perform the first data analysis, then distributes the processing work to the related fog nodes for further analysis. The cloud computing tier manages the actions that need to be performed by the health monitoring system. Finally the health monitoring system consists of the region, the institution, the clinical department, and the individual doctor, nurse, or patient.

In this experiment several test runs and are simulated for five different configurations of monitoring devices. Five configurations are considered, config1, config2, config3, config4, and config5, they each have by 4, 8, 16, 32, and 64 monitoring devices, respectively. This means that each configuration will give different results during the simulation process. In Table 2, Fig. 2, and Fig. 3 we see the results of the simulation experiment for each configuration both with and without fog layer. We see that applications with the fog layer outperforms the applications using only the cloud layer.

TABLE II. COMPARISON OF PERCENTAGES

Physical Topology	Average Latency (ms)		Network Usage (KBs)	
	Cloud only	With fog layer	Cloud only	With fog layer
Config1	210.38	8.47	130	12
Config2	210.78	8.47	351	22
Config3	211.57	8.47	672	53
Config4	1283.86	8.47	1061	98
Config5	3225.91	8.47	1102	189

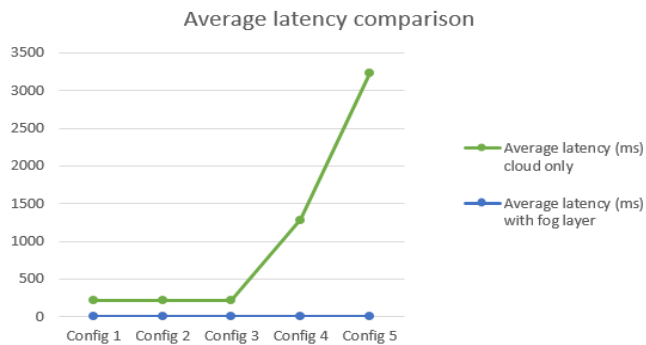


Fig. 2. Average latency comparison.

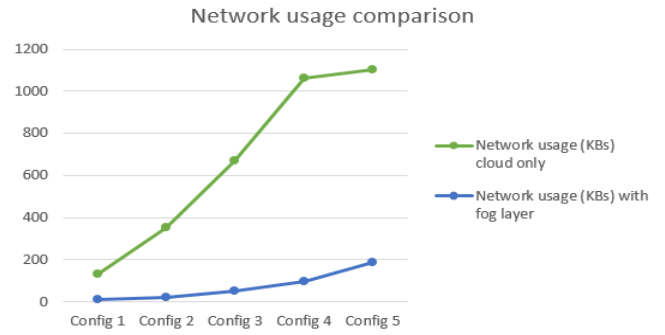


Fig. 3. Network usage comparison.

Reference [16] simulate different configurations for different fog nodes. The application that is being analyzed is an healthcare application that processes heart patients’ data. Considering the number of the requests lunched by the patients, network usage, and latency is being compared for architectures with and without fog computing. Table 3 and Fig. 4 displays the results of this experiment. The study has shown that fog assisted application architecture manages the data of heart patients much more efficiently.

TABLE III. PERFORMANCE COMPARISON

Environment	Average Network Usage Time (s)	Average Latency (s)
Cloud Computing	84.58	24.33
Fog Computing	23.36	8.3

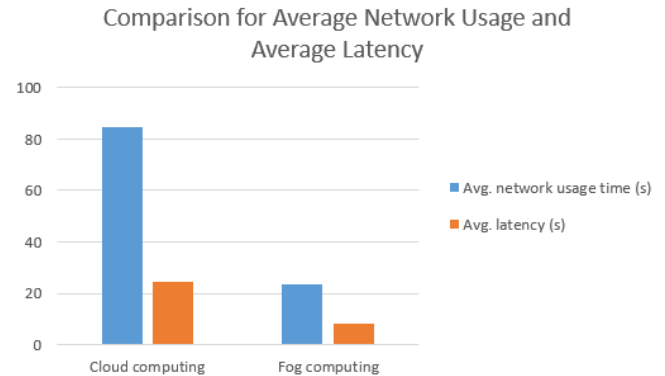


Fig. 4. Comparison for Average Network Usage and Average Latency.

P.H. Vilela et al. [17] test a fog assisted approach under two different experimental scenarios. The first scenario consists of only the cloud layer, whereas the second scenario includes the fog layer. In the proposed healthcare application the sensors include environmental sensors, i.e. sensors measuring room’s humidity, noise level, and temperature, and medical sensors, i.e. sensors measuring body temperature, blood pressure, and heart beat rate. After 30 experiments, the Central Limit Theorem was used to obtain an average. The study concludes that for one minute data transfer time there is 0.25 seconds of delay. On the other hand, as the number of the sensors increase the network usage decreases more than by half if a fog layer is used.

In [18] authors proposes an energy – efficient strategy which allocates the incoming tasks according to the remaining CPU capacity and energy consumption. In other words, this approach takes into account the distance of each fog node to

the sensor. Moreover the proposed strategy tries to use the fog nodes efficiently, and ensures it is neither underused nor overloaded. In this case study the energy efficiency of the proposed strategy is evaluated by remotely monitoring patients with diabetes. For simplicity the patients' were equipped only with the blood glucose sensor. Four different configurations with different workloads were considered. The configurations consist of 2 fog devices and 4 smartphones, 4 fog devices and 4 smartphones, 4 fog devices and 8 smartphones, 4 fog devices and 16 smartphones respectively. The results show that the use of fog layer significantly reduces the latency, particularly in the third and the fourth configurations, i.e. the configurations with the higher number of sensors. The same results hold for the network usage. Using fog nodes reduces the amount of data transmitted over the network, decreasing the network usage. Although this is true for all four configurations, the proposed strategy is particularly better in configurations three and four, i.e. the configurations with heavier workloads.

VI. CONCLUSION

With the most important aspects being discussed it is important to remember that healthcare industry is still in its infancy stage and is open for improvement. With this in mind, it has to be appreciated the long way it has come. For now fog computing is considered as the best method to rely on because it uses shared resources which affects its performance in a way that meets the requirements of healthcare IoT applications. Nevertheless the applications should be designed very carefully because they are time-sensitive but for more complex tasks more fog nodes can be needed. Though scalability is one of the most important advantages of fog computing, increasing the number of fog nodes increases the probability of failure.

It is safe to say that fog computing is the most suitable technology for time-sensitive applications and particularly for healthcare applications. All four studies discussed in this paper demonstrates clearly that fog computing outperforms cloud computing with respect to latency and network usage.

Keeping this in mind one should not forget that outsourcing data analytics fully to the edge of network may result in unwanted consequences arising from limited computational capacity of the edge nodes. Many applications require both fog localization and cloud globalization to operate in the best manner. Therefore, the distribution of functions between the cloud and the fog nodes is a crucial factor.

REFERENCES

- [1] S. Kunal, A. Saha, and R. Amin, "An overview of cloud-fog computing: Architectures, applications with security challenges," A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [2] T. H. Luan, et al., "Fog computing: focusing on mobile users at the edge," <https://arxiv.org/abs/1502.01815>, 2018.
- [3] P. Hu, S. Dhelim, H. Ning, and T. Qiu, "Survey on fog computing: architecture, key technologies, applications and open issues," *Journal of Network and Computer Applications*, vol. 98, pp. 27-42, 2017.
- [4] Cisco Systems (2016). "Fog computing and the internet of things: Extend the cloud to where the things are," www.cisco.com, 2018.
- [5] G. Manogaran., "A new architecture of Internet of Things and big data ecosystem for secured smart healthcare monitoring and alerting system," *Future Generation Computer Systems*, <https://doi.org/10.1016/j.future.2017.10.045>, 2017.
- [6] A. A. Mutlag, M. K. A. Ghani, N. Arunkumar, M. A. Mohammed, and O. Mohd, "Enabling technologies for fog computing in healthcare IoT systems," *Future Generation Computer Systems*, vol. 90, pp. 62-78, 2019.
- [7] C. S. Nandyala, and H. K. Kim, "From cloud to fog and IoT-based real-time U-Healthcare monitoring for smart homes and hospitals," *International Journal of Smart Home*, vol. 10, no. 2, pp. 187-196, 2016.
- [8] M. Ahmadi, B. B. Rad, M. O. Thomas, and B. A. Onyimbo, "A shift in technological paradigm: cloud computing to fog computing," *Journal of Engineering Science and Technology, ICCSIT*, pp. 216-228, 2018.
- [9] S. Yi, C. Li, and Q. Li, "A Survey of fog computing: Concepts, applications and issues," *Proceedings of the 2015 Workshop on Mobile Big Data, ACM*, pp. 37-42, 2015.
- [10] L. M. Vaquero, and L. Rodero-Merino, "Finding your way in the fog: Towards a comprehensive definition of fog computing," *ACM SIGCOMM Computer Communication Review, ACM*, vol. 44, no. 5, pp. 27-32, 2014.
- [11] I. Stojmenovic, and S. Wen, "The fog computing paradigm: Scenarios and security issues," *Federated Conference on Computer Science and Information Systems, IEEE*, pp. 1-8, 2014.
- [12] S. B. Baker, W. Xiang, and I. Atkinson, "Internet of Things for smart healthcare: technologies, challenges, and opportunities," *IEEE Access*, 5, 26521-26544, 2017.
- [13] F. A. Kraemer, N. Tamkittikhun, A. E. Braten, and D. Palma, "Fog Computing in Healthcare – A Review and Discussion," *IEEE Access*, 5, pp. 9206 – 9222, 2017.
- [14] R. Mahmud, F. L. Koch, and R. Buyya, "Cloud – Fog Interoperability in IoT – enabled Healthcare Solutions," *In ICDCN '18*, pp. 4–7, 2018.
- [15] A. Paul, H. Pinjari, W. H. Hong, H. C. Seo, and S. Rho, "Fog Computing – Based IoT for Health Monitoring System," *Journal of Sensors*, 2018.
- [16] S. S. Gill, R. C. Arya, G. S. Wander, and R. Buyya, "Fog – Based Smart Healthcare as a Big Data and Cloud Service for Heart Patients Using IoT," *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI)*, pp. 1376 – 1383, 2018.
- [17] P. H. Vilela, J. J. P. C. Rodrigues, P. Solic, K. Saleem, and V. Furtado, "Performance evaluation of a Fog – assisted IoT solution for e-Health applications," *Future Generation Computer System*, vol. 97, pp. 379 – 386, 2019.
- [18] M. E. E. Mahmoud, et al. "Towards energy – aware fog enabled cloud of things for healthcare," *Computer and Electrical Engineering*, vol. 67, pp. 58 – 69, 2018.

Secure ECash Payment Method Based on Pseudo-Random Functions in Centralized and Decentralized Systems

Stefan Andonov
stefan.andonov@finki.ukim.mk

Jovana Dobрева
jovana.dobрева@students.finki.ukim.mk

Keti Isajloska
keti.isajloska@students.finki.ukim.mk

Lina Lumburovska
lina.lumburovska@students.finki.ukim.mk

Vesna Dimitrova
vesna.dimitrova@finki.ukim.mk

*Faculty of Computer Science & Engineering
Ss Cyril and Methodius, University of Skopje
Skopje, North Macedonia*

Abstract—One of the first online payment systems was eCash, which was discovered in the previous century and is still used worldwide. In this communication system, we are talking about more participants and contributors, but the main characters are the buyer and the seller. Their identity should stay anonymous and that is what this protocol can afford. Taking into consideration its advantages and disadvantages, the system works in real life centralized and decentralized systems and is spreading its usage to a higher level. Maintaining secure protocol must be obtained, so we discussed a few methods in order to achieve this property. We have also shown some basic principals for good implementation and what should be fulfilled in order to provide the best version of the eCash payment system. Basically, starting from the beginnings of eCash, we studied its implementation and design and then presented its security aspects and how this system works in real life examples in both centralized and decentralized systems.

Keywords – eCash payment system, digital cash, pseudo-random functions, anonymity, (de)centralized systems

I. INTRODUCTION

The emergence of electronic payment systems (EPS) revolutionized the way we buy and sell goods and services. As a result, we can say that global e-commerce is growing at an unprecedented pace. ECash is one of the first online payment systems. It is developed by the company DigiCash. Since it is based on the real money principle, it has changed the world of financial transactions. But, it seems unlikely that virtual money will ever completely replace conventional money. In comparison with conventional cash where the cash is spent many times by many people i.e it is completely transferable, eCash prevents double spending; it can be transferable only until it is spent. eCash payment system is mainly focused on electronic money anonymity assurance, and what is important is that the buyer and the seller must have an account opened at the same bank. [?]

A. Types of ECash

- Network type - supporting the electronic payment system using the value that is transmitted and stored to the PC from cyber bank and
- Smart card type - plastic cards embedding chips which involve the value and user's information [?]

B. How Does ECash Work?

In order to use eCash, every user needs to have an account. The bank provides and marks eCash in lieu of the user's conventional cash. Some protocols have been developed for ensuring that the system works seamlessly concerning account opening, withdrawal and payments. The end-user can download eCash from their bank account, and store them on a local hard drive. The same software provides taking amounts from their eCash wallet and adding it to the other's wallet. In order to verify the transaction, the eCash must go through the eCash bank. Transactions don't incur a fee except for a small amount charged by the eCash company. [?]

C. Advantages of ECash

- **Anonymity** - eCash system provides usage of the blind signatures in coin generating. In that way, the electronic money is impossible to trace them.
- **Security** - eCash uses secure protocol. The system uses the public cryptographic keys RSA that assure both the digital and blind signatures. In that way eCash is secure against the popular eaves-dropping attacks. The coin protection in the local machine can be improved by providing crypting and passwords.
- **Low cost** - in comparison to bank transactions which require huge amounts of infrastructures, eCash can use basic services such as the Internet to make the same transaction online, so it brings down the cost of the transaction.

- **Long Distance Transactions** - for long distance transactions, sending money with physical cash can be very expensive (paying fees), but eCash can be sent without too much of a hassle. [?]

D. Disadvantages of ECash

- **Spent coin database dimension** - since eCash system has very large database dimension for the signatures, and it is hard to manage with – this problem has been solved by using marked electronic tokens.
- **Standard** - many companies offer complete property on eCash so the system is not a standard one.
- **Not Traceable** - because eCash uses the Internet, traceability is getting difficult. This provides anonymity. The bad thing is that it can be susceptible to money laundering.
- **Forgery** - eCash system can be susceptible to forgery. There is a risk of breaking into the system, and making inappropriate actions - generating more coins on an inappropriate way - not paying anything to earn that cash. [?]

II. IMPLEMENTATION AND DESIGN OF ECASH

A. General Information About Good Implementation

In the following section, we will describe what are the key issues in order to provide good implementation of the eCash payment system. General implementation of the system is based on real money principle due to its primary usage described in the introduction part. Basically, this system uses cryptography with public keys in order to assure both the digital and the blind signatures. If we take a scenario from everyday life, for example a buyer and a seller trying to make a transaction or transfer money, the main point of this payment is keeping money anonymity assurance, which means both of them must have an account opened at the same bank as mentioned before. There are some essential requirements that must be accomplished for a successful implementation starting with its security. We researched more ways for establishing security, for example pseudo-random functions, more of which will be explained in the next section. [?]

We can discuss good implementation if the scalability of the system is at high point, because a stable system may support bigger number of users. What is more, if the system has an ability to transfer fast enough, the coins can be transferred directly to the user, without necessary verification of the coin issuer. Consequently, it leads to offline operability because the transfer is executed between two parts: the buyer and the seller, which means again there is no need of the coin issuer. It is one of the reasons why the security must be at the highest level and even when the transmission is not watched, the communication may happen without any problems. Since, the user interface is not a cryptographic problem and in most cases the user is a non-technical person, the system must be practical and easy to use. Last but not least, efficiency of the system is an important factor for all of the above, including the operations adding and

deleting users from the system which also contribute to a better system implementation. [?]

B. Design of the Payment System

The system is designed with few participants. As mentioned, the transfer has an issuer (issues coins), user (uses e-coins to buy or sell merchandise), payer (uses e-coins to buy merchandise), payment beneficiary (receives e-coins in order to sell merchandise) and certification authority (certificates the public keys of the participants). The design of the whole protocol is described graphically in Figure 1 [?], where we can see the process in two directions which means it is a constant process and each element is connected and depends on others.

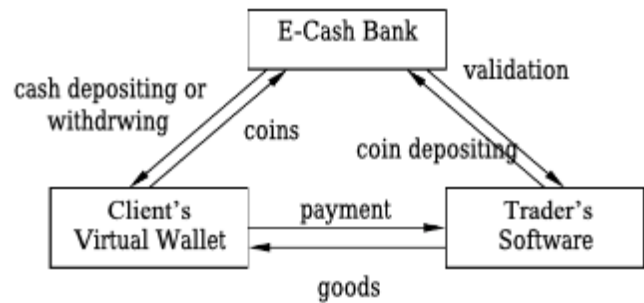


Fig. 1. ECash payment system [?]

The coins in eCash are pairs of two integers, first coordinate is the serial number of the coin and the second one is its value obtained during a calculation. For instance, the bank can use a private key from RSA algorithm to sign the second coin when the second value is being computed. If a is the first coordinate, $f(a)$ is the second one and f is a hash function.

Before the bank signs the coin, the user must prepare it. The user prepares a demand, for example 50 coins, 50\$ for each. He requires the following information: sum (50\$), serial number (the first coordinate) and identification number (11,12,13...), so that the serial number must be different for each coin. The identification number is a combination of two parts which are being generated using a different secret protocol. When the bank receives a request for transfer, it receives prepared money and uses the cut-and-choose protocol in order to open 40 from 50 coins and must verify if the sum is the same, but different serial number and valid identification number. After this, the next step is the signing process using the blind signature, which is introduced as the bling factor r . The bling factor is a random integer number which can be multiplied in coin before the bank signs it. After signing, the user can eliminate this random number. Instead of sending only $f(a)$, the user sends $f(a)*r$ to the bank. When the bank signs the coin, only the person knows what is the value after the factor r will be eliminated. [?]

C. Simple Example for Its Usage

The user A has some money and he wants that money to be signed from the user B using the blind signature. The user B has the public key e , the private key d and a public

module n . The user A selects a random number k from 1 to n . After this, the user A blinds the value a , computing $t = a * k^e(modn)$. The user B signs t with his private key d , $t^d = (a * k^e)^d(modn)$. The user A can reveal the money when the previous result is divided by k . When the process is over, the user A has the money signed by user B , which was his primary goal and in this case the user B does not know what he or she has signed. The formula for this example is:

$$\frac{f^d}{k} = \frac{(a * k^e)^d(modn)}{k} = \frac{a^d * k(modn)}{k} = a^d(modn) \quad (1)$$

The process shown on Figure 2 [?] is described in five steps

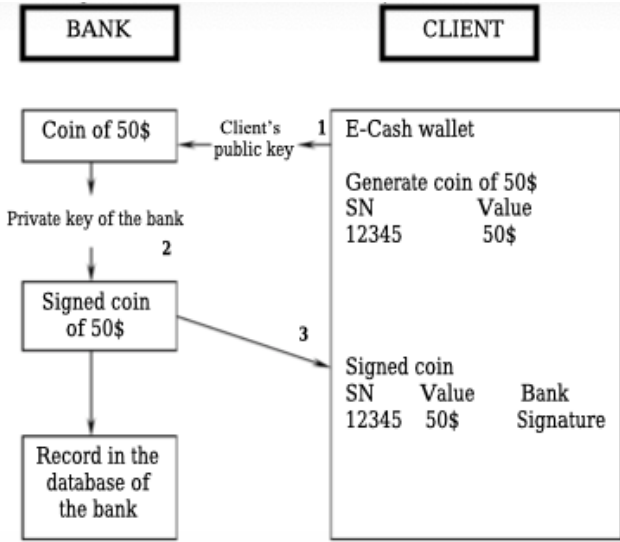


Fig. 2. Example for using eCash system [?]

and all in all it is how the spending of eCash works.

- 1) the seller requires the payment – if the buyer agrees then the eCash coins are selected and erased from the buyer, invalidating the numerical series. Then, each coin is sent to the seller
- 2) the seller sends the coins to the bank to see if the coins were spent another time
- 3) the bank verifies the signature and informs the seller if the coins are valid
- 4) if the coins are valid, the seller receives a confirmation and the value of the coins is transferred in the seller's account
- 5) the goods are transferred to the buyer.

III. SECURITY ASPECTS OF ECASH

A. Anonymous Compact eCash

Anonymous compact eCash was suggested by Camenisch, Hohenberger and Lysyanskaya [?], which was informally based on the following idea. Let N be the amount a user withdraws from a wallet. A consumer must then basically disclose $F_s(i)$, where F is a pseudo-random function with secret key s to spend the i -th coin and proof, in a zero-knowledge way, that it's well-formed. In other words, this

must be certified and the serial number must be generated using F_s on the input belonging in the interval of $[1, N]$. All these proofs can be mounted easily in several settings. Anonymity stems from the zero-knowledge properties of the proofs and the properties of the pseudo-random function, since it is difficult to determine whether $F_s(i)$ and $F_s(j)$ were generated under the same secret key s .

B. Compact / divisible eCash systems

A system offering efficient storage is called compact and a system supporting both efficient storage and spending is called divisible [?]. It leads for more formal approach, turning the traditional concept into a structured architecture. A user receives a certificate on some hidden value s during a withdrawal that will be used as a seed for a pseudo-random function (PRF) F , thus specifying the serial numbers of the N coins as $F_s(i)$ for $i \in (1, N]$. For more than a decade, the above issue remained undetected, while all compact / divisible eCash systems are based on the same intuition. For example let we have €8.75 investment in case all of the coins are of the smallest possible size. That means the consumer no longer has an € 10 coin but has 1000 coins of €0.01. Such a device can manage any amount without change but must have an effective means of stocking and spending hundreds of coins at once.

C. Generic Frameworks

In this paper we introduce the first frameworks for divisible eCash systems that only use constrained PRFs and very standard cryptographic primitives. We use the proofs from the book *Advances in Cryptology* [?] for the security of the global constructions assuming that each of the building blocks achieve some properties that are identified. It presents this complex primitive in a new light, highlighting its strong relations with constrained PRFs.

D. Pseudo-Random Functions

Starting with Camenisch's work [?] defining the serial numbers as outputs of a PRF, the specifications on a divisible eCash framework are formalized as properties that the PRFs must achieve. Actually, the main requirements is that the serial numbers can be revealed by batches, which means that it must be possible to reveal some element k_s that allows to compute $F_s(i)$, for every i of S , where $S \in [1, N]$ and does not provide any information on the other serial numbers, i.e. on the outputs of the PRF outside S . This exactly matches the definition of constrained PRF. There are also several requirements that the restricted key k_s must implicitly fulfill for anonymity to hold, and especially non-linkability of transactions: different restricted keys created from the same master key must be unlinkable, which also requires k_s to hide any information on the subset S (besides its cardinality, which will reflect the amount).

E. Security model

The security models are basically based on the interests of the consumer and the bank. Indeed, the former must be

able to anonymously invest their coins without being falsely accused of fraud, while the latter must be able to detect fraud and identify the perpetrators. This is formally defined by three security properties: anonymity (user spendings are anonymous, even with respect to the bank), exculpability (honest users cannot be falsely accused, even by the bank) and traceability (an author of overspending should be traced back). The framework [?] makes use of three standard cryptographic primitives, namely digital signature, commitment scheme and non-interactive zero-knowledge (NIZK) proofs along with their respective security properties.

- digital signature

Given a security parameter $\lambda \in N$, the algorithm generates a digital signature key pair (pk, sk) . To generate a private key for some user's identity ID, the PKG generates a fresh digital signature key pair.

- 1) $\text{Keygen}(1^\lambda)$: on input a security parameter λ , this algorithm outputs a pair of signing and verification keys (sk, pk)
- 2) $\text{Sign}(sk, m)$: on input the signing key sk and a message m , this algorithm outputs a signature σ
- 3) $\text{Verify}(pk, m, \sigma)$: on input the verification key pk , a message m and its alleged signature σ , this algorithm outputs 1 if σ is a valid signature on m under pk , and 0 otherwise

- commitment scheme

- 1) $\text{Keygen}(1^\lambda)$: on input a security parameter λ , this algorithm outputs a commitment key ck that specifies a message space M , a randomizer space R along with a commitment space C
- 2) $\text{Commit}(ck, m, r)$: on input ck , an element $r \in R$ and a message $m \in M$, this algorithm returns a commitment $c \in C$

- zero-knowledge proof systems

- 1) completeness: if the statement is true, the prover should be able to convince the verifier.
- 2) soundness: a malicious prover should not be able to convince the verifier if the statement is false.
- 3) zero-knowledge: a malicious verifier learns nothing except that the statement is true

F. First Divisible eCash System Secure in the Standard Model

Finally, extensive evidence is given for the structures to demonstrate that the overall construction protection usually keeps each of the building blocks under the defense. In concrete terms, this means that a safe divisible eCash framework can be designed for any setting by essentially designing a restricted PRF which achieves some simple properties. To highlight this point, the first divisible eCash method in the standard model is protected by using this structure.

IV. REAL LIFE EXAMPLES OF ECASH PAYING SYSTEMS

Although eCash as a concept was invented in the previous century, it is still a huge inspiration for the modern electrical paying systems. In this part we will give an overview on

some of the currently most used paying systems based on the eCash system. From 2015 the term eCash is used for the digital cash that can be transferred between entities and can be stored on electronic cards or some alternative online mobile or web platforms. [?]

Nowadays, the electronic payment systems are a very important part of our lives and we use them all the time. Based on the fact whether there is a central authority that controls the money flow, the payments systems can be classified as centralized or decentralized.

The centralized payment systems have a central authority (a bank, a nation) and therefore they are regulated by laws in specific countries. The centralized systems support cash transfer protocols (which have to be authorized by a bank), but they do not support in general the concept of digital currency. Also, nowadays the centralized systems are moving into the direction of digitalization of the transactions via creation of digital wallets where people can store different types of funds and make contactless transactions through one or multiple mobile devices. The first example of a digital wallet was the Mondex who introduced the so called electronic purse. Some of the most popular digital wallets that are being used nowadays are: PayPal, Google's Wallet, Apple Pay, Venmo, eWallet etc. Some of the digital wallets can work with digital currencies that are part of the decentralized systems. [?]

The decentralized payments systems do not have a central authority and mainly they are not regulated by laws. The main asset in the decentralized systems are the cryptocurrencies. The cryptocurrencies are dependent on the digital signatures for asset transfer, on peer-to-peer networkings and on proof-of-work and proof-of-stake schemes for management of the payment systems. The first published cryptocurrency was Bitcoin which is based on blockchain technology. After that over 6000 altcoins (alternative cryptocurrencies to Bitcoin) have been released. [?]

ACKNOWLEDGMENT

This work was partially financed by the Faculty of Computer Science and Engineering at the Ss. Cyril and Methodius University, Skopje, North Macedonia.

REFERENCES

- [1] Baddeley, Michelle. "Using e-cash in the new economy: An economic analysis of micro-payment systems." *Journal of Electronic Commerce Research* 5.4 (2004): 239-253.
- [2] Fera, Leah, et al. "Digital cash payment systems." Dec 6 (1996): 1-21. APA
- [3] <https://due.com/blog/defining-ecash-and-calculating-its-benefits/>
- [4] Jing, Yang. "On-line Payment and Security of E-commerce." *Proceedings. The 2009 International Symposium on Web Information Systems and Applications (WISA 2009)*. Academy Publisher, 2009.
- [5] Srivastava, Lara and Robin Mansell. "Electronic Cash and the Innovation Process: A User Paradigm." (1998).
- [6] Marius, Popa, Calugaru, Adrian. (2009). On-line Payment System Survey-eCash. *Journal of Mobile, Embedded and Distributed Systems*. 1. 95-103.

- [7] Boldyreva, Alexandra, and Daniele Micciancio, eds. *Advances in Cryptology—CRYPTO 2019: 39th Annual International Cryptology Conference*, Santa Barbara, CA, USA, August 18-22, 2019, Proceedings. Part III. Vol. 11694. Springer, 2019.
- [8] Camenisch, Jan, Susan Hohenberger, Anna Lysyanskaya. "Compact e-cash." *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, Berlin, Heidelberg, 2005.
- [9] <https://due.com/ecash/>
- [10] <http://nearfieldcommunication.org/payment-systems.html>
- [11] Böhme, Rainer, et al. "Bitcoin: Economics, technology, and governance." *Journal of economic Perspectives* 29.2 (2015): 213-38.
- [12] Bourse, Florian, David Pointcheval, and Olivier Sanders. "Divisible e-cash from constrained pseudo-random functions." *International Conference on the Theory and Application of Cryptology and Information Security*. Springer, Cham, 2019.
- [13] Camenisch, Jan, et al. "Oblivious PRF on committed vector inputs and application to deduplication of encrypted data." *International Conference on Financial Cryptography and Data Security*. Springer, Cham, 2019.

Scanning of services based on E-Governance Macedonia 2020

Boshko Kitanov, Gzim Ibraimi, Marjan Gusev

Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering

Skopje, Republic of North Macedonia

kitanov93@icloud.com, gzim.ibraimi@gmail.com, marjan.gushev@finki.ukim.mk

Abstract—This article presents the activities and the mechanisms of scanning the effectiveness of e-Government services in Republic of North Macedonia. The goal is to compare the results with other European Union members and Western Balkan's countries. The used methodology for scanning the effectiveness of e-Government services is adopted by the European Union, but since North Macedonia is not a member of the Union, the scanning was not conducted within their earlier measurement reports. These results will help the e-Government institutions in North Macedonia to set priorities and improve their services. Finally, the Government can sooner implement the proposed development standards and integrate in the European e-Government services framework, since the benchmark score is below EU average.

Index Terms—e-government, e-governance, administration, e-services, benchmark

I. INTRODUCTION

E-government is generic term for web-based services of the local, public and federal government institutions [1]. Based on our analysis we can conclude that Republic of North Macedonia joined the trends of using the information and communication.

The scanning of performances, the influence and transparency in the government in the context of e-government comes from the citizens who want to have better services from the public institutions, more effective process when request services, or even making decisions within the processes. The scanning of e-government in Republic of North Macedonia is based on three priorities from the Benchmark in EU for 2020 (a) modernizing public administration using digital keys; (b) enabling citizens and businesses with cross-border interoperability, and (c) facilitating digital interaction between administrations and high quality public service businesses based on the following categories: business services, citizen services and government services[2][3].

This process points out if the goals overlap the implementation of Information and communications Technology in the government of Republic of North Macedonia. Following this importance, the scanning should result increment of the effectivity in the internal organization in the institutions, identifications of the level of implementation of e-services and citizen satisfaction with the services.

The main goal of this document is to gather information of the actual state of the presence of information technology in various fields that are important for the citizens of Republic of North Macedonia.

The paper follows the next structure. Section II presents the used methods and Section III the results from the 2019 benchmarking. The obtained results are discussed in Section IV and conclusions presented in Section V.

II. METHODS

The methods are specified in the 2019 eGovernment Benchmark Insight Report and Background Report [3]. The scanning data are collected from public available data and in the first phase, the requests have been sent to 10 municipalities, 3 ministries for measuring of ICT in the internal organization of the institutions. The findings from the collected data are analyzed in Section III.

III. RESULTS

A. Our results

All institutions reported that they have internet connection, they are interconnected to departments parts of the institutions and from data security level all of them have installed Antivirus software, which we consider that the level of data security is low. The data that we received based on online portals in the second phase are the data that give us clarity how the process of release of one service is conducted . For instance:

- business services like company registration or removing data from the Central Register can be handled very easy online through the web site [4].
- The annual tax return allows all citizens to send all the documents like tax reports, requests, analysis, balance sheets etc., nonstop from everywhere with the digital certificate.
- The request for environmental permit is not well digitalized and designed, on their website there is only data and steps how they one can conduct a request on the counters [5].
- E-banking also did positive impact to companies like paying current receipts or invoices online and with mobile apps.
- The citizen services have also increased their online presence. The birth certificate is one of the services that can be completed online.
- Through their web portal [6] , the citizens can apply for birth certificate, to pay administrative taxes and other

Results summary									
Country	Benchmark	Biennial average (2017 + 2018)	Average (2018)	Average (2017)	Average (2016)	Business Start-up (2018)	Family (2018)	Losing and Finding a Job (2018)	Studying (2018)
MK	OVERALL AVERAGE	37.3	36.9	.	.	51.8	36.6	33.5	25.5
MK	USER CENTRIC GOVERNMENT	61.0	61.1	.	.	73.3	65.5	55.9	49.5
MK	TRANSPARENT GOVERNMENT	42.8	42.9	.	.	55.4	44.3	44.7	27.1
MK	CROSS-BORDER MOBILITY	40.5	40.5	.	.	55.5	.	.	25.5
MK	KEY ENABLERS	5.8	5.8	.	.	23.2	0.0	0.0	0.0

Fig. 1. Average values of EU Benchmark for North Macedonia

payments online. After this they can choose from which unit, they want to pick up their certificate.

- The services for unemployed people are not fully online. If they want to register to the employment agency, they can only read the terms and conditions that this agency offers through their website.
- In some universities it is possible college enrollment and total evidence of the semester online [7].

Public procurement takes a lot of time of handling in the total documentations. E-procurement is the portal through which companies can follow the public tenders, to apply online and to follow the total procedure through the portal. [8] The custom declarations and custom documentations can also be handled online. [9] The usage of the Information and communication Technology in the public administration is global trend in the theory and practice known as process of e-government development.[10] Based on the EU Benchmarking in Republic of North Macedonia is measured with these indicators: User centric government, Transparent government, Citizen Mobility, Business Mobility, and Key Enablers.

B. 2019 Benchmarking results

According to the EU, benchmarking North Macedonia is measured by these indicators: USER CENTRIC GOVERNMENT (User - Online availability, User – Usability, User - Mobile Friendliness), TRANSPARENT GOVERNMENT (Service delivery, Public organizations, Personal data), CITIZEN MOBILITY (Online availability, Usability, eID Cross Borders, eDocuments Cross Borders), BUSINESS MOBILITY (Online availability, Usability, eID Cross Borders, eDocuments Cross Borders), KEY ENABLERS (eID, eDocuments, Authentic sources, Digital Post). Evaluating each indicator in a scalar way yields the following results.

Figure 1 shows that North Macedonia achieves poor performance in all indicators. The weakest indicator is the digital certificates (signatures, authentication), which is in the first steps of implementation in our country.

The table shows that interoperability across borders is slightly more developed but still below the average of other countries in the region. For interoperability North Macedonia has a national framework for all public services offered to citizens, the business sector and government authorities. Services to citizens is more developed where North Macedonia has an average of 61.0, but compared to other U countries, North Macedonia is below the average of those countries, which means services cannot yet be fully implemented electronically. All services are not yet electronically automated,

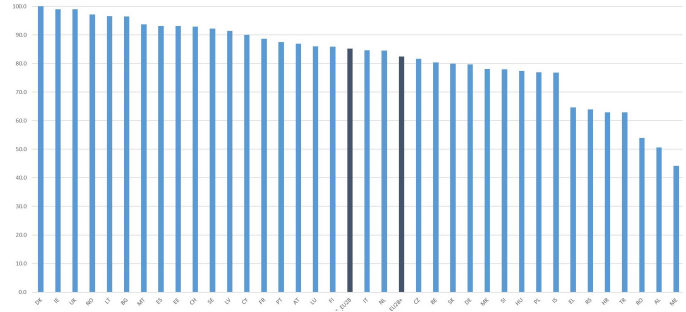


Fig. 2. Rank of business e-Gov services in 2019, (MKD close to the average)

and we should contact the institution for more details or to complete the full procedure. The only full service available electronically is tax reporting by the Public Revenue Office-PRO.

Compared to other EU countries and EU candidate countries, North Macedonia is ranked (Figure 2) according to the following indicators: Online availability, Usability, eID Cross Borders, eDocuments Cross Borders [11].

Figure ?? describes all electronic services offered at the state level broken down by categories of business, family, work and study. Different services are grouped that are related to that category

The way in which the procedure for completing and closing a service is performed, the measurement is based on whether the service is fully electronic or not. According to the measurements and research of all electronic services it is concluded that North Macedonia has not implemented fully online electronic services that covers only certain service. According to this North Macedonia is ranked as second last compared to all European countries (Figure 4.

The next indicator that is important for the initial phase of e-services is pre-filled forms or partially filled forms as a facilitating step for citizens to obtain e-services. Related to this indicator North Macedonia had a positive increasing on implementing, but development and expansion did not go as planned, and by the end of 2019 North Macedonia is at the bottom of the list compared to other European countries (Figure ??).

IV. DISCUSSION

By the end of the decade, the Government of the Republic of North Macedonia is committed to establish an environment that will take advantage of the ICT industry and create an advanced information society.

Country	Life Event	Label service	URL	Service Provider	Automated	Applicable	Geo level
MK	Business	10.1 Register your company	Not applicable	.	.	Not Applicable	.
MK	Business	10.2 Register employee before	https://e-rabota.avrm.gov.mk/Default.aspx	Employment Service Agency	Not Automated	Applicable	National
MK	Business	10.3 Tax related obligations	www.ujp.gov.mk	Public Revenue Office	Not Automated	Applicable	Regional
MK	Business	10.4 Obligations related to social	http://www.piom.com.mk/	Pension and Disability Insurance	Not Automated	Applicable	Local
MK	Business	10.5 Obligations regarding regist	www.stat.gov.mk	State Statistical Office	Not Automated	Applicable	Regional
MK	Business	10.6 Obligations related to work	http://www.mtsp.gov.mk/content/pdf/trud_2017/pravilnici/%D0%97%D0%B0	Ministry of Labour and Social Pol	Not Automated	Applicable	National
MK	Business	10.6 Obligations related to work	http://www.mtsp.gov.mk/zekoni.nsp	Ministry of Labour and Social Pol	Not Automated	Applicable	National
MK	Business	10.7 Obligations related to tra	Not applicable	.	.	Not Applicable	.
MK	Business	11.1 Find out if you need to re	http://www.economy.gov.mk/	Ministry of Economy	Not Automated	Applicable	Local
MK	Business	11.1 Find out if you need to re	http://www.moep.gov.mk/?page_id=901	Ministry of Environment and Phys	Not Automated	Applicable	Local
MK	Business	11.1 Find out if you need to re	http://www.mzsv.gov.mk/?q=node/1127	Ministry of Agriculture, Forestry	Not Automated	Applicable	Local
MK	Business	11.2 Submit an application for	http://www.economy.gov.mk/	Ministry of Economy	Not Automated	Applicable	Local
MK	Business	11.2 Submit an application for	http://www.moep.gov.mk/?page_id=901	Ministry of Environment and Phys	Not Automated	Applicable	Local
MK	Business	11.2 Submit an application for	http://www.mzsv.gov.mk/?q=node/1127	Ministry of Agriculture, Forestry	Not Automated	Applicable	Local
MK	Business	1.1 Obtaining information abou	http://rabotaimoznosti.mk/samovrabotuvanje/?lang=mk	Ministry of Labour and Social Pol	Not Automated	Applicable	National
MK	Business	1.1 Obtaining information abou	http://www.avrm.gov.mk/	Ministry of Labour and Social Pol	Not Automated	Applicable	National
MK	Business	1.1 Obtaining information abou	http://www.economy.gov.mk/	Ministry of Economy	Not Automated	Applicable	National
MK	Business	1.2 Setting up a business plan	http://rabotaimoznosti.mk/samovrabotuvanje/?page_id=148&lang=mk	Ministry of Labour and Social Pol	Not Automated	Applicable	National
MK	Business	1.2 Setting up a business plan	http://www.avrm.gov.mk/	Ministry of Labour and Social Pol	Not Automated	Applicable	National
MK	Business	1.2 Setting up a business plan	http://www.economy.gov.mk/	Ministry of Economy	Not Automated	Applicable	National
MK	Business	1.3 Explore financial possibili	http://www.economy.gov.mk/	Ministry of Economy	Not Automated	Applicable	National
MK	Business	2.1 Confirm general manager	Not applicable	.	.	Not Applicable	.
MK	Business	2.2 Confirm activity-specific g	Not applicable	.	.	Not Applicable	.
MK	Business	3.1 Obtain certificate of no out	http://www.ujp.gov.mk/m/plakjanje/category/615	Public Revenue Office	Not Automated	Applicable	National
MK	Business	3.2 Obtain character referenc	http://sud.mk/wps/portal/osbitola/sud/!ut/p/z1/04_Sj9CPykssy0xPLMnMz0vMA	Primary Court Bitola	Not Automated	Applicable	Local
MK	Business	3.2 Obtain character referenc	http://sud.mk/wps/portal/oskumanovo/sud/!ut/p/z1/04_Sj9CPykssy0xPLMnMz0vMA	Primary Court Kumanovo	Not Automated	Applicable	Local
MK	Business	3.2 Obtain character referenc	http://sud.mk/wps/portal/osprilep/sud/!ut/p/z1/hZBCslwEES_SHY0scZjCpRQ	Primary Court Prilep	Not Automated	Applicable	Local
MK	Business	3.2 Obtain character referenc	http://sud.mk/wps/portal/oskopje1/sud/za-sudot/opsti-podatoci/!ut/p/z1/JZB	Primary Court Skopje 1	Not Automated	Applicable	Local
MK	Business	3.3 Obtain certificate of no out	www.fzo.com.mk	Health Insurance Fund	Not Automated	Applicable	Local
MK	Business	3.3 Obtain certificate of no out	www.piom.org.mk	Pension and Disability Insurance	Not Automated	Applicable	Local
MK	Business	3.4 Obtain certificate from bar	http://www.crm.com.mk/DS/	Central Register	Not Automated	Applicable	National
MK	Business	4.1 Fill in standard form for re	http://www.crm.com.mk/DS/	Central Register	Not Automated	Applicable	National
MK	Business	4.2 Register company name	http://www.crm.com.mk/DS/	Central Register	Not Automated	Applicable	National
MK	Business	4.3 Register domicile of busin	http://e-submit.crm.com.mk/eFiling/informativni-sodrzini/podnesete-prijava-v	Central Register	Not Automated	Applicable	National
MK	Business	4.4 Formal validation of signat	Not applicable	.	.	Not Applicable	.
MK	Business	5.1 Register with Commercial	Not applicable	.	.	Not Applicable	.
MK	Business	5.2 Register with central / regi	Not applicable	.	.	Not Applicable	.
MK	Business	5.3 Register with Trade Regist	http://e-submit.crm.com.mk/eFiling/redefault.aspx	Central Register	Not Automated	Applicable	National
MK	Business	6.1 Register with Trade Assoc	http://chamber.mk/en/	Macedonian Chambers of Comm	Not Automated	Applicable	National
MK	Business	7.1 Obtain tax identification c	www.ujp.gov.mk	Public Revenue Office	Automated	Applicable	Regional
MK	Business	7.2 Obtain VAT collector num	www.ujp.gov.mk	Public Revenue Office	Not Automated	Applicable	Regional
MK	Business	8.1 Register with Social Secur	http://www.avrm.gov.mk/	Employment Service Agency of t	Not Automated	Applicable	Local
MK	Business	8.2 Register with mandatory p	http://www.piom.com.mk/	Pension and Disability Insurance	Not Automated	Applicable	Local
MK	Business	8.3 Register with compulsory	http://www.fzo.org.mk/default-mk.asp	Health Insurance Fund	Not Automated	Applicable	Local
MK	Business	8.4 Register with mandatory c	Not applicable	.	.	Not Applicable	.
MK	Business	9.1 Publish registration in Offi	Not applicable	.	.	Not Applicable	.

Fig. 3. E-Business services in North Macedonia according to EU 2020 e-Government benchmark

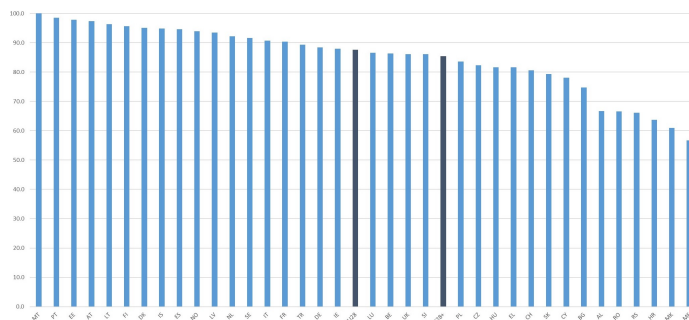


Fig. 4. Rank of completely online services (MKD is second last)

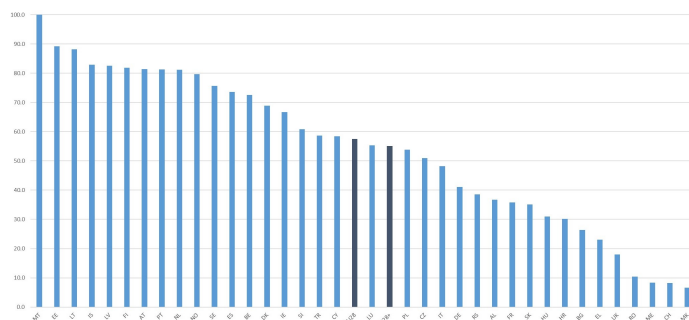


Fig. 5. Rank of pre-filled online forms (MKD is last)

In regards of the progress of e-Government sophistication of public services, North Macedonia is surrounded by good practices in developed countries, as well as in fellow neigh-

bors' governments. These practices provide a guide to how public services can strive towards growth in the near future, because growth in Macedonia's e-government sophistication has been detected since 2001 in research papers, but not so significant as other e-government's efforts [2] [12][13].

The first measurement in a 2004 study has reported that the average online sophistication in North Macedonia is 9 percents, in March 2006 the average increased to 32.75 percents, and in March 2007 it was 50 percents. The 2018 average score is 77 percents by 2010 benchmarks, so a significant increase is detected in all public services [2].

The National Short-term ICT Strategy covers the period up to 2017 [14] and is the first step towards creating a long-term strategy. This should provide a good basis for a general National ICT Strategy (by 2020), which will help the Republic of North Macedonia to advance the development of the information society and create an advanced skills ICT society.

Based on the above we can conclude that North Macedonia has made progress in introducing ICT in the business sector (Figure 6) and leads in the implementation of e-government services for the business sector in relation to Western Balkan countries.

When it comes to fully online (electronic) services there is no progress like other countries in the region. Processes of e-services have started but are still in the development phase and the process of introducing fully electronic services has been slowed down (Figure 7).

Pre-filled e-forms is the first step of e-government as an

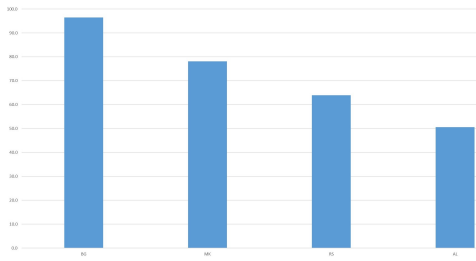


Fig. 6. Rank of pre-filled online forms for business sector in WB countries

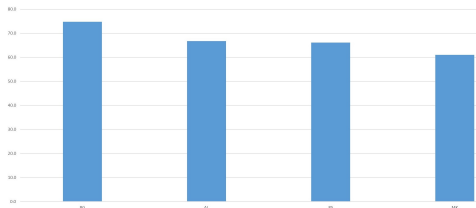


Fig. 7. Rank of fully online services in WB countries

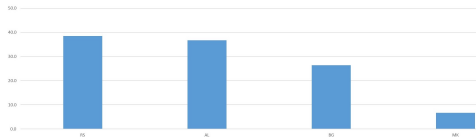


Fig. 8. Implementation of e-services in WB countries

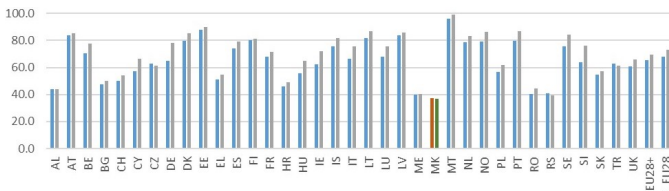


Fig. 9. Comparison of EU 2020 e-Government Benchmark for 2018

incentive for using these e-services. Although North Macedonia was a leader in the introduction of ICT in the region of high-speed broadband Internet, which was expected implementing electronically of almost all services, there were many contentious issues in the implementation and realization of the processes which indicated pausing or delaying of those processes. This makes North Macedonia the last in the region regarding the implementation of e-services. Figure 8 shows that even at the first most important step, introducing electronic forms for offered services, North Macedonia ranks behind its neighbors like Albania, Serbia and Bulgaria.

North Macedonia is below the average of other European countries (Figure 9). Although measures have been taken in recent years to develop ICT at the state level and to develop e-government as a separate segment, the graph shows us that we are still at an early stage and should work more seriously by applying the good practices of other countries that have already passed this process.

V. CONCLUSION

Based on the research, we can conclude that e-government remains a key issue for government institutions both at the central and local levels as using e-government services is proven to reduce costs for citizens, improve administrative efficiency and increases transparency of government institutions.

At the same time, we are taking into a consideration the benefits that citizens have using the Internet. The government must regulate these aspects so they would be more accessible, more user -friendly to citizens, business sector, the non-governmental sector and the public sector.

As we expect the increase of computer literacy at young population group in the future and becoming familiar with the techniques and methods of using contemporary ICT, we hope that the implementation of the e-government concept will evolve with a fast-growing dynamic.

Considering that introduction of IT society in North Macedonia directly depends on the level of development of communication infrastructure and technologies and the level of usage of services, two strategies presents strong initiator of balanced economic process which leads to establishment of IT society in Republic of North Macedonia.

According to the analysis of the indicators and results presented, we can say the denial of all these processes are related to digital division among the citizens of North Macedonia as a factor that contributes in it. In order to eliminate this factor, we think that free short training should be organized and provided on how to implement an e-service with the citizens through a mobile team and target the settlements away from the capital.

REFERENCES

- [1] S. Janeska-Sarkanjec, "Modeli na e-upravljanje," in *Modeli na e-upravljanje*, 2012, pp. 106–111.
- [2] I. Akjimoska, A. Antikj, and M. Gusev, "Online sophistication of e-government services in north macedonia," 2019.
- [3] European Commission. *egovernmentbenchmark2019results*. [Online]. Available: <https://ec.europa.eu/digital-single-market/en/news/egovernment-benchmark-2019-trust-government-increasingly-important-people>
- [4] Central Register of the Republic of North Macedonia. Online registration for businesses. [Online]. Available: <https://e-submit.crm.mk/eFiling/e-podnesi-sam/-.aspx>
- [5] Ministry of Environment and Physical Planning. Environment permits. [Online]. Available: <http://www.moep.gov.mk>
- [6] Office for Management of registers of births, marriages and deaths. Personal documentation. [Online]. Available: <https://e-portal.uvmk.gov.mk/external/portalHomePage>
- [7] Employment Service Agency of the Republic of North Macedonia. Employment services. [Online]. Available: <https://av.gov.mk>
- [8] Public Procurement Bureau. Procurement services. [Online]. Available: <https://www.e-nabavki.gov.mk/PublicAccess/Home.aspx#/home>
- [9] Ministry of Finance of Republic of North Macedonia. Cross border interoperability. [Online]. Available: <https://trader.customs.gov.mk/myAccount-ui/protected/welcome.htm>
- [10] USAID, MIOA, Metamorphosis, "Osnovi i razvoj na e-vlada," 2015.
- [11] European Commission, "Digital government factsheet, republic of north macedonia," 2019.
- [12] M. Gušev, K. Kiroski, M. Kostoska, and K. Budimovski, "Growth rate categorization of e-government development," 2012.
- [13] K. Kiroski, M. Kostoska, M. Gusev, and S. Ristov, "Growth rate analysis of e-government development," in *Proceedings of the Fifth Balkan Conference in Informatics*, 2012, pp. 106–111.
- [14] Ministry of Information Society and Administration, "Nacionalna kratkoročna ikt strategija," 2015.

PubSub implementation in Haskell with formal verification in Coq

Boro Sitnikovski*, Biljana Stojcevska*, Lidija Goracinova-Ilieva*, Irena Stojmenovska†

*Faculty of Informatics, UTMS Skopje

buritomath@gmail.com, {b.stojcevska, l.goracinova}@utms.edu.mk

†School of Computer Science and IT, UACS Skopje

irena.stojmenovska@uacs.edu.mk

Abstract—In the cloud, the technology is used on-demand without the need to install anything on the desktop. Software as a Service is one of the many cloud architectures. The PubSub messaging pattern is a cloud-based Software as a Service solution used in complex systems, especially in the notifications part where there is a need to send a message from one unit to another single unit or multiple units. Haskell is a generic typed programming language which has pioneered several advanced programming language features. Based on the lambda calculus system, it belongs to the family of functional programming languages. Coq, also based on a stricter version of lambda calculus, is a programming language that has a more advanced type system than Haskell and is mainly used for theorem proving i.e. proving software correctness. This paper aims to show how PubSub can be used in conjunction with cloud computing (Software as a Service), as well as to present an example implementation in Haskell and proof of correctness in Coq.

Index Terms—cloud computing, Software as a Service, PubSub, Haskell, Coq

I. INTRODUCTION

A cloud can be both software and infrastructure. It can be an application accessed via the Internet or a server provided when needed. If a service can be accessed by a device, regardless of the operating system of that device, then that service is cloud-based [1].

Typically, three criteria are defined as labels whether a particular service is a cloud service [1]:

- the service is available through a web browser or web services API,
- zero capital spending is needed to get started,
- payment is required only for those services that are used.

The PubSub (publish-subscribe) pattern allows for easy message transfer to specific channels.

Haskell and Coq are programming languages designed with an aim to accomplish software correctness, based on a typed version of the lambda calculus system [2], [3], [4].

II. ARCHITECTURE

A. Cloud computing

We can categorize most cloud-computing projects in three basic categories:

- projects that deploy services for multiple applications or clients,
- projects that are single, standalone cloud applications,

- cloud-based service provider projects (e.g. Google).

Besides these project categories, there are three main cloud-based architectures [5]:

- SaaS - software as a service (e.g. Google Apps, Salesforce, Dropbox, games such as World of Warcraft)
- PaaS - platform as a service (e.g. Windows Azure, Heroku, Google App Engine, WordPress web site development platform)
- IaaS - Infrastructure as a Service (e.g. Google Cloud Platform, Amazon Web Services, Microsoft Azure, etc.)

B. Software as a Service

SaaS refers to software hosted on the cloud in a central location (Figure 1) [1]. Typically, this architecture consists of web-based software, but it is not limited to.

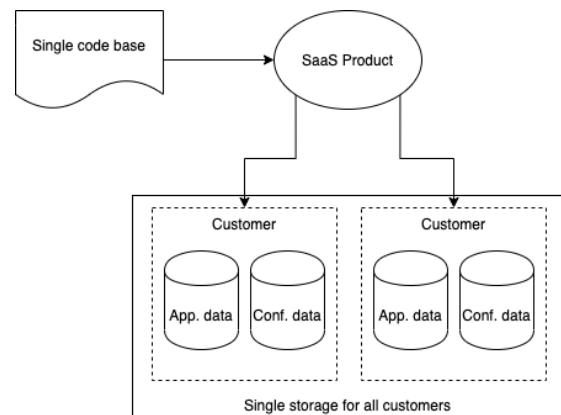


Fig. 1. SaaS architecture

SaaS applications are accessed through a client such as a web browser [1]. This architecture applies to many business applications, including Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), Office software, messaging software, etc [5]. SaaS is involved in the strategy of almost all leading software companies - Amazon, Google, Microsoft [5].

C. PubSub service

The PubSub pattern is a messaging pattern where publishers can send messages to specific channels to which

subscribers are subscribed [6]. The commands `SUBSCRIBE`, `UNSUBSCRIBE`, and `PUBLISH` implement this pattern. This separation of publishers and subscribers provides better scalability to the service [6].

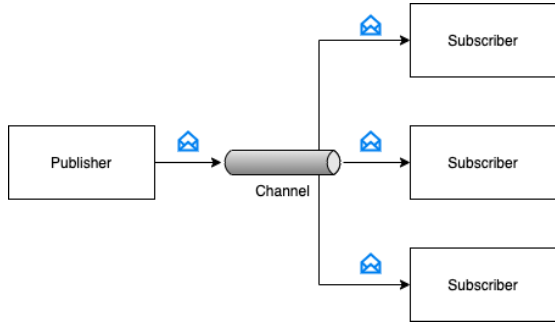


Fig. 2. PubSub architecture

There are three basic components to understanding a PubSub messaging scheme, as shown in Fig. 2:

- Publisher - Publishes messages to the communication infrastructure
- Subscriber - Subscribes to a specific channel/category of messages
- Information Infrastructure (Channel) - Handles subscriptions and receives publisher's messages

III. IMPLEMENTATION IN HASKELL

In our implementation of PubSub, we use the Haskell programming language. Haskell is an advanced functional programming language. The development of Haskell is rooted in mathematics and computer science research [2].

A. Subscription module

This module represents the business logic functions of the PubSub architecture. Namely, each subscription is represented as a pair of Channel and Connection. Further, all such subscriptions are merged into a single list of subscriptions.

The function `getHandlesByCh` takes a Channel and a list of Subscriptions and then returns a filtered list of subscriptions such that the channel is matched. The Haskell's built-in function `filter` [2] will be used by the `publish` command.

```

getHandlesByCh :: Channel -> [Subscription a]
               -> [Subscription a]
getHandlesByCh c = filter (\x -> c == fst x)
  
```

The function `removeHandleByCon` takes a connection and a list of Subscriptions and then returns a filtered list of subscriptions such that the selected connection is not contained. This will be used by the `quit` command.

```

removeHandleByCon :: (Eq a) => Connection a
                  -> [Subscription a] -> [Subscription a]
removeHandleByCon h = filter (\x -> h /= snd
                               x)
  
```

The function `removeSubscription` takes a channel and a connection and a list of Subscriptions and then returns a filtered list of subscriptions such that the selected connection and channel are not contained. This will be used by the `unsubscribe` command.

```

removeSubscription :: (Eq a) => Channel ->
                    Connection a -> [Subscription a] ->
                    [Subscription a]
removeSubscription c h = filter (\(x, y) ->
                                not (x == c && y == h))
  
```

The function `addSubscription` takes a channel and a connection and a list of Subscriptions and then returns a list such that the selected connection and channel are contained. This will be used by the `subscribe` command.

```

addSubscription :: Channel -> Connection a ->
                [Subscription a] -> [Subscription a]
addSubscription c h s = (c, h) : s
  
```

The function `hInSubscription` checks if a given channel/subscription is contained into a list of subscriptions.

```

hInSubscription :: (Eq a) => Channel ->
                  Connection a -> [Subscription a] -> Bool
hInSubscription c h = any (\(x, y) -> x == c
                               && y == h)
  
```

The next function is a helper function for the `subscribe` command.

```

subscribe ch h s = if not (hInSubscription ch
                          h s) then addSubscription ch h s else s
  
```

Conversely, the following function is for the `unsubscribe` command.

```

unsubscribe ch h s = if hInSubscription ch h
                    s then removeSubscription ch h s else s
  
```

Finally, the last function `publish` uses the condition the types `t` and `m` to be `Foldable` (to be iterated) and `Monad` (to be chain-operated, for example in IO) respectively. This covers a more general case.

However, in this specific case (PubSub), `mapM` iterates through the list of subscriptions and perform the function `f` on each subscription individually. In the PubSub implementation, `f` is a function that writes to an IO object (`hPutStrLn`). This generalization, even-though complex, is necessary to easily prove correctness, since in Coq, IO does not exist as a concept.

```

publish :: (Traversable t, Monad m) => (t2 ->
                                         m b) -> t (a, t2) -> m (t b)
publish f = Data.Traversable.mapM (\(_, y) ->
                                   f y)
  
```

B. Main module

This module represents the entry-point of the program, where the logic to accept new clients is initialized.

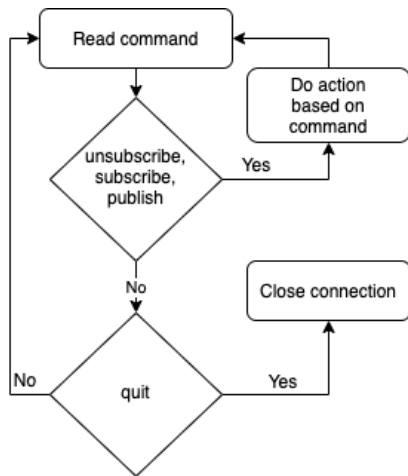


Fig. 3. PubSub algorithm

The main module accepts new connections indefinitely, launching a new thread for every connection. These threads use the algorithm described in Fig. 3.

The algorithm handles different commands passed through the communication channel. For a given command, list of subscriptions, and a connection, the algorithm updates the list of subscriptions and returns a boolean result, representing the success of the command's execution. Accepted commands are `unsubscribe`, `subscribe`, `publish` and `quit`. This algorithm will be executed recursively until the client closes the connection.

C. Running example

In this subsection, we will demonstrate the usage of our application through the client program `telnet`. We will create two connections, subscribed to channel 1 and 2 respectively, and then we will send messages to these subscribers through a third connection.

First connection:

```
$ telnet localhost 123 import
Connected to localhost.
Write 'publish <ch> <msg>' to publish,
      'subscribe <ch>' to subscribe.
> subscribe 1
```

Second connection:

```
$ telnet localhost 123
Connected to localhost.
Write 'publish <ch> <msg>' to publish,
      'subscribe <ch>' to subscribe.
> subscribe 2
```

Third connection:

```
$ telnet localhost 123
Connected to localhost.
Escape character is '^]'.
Write 'publish <ch> <msg>' to publish,
      'subscribe <ch>' to subscribe.
> publish 1 hey there
```

```
> publish 2 hello!
```

The first and the second connection will receive `[1] hey there` and `[2] hello!` respectively.

IV. FORMAL PROOF OF CORRECTNESS IN COQ

Coq is a programming language used as an interactive theorem prover that enables expressing mathematical definitions [3]. By mechanically validating evidence of mathematical claims, Coq helps in producing a certified program. It is based on the theory of Calculus of Inductive Constructions, a system in the lambda calculus family [4]. Haskell is also based on a weaker version of lambda calculus, hence there is a strong connection between these two programming languages.

Coq's initial release was done in 1989 by INRIA [3]. This programming language has support for representing dependent types [7]. These types are precisely what enables us to write mathematical proofs [8].

Haskell has no support for dependent types, so the first challenge that we face is how to prove some basic property of our programming code (written in Haskell) in Coq. For this, we use an existing tool called `hs-to-coq` [9] that allows us to convert Haskell code to Coq. The reverse conversion is already supported in Coq itself. Using this tool, we get the `Subscription.v` (Coq source code) file.

Semantically, this auto-generated file from `hs-to-coq` contains the same functions that we have defined earlier in Haskell. The only difference is the syntax, where the Coq syntax is used instead of Haskell. At this point, we can start using the power of Coq.

Thus, we create the following `Proofs.v` file that has the following content:

```
Require Import Prelude.
Require Import Subscription.
Require Import Proofs.GHC.Base.
Require Import Data.Semigroup.
```

Now we can finally work on the proof itself. We will prove a few simple properties for this paper's argument. We first define a list of subscriptions to use in our proofs:

```
Definition subs := addSubscription #1 &"fh01"
nil.
```

We then proceed showing that

```
length(getHandlesByCh 1 subs) = 1
```

```
Lemma example_1 : GHC.List.length
  (getHandlesByCh #1 subs) = 1%Z.
```

```
Proof.
  auto.
Qed.
```

We will explain the syntax used in the proof briefly. The interested reader can look at the details in [10].

In the code above we have proved the lemma named `example_1` which states that the length of the evaluation of `getHandlesByCh 1 subs` is 1. We begin the proof using the `Proof.` command. What follows afterward are

commands called tactics. These are macro commands that use the lambda calculus rules to simplify formulas. In this case, by using the `auto` tactic, Coq can mechanically prove the claim, because the claim is simple enough. If the claim was a bit more complex, we would have to use different tactics.

We run Coq and it executes the code mechanically, returning no errors meaning the proof is complete.

```
Lemma example_2 : GHC.List.length
  (getHandlesByCh #2 subs) = 0%Z.
Proof.
  auto.
Qed.
```

The lemma `example_2` is similar to `example_1`. It claims that the subscription of channel 2 does not exist in the list, namely that the length of such a list is 0.

```
Lemma example_3 : getOption (publish (fun y =>
  GHC.Base.return_ 1) subs) = Some (1 ::
  nil).
Proof.
  auto.
Qed.
```

The lemmas `example_3` and `example_4` prove that the second argument of `publish` has an effect on the output of the subscriptions. This is as expected, since this is how we defined `publish` in Haskell earlier.

```
Lemma example_4 : getOption (publish (fun y =>
  GHC.Base.return_ y) subs) = Some (&"fh01"
  :: nil).
Proof.
  auto.
Qed.
```

With the following lemma we will prove the fact:

$$\forall l, hInSub(1, "fh01", (addSub, 1, "fh01", l))$$

That is, for all lists `l`, upon which a subscription `1` with `"fh01"` is added, the function `hInSubscription` returns `true`.

We further take a brief look at the kinds of errors that Coq may return.

```
Lemma example_5 : forall l : list
  (Subscription Base.String),
  hInSubscription #1 &"fh01"
  (addSubscription #1 &"fh01" l) = true.
Proof.
Qed.
```

The example given above results with the following error from Coq:

```
Error: (in proof example_5): Attempt to save
  an incomplete proof
```

It warns us that the proof is not complete. We can use the command `Show Existentials` to see the current state of the proof:

```
Existential 1 =
?Goal : [
  |- forall l : list (Subscription
    String),
    hInSubscription 1%Z &"fh01"
      (addSubscription 1%Z
        &"fh01" l) =
      true]
```

This proof is also simple enough for Coq, so we can use `auto` to complete it as well.

CONCLUSIONS

Moving to the cloud is one of the current challenges in enterprises. This technology provides a new "on-demand" paradigm for information and communication technologies.

The advantages are cheaper systems, access from countless devices, centralization of data. The main disadvantage is that an Internet connection is required.

PubSub is just one of the many SaaS possibilities. As the usage of such systems continues to rise, formally proving correctness according to specifications is crucial to the system functioning and accuracy as expected. The solution presented in this paper demonstrates how Haskell in conjunction with Coq can be applied to perform this vital step in the process of cloud-based software development.

REFERENCES

- [1] G. Reese, *Cloud Application Architectures*, O'Reilly, 2009.
- [2] B. O'Sullivan, D. Stewart, and J. Goerzen, *Real World Haskell*, O'Reilly, 2008.
- [3] INRIA, "The Coq Proof Assistant". [Online]. Available: <https://coq.inria.fr/> (Accessed Jan. 2020)
- [4] A. Church, "An Unsolvability Problem of Elementary Number Theory," *American Journal of Math.*, 1936.
- [5] G. Shroff, *Enterprise Cloud Computing*, Cambridge University Press, 2010.
- [6] Redis, "Pub/Sub". [Online]. Available: <https://redis.io/topics/pubsub> (Accessed Dec. 2019)
- [7] A. Chlipala, "An Introduction to Programming and Proving with Dependent Types in Coq," *Journal of Formalized Reasoning Vol. 3, No. 2*, 2010.
- [8] B. Sitnikovski, *Gentle Introduction to Dependent Types with Idris*, Leanpub/Amazon KDP, 2018.
- [9] A. Spector-Zabusky, J. Breitner, C. Rizkallah, and S. Weirich, "Total Haskell is Reasonable Coq," *Proceedings of the 7th ACM SIGPLAN International Conference on Certified Programs and Proofs*, 2018.
- [10] INRIA, "The Coq Reference Manual". [Online]. Available: <https://coq.inria.fr/> (Accessed Jan. 2020)

ISBN 978-608-4699-10-1



9 786084 699101