Technical University of Košice

Faculty of Electrical Engineering and Informatics

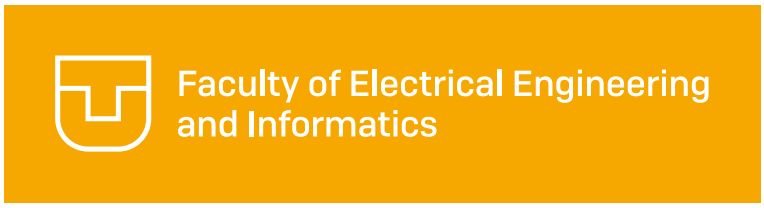
SCYR

22nd Scientific Conference of Young Researchers
Proceedings from Conference

ISBN 978-80-553-4061-6

2022

Sponsors


Faculty of Electrical Engineering
and Informatics


SES
člen
ZSVTS


SIEMENS
Healthineers

SCYR 2022: 22nd Scientific Conference of Young Researchers
Proceedings from Conference

Published: Faculty of Electrical Engineering and Informatics
Technical University of Košice
Edition I, 240 pages, number of CD Proceedings: 50 pieces

Editors: Prof. Ing. Alena Pietriková, CSc.
Assoc. Prof. Ing. Emília Pietriková, PhD.

Acknowledgement:
This conference and proceedings was supported by Grant KEGA no. 017TUKE-4/2020 "Implementation of Advanced Methods of Scientific Work in the Context of Rebuilding Engineering and Doctoral Studies in the Study of Smart Electronics".

ISBN 978-80-553-4061-6

Scientific Committee of SCYR 2022

General chair: Prof. Ing. Liberios Vokorokos, PhD.

Editorial board chairman: Prof. Ing. Alena Pietriková, CSc.

Committee Members & Reviewers:
Prof. Ing. Roman Cimbala, PhD.
Prof. Ing. Kristína Machová, PhD.
Prof. Ing. Ján Paralič, PhD.
Prof. Ing. Daniela Perduková, PhD.
Prof. Ing. Alena Pietriková, CSc.
Prof. Ing. Jaroslav Porubän, PhD.
Prof. RNDr. Jana Tóthová, PhD.
Assoc. Prof. Ing. František Babič, PhD.
Assoc. Prof. Ing. Jaroslav Džmura, PhD.
Assoc. Prof. Ing. Ján Genči, PhD.
Assoc. Prof. Ing. Ján Papaj, PhD.
Ing. Ján Majoroš, Siemens Healthineers Košice

Organizing Committee of SCYR 2022

Members: Prof. Ing. Alena Pietriková, CSc.
Ing. Ivana Olšiaková
Assoc. Prof. Ing. Emília Pietriková, PhD.

Contact address: Faculty of Electrical Engineering and Informatics
Technical University of Košice
Letná 9
040 01 Košice
Slovak Republic

Foreword

Dear Colleagues,

SCYR (Scientific Conference of Young Researchers) is a scientific event focused on exchange of information among young researchers from Faculty of Electrical Engineering and Informatics at the Technical University of Košice – series of annual events that was founded in 2000. Since 2000, the conference has been hosted by FEEI TUKE with rising technical level and unique multicultural atmosphere. The 22nd Scientific Conference of Young Researchers (SCYR 2022) was held on April 8, 2022. Due to COVID-19 pandemics, the conference was held online. The mission of the conference, to provide a forum for dissemination of information and scientific results relating to research and development activities at the Faculty of Electrical Engineering and Informatics, has been achieved. Approx. 70 participants, mostly by doctoral categories, were active in the conference.

Faculty of Electrical Engineering and Informatics has a long tradition of students participating in skilled labor where they have to apply their theoretical knowledge. SCYR is an opportunity for doctoral and graduating students to train their scientific knowledge exchange. Nevertheless, the original goal is still to represent a forum for the exchange of information between young scientists from academic communities on topics related to their experimental and theoretical works in the very wide spread field of a wide spectrum of scientific disciplines like informatics sciences and computer networks, cybernetics and intelligent systems, electrical and electric power engineering and electronics.

Traditionally, contributions can be divided in 2 categories:

- Electrical & Electronics Engineering
- Computer Science

with approx. 70 technical papers dealing with research results obtained mainly in the University environment. This day was filled with a lot of interesting scientific discussions among the junior researchers and graduate students, and the representatives of the Faculty of Electrical Engineering and Informatics. This Scientific Network included various research problems and education, communication between young scientists and students, between students and professors. Conference was also a platform for student exchange and a potential starting point for scientific cooperation. The results presented in papers demonstrated that the investigations being conducted by young scientists are making a valuable contribution to the fulfillment of the tasks set for science and technology at the Faculty of Electrical Engineering and Informatics at the Technical University of Košice.

We want to thank all participants for contributing to these proceedings with their high quality manuscripts. We hope that conference constitutes a platform for a continual dialogue among young scientists.

It is our pleasure and honor to express our gratitude to our sponsors and to all friends, colleagues and committee members who contributed with their ideas, discussions, and sedulous hard work to the success of this event. We also want to thank our session chairs for their cooperation and dedication throughout the entire conference.

Finally, we want to thank all the attendees of the conference for fruitful discussions and a pleasant stay in our event.

Liberios VOKOROKOS
Dean of FEEI TUKE

April 8, 2022, Košice

Contents

Evaluation of 64x100 Gb/s DWDM system with DQPSK FRFT modulation

[1]*Norbert ZDRAVECKÝ (2nd year)*
Supervisor: [2]Ľuboš OVSENÍK

[1,2]Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

[1]norbert.zdravecky@tuke.sk, [2]lubos.ovsenik@tuke.sk

***Abstract*— The main goal of this article is evaluation of our DWDM system, which was designed in the OptiSystem program. 64-channel DWDM system with different optical modulations and center frequency of 193.1 THz. The scheme also includes a nonlinear phenomena affecting the transmission. The theory of DPSK and DQPSK modulation is described. We will evaluate BER and Q factor parameters for 100 Gb/s systems with DQPSK FRFT (Fractional Fourier Transform) modulation.**

***Keywords*— BER, DPSK, DQPSK, DWDM, FRFT, Q-factor.**

I. Introduction

Appropriate modulation is needed to adapt the signal so that it can be transmitted over the optical fiber. Achieving high transmission speed and large information capacity of transmission systems is associated with the use of appropriate modulation formats. Each modulation is performed differently and is suitable for certain types of transmissions [1][2]. For this reason, it is necessary to describe of advanced modulation formats that are part of every process of optimizing modern high-speed optical transmission networks. The basic goal of modulation is to adapt the information signal to a form that can be transmitted in the information channel. The modulation has the task of reducing the chromatic dispersion, which causes a different rate of the individual frequency components contained in the optical pulse. The purpose of modulation is to prevent the formation of problematic clusters of logical zeros or logical ones that cause problems with clock detection. The main comparison between WDM (Wavelength Division Multiplex) and DWDM (Dense Wavelength Division Multiplex) optical system is that DWDM has a greater overall capacity, so DWDM spaces the wavelength more closely than WDM. In WDM channel spacing reduces to 1.6 nm (200 GHz) while in DWDM channel spacing is less than 200 GHz.

II. Comparison of DWDM Research

The development of DWDM systems is still advancing at high speed. Modulations in optical systems are being studied by scientists around the world. Gill and M. Singh evaluated 64x20 Gb/s DPSK system with 210 km optical fiber. Conclusion from this work is that maximum Q factor is 15.1 dB and minimal BER is 1.04e-09 [5]. S. Vats and V. Kakar experimented with 64x40 Gb/s DWDM with RZ (Return to Zero) modulation. The result from this work is that maximum Q factor and minimal BER (Bit Error Rate) is at power 2 dBm [6].

III. Modulation Methods

DPSK (Differential Phase-Shift Keying) is a modulation that records changes in binary flow. Based on these changes, the properties of the transmitted signal have changed. The demodulator records changes in the phase of the incoming signal. DPSK differs from PSK (Phase-Shift Keying) according to the following rules. In the case of a logical one, the phase remains unchanged. At logic zero, the phase changes by π [9]. In differential phase modulation, the characteristic phase values do not express the magnitude of the carrier phase, but the magnitude of the phase change compared to the previous interval. DPSK modulation is an ideal choice when you create a DWDM optical network [10][11].

QPSK (Quadrature Phase-Shift Keying) is a multistate system that uses four signal elements, expressed by a carrier with a different initial phase. Each signal element is represented by a signal bit pair [3][7]. QPSK modulation is created from two-state binary phase keying. It is created using two carrier waves that have the same frequency. In this case, the frequency is phase shifted by $\pi/2$ [4]. DQPSK (Differential Quadrature Phase Shift Keying) is a differential quadrature phase keying that uses four signal elements that are expressed by a carrier with a different initial phase. Each signal element contains one bit pair. The bit combinations are 00, 01, 10 and 11.

IV. Design of 64 channel DQPSK FRFT

Transmitter frequency is 193.1 THz. Spacing frequency is 200 GHz. Power of laser is 0 dBm. EDFA (Erbium Doped Fiber Amplifier) amplify gain is 10 dB. Optical fiber is 60 km long. BER analyzer in system examines BER function. The Q factor is a function of the OSNR (Optical Signal to Noise Ratio) and represents the tolerance of the system in dB. BER and Q factor are close parameters.

Table I shows first channel result of our system. On Fig. 1 is shown scheme of a 64x100 Gb/s system that was simulated in Optisystem. On Fig. 2 is shown DQPSK FRFT (Fractional Fourier Transform) subsystem simulation scheme in Optisystem.

TABLE I RESULTS

Parameters	Channel 1
Q factor	11.1 dB
Minimal BER	4.305e-29
Eye height	1.52549e-5

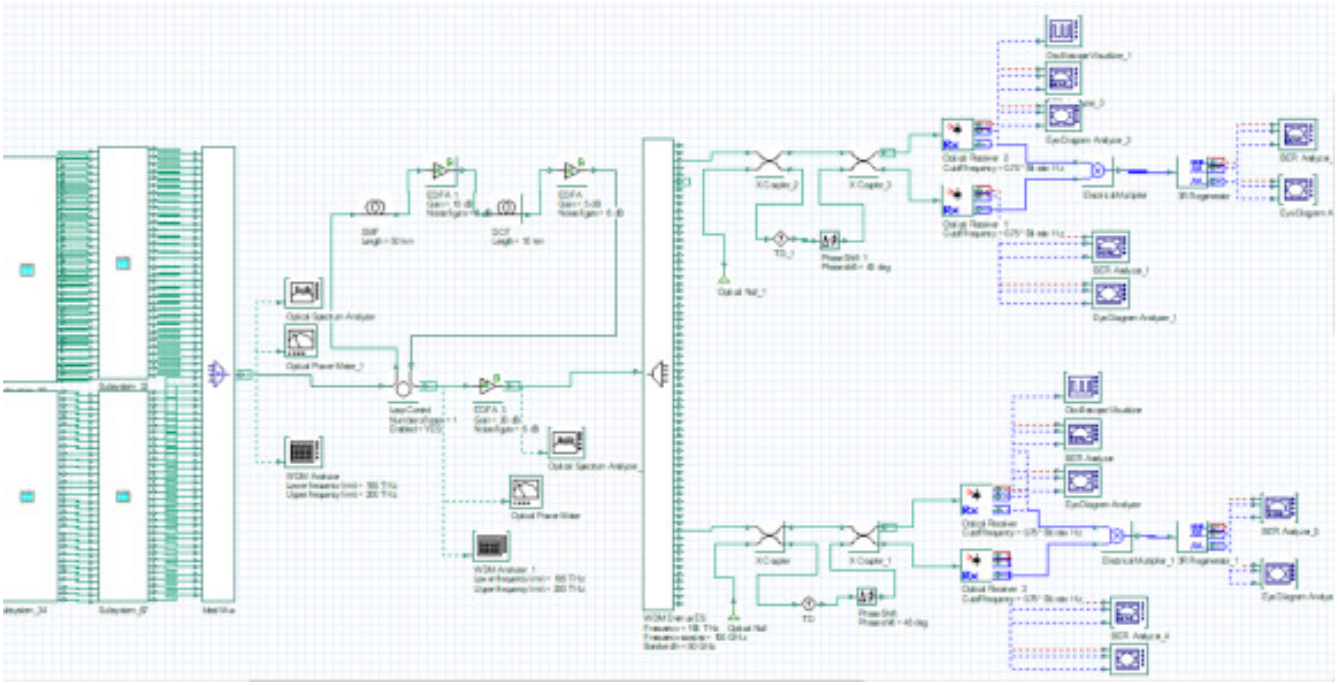


Fig. 1 General diagram of the 64-channel system from the OptiSystem program

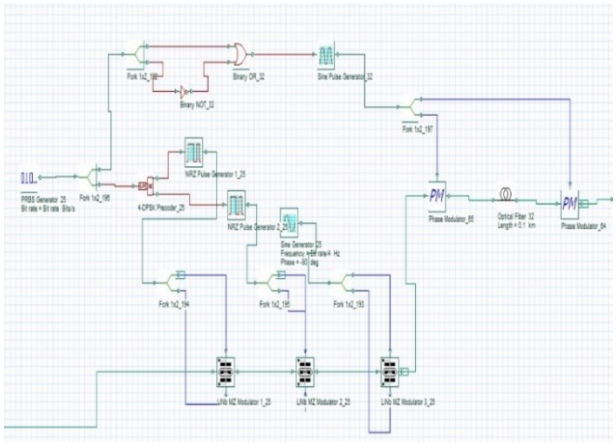


Fig. 2 DQPSK FRFT modulation scheme in Optisystem

V. Future Work

Our future work is defined by dissertation thesis:

1) Design of high-capacity Ultra-DWDM systems for the investigation of nonlinear effects (SPM, XPM and FWM) for transmission speed above 40 Gb/s and 100 Gb/s in fully optical transmission networks.
2) Design of implementation of modified advanced optical modulation formats to reduce the impact of nonlinear phenomena for Ultra-DWDM optical transmission systems.
3) Optimization of optical link with the optical amplifier (EDFA) transmission paths in high-capacity Ultra-DWDM optical transmission systems.

Acknowledgment

This work was supported by the research project FEI-2022-84 "Data Processing Techniques in High Speed Transmission Systems".

References

[1] AGRAWAL, Govind P. Optical Fiber Communications: Optics and Photonics. 6th Edition. Cambridge: Academic Press, 2019. ISBN 9780128170427

[2] WINZER, Peter J. Energy-Efficient Optical Transport Capacity Scaling Through Spatial Multiplexing. IEEE Photonics Technology Letters. 2011, 23(13), 851 - 853. ISSN 1041-1135. Available on.: doi:10.1109/LPT.2011.2140103

[3] HUSZANÍK, Tomáš, Ján TURÁN and Ľuboš OVSENÍK. On the Impact of Fiber Nonlinear Effects on the CP-DQPSK Modulated Ultra-DWDM System. Acta Electrotechnica et Informatica. 2019, 19(4), 21-28. ISSN 1335-8243. Available on: doi:10.15546/aeei-2019-0026

[4] CHANDRA, Aniruddha and Chayanika BOSE. Series solutions for pi/4-DQPSK BER with MRC. International Journal of Electronics. 2011, 99(3), 391-416. Available on: doi:10.1080/00207217

[5] GILL, M. S., S. DEWRA (2017). Analysis of DWDM System Using DPSK Modulation Technique with Raman–EDFA Hybrid Optical Amplifier. Journal of Optical Communications, 38(4). doi:10.1515/joc-2016-0056

[6] VATS, Saurabh and Vidur KAKAR. Performance Analysis of 64-Channel DWDM System Using EDFA. International Journal of Innovative Research in Computer Science & Technology. 2015, 3(2), 79-82. ISSN 2347-5552.

[7] HUSZANÍK, Tomáš, Ján TURÁN and Ľuboš OVSENÍK, "Evaluation of CP-DQPSK modulated DWDM system with highly nonlinear fiber in C band," Radioelektronika 2019 : 29th International conference. – Pardubice, p. 304-307. ISBN 978-1-5386-9321-6..

[8] HUSZANÍK, Tomáš, Ján TURÁN and Ľuboš OVSENÍK, "Performance Analysis of Optical Modulation Formats for 10 Gbit/s DWDM System," Carpathian Journal of Electronic and Computer Engineering, vol. 10, no. 2, 2017, pp. 3-8. ISSN 1844 – 9689.

[9] KAUR A., S. DEWRA. "Comparative Analysis of Different Modulation Techniques in Coherent Optical Communication System," International Journal of Innovative Research in Computer and Communication Engineering, vol. 3, issue no 8, p. 7193 – 7200, 2015. ISSN: 2320- 9801.

[10] HAN, Sungmin, Jaeseok LEE, Taesoon KWON and Ji-Woong CHOI. Performance analysis on DPSK modulation using symbol repetition and interleaving. Communication systems. 2018, 31(11), 1-16. Available on: doi:10.1002/dac.3589

[11] HUSZANÍK, Tomáš, Ján TURÁN and Ľuboš OVSENÍK, "Experimental Simulation of Coherent 40 GBPS 2-DPSK DWDM Long-haul Fiber Optical System with Counter-directional EDFA," Acta Technica Napocensis, vol. 59, no. 2, p. 5-8, 2018. ISSN 1221-6.

Design of a secondary RF channel for a hybrid FSO/RF system

[1]*Maroš LAPČÁK (2nd year)*
Supervisor: [2]Ľuboš OVSENÍK

[1,2]Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

[1]maros.lapcak@tuke.sk, [2]lubos.ovsenik@tuke.sk

***Abstract*— This publication deals with the design of a backup RF line for a hybrid FSO/RF system. This system consists of a primary FSO line and a secondary RF line. Both communication lines have different environmental influences on their functionality. In the case of FSO, it is mainly fog and in the case of the RF line, it is rain and mutual interference from users. Therefore, it is necessary to properly design and configure a backup RF line to operate in the unlicensed band and to meet all the conditions for proper functionality.**

***Keywords*—FSO, horn antenna, hybrid FSO/RF, RF link.**

I. Introduction

Optical networks can achieve very high transmission speeds. Another advantage is the provision of secure and broadband symmetric transmission. When comparing optical networks with other networks, we can say that the optical network is not dependent on limited and regulated spectrum and frequency licensing, such as radio networks. However, the condition for transmission is direct visibility between the transmitter and receiver. For optical networks, a laser or LED can be used as a transmitter and an optical detector can be used as a receiver. Therefore, optical systems operate in full duplex mode. A disadvantage of FSO (Free Space Optics) networks is the dependence on the effects of the atmospheric transmission environment (smog, fog (especially), snow, etc.) [1]. As mentioned above, FSO junctions mainly suffer from fog attenuation. The RF connection for FSO/RF hybrid systems should therefore meet the following requirements:

- availability during the period of loss of the FSO connection, i.e., during the fog event,
- bandwidth comparable to an FSO link to provide similar performance,
- operation in the frequency band without a license.

As it turns out, these requirements are contradictory. High carrier frequencies must be used to achieve high bandwidths. These frequencies, on the other hand, are strongly affected by precipitation and fog, although the attenuation caused by fog is much lower than in the case of FSO connections. On the other hand, lower frequencies are not affected by fog at all, because unlicensed systems in the ISM bands (2.4 or 5 GHz), are used for IEEE 802.11a/b/g compatible wireless LAN devices, are only affected by rain or multi-user interference [2]. Hybrid FSO/RF link is considered as alternative and cost competitive solution for high-speed, point-to-point wireless communication. They combine the advantages of both links. FSO links provide gigabit data rates and low system complexity, but there are data losses mainly due to fog and scintillation. RF links have lower data rates compared to the FSO links, but their main advantage is their independence from the weather.

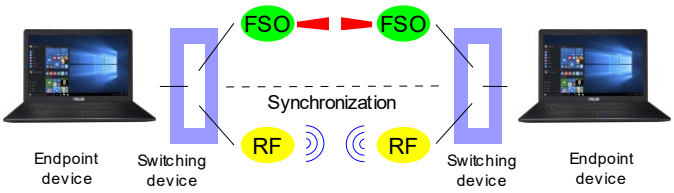


Figure 1 The principle of operation of the Hybrid FSO/RF system.

II. Theory

The horn antennas belong to the group of aperture antennas and are considered a characteristic example of this group. These antennas consist of two main parts that are waveguide and horn. The shape of the waveguide depends on the shape of a horn. The horn can have either a cone or shape of the needle. This type of antenna also allows various power options or using a waveguide or a coaxial driver if the antenna is adapted to it. The positive properties of these antennas include a large bandwidth as suitably adapted horn antennas can have bandwidth of up to 1 GHz. Another advantage is the high directionality of these antennas, but also a high level of efficiency. This means that the horn antennas achieve high gain and due to a small radiation angle, this type of antenna is ranked among directional antennas. Thanks to these features, the horn antennas are used in multiple areas. For example, they are used as standards for calibration, as well as resources for large parabolic reflector antennas or such short-range radar systems [3] – [9].

III. Design of a horn antenna with a working frequency of 2.4 GHz

Table 1 shows the dimensions of the horn antenna with a 2.4 GHz working frequency. The power supply is realized by using an SMA (SubMiniature version A) connector for connecting the coaxial cable. A monopole antenna is used as an antenna emitter. The horn skeleton serves as a ground surface and for subsequent radiation antenna. The antenna skeleton is formed of 1mm thick aluminum. The emitter consists of a copper cable with 1mm radius. In Figure 2 there are several openings and ribs in the horn of this antenna, which further improve the radiation of the antenna but make the manufacturing process more demanding. These holes and

ribs are different depending on the side on which they are located.

Table 1 Horn antenna parameters with 2.4 GHz working frequency.

Parameters	Description
Antenna type	Horn antenna
Work frequency	2.4 GHz
Dimensions of waveguide	10.5 x 4.9 x 4.9 cm (width x height x length)
Dimensions of horn	48.33 x 50.4 x 40.4 cm (width x height x length)
Used wire	Aluminum, copper
Power type	SMA Connector for Coaxial Driver Connection

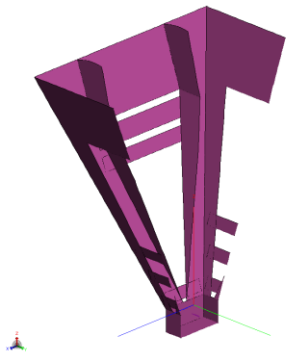


Figure 2 Section of a horn antenna with a working frequency of 2.4 GHz in 3D display in the FEKO package.

From Figure 2 we can see that the ribs are only found on the sides of the horn that are parallel to the X axis. On these sides of the horn parallel to the X axis we can see four holes that are trapezoidal shape. The various sides of these holes have the same angle as the horn itself, i.e., their height is enough to describe these holes. The height of the holes on this side is 1.5, 27, 1.5 and 0.6 cm. The heights of the openings on the sides of parallel to the axis of Y are 3, 3.5, 1.5 and 22.6 cm. Single holes are sorted from below.

Based on Figure 3, where the VSWR (Voltage Standing Wave Ratio) values are shown, we can see that this horn antenna is suitably impedance matched for the 2.4 GHz frequency and at the same time for the whole unlicensed frequency band available in Slovakia. The lowest VSWR value is 1.08 at 2.43 GHz. This minimum value is located almost in the middle of the unlicensed frequency band we want to use for this antenna. Based on the appearance of the curve in the chart, it is possible to see a wide range of this horn antenna, as VSWR is located below 2 throughout the frequency range from the frequency of 2.151 GHz to a frequency of 2.667 GHz.

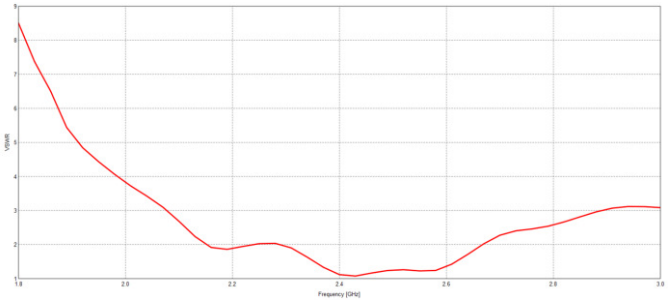


Figure 3 VSWR values of a horn antenna with a working frequency of 2.4 GHz in the FEKO package shown in a linear graph.

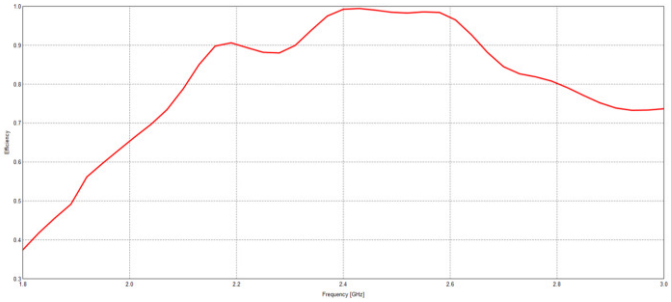


Figure 4 Efficiency of a horn antenna with a working frequency of 2.4 GHz in the FEKO package.

The results of this antenna in terms of impedance matching shown in Figure 3 indicated that the antenna could achieve high efficiency due to the excellent impedance matching of this antenna. As can be seen from Figure 4, the antenna achieves a high level of efficiency. The maximum efficiency level is 99.4% at a frequency of 2.43 GHz. As mentioned above, the antenna can be considered broadband, as the efficiency level shown in Figure 4 is at 88% in the same frequency range indicated in this section.

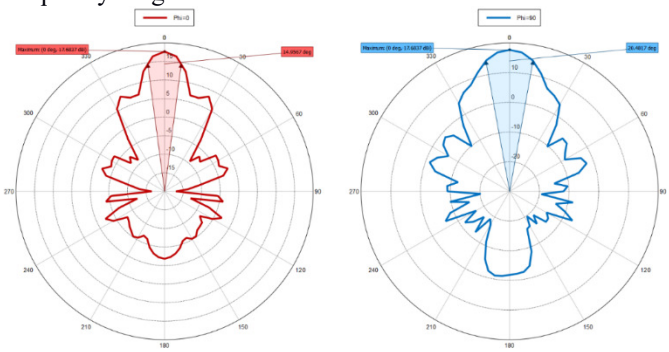


Figure 5 Radiation characteristics of a horn antenna with a working frequency of 2.4 GHz in the FEKO package in a polar graph at sections with angles $\phi = 0°$ and $\phi = 90°$.

On the radiating characteristics shown in Figure 5 at angles $\varphi = 0°$ and $\varphi = 90°$, we can see the excellent directions of this antenna. The efficiency of this antenna at the working frequency achieves a level of more than 99%, so it is not necessary to indicate the value of the direction, as the maximum gain of this antenna is almost the same as the level of direction. In the case of angle $\varphi = 0°$, which is marked with red, the antenna achieves the maximum yield of 17.6837 dBi and the radiation angle is at 14.95°. The second part of Figure 5, which is displayed by a blue color, shows the radiation characteristics when the angle $\varphi = 90°$, wherein the antenna reaches a maximum yield of 17.6837 dBi and the radiation angle at 20.48°. The width of the radiation lobe is determined between two points in which the gain is 3 dBi less than in the case of maximum gain. The maximum gain level is

the same at both cuts by radiating characteristics. This means that there is no movement of the maximum radiation antenna. In Figure 5 we can also see the incurred lateral lobes, which do not have a big impact when radiating, as gain in these places is more than 10 dBi lower than maximum gain.

IV. Conclusion

The result is the design and analysis of a horn antenna with a working frequency of 2.4 GHz. As can be seen from the simulations, this antenna provides excellent parameters such as impedance matching and efficiency and can be used directly as a backup RF line. Its advantage is high efficiency, high gain, and small beam angle. All these benefits are required for a hybrid FSO/RF system. One of the disadvantages of this antenna is its size, which greatly complicates its real use for our system.

Acknowledgment

This work was supported by the research project FEI-2022-84 "Data Processing Techniques in High-Speed Transmission Systems".

References

[1] R. Pužmanová, “Bezdrátová optika a její možnosti,” 2006, available at: https://www.dsl.cz/clanky/533-bezdratova-optika-a-jeji-moznosti.

[2] O. Ugweje, Radio Frequency and Wireless Communications. *Internet Encyclopedia*, 2004. 72 p. DOI: 10.1002/047148296X.tie151.

[3] A. Z. Elsherbeni, C. J. Reddy, P. Nayeri, “Antenna Analysis and Design using FEKO Electromagnetic Simulation Software,” India: SciTech Publishing, 2014, 245 p., ISBN: 978-1-61353-205-8.

[4] J. R. Hallas, “Basic Antennas: Understanding Practical Antennas and Design,” Connecticut: Amer Radio Relay League, 2009, 183 p. ISBN: 9780872599994.

[5] J. Carr and G. Hippisley, “Practical Antenna Handbook,” New York: McGraw-Hill/TAB Electronics, 2011, 784 p., ISBN: 9780071639583.

[6] P. Bevelacqua, “List of Antennas,” 2016, Available at: https://www.antenna-theory.com/antennas/main.php

[7] Y. Lo, S. Lee, “Antenna Handbook,” Massachusetts: Morgan Kaufmann Publishers, Van Nostrand Reinhold, 1993, 830 p., ISBN: 0442015933.

[8] W. L. Stutzman and G. A. Thiele, “Antenna Theory and Design,” New Jersey: Wiley, 2012, 843 p., ISBN: 0470576642.

[9] T. S. Bird, “Fundamentals of aperture antennas and arrays: from theory to design, fabrication and testing,” New Jersey: Wiley, 2016, 448 p., ISBN: 9781119127451.

Decision Models Interpretability for the Domain Experts

Viera ANDERKOVÁ (2nd year)
Supervisor: František BABIČ

Department of Cybernetics and Artificial Intelligence, FEI TU of Košice, Slovak Republic

viera.anderkova@tuke.sk, frantisek.babic@tuke.sk

***Abstract*— This paper presents an overview of open problems in the domain of interpretability of machine learning models and possible solutions in several application areas. It points out that interpretable models are important and necessary for the end-user to understand the model more quickly and then make the right decision. The paper also describes our experiments and results-focused on traffic accident analysis, data-based learning analytics, and medical diagnostics support. Based on them in combination with obtained knowledge, we have defined the goal of the dissertation thesis and possible ways to achieve this goal.**

***Keywords*—machine learning, interpretability, decision-making**

I. INTRODUCTION

According to Alpaydin [1], machine learning (ML) and data acquisition help create tools and solutions supporting making decisions and even acting on our behaviour. These solutions understand our habits and interests and navigate us to the products or services that are most useful to us and our decisions. ML can help generate predictive models, create early prediction models, detect hidden patterns and relationships, and draw conclusions in different domains. In the field of law, it is possible to automatically filter unsolicited e-mail (spam) of e-mail communication using ML methods [2].

Another exciting domain where ML algorithms are used is traffic and traffic accidents. Dogru et al. use random forest (RF), support vector machine (SVM), and artificial neural networks (ANN) for detecting accidents on a highway using microscopic vehicle information, where vehicles exchange information to improve mobility and safety [3]. Cheng et al. tried to design an RF prediction model to predict the number of traffic accidents [4]. Haynes et al. focused on the factors affecting the number of road accident-related deaths [5]. They used ML algorithms such as Decision Tree (DT), RF, k - nearest neighbour (k-NN), and Gaussian. Bahiru et al. examined various classification algorithms (ID3, J48, CART, Naïve Bayes) to understand the most serious accident factors and predict road accidents' severity [6].

In the field of education and learning activities is necessary to examine the individual processes, such as collaborative education for students and e-learning, whether the level of knowledge, goal, and student's learning style, motivation, and activity play a central role or not [7]. These problems can be solved with Learning Analytics (LA). Gouripeddi et al. examined supporting vector regression and graphical representation in the freely available dataset and provided visualization of attributes and some methods of ML in creating predictive models to raise learning approaches [8]. Hussain et al. focused on predicting student involvement during the virtual learning environment course and compared research performance using DT, Gradient-boosted trees, CART, and Naïve Bayes classifier [9]. Lu et al. applied big data approaches to education and LA to early prediction of students' final academic performance in a blended Calculus course [10].

ML can be used for various medical or healthcare applications. Raturi and Kumar used the DT model of the leading causes of diabetes [11]. Ghiasi et al. focused on examining coronary artery disease, one of the most cardiovascular disease (CAD) [12]. They developed several CART models with different input parameters that were not selected randomly. Kaur and Bawa developed an ES that could predict whether or not a patient was drug-addicted or prone to drugs [13]. They created the ES in WEKA and used the ID3, DT algorithm. Pourhomayoun and Shakibi looked at predicting mortality risk in patients with COVID-19 using ML to help make medical decisions [14].

While analysing existing solutions in various areas, we found that using ML to create the decision-making models and the form of its interpretation in each domain is necessary. In our research, we special focused on the medical domain and healthcare because it is one of the most critical domains that can concern human life because wrong decisions can have fatal consequences.

A. Importance of interpretable models

The correct interpretation can improve models' readability, understandability, and deployment in practice. In recent years, several studies have focused on understanding the behaviour of analytical models, thus ensuring trust in the predictions of individual models [15]–[17]. End-users can make decisions based on these reliable predictions. Luna et. al aimed to create an end-user design and to scientifically compare the usability of standard and new interfaces in terms of efficiency (measured as clicks and time to complete a task), effectiveness (measured by rewriting and task completion rates), and user satisfaction (evaluated using the system usability scale (SUS)) [17].

Subsequently, Biran and Cotton found that the key role of an intelligent or DSS is to explain the decisions, recommendations, or predictions that the models have made [18]. They described that the explanation is also related to interpretability. Thus "systems are interpretable if their operations can be understood by a human, either through a produced explanation." They also point out that explanation is often difficult, as most models are challenging to interpret. Murdoch et al. define Interpretable ML as the extraction of

relevant knowledge from the ML model regarding the relationships either contained in the data or learned by the model [19]. Rudin points out the apparent differences between interpretable and explainable ML [20]. Explainable ML usually focuses on deep learning and NN, presenting what a node represents and its importance to performance models. In contrast, interpretability is the ability to determine the cause and effect of ML model by domain experts in practice.

Based on acquired knowledge, we have found the following open problems:

- interpretable ML models,
- evaluation of interpretable models by end-users (domain experts),
- evaluation of the quality of individual explanations.

II. Our preliminary results

During the 2nd year of doctoral studies, we published several articles in which we focused on the identified open problems.

A. Data-Based Road Accidents Analysis

This paper describes the experiments based on 1,2 mil. dataset about traffic accidents in England, Scotland, and Wales [21]. This data can provide important knowledge to support relevant decision-makers or processes. We focused not on the prediction models and their accuracy but on their explanations for the end-users with limited knowledge from data mining, machine learning, or artificial intelligence. For this purpose, we have improved the generated decision models with selected interpretable methods like The Local Interpretable Model-Agnostic Explanation (LIME) and SHap Additive exPlanations (SHAP) values. The final interpretations show which features and to what extent they involve in each type of accident.

We learned that interpretable methods for ML decision models bring more straightforward and understandable visualizations.

B. Learning Analytics

This paper presents a comprehensive data analysis and early prediction inspired by publicly available data on courses, students, and learning outcomes. We focused on generating predictive models within typical ML algorithms, comparing these algorithms based on accuracy, return, f1-score, and ROC. All generated models achieve relatively high metrics values, but we consider the GBM model the best for a data sample of 80% because it reached the smallest number of misclassified examples for target class 1. The results are promising based on related work in this domain, e.g., around 5% higher accuracy. The paper will be published in March by reason of postponement of the conference SAMI 2022.

Performed experiments confirmed our hypothesis that understandable models and extracted behaviour patterns are very important for the teachers and continuous improvement of the educational process.

C. Data-Based Medical Diagnostics

This paper aimed to identify the clusters of physical frailty and cognitive impairment diseases in the population with primary care over 60 years of age [22]. We worked with a sample of 263 patients and performed several experiments using supervised latent class analysis (LCA); the differences were assessed by using a multinomial logistic regression models. A comorbidity pattern that may distinguish the clusters depends on the degree of development of cardiometabolic disorders in combination with advancing age.

This collaboration in the international team of authors brought us experience with communication and interaction with a medical expert. It was necessary to generate many results, which did not always lead to the achievement of the goal.

This fact motivated us to consider the possibility of how to make this phase in the analytical process more effective for the domain experts. Although we have primarily focused on cluster analysis and regression in this paper, our research group have experience with other ML algorithms supporting medical diagnostics [23]–[28].

III. Conclusion

Finally, we learned working with huge data samples, applied several ML algorithms with promising results, and investigated the area of interpretable ML models. Based on the results and obtained experiences, we have specified the hypotheses of my dissertation. Therefore, the main aim is a semi-automatic evaluation of the decision models generated by the ML methods from the point of their interpretability. This goal will be achieved through the: analysis of existing approaches supporting the interpretability of decision models generated by the selected ML methods (done); design of a decision models quality evaluation system containing selected metrics to reduce the time and effort required for domain expert evaluation; experimental verification and evaluation of the proposed system in the process of medical diagnostics.

In future research, we will be focused on interpretable ML models for end-users in the medical domain and how to evaluate the interpretability and quality of these models based on suitable metrics.

Acknowledgment

The work was supported by The Slovak Research and Development Agency under grant no. APVV-17-0550.

References

[1] E. Alpaydin, *Machine learning: The new {AI}*. London, England: MIT Press, 2016.

[2] H. Surden, “Machine learning and law,” *Washingt. Law Rev.*, vol. 89, no. 1, pp. 87–115, 2014, [Online]. Available: https://digitalcommons.law.uw.edu/cgi/viewcontent.cgi?article=4799&context=wlr#page=5&zoom=100,0,798.

[3] N. Dogru and A. Subasi, “Traffic accident detection using random forest classifier,” in *2018 15th Learning and Technology Conference (L T)*, 2018, pp. 40–45, doi: 10.1109/LT.2018.8368509.

[4] R. CHENG, M.-M. ZHANG, and X.-M. YU, “Prediction Model for Road Traffic Accident Based on Random Forest,” *DEStech Trans. Soc. Sci. Educ. Hum. Sci.*, Feb. 2019, doi: 10.12783/dtssehs/icesd2019/28223.

[5] S. Haynes, P. Estin, S. Lazarevski, M. Soosay, and ah-lian Kor, “Data Analytics: Factors of Traffic Accidents in the UK,” 2019, pp. 120–126, doi: 10.1109/DESSERT.2019.8770021.

[6] T. K. Bahiru, D. Kumar Singh, and E. A. Tessfaw, “Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity,” in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2018, pp. 1655–1660, doi: 10.1109/ICICCT.2018.8473265.

[7] M. Goyal, D. Yadav, and A. Choubey, “E-learning: Current State of Art and Future Prospects,” no. c, 2012, [Online]. Available: https://core.ac.uk/download/pdf/25833926.pdf.

[8] P. S. Gouripeddi, R. Gouripeddi, and S. P. Gouripeddi, “Toward Machine Learning and Big Data Approaches for Learning Analytics,” in *2019 IEEE Tenth International Conference on Technology for Education (T4E)*, 2019, pp. 256–257, doi: 10.1109/T4E.2019.000-6.

[9] M. Hussain, W. Zhu, W. Zhang, and S. M. R. Abidi, "Student engagement predictions in an e-learning system and their impact on student course assessment scores," *Comput. Intell. Neurosci.*, vol. 2018, 2018.

[10] O. H. T. Lu, A. Y. Q. Huang, J. C. H. Huang, A. J. Q. Lin, H. Ogata, and S. J. H. Yang, "Applying Learning Analytics for the Early Prediction of Students' Academic Performance in Blended Learning," *J. Educ. Technol. Soc.*, vol. 21, no. 2, pp. 220–232, 2018, [Online]. Available: http://www.jstor.org/stable/26388400.

[11] R. Raturi and A. Kumar, "An Analytical Approach for Health Data Analysis and finding the Correlations of attributes using Decision Tree and W-Logistic Modal Process," 2019.

[12] M. M. Ghiasi, S. Zendehboudi, and A. A. Mohsenipour, "Decision tree-based diagnosis of coronary artery disease: CART model," *Comput. Methods Programs Biomed.*, vol. 192, p. 105400, 2020, doi: https://doi.org/10.1016/j.cmpb.2020.105400.

[13] S. Kaur and R. K. Bawa, "Implementation of an expert system for the identification of drug addiction using decision tree ID3 algorithm," *Proc. - 2017 3rd Int. Conf. Adv. Comput. Commun. Autom. (Fall), ICACCA 2017*, vol. 2018-Janua, pp. 1–6, 2018, doi: 10.1109/ICACCAF.2017.8344677.

[14] M. Pourhomayoun and M. Shakibi, "Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making," *Smart Heal.*, vol. 20, no. March, 2021, doi: 10.1016/j.smhl.2020.100178.

[15] S. S. Ajibade and A. Adediran, "An overview of big data visualization techniques in data mining," *Int. J. Comput. Sci. Inf. Technol. Res.*, vol. 4, no. 3, pp. 105–113, 2016.

[16] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, 2018, pp. 80–89.

[17] D. R. Luna, D. A. Rizzato Lede, C. M. Otero, M. R. Risk, and F. González Bernaldo de Quirós, "User-centered design improves the usability of drug-drug interaction alerts: Experimental comparison of interfaces," *J. Biomed. Inform.*, vol. 66, pp. 204–213, 2017, doi: https://doi.org/10.1016/j.jbi.2017.01.009.

[18] O. Biran and C. Cotton, "Explanation and Justification in Machine Learning: A Survey," *IJCAI-17 Work. Explain. AI*, pp. 8–13, 2017, [Online]. Available: http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf.

[19] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 44, pp. 22071–22080, Oct. 2019, doi: 10.1073/pnas.1900654116.

[20] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.

[21] V. Anderková and F. Babič, "Better understandability of prediction models: a case study for data-based road safety management system," in *2021 IEEE 21st International Symposium on Computational Intelligence and Informatics (CINTI)*, 2021, pp. 155–160.

[22] S. Bekić, F. Babič, V. Pavlišková, J. Paralič, T. Wittlinger, and L. T. Majnarić, "Clusters of physical frailty and cognitive impairment and their associated comorbidities in older primary care patients," *Healthc.*, vol. 9, no. 7, 2021, doi: 10.3390/healthcare9070891.

[23] F. Babič, L. Majnarić, A. Lukáčová, J. Paralič, and A. Holzinger, "On patient's characteristics extraction for metabolic syndrome diagnosis: predictive modelling based on machine learning," in *International Conference on Information Technology in Bio-and Medical Informatics*, 2014, pp. 118–132.

[24] F. Babič, J. Olejár, Z. Vantová, and J. Paralič, "Predictive and descriptive analysis for heart disease diagnosis," in *2017 federated conference on computer science and information systems (fedcsis)*, 2017, pp. 155–163.

[25] D. Pella *et al.*, "The possible role of machine learning in detection of increased cardiovascular risk patients--KSC MR Study (design)," *Arch. Med. Sci.*, vol. 16, no. 1, 2020.

[26] F. Babič, L. Pusztová, L. T. Majnarić, and others, "Mild Cognitive Impairment Detection Using Association Rules Mining," *Acta Inform. Pragensia*, vol. 9, no. 2, pp. 92–107, 2020.

[27] J. Rokošná, F. Babič, L. T. Majnarić, and others, "Cooperation Between Data Analysts and Medical Experts: A Case Study," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 2020, pp. 173–190.

[28] L. Pusztová, F. Babič, J. Paralič, and Z. Paraličová, "How to Improve the Adaptation Phase of the CBR in the Medical Domain," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 2019, pp. 168–177.

Design of energy source models for a microgrid system

[1]*Róbert Štefko (2nd year)*
Supervisor: [2]Michal Kolcun

[1,2]Dept. of Electric Power Engineering, FEI TU of Košice, Slovak Republic

[1]robert.stefko@tuke.sk, [2]michal.kolcun@tuke.sk

***Abstract*— Microgrid systems are much more complicated, and complex compared to the current system. Although microgrid systems are already in existence their development is still in its beginning stages. To accelerate the research and development of these new microgrid systems, it is necessary to test and refine these systems in complex simulations to achieve the desired goal.**

***Keywords* — microgrid, renewable energy sources, protection system, simulations, design.**

I. Introduction

The main issue for microgrid systems is the appropriate design of power supplies to suit islanded operation under all conditions. This is also the main advantage of the microgrid over the current system, ensuring an uninterrupted power supply to consumers and a significant increase in reliability.

When compiling the energy mix of resources, it is necessary to consider several factors, the main role being played by the daily load diagram and the selection of energy sources in terms of CO_2 minimization, risk minimization, or balancing energy costs [1].

For the full application of microgrid systems in practice, the development of new technologies is still needed. These new technologies will not only ensure more efficient power generation but also faster and more reliable control and protection systems that communicate with each other.

The continuous growth in electricity demand makes it necessary to solve the problem with microgrid systems, even without islanded operation, which opens new possibilities for control and protection systems to ensure sufficient power.

II. Microgrid

A key feature of the microgrid system is the seamless transition from island operation to grid operation while leveraging the island mode auto-control capability. In the event of a fault, local microgrid systems can increase system reliability by switching to islanded mode as local resources and renewable energy sources (RES) continue to power the microgrid system. The benefits of microgrids eventually eliminate the technical challenges of controlling and protecting microgrid systems. A significant challenge in microgrid systems is the design of an appropriate energy mix of local energy sources, RES and proper design can help solve the control and system protection problems of the power system [2].

The problem with applying these structures in practice is several, but the essential problems are with a suitable mix of energy sources for such systems and the transmission infrastructure itself, which is undesigned for such changes. The key will be to determine the size of the area that such a microgrid system will control [3].

III. Design of power sours for microgrid

The following section suggests some prospective energy sources that should be considered when designing the mix of sources for a microgrid system.

The following models use the Simscape Power System library (v7.5, MathWork, CA, USA) in the Simulink (v9.10.0.1649659, MathWork, CA, USA). For the following models, we will consider photovoltaic stations, battery storage systems, thermal power plants, and hydropower plants, which would be the most suitable use for Slovakia [3].

A. Model of Photovoltaic station and battery storage system

Fig. 1 shows perspective places on the territory of the Slovak Republic for the location of photovoltaic stations, where the greatest perspective is in the southern part of Slovakia, while we can get the most electricity from photovoltaic stations in the vicinity of Komarno and Nitra [3].

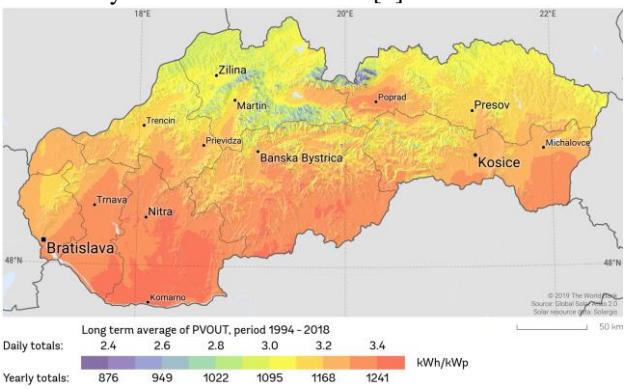


Fig. 1. The photovoltaic power potential of Slovak Republic [4].

The model of the photovoltaic (PV) station was designed for an output of 100 kW. The PV array has seven modules per string connected in series, which are connected in thirty-five strings in parallel. The model of the battery storage system was based on the same design as the replica of the

photovoltaic station as shown in Fig. 2 and Fig. 3. The battery module can also be added in a model of PV station [3].

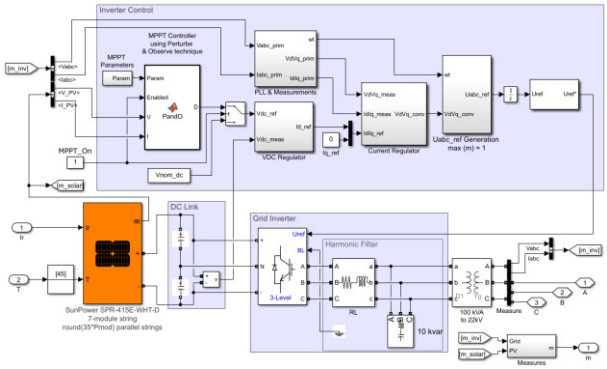


Fig. 2. Schematic model of the photovoltaic station [3].

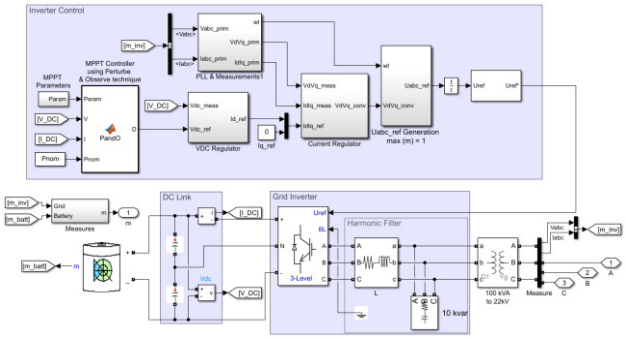


Fig. 3. Schematic model of battery storage system [3].

B. Model of Hydropower Plant

As shown in Fig. 4 current hydropower plants on the territory of the Slovak Republic, where the most power plants are placed on the Váh River. The largest hydroelectric power plants in Slovakia are Gabčíkovo on the Dunaj river and Pumped-storage hydropower plants the Čierny Váh and the Liptovská Mara [3].

Fig. 4. Schematic model of hydropower plant [5].

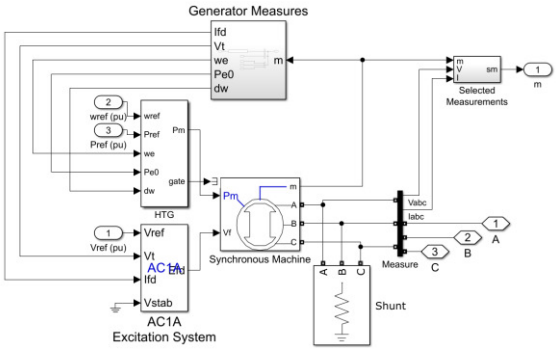


Fig. 5. Schematic model of hydropower plant [3].

The design of the hydropower model was much simpler than the design of the PV station or the battery system as shown in Fig. 5. This also follows from the number of modules used and consequently their settings. The model of the hydropower plant was designed for an output of 125 kW [3].

C. Model of Thermal Power Plant

The model of the thermal power plant was based on the same design as the replica of the hydropower plant. This model can also be used as biogas or biomass power plant after some modifications. The model of the thermal power plant was designed for an output of 105 kW [3].

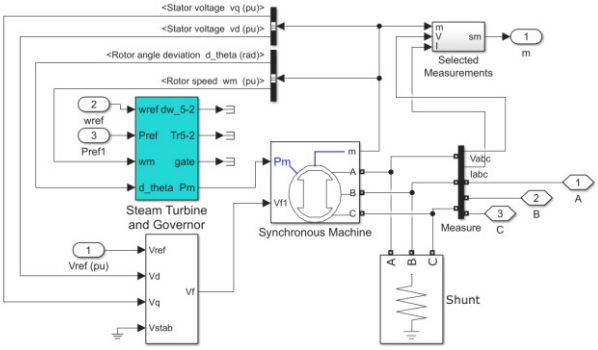


Fig. 6. Schematic model of thermal power plant [3].

IV. Next Research Steps

The next course of action on this issue will be to design a suitable network topology for the microgrid system and provide power for the network with a suitable mix of sources. When designing the topology, the size of the microgrid will be critical, as well as the placement of power supplies and ensuring uninterrupted power throughout all day. The next progress will be to deploy the protection and fault management system into the microgrid system and test the functionality of the system.

Acknowledgment

This work was supported by the Slovak Research and Development Agency under contract No. APVV-19-0576 and the Ministry of Education, Science, Research and Sport of the Slovak Republic and the Slovak Academy of Sciences VEGA 1/0757/21.

References

[1] Kadar, P. Power mix optimization on risk base. In Proceedings of the 2015 18th International Conference on Intelligent System Application to Power Systems (ISAP), Porto, Portugal, 11–16 September 2015; pp. 1–5.

[2] Štefko, R.; Čonka, Z.; Kurimský, J.; Kolcun, M. Špecifikácia vplyvov nepriaznivo pôsobiacich na stabilitu prevádzky ES SR a ich eliminácia. In Elektroenergetika: International Scientific and Professional Journal on Electrical Engineering: Medzinárodný vedecký a odborný časopis pre elektroenergetiku; Košice, Slovakia; Technical University of Košice; Vol. 13; No. 1 (2020); pp. 15–19; ISSN: 1337–6756.

[3] Štefko, R.; Čonka, Z.; Kolcun, M. Case Study of Power Plants in the Slovak Republic and Construction of Microgrid and Smart Grid; In Appl. Sci. (MDPI); 2021; No. 11:5252; [CrossRef].

[4] TheWorld Bank. "Global Solar Atlas 2.0", Solar Resource Data: Solargis; 2019; Available online: https://solargis.com/maps-and-gis-data/download/slovakia (accessed on 4 March 2021).

[5] Slovak Renewable Energy Agency, "Obnovitel'né Zdroje Energie". Available online: http://skrea.sk/obnovitelne-zdroje-energie/ (accessed on 3 March 2021).

Comparison of liquid insulating materials in the alternating electric field using dielectric spectroscopy

[1]*Peter HAVRAN (3rd year)*
Supervisor: [2]Roman CIMBALA

[1,2]Department of Electric Power Engineering, FEI TU of Košice, Slovak Republic

[1]peter.havran@tuke.sk, [2]roman.cimbala@tuke.sk

***Abstract*—The essence of this article is to compare insulating oils for high voltage applications in terms of dielectric losses and to point out the dynamic change of polarization and conductivity processes in an alternating electric field. As the current global demands lie in the replacement of mineral oil, in this research we present a potential replacement in the form of hydrocarbon oil based on the results of dielectric spectroscopy analysis.**

***Keywords*—complex electric modulus, dielectric spectroscopy, dissipation factor, insulating liquids.**

I. Introduction

The development of materials is advancing, which also applies to liquid insulating materials. Mineral oil produced based on non-renewable petroleum products has long been used in power transformers due to its low cost and relatively high rate of heat dissipation. Therefore, based on its biodegradability, global researchers are focusing on alternative, electrical insulating liquids [1][2].

In recent years, it is possible to register the development of hydrocarbon transformer oil, produced based on GTL (Gas to Liquid) technology, which is obtained by converting natural gas into liquid waxy hydrocarbons using the Fischer-Tropsch process. The absence of sulfur and the negligible amount of aromatic and unsaturated hydrocarbons, provide GTL with excellent properties compared to the mineral oils used in operation [3][4][5].

II. Experiment

Experimental measurements were performed on two new, ageless samples of liquids, Mogul TRAFO CZ-A (MO) and Shell DIALA S4 ZX-1 (SD) under laboratory conditions. MO is an inhibited mineral oil with an oxidation inhibitor. SD is a hydrocarbon oil produced based on GTL technology with low content of aromatic and unsaturated substances. Measuring setup contains the Tettex AG Zürich electrode system (2903a) and an IDAX 300 measuring device, which was connected to a PC. In the test procedure, three alternating electric field intensities of 0.5, 5, and 50 kV/m were applied to the oil sample in the frequency range 0.1 mHz – 3 kHz. The measured data were transferred to a PC, which were analyzed by dielectric spectroscopy. For the analysis of experimental samples, were chosen the dielectric parameters, the complex electric modulus M^* [6] and the complex impedance Z^*.

A. Analysis of Relaxation Processes

According to the Cole-Cole diagram, the complex electric modulus is shown in Fig. 1. All the continuous curves show the character of a waveform composed of two polarization processes with different relaxation times τ_M. The ideal Debye relaxation curves were plotted for both polarization processes (dotted and dashed curves) with plot parameters, where M_s is the static modulus, M_∞ is the optical modulus, and ω is the angular frequency. We can say that with increasing intensity of the alternating electric field in the measured frequency band of the oil MO, there is a more excellent approximation of the Debye behavior.

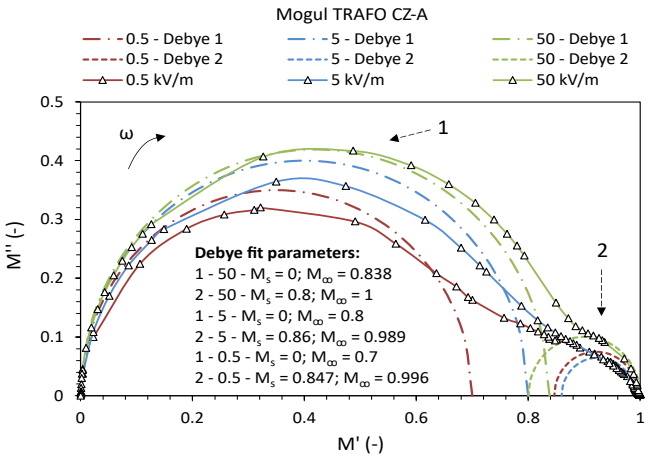


Fig. 1. Cole-Cole diagram of a complex electric modulus M^* of MO oil

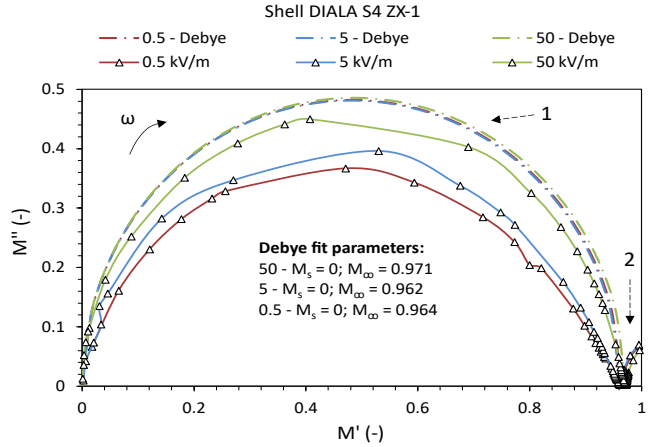


Fig. 2. Cole-Cole diagram of a complex electric modulus M^* of SD oil

In Fig. 2, we point to the presence of one polarization process with a capture of the beginning of the developing second polarization process. The obtained parameters α (1 Debye = 0), indicate a more significant distribution of relaxation times with a decimal decrease of the applied intensity of the alternating electric field. The relaxation times of the polarization processes decrease with increasing intensity of the alternating electric field in both oils. It means that a higher electric field intensity accelerates changes in the rotation of the electric dipoles behind changes in the alternating electric field.

B. Analysis of Conductivity Processes

To test the electrical insulating character of MO and SD oils, we present graphs of complex impedance Cole-Cole in Fig. 3. As the intensity of the alternating electric field increases, the impedance in the complex plane decreases and the electrical conductivity increases. In addition to the Cole-Cole impedance curves, ideal semicircles (IS) were plotted. The ideal Debye formalism (ideal semicircles IS) represents pure unidirectional conductivity. It is clear that the distribution of free charges in the fluids under an alternating electric field is not purely unidirectional. We attribute this dispersion to the superimposed conductivity of direct current and alternating current.

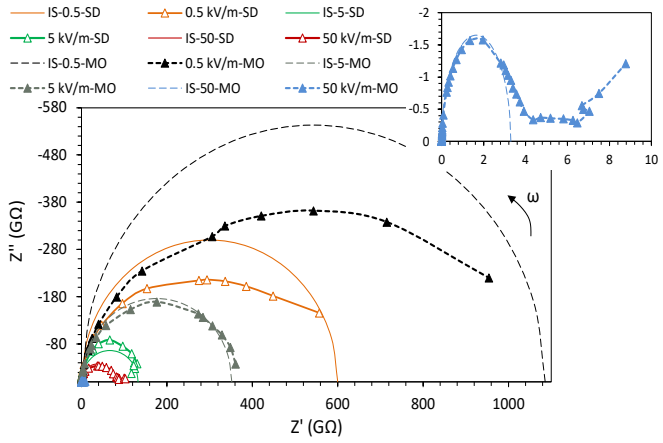


Fig. 3. Cole-Cole diagram of the complex impedance Z^* of the oils

C. Analysis of Dielectric losses

We found that the relaxation processes of MO mineral oil showed more characteristical parameters in terms of approach to the characteristics of Debye than hydrocarbon oil SD. Fig. 4 shows the dependence of dielectric losses in the frequency band.

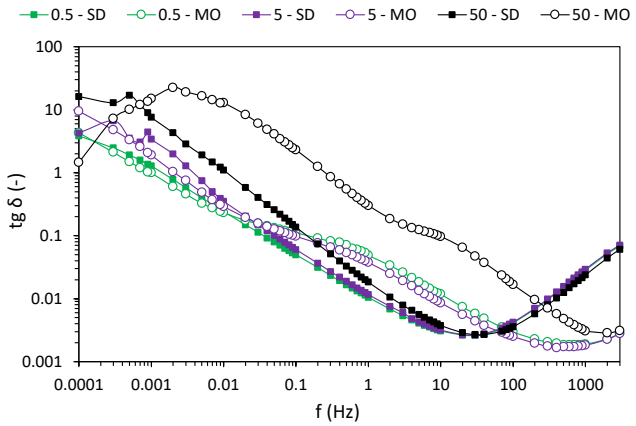


Fig. 4. *Tg δ* of electrical insulating oils in the frequency band

The paradox is that in terms of dielectric losses of electrical insulating oils at mains frequency (50 and 60 Hz), SD hydrocarbon oil achieves lower dielectric losses at all times applied intensities of the alternating electric field. This work aimed to find out the initial state of classic mineral oil and a unique hydrocarbon oil based on liquefied natural gas and compare them.

III. Conclusion and Further Research

MO mineral oil causes less charge absorption and greater distribution of free charges in the liquid. MO is more homogeneous in terms of polarization processes. MO causes higher dielectric losses around the mains frequency. This fact is crucial from practical use and points to the progressiveness and fundamental difference of SD hydrocarbon oil in the operation and cooling of power transformers. The general requirement is to reduce the environmental impact, which corresponds to the reduction in heating into the atmosphere caused by the inefficient operation of power transformers, which has a slight contribution to climate change. These facts are related to dielectric losses of liquid insulation in power transformers. This research provides a reasoned proposal to replace mineral oil with SD hydrocarbon oil.

Future research will focus on the dielectric properties of hydrocarbon oil SD enriched with impurities based on nanoparticles of magnetite Fe_3O_4 and fullerene C_{60}, which will analyze its diagnostic parameters at different stresses to describe the progressive material in more detail.

Acknowledgment

This research was funded by the Ministry of Education, Youth and Sports within the project VEGA 2/0011/20 and 1/0154/21 and the Slovak Agency for Research and Development based on contracts no. APVV-15-0438, APVV-17-0372, and APVV-18-0160.

References

[1] A.J. Amalanathan, R. Sarathi, N. Harid, H. Griffiths, "Investigation of the Effects of Silver Sulfide on the Dielectric Properties of Mixed Insulating Liquid," Proc. 2020 Inter. Sym. Elec. Insul. Mater., Piscataway, 2020, pp. 343-346.

[2] V. Mentlik, P. Trnka, J. Hornak, P. Totzauer, "Development of a Biodegradable Electro-Insulating Liquid and Its Subsequent Modification by Nanoparticles," Energies, vol. 11, 508, 2018.

[3] T. Münster et al., "Investigation on the impregnation characteristics of a new GTL based synthetic insulating fluid," Proc. 2017 IEEE 19th Inter. Conf. Diel. Liq., Piscataway, 2017, pp. 1-4.

[4] P. Havran et al., "Dielectric Properties of Electrical Insulating Liquids for High Voltage Electric Devices in a Time-Varying Electric Field," Energies, vol. 15, 391, 2022.

[5] H. Yu, Q. Liu, Z. Wang, C. Krause, A. Hilker, "Effects of Electric field Uniformity on Streamer and Breakdown Characteristics in a Gas-to-Liquid Oil under Positive Lighting Impulse," Proc. 2020 IEEE Inter. Conf. High Volt. Engin. App., Piscataway, 2020, pp. 1-4.

[6] F. Tian, Y. Ohki, "Electric modulus powerful tool for analyzing dielectric behavior" IEEE Trans. Dielec. Electr. Insul., vol. 21, 2014.

High-frequency Signal Injection-based Sensorless Control of PMSM

[1]*Viktor Petro (2nd year),*
Supervisor: [2]*Karol Kyslan*

[1,2]Dept. of Electrical Engineering and Mechatronics, FEI TU of Košice, Slovak Republic

[1]viktor.petro@tuke.sk, [2]karol.kyslan@tuke.sk

***Abstract*—In this article, a simple but effective approach for sensorless control of permanent magnet synchronous motor (PMSM) is presented. The main drawback of conventional methods for sensorless control is the usage of digital filters which affect the bandwidth of the PI controllers. The method introduced in this paper does not need digital filters. The simulation results of the proposed method with permanent magnet synchronous machines show its high performance for sensorless control.**

***Keywords*—Sensorless control, high-frequency signal injection, position estimation**

I. Introduction

Permanent magnet synchronous motor (PMSM) is widely used in the industry. It is preferred due to its high efficiency, high torque density, and maintenance-free operation. The field-oriented control (FOC) is usually adopted for high-performance motor control from zero to nominal speed range. For appropriate motor control, the exact position of the rotor magnetic field is essential. Therefore, mechanical sensors such as the incremental rotary encoder (IRC) or resolver are used. However, the mechanical sensor negatively affects the robustness of the system due to its sensitivity to electromagnetic noise. Furthermore, they are expensive, unreliable, and require mounting space. Therefore, different approaches for the elimination of the mechanical sensor have been researched for the last two decades. These approaches are referred to as sensorless control [1].

In general, there are sensorless control approaches that are applicable either for high speed or the low-speed region. Usually, approaches used for high-speed operation are based on the back electromotive force (back-EMF) estimation. These approaches fail to observe the rotor position at low and zero speed regions since the amplitude of the back-EMF is small or zero. Therefore, approaches based on high frequency (HF) signal injection into the stator windings are used in this speed region. Here, the rotor position is observed using the magnetic saliency of the motor. This article mainly focuses on HF signal injection-based methods applicable for low-speed and standstill motor operation [2], [3].

II. HF signal injection based sensorless control

HF signal injection can be provided in two ways:

- HF signal is superimposed onto the fundamental-frequency component of the stator voltage,
- HF signal is injected while the motor control is interrupted.

In the first approach, sinusoidal or square wave HF voltage is usually superimposed onto the fundamental-frequency component of the voltage applied to the motor windings. Therefore the induced current in the motor windings also contains HF current components. These HF components need to be filtered out before the current could be used as feedback for the current loop controller. Furthermore, the HF current components carry information about the rotor position. It is important to note, that HF components need to be separated from the fundamental frequency current before the extraction of the position. Therefore, a high-pass or band-pass filter is necessary for the signal demodulation process. Using digital filters, either in the current loop or in the position estimation process decreases the bandwidth of the controllers and causes deterioration in the dynamic performance of the drive [4]. To overcome this drawback, approaches, where the high-frequency signal is injected while the motor control is interrupted were developed. No HF component is injected when the motor control period is enabled and the sampled current can be directly used as feedback for the current controller. Besides, when the injection period is enabled, the motor control is interrupted, therefore the sampled current in this period is directly linked to the rotor position and the digital filters are not necessary anymore. In this article, an approach where the HF signal is injected during the motor control interruption will be presented.

III. HF pulse signal injection-based method

In this method, the HF voltage will be injected between two FOC periods. Furthermore, the injected voltage will be applied in the estimated $\hat{d}\hat{q}$ axis, often referred to as $\gamma\delta$ reference frame. The currents are sampled at the beginning of each switching period as shown in Fig. 1. In this case, the switching period is set to 40 kHz (25 µs) and the overall control period will consist of three switching periods. In the first switching period, the applied voltage vector **U** will be created using the output from the PI controllers in the FOC. In the second and third switching period a positive and negative pulse voltage will be applied either in the estimated $\hat{d}$ (γ) or $\hat{q}$ (δ) axis. For the current variation in $\alpha\beta$ reference frame stands [3]:

$$\Delta\mathbf{i}_{\alpha\beta} = (c_1 + c_2 e^{j2(\theta_e - \theta_u)})\mathbf{u}_{\alpha\beta}\Delta T, \tag{1}$$

where $c_1 = \frac{\Sigma L}{\Sigma L^2 + \Delta L^2}$, $c_2 = \frac{-\Delta L}{\Sigma L^2 + \Delta L^2}$, where $\Sigma L = \frac{L_d + L_q}{2}$, $\Delta L = \frac{L_d - L_q}{2}$ and L_d, L_q are the direct and quadrature axis inductances respectively, θ_e is the actual electrical rotor position, θ_u is the voltage angle in the $\alpha\beta$ reference frame

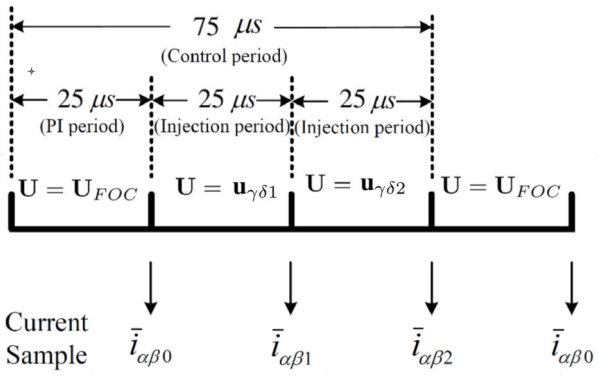


Fig. 1. The sequence of the FOC and voltage impulse injection within one control period.

and $\mathbf{u}_{\alpha\beta} = U_m e^{j\theta_u}$, where U_m is the amplitude of the voltage vector $\mathbf{u}_{\alpha\beta}$. Since the pulse voltage is injected in the estimated $\hat{d}\hat{q}$ ($\gamma\delta$) axis, it is convenient to transform (1) into the estimated rotor reference frame. For this purpose the estimated rotor position $\hat{\theta}_e$ is introduced:

$$\Delta\mathbf{i}_{\gamma\delta} = \Delta\mathbf{i}_{\alpha\beta} e^{-j\hat{\theta}_e} = (c_1 + c_2 e^{j2(\theta_e - \hat{\theta}_e - \hat{\theta}_u)})\mathbf{u}_{\gamma\delta}\Delta T, \quad (2)$$

where $\mathbf{u}_{\gamma\delta}$ is the voltage vector represented in the estimated $\hat{d}\hat{q}$ ($\gamma\delta$) axis and $\hat{\theta}_u = \theta_u - \theta_e$ is the voltage angle in the estimated reference frame, and ΔT is the duration of one switching period. The pulse voltages applied in the two injection periods in the estimated $\hat{q}$ (δ) axis can be expressed as:

$$\begin{aligned} \mathbf{u}_{\gamma\delta 1} &= U_m e^{j\hat{\theta}_{u1}} \approx U_m e^{j\pi/2} - \Delta\mathbf{u}, \\ \mathbf{u}_{\gamma\delta 2} &= U_m e^{j\hat{\theta}_{u2}} \approx U_m e^{j-\pi/2} - \Delta\mathbf{u}, \end{aligned} \quad (3)$$

where $\mathbf{u}_{\gamma\delta 1}$, $\mathbf{u}_{\gamma\delta 2}$ denote the first and second injection period respectively, $\hat{\theta}_{u1}, \hat{\theta}_{u2}$ is the injected voltage angle in the estimated $\hat{d}\hat{q}$ ($\gamma\delta$) axis and $\Delta\mathbf{u}$ is the inverter voltage error. Substituting (3) into (2) leads to:

$$\begin{aligned} \Delta\mathbf{i}_{\gamma\delta 1} &= \Delta T c_1 (U_m e^{j\pi/2} - \Delta\mathbf{u}) + \Delta T c_2 e^{j2\tilde{\theta}_e} e^{-j\hat{\theta}_{u1}} U_m, \\ \Delta\mathbf{i}_{\gamma\delta 2} &= \Delta T c_1 (U_m e^{-j\pi/2} - \Delta\mathbf{u}) + \Delta T c_2 e^{j2\tilde{\theta}_e} e^{-j\hat{\theta}_{u2}} U_m, \end{aligned} \quad (4)$$

where $\tilde{\theta}_e = \theta_e - \hat{\theta}_e$ is the rotor position observation error. The position error can be linked to the real part of the current variations as follows:

$$\mathrm{Re}(\Delta\mathbf{i}_{\gamma\delta 1} - \Delta\mathbf{i}_{\gamma\delta 2}) = 2k \sin(2\tilde{\theta}_e) \approx 4k\tilde{\theta}_e, \quad (5)$$

where $k = \Delta T c_2 U_m$. To obtain the estimated rotor position $\hat{\theta}_e$ and speed $\hat{\omega}_e$ a phase locked loop (PLL) might be adopted, according to Fig. 2 [5]. By regulating the current variation expressed by (5) to zero, the observation error $\tilde{\theta}_e$ will converge to zero and the estimated $\hat{d}\hat{q}$ ($\gamma\delta$) reference frame will align with the real dq reference frame.

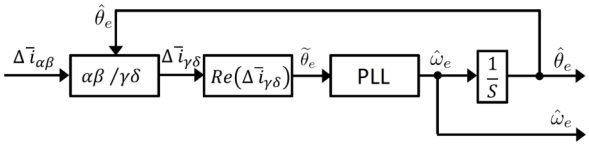


Fig. 2. The coordinate transformation and PLL used for position and rotor speed observation during HF pulse signal injection [6].

IV. Simulation Results

Simulations were carried out using the Matlab/Simulink environment. In the simulation, the switching frequency was set to 40 kHz, and the amplitude of the injected voltage signal was set to $U_m = 40$ V. The simulation results are presented in Fig. 3. The peak error in the speed observation did not exceed 0.7 rad (6.7 rpm) whereas the peak electrical rotor position error was 0.0007 rad (0.4 electrical degrees). The full load was applied in $t = 0.4$ s. From the presented simulation results it can be seen, that the sensorless control has good stability and the speed and position observation error is very low either during transient state or steady-state.

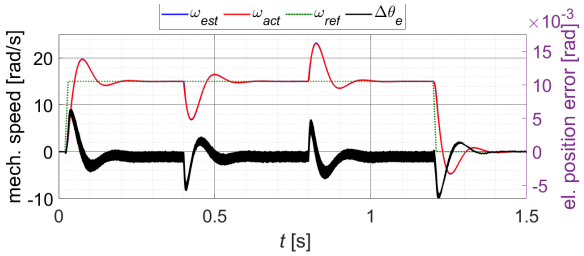


Fig. 3. Simulation results of sensorless control using the HF pulse signal injection-based method.

V. Conclusion

In this article, a sensorless approach based on HF pulse signal injection is presented. The solution offers high-performance motor control in the full load range. Since the digital filters are omitted the bandwidth of the controllers is not affected. Even though the frequency of the field-orientated control loop is decreased by 1/3 the quality of the sensorless control is good and reliable. Future work will consist of experimental verification on a test bench.

Acknowledgment

This work was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic under the project VEGA 1/0493/19 by Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovakia under Grant FEI-2022-86: *Experimentálne overenie bezsnímačového riadenia SMPM s integrovaným meničom.*

References

[1] Y. Hosogaya and H. Kubota, "Position estimating method of ipmsm at low speed region using dq-axis current derivative without high frequency component," in *2013 IEEE 10th International Conference on Power Electronics and Drive Systems (PEDS)*, 2013, pp. 1306–1311.

[2] J. Agrawal and S. Bodkhe, "Experimental study of low speed sensorless control of pmsm drive using high frequency signal injection," *Advances in Electrical and Electronic Engineering*, vol. 14, 03 2016.

[3] G. Xie, "Design of parameter independent, high performance sensorless controllers for permanent magnet synchronous machines," Ph.D. dissertation, Aalborg University, Denmark, 2015.

[4] W. Gaolin, Z. Guoqiang, and X. Dianguo, *Position Sensorless Control Techniques for Permanent Magnet Synchronous Machine Drives*. Singapore: Springer Verlag, 2020.

[5] G. Wang, Z. Li, G. Zhang, Y. Yu, and D. Xu, "Quadrature pll-based high-order sliding-mode observer for ipmsm sensorless control with online mtpa control strategy," *IEEE Transactions on Energy Conversion*, vol. 28, no. 1, pp. 214–224, 2013.

[6] G. Xie, K. Lu, S. K. Dwivedi, J. R. Rosholm, and F. Blaabjerg, "Minimum-voltage vector injection method for sensorless control of pmsm for low-speed operations," *IEEE Transactions on Power Electronics*, vol. 31, no. 2, pp. 1785–1794, 2016.

Influence of the plasma treatment on the surface's wettability

[1]*Daniel DZIVY (3rd year)*
Supervisor: [2]Alena PIETRIKOVA

[1,2]Dept. of Technologies in Electronics, FEI TU of Kosice, Slovak Republic

[1]daniel.dzivy@tuke.sk, [2]alena.pietrikova@tuke.sk

***Abstract*— This paper is focused on the tasks and obtained results in the previous year of the post gradual study. The work is focused on the possibility of the real-time contact angle's measurement of the melted solder balls by using the developed laboratory equipment, which allows for the measurement of the spreading rate and wetting coefficient as well. The article describes this developed equipment in detail. This paper also describes the improvement of the solder flux activity when using the plasma treatment.**

***Keywords*—Surface finishes, solder joint, solder ball, contact angle, plasma treatment.**

I. Introduction

Nowadays, modern trends in electronics are aimed at the evaluation of the wetting and for the evaluation of the contact angle of the melted solders. However, today's modern measuring devices do not allow us to measure the contact angle quickly and cheaply. The wetting balance test is one of the most popular methods for the evaluation of wetting, where the sample is inserted into the melted solder alloy and the wetting force is measured. This type of test can give us the result that the tested sample can be wetted or not. Another type of test is the drop shape analysis. This test consists of melting the solder on the tip and putting the melted drop on the substrate. These types of measurements do not allow to measure contact angle in real-time. Our developed laboratory equipment allows measuring the contact angle of the melted solder ball in real-time and spreading rate in real-time as well. The main task of our measurement is to prove, that the plasma treatment positively affects the performance of the solder flux [1]-[4].

II. Initial Status

During the third year of my PhD study, I was analyzing the surface finishes of the PCB from the wetting point of view. Also, I advised with the plasma treatment to ensure the better wettability of the surfaces. The analysis of the plasma influence on the surface free energy that depends on the surface finish of the PCB and the plasma was realized as well. My work was aimed at the evaluation of the contact angles of the melted solder balls to compare the solder finishes. Another part of my work was aimed at testing a newly developed surface finish based on SnAg alloy. The tests consisted of an evaluation of the electrical and mechanical properties with the main accent on the wettability of the surface and on the electrical resistivity of the solder joints. Our further research leads us to develop a new type of measuring equipment for evaluation of the contact angles and spreading rate of the melted solder balls. This type of equipment allows for comparing the surface finishes. The aim of our work is focused on plasma treatment. Our experiments have shown that the plasma treatment improves the flux activity by lowering the contact angle of the melted solder balls.

III. Solved Tasks in The Previous Year

Tasks which are summarized in the following section were solved in the last year of postgraduate study.

A. *Real-time measurement of the melted solder balls' contact angle in laboratory conditions*

The aim of this task was to evaluate the influence of two types of solder alloys and the plasma treatment used before the application of the solder flux on the contact angle of the melted solder balls. For this purpose, laboratory equipment for the real-time measurement of the contact angle was developed. It was able to measure the contact angle of the melted solder balls and the spreading rate as well. The laboratory equipment of our real-time contact angle equipment has exactly controlled heating. This equipment includes moreover an optical part that provides the changes of the contact angle of the gradually melting solder balls from a side view in real-time. The principle of this method was verified by comparison of the pure copper substrate with substrate treated with flux or combination plasma – flux. Our developed laboratory equipment for the real-time contact angle measurement consists of two parts: optical and heating part, as can be seen in the Fig. 1. The heating part is based on the resistance wire used for the heating inserted in the aluminum thermal block to ensure the stability of the melting process. The heating part was controlled by the microcontroller Atmega 328P using a switching relay and the K-type thermocouple for measuring the substrate's temperature.

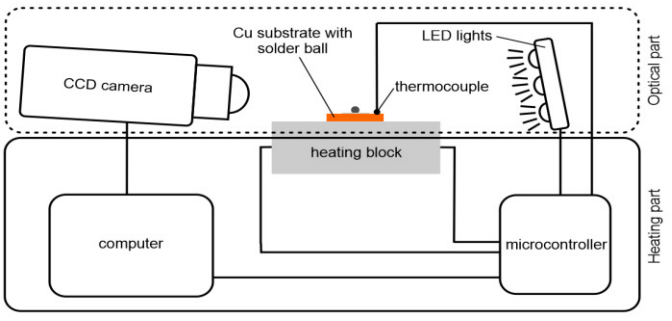


Fig. 1 Schematic view of the optical part and of the heating part of real time contact angle equipment

The contact angle (θ), height (h), and diameter (d) of the solder ball were measured using the CCD camera placed from the side view (Fig. 2). Recorded video was evaluated using the computer. The ImageJ software allows us to measure the contact angle with accuracy ±1° using the θ/2 – method, a widely used technique for measuring the contact angle of the solder balls. It is assumed that the solder ball's contour is part of a sphere and gravity force is neglected due to the low volume of the solder. The contact angle can be counted using the height, and diameter of the solder balls by the spherical approximation.

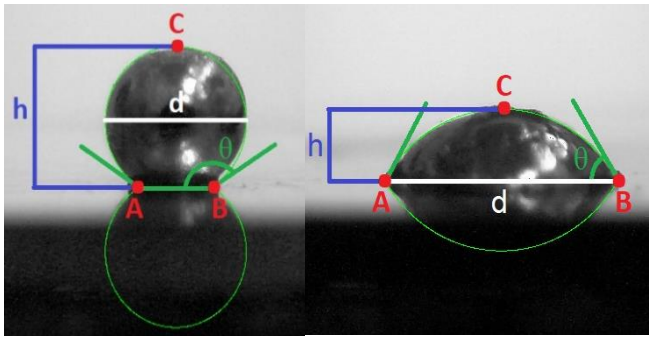


Fig. 2 Principle of optical evaluation: side view from the camera on solder ball (left) before melting (right) after melting

B. *Influence of the plasma treatment and its effectiveness on the wettability of the solder*

The aim of this task was to prove that the plasma treatment influences the solder flux activity by lowering the contact angle of the melted solder balls. This research was done by using the developed laboratory equipment for the evaluation of the contact angles. The contact angle of the melted SAC solder is 143° (without fluxing) after 50 seconds since melting started on a pure copper substrate as can be seen in the Fig 3. For the SnPb solder, the contact angle after 50 seconds without fluxing is 142°. Just fluxing of the substrate decreases the contact angle for both solders (46° for the SAC and 24° for the SnPb solder). The influence of the plasma treatment was proved for both types of solders. Plasma treatment improved the contact angle of the SnPb solder to 18°. For the SAC solder, improvement was almost identical, and the contact angle stabilized at 39°.

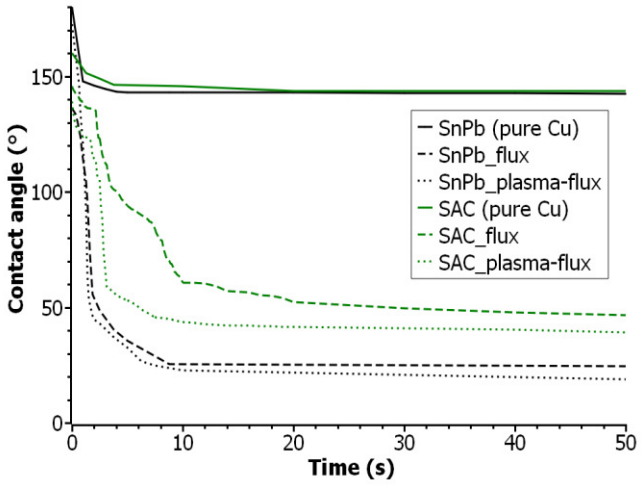


Fig. 3 Dependency of melted solder balls' contact angle on real time

This task was also aimed at the evaluation of the effectiveness of the plasma after treatment. Comparison of the samples treated with the flux in combination with the plasma and only flux is shown in the Fig. 4. As can be seen, the surface cleanliness after plasma treatment lasted approximately from 3 – 4 hours. However, the effect of plasma does not change even after 8 hours.

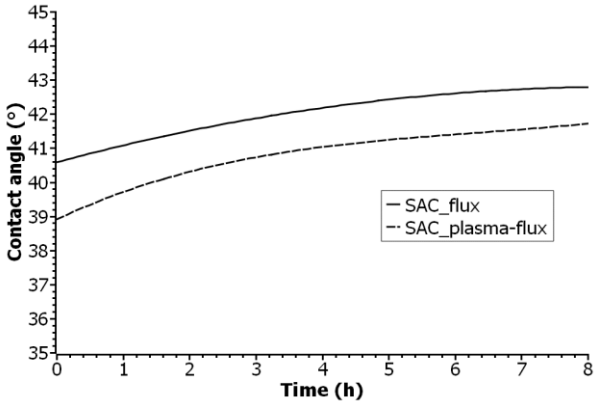


Fig. 4 Comparison of contact angle in terms of atmosphere exposure duration for SAC solder for Cu samples with and without combination plasma- flux treatment

IV. Future Work

In future work, we will focus on the mechanical properties of the solder balls with the main accent on their shear strength. My work will be also focused on completing the dissertation thesis, which is aimed at the analysis of the surfaces' quality and their wettability.

V. Conclusion

In this article, real-time contact angle measurement of solders in laboratory conditions was presented. This type of measurement allowed to research the influence of the plasma treatment on the contact angle of the melting solder balls. The results showed that the flux impact is even more efficient if the samples' surface is also treated with plasma. Similarly, the results of experiments pointed to the positive effect of plasma treatment and the effectiveness of flux in the case when thermal processing occurs after several hours. A decrease of the contact angle after applying plasma-flux treatment is more noticeable in the contrast to no flux or flux treatment only. The effect of plasma treatment lasted even after 8 hours after application, and it can be recommended in the case when a solder paste cannot be deposited on a PCB immediately even after 3 hours.

References

[1] WANG, H., GAO, F., MA, X., QIAN, Y.: Reactive wetting of solders on Cu and Cu6Sn5/Cu3Sn/Cu substrates using wetting balance, *Scripta Materialia*, vol. 55, pp. 823–826, 2006.

[2] WANG, J., WU, Y., CAO, Y., LI, G., LIAO, Y.: Influence of surface roughness on contact angle hysteresis and spreading work, *Colloid and Polymers Science*, vol. 298, pp. 1107–1112, 2020.

[3] WEDI, A., BAITHER, D., SCHMITZ, G.: Contact angle and reactive wetting in the SnPb/Cu system, *Scripta Materialia*, vol. 64, pp. 689–692, 2011.

[4] SIEWIOREK, A., KUDYBA, A., SOBCZAK, N., HOMA, M., HUBER, Z., ADAMEK, Z., WOJEWODA-BUDKA, J.: Effects of PCB Substrate Surface Finish and Flux on Solderability of Lead-Free SAC305 Alloy, *Journal of Materials Engineering and Performance*, vol. 22, pp. 2247–2252, 2013

Cloud based system for freezing of gait cueing using artificial intelligence

[1]*Pavol ŠATALA (3rd year)*
Supervisor: [2]Peter BUTKA

[1,2]Department of Cybernetics and Artificial Intelligence, FEI TU of Košice, Slovak Republic

[1]pavol.satala@tuke.sk, [2]peter.butka@tuke.sk

***Abstract*—Term "Cloud computing" is becoming more and more popular, nowadays. It brings many advantages as scalability or remote management and updates. This paper offers an overview of theoretical background of cloud computing, as well as proposal of system using mobile device with cloud-based FoG recognition to help patients.**

***Keywords*—cloud computing, freezing of gait, smart devices**

I. Introduction

Large improvements in performance, availability and popularity of cloud-based systems appeared in the last few years. It caused, that they began to gain popularity among AI developers.

Presented article shows possibilities of usage a cloud-based freezing of gait detection to help people suffering Parkinson's disease. The paper consists of five chapters. The second chapter provides overview of cloud-based systems and cloud computing. It says a little about term "cloud" and explains three main types of cloud services. It is followed by chapter number three, which describes what freezing of gait is. Later the paper offers chapter describing different approaches to FoG detection. Finally, it presents suggested architecture and usage of smart devices in FoG detection in the fifth chapter.

II. Cloud computing

Phrase "cloud computing" firstly appeared in 1996 in Compaq internal document [1]. Before it, cloud symbol has been used to represent a computer network in two internet predecessors: ARPANET and CSNET [2, 3]. Later, the term "cloud" started to be used in schemesm where it was not important to describe the physical involvement of systems in detail. To refer a platform for distributed computing the term was used in 1993, when Apple sign-off a contract with General Magic company. General Magic and AT&T used it in describing their technology [4].

Nowadays, we can talk about cloud computing as a server-based systems providing to customers by external organizations. One of the main advantages of cloud computing is that it offers on demand scalability and flexibility of information communication technology and resources (i.e., computing, storage and network) [5–7]. Based on the different level of abstraction, cloud computing offers three main types of services [8]:

Infrastructure as a service (IaaS)

The IaaS lets companies to build their systems, on rented resources like servers, firewalls, or datacentres. It allows to build large server-based application without need to build own datacentre [5].

Platform as a Service (PaaS)

Provides network infrastructure together with operation system and other management tools.

Software as a Service (SaaS)

Offers specific software applications, distributed over the internet.

In addition to three main types, we can talk also about (Mobile) Backend as a Service ((M)BaaS). (M)BaaS is type of cloud computing, which offers services for customer applications in form of API (i.e.: Firebase services) [9]. It can provide basic functionalities like crash analysis, user management or notifications.

III. Freezing of gait

Freezing of gait is one of the most common symptoms of Parkinson's disease. It affects about 30% of all patients [10]. It is clinically characterized by sudden, brief episodes of inability to produce effective forward stepping that typically occur during gait initiation or turning while walking [11]. The Freezing of gait events often leads to falls, what significantly decreases patients' quality of life. The falls may lead to serious injuries [12]. Instead of existence of several pharmacological strategies to manage freezing of gait phenomena, the gait problems remain persistent in patients suffering from Parkinson's disease [13]. An audition and visual cueing, looks like to be an effective way to improve patients' ability to walk.

IV. Freezing of gait detection

Moore et al. was one of the first works dealing with freezing of gait detection [15]. He made an experiment with eleven patients suffering by idiopathic Parkinson's disease. The 46 Freezing of gait episodes were recorded. Four patients does not show any freezing event during the experiment. The patients walked up to 100m (based on their ability to walk). The patients' motion has been recorded by 3 axis accelerometers placed on patient's ankle. The acceleration has been recorded in frequency of 100Hz. The work finds out the

differences in dominant frequencies in FoG event and normal gait. Freezing index has been defined as a division of the square of the area under the power spectra of 'freeze 'band (3 - 8Hz) by the square of the area under the spectra in the 'locomotor 'band (0.5 - 3 Hz). They successfully detected 78% of FOG events using thresholding of the Freezing index. Twenty percent of stand events were incorrectly labelled as FOG by global trash. Lima et. al. [16] brings detailed review of existing works focusing to freezing of gait detection using wearable devices. The 27 works has been selected from PubMed and Web of Science databases. Four of them focuses to falls and 23 to freezing of gait.

V. Presented solution

In this work we present a solution, which uses mobile devices to detect freezing of gait. The mobile devices are widely spread among the people. They offer user friendly and cheap alternative for specialised medical sensors. Based to Android minimal hardware requirements (and iPhone hardware specifications) we can say that all currently used devices are equated by three axis accelerometer. However, a limitation of mobile devices, mainly in low end cheap devices, may be limited performance or limited battery life when running high performance. To overcome this limitation, we decided to run FoG detection algorithm in cloud. This decision also brings another important advantage: simple updates and improvements of recognition algorithm, without needs to release a new app version. The illustration of high-level system architecture can be seen on the following figure 1.

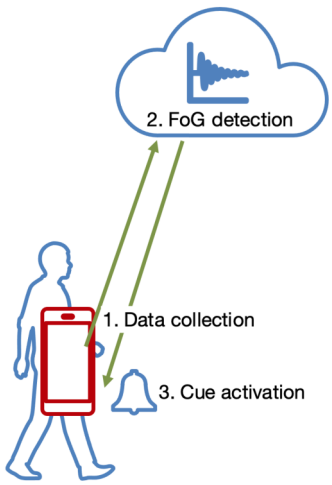


Fig. 1. High level system architecture

VI. Cueing

An experiment with eleven patients has been done to ensure the usability of wearable devices to cue the FoG. Patients, who's agreed to take part in the experiment, were asked to walk on the test trial and medical specialist count the number of FoG events. The experiment showed that acoustic aids can decrease time of walk up to 23% and vibration aids up to 11%. The next goal is to be allow system to provide on-demand cueing. On-demand cueing may provide better cueing performance [14]. FoG detection algorithm is needed to provide on-demand cueing. First experiments, using neural network show ability to use it to detection. The goal is to improve this algorithm.

VII. Conclusion

The article summarised the base benefits and services of cloud computing. It brings the overview of freezing of gait problematics. It also presented various ways used to detect freezing of gait using algorithms of artificial intelligence. It proposes a system combining the cloud based recognition system with mobile smart device used as sensor for data collection. Finally, performed works shows that using wearable devices and cloud-computing may be correct way to achieve required goal.

Acknowledgment

The This work was supported by Slovak VEGA research grant No. 1/0685/21 and Slovak APVV research grant under contract No.APVV-17-0550.

References

[1] A. Regalado, "Who Coined 'Cloud Computing'?" in Technology Review. MIT, 31 July 2013.
[2] "Internet History of 1970s | Internet History | Computer History Museum". www.computerhistory.org
[3] "National Science Foundation, "Diagram of CSNET," 1981".
[4] AT&T (1993). "What Is The Cloud?". Archived from the original on 2021-10-27. Retrieved 2017-10-26. "You can think of our electronic meeting place as the Cloud."
[5] ArmbrustM,FoxA,GriffithR,JosephAD,KatzR,KonwinskiAetal.Aviewof cloudcomputing.Communications of the ACM 2010; 53: 50–58.
[6] Lenk A, Klems M, Nimis J, Tai S , and Sandholm T, "What's inside the Cloud? An architectural map of the Cloud landscape," In Software Engineering Challenges of Cloud Computing, 2009. CLOUD '09. ICSE Workshop on, 2009, pp. 23–31.
[7] Louridas P. Up in the Air: moving your applications to the cloud. Software, IEEE 2010; 27: 6–11.
[8] M. A. Chauhan, M. A. Babar, B. Benatallah, Architecting cloud-enabled systems: a systematic survey of challenges and solutions. Software Practice and Experience. (2016)
[9] A. Williams,. "Kii Cloud Opens Doors For Mobile Developer Platform With 25 Million End Users". TechCrunch, 16th October 2012.
[10] Macht, M.; Kaussner, Y.; Möller, J.C.; Stiasny-Kolster, K.; Eggert, K.M.; Krüger, H.P.; Ellgring, H. Predictors of freezing in Parkinson's disease: a survey of 6,620 patients. Movement disorders 2007, 22, pp. 953–956.
[11] Nutt, J.G.; Bloem, B.R.; Giladi, N.; Hallett, M.; Horak, F.B.; Nieuwboer, A. Freezing of gait: moving forward on a mysterious clinical phenomenon. The Lancet Neurology 2011, 10, 734–744.
[12] Gao, C.; Liu, J.; Tan, Y.; Chen, S. Freezing of gait in Parkinson's disease: path-ophysiology, risk factors and treatments. Translational neurodegeneration 2020, 9, 1–22.
[13] Spaulding, S.J.; Barber, B.; Colby, M.; Cormack, B.; Mick, T.; Jenkins, M.E. Cueing and gait improvement among people wit Parkinson's disease: a meta-analysis. Archives of physical medicine and rehabilitation 2013, 94, 562–570.
[14] Velik, R. Effect of on-demand cueing on freezing of gait in Parkinson's patients. International Journal of Biomedical Engineering 2012, 6.
[15] Moore, S.; MacDougall, H.; Ondo, W. Ambulatory monitoring of motor fluctua-tions in Parkinson's disease. Journal of Neuroscience Methods 2008, 167, 340.
[16] De Lima, A.L.S.; Evers, L.J.; Hahn, T.; Bataille, L.; Hamilton, J.L.; Little, M.A.; Okuma, Y.; Bloem, B.R.; Faber, M.J. Freezing of gait and fall detection in Parkinson's disease using wearable sensors: a systematic review. Journal of neurology 2017, 264, 1642–1654.

An overview of compressed sensing and sparse signal recovery algorithms

[1]*Jozef KROMKA (1st year)*
Supervisor: [2]Ján ŠALIGA

[1,2]Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

[1]jozef.kromka@tuke.sk, [2]jan.saliga@tuke.sk

***Abstract*—Compressed sensing (CS), in recent years, gained attention in Electrical Engineering, applied mathematics, statistics, and computer science as a new method of acquiring signals by using sub-Shannon sampling frequency. In this article, the basic theory of signal sparsity and CS is presented. Later an Example of performing CS with a periodic signal is shown as well as applications of CS in signal processing. In the end, sparse signal recovery algorithms used in CS are explained.**

***Keywords*—Compressed sensing, Compression, Orthogonal transform, Sparse representation, Sparse signal recovery algorithms**

I. Introduction

In the digital world, signal and data compression is a rapidly growing element of the signal processing field. There are several ways to compress the signals. In recent decades, orthogonal transforms were used the most amongst other ways to compress the signal. The most popular orthogonal transforms are Fourier [1] and Wavelet [2] transforms. They can achieve a good compression ratio [3], but their computation is often very resource-demanding. This is because the transformation is performed by using many multiplications. For a modern computer, this is not an issue but for remote sensors, low-energy, or time-critical applications additional compute competitivity can present a big issue. CS is a novel approach to compress signals by acquiring only a fraction of the signal. It was originally proposed in [4] and [5], [6]. CS assumes, that the signal can be represented by using only a few suitable basis functions. A more detailed explanation will be presented in the next chapter. CS can be used in image processing [7], medical imaging [8], ECG sensing [9], radar technology [10], and others [11].

II. Signal sparsity and compressed sensing

Nowadays, Signal sparsity and CS is often discussed topic in the signal compression field. In this chapter, the sparsity of signals, as well as the basic theory of CS will be explained.

A. Signal sparsity

CS can be used to compress any type of signal which can be represented by K nonzero coefficients, where K is much smaller than the signal length. Signal that can be represented by K nonzero coefficients where K << N is called K-sparse. Usually, the basis on which the signal can be represented by a small number of K nonzero coefficients is the orthogonal transform basis. Then the sparse representation can be obtained according to (1).

$$f = \psi x \tag{1}$$

Where f is the signal, $\psi \in \mathbb{R}^{NxN}$ is the transform basis matrix and x is the K-sparse representation of the signal. The signal can be k-sparse on a dictionary basis as well, but in that case, the dictionary must be generated first. The main advantage of the orthogonal transform basis is that it can be generalized for most 1D or 2D signals. The disadvantage is that it cannot be used for wideband signals. Dictionary can be used and utilized for almost any type of signal but on the other hand, it cannot be generalized for more than one type of signal [12].

B. Compressed sensing

Typical compression of the signals is usually performed by acquiring the signal according to the sampling theorem. This signal is then transformed using the basis matrix ψ and only K coefficients are stored. In CS if the signal is K-sparse then it is possible to reconstruct the signal from M measurements where K < M << N. The measurements are not performed according to the sampling theorem, but the signal is measured at random positions. The random measurements are represented by the sensing matrix $\phi \in \mathbb{R}^{MxN}$. The measured signal is called y and it is represented by (2).

$$y = \phi f \tag{2}$$

The sensing matrix and basis matrix must be incoherent, which means that the rows of the sensing matrix cannot sparsely represent the columns of the basis matrix and vice versa. If all requirements are achieved then by inserting (1) into (2) it is possible to get the main idea of the compressed sensing represented by (3) [13], [14].

$$y = \phi\psi x \tag{3}$$

This equation can be also demonstrated by Figure 1.

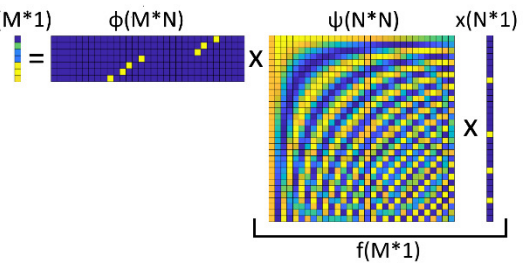


Figure 1. Visualized equation of Compressed sensing

Equation (3) represents a system of linear equations, where there are more unknown variables than the equations. This system of linear equations is also called an underdetermined system of equations. This means that there are infinitely many solutions to this system of equations. In CS, only the K-sparse solutions are interesting. There are a lot of algorithms and methods of obtaining the K-sparse solution to the underdetermined system of linear equations. An overview of these algorithms and methods will be presented in the next chapter.

In practice, CS is performed by sampling the input signal in random intervals or multiplying it by the sensing matrix ϕ. Compressed signal y is then stored in memory or transmitted to the device where it will be reconstructed. Since the reconstruction process is complex and a lot of computational power is needed, the signal is usually reconstructed using a computer. The reconstruction consists of two steps. The first step is to find a K-sparse solution to an underdetermined system of linear equations by using one of the existing algorithms. The second step is to reconstruct the signal from the obtained K-sparse representation. This step is performed by multiplying the basis matrix ψ with the K-sparse representation x according to equation (1). The general way of implementing CS is shown in Figure 2. [15].

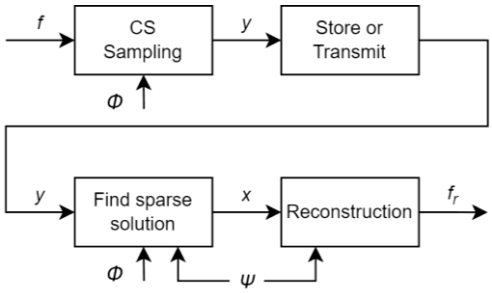


Figure 2. Implementation of Compressed sensing

C. An example of compressed sensing

Implementation of CS can be explained by using the following example. Suppose there is a signal f with the length of 256 samples, which consists of four spectral coefficients in the Cosine Fourier domain according to Figure 3. The sampling frequency of the signal is 360 Hz.

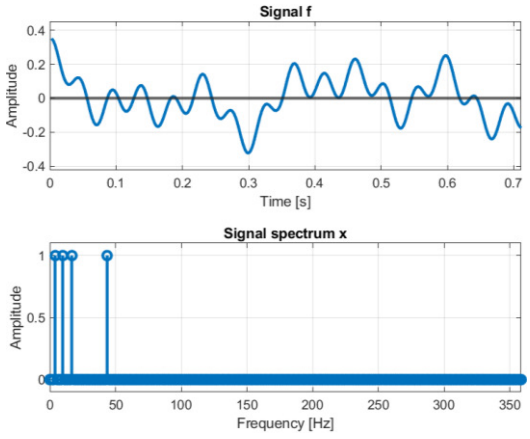


Figure 3. Example signal f with its spectrum in Cosine Fourier domain

Since this signal consists of four spectral coefficients which is much less than the signal length, we can say this signal is K-sparse where K equals four. Then we need to choose how many samples we will acquire from the signal f. In this example, we can choose M equals sixteen which is more than K but much less than the signal length. Then the compressed signal f with the sensing matrix ϕ will look according to Figure 4.

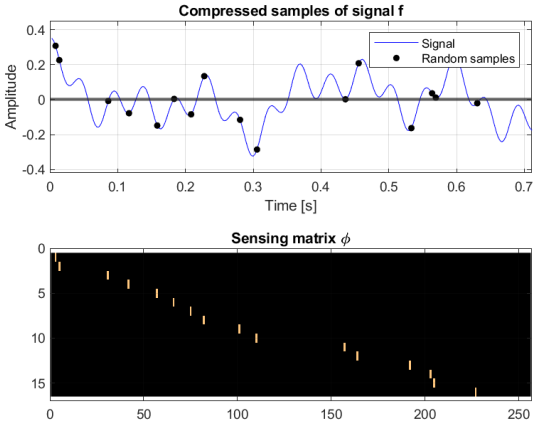


Figure 4. Compressed samples of signal f with the Sensing matrix ϕ

Then these data are supplied to the algorithm which will find the K-sparse solution of this undetermined system of linear equations. Since the signal is sparse in the Cosine Fourier domain, we also need to supply the basis matrix which in this case will be a Discrete cosine transform matrix. After the algorithm finds the K-sparse solution, we can simply reconstruct the signal by using equation (1). The results after the reconstruction of the signal are shown in Figure 5.

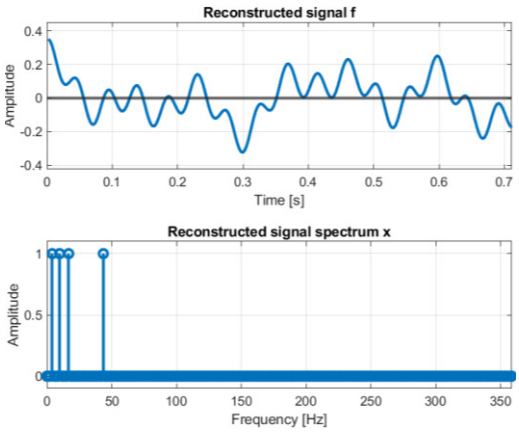


Figure 5. Reconstructed signal f with reconstructed signal spectrum in Cosine Fourier domain

In practice, the signal is not perfectly sparse as in the example. In that case, depending on the signal the results after reconstruction can vary. One of the problems of CS is to find which signals can be reconstructed by CS. Another problem is to find a suitable basis matrix for the signal to acquire the best results after reconstruction. These problems will be described in the following section.

D. Application of compressed sensing

When the CS theory was developed, the main goal was to find a way of sub-Shannon signal sampling. Throughout the years, CS has already found use in many fields of signal processing and signal compression.

In image processing, it has been used to reduce the number of measurements, to save energy or memory without significant loss of information. It has been used in [7] together with a convolutional neural network for that purpose. It can be also used, as demonstrated in [13], to implement face recognition or pattern recognition in general. In image

processing, CS was also used in seismic imaging, parallel imaging, subwavelength imaging, thermoacoustic imaging, microwave imaging, single-pixel camera, and video processing [16].

Medical imagining as part of the Image processing field also benefited from the development in the CS theory. Particularly it found use in magnetic resonance imagining (MRI) [8]. MRI is a costly and time-consuming procedure if it is done by using the Shannon sampling frequency. Since the MRI images are sparse in Fourier and Wavelet domains, CS can be utilized to lower acquired samples. This can significantly decrease the cost and time of the procedure [16].

Another area of medicine where CS can be used is electrocardiogram (ECG) sensing. In [9] it was used to decrease power consumption in remote health monitoring systems which sense the ECG of a patient. To perform CS on ECG signals, usually, wavelet or dictionary bases are used for reconstruction. To accurately measure key details from the signal, often some form of detector is used which would make more measurements at the required part of the signal. In general, it is possible to obtain high compression for ECG signals by using CS. A similar approach was also used for Electroencephalographic (EEG) or neural signals [16].

CS was also used in radar imaging [10] as well as in different radar frameworks like multistatic, bistatic, and monostatic radar. By using CS it is possible to identify a target and many other properties of the target from incoherent measurements [16].

Communication systems benefited from applications of CS as well. CS found use in sparse channel estimation, wireless sensor networks, ultra-wideband systems, cognitive radio, array signal processing, multiple access scheme, network tomography, multimedia coding, and communication as well as in information security. The last mentioned is very interesting because it uses the sensing matrix as a key which would be known to the encoder and decoder [16].

There are many other applications of CS. Since most of the natural signals are sparse, CS can be used widely for many types of data and signals. More applications of CS can be found in [11], [14], [16].

III. Sparse signal recovery algorithms

Since CS is based on assumption that the signal is sparse then it is possible to solve an undetermined number of the equation by searching for a sparse solution. There are many ways of finding a sparse solution to an undetermined system of linear equations. There are two main types of algorithms for sparse signal recovery used in CS. The first type is the l_1 minimization algorithms, and the second type is greedy algorithms. In this chapter most used algorithm for sparse signal recovery will be explained.

A. l_1 minimization algorithms

Since CS works with sparse signals, then the sparsest solution to the undetermined linear equation system can be solved by using the l_0 minimization. This minimization is described with the following equation (4).

$$\min_{x\in\mathbb{R}^n} \|x\|_0, \qquad subject\ to\ y = Ax \tag{4}$$

Where A can be obtained by (5).

$$A = \phi\psi \tag{5}$$

However, this problem cannot be solved because the problem is NP-hard. That is why the l_1 minimization is used to overcome such a complex problem. l_1 minimization can be described by equation (6).

$$\min_{x\in\mathbb{R}^n} \|x\|_1, \qquad subject\ to\ y = Ax \tag{6}$$

It was proven that under some conditions, this problem is equivalent to the l_0 minimization problem. The main advantage is that solving the l_1 minimization problem is that this problem is convex and feasible to solve. On the other hand, it is not guaranteed it will get the correct results [17]. This is also one of the problems which the research about CS is dealing with.

B. Greedy algorithms

The greedy algorithms are sparse signal recovery algorithms that use iterative methods to solve the l_1 minimization problem. These algorithms are iteratively approximating the results until the error of approximation is lower than a certain value. These algorithms are called "Greedy" because they are trying to minimize the sparse vector as much as possible [18]. There exists a lot of greedy algorithms which can solve the l_1 minimization problem. In this section, only the three most used will be explained.

The first algorithm is called Orthogonal matching pursuit (OMP) [19]. This algorithm is trying to minimize the l_1 norm by gradually approximating the K-sparse representation of the signal. The formal transcription of OMP is shown below.

Orthogonal matching pursuit

Input: Compressed sensing matrix A, measurement vector y
Initial values: $\hat{x}_0 = 0, r = y, T_0 = \emptyset, i = 0$
Iteration step: Repeat until the stopping criterion is met
- $i := i + 1$
- $T_i \leftarrow T_{i-1} \cup supp\ H_1(A^T r)$ // Add the largest residual entry to the support
- $\hat{x}_i|T_i \leftarrow A^{\dagger}_{T_i} y$ // Update the estimate of the signal
- $r \leftarrow y - A\hat{x}_i$ // Update the residual of the measurements

Output: $\hat{x}_i$

Where A is a compressed sensing matrix, y is a compressed measurement vector. $\hat{x}$ is the K-sparse representation of a signal, T is the support, $A^{\dagger}$ is a pseudo-inverse matrix of A. The OMP also makes use of the hard thresholding operator $H_k(x)$. The result of this hard thresholding operator is equal to x on the k entries of x with the largest magnitude and zero otherwise.

The function of the OMP algorithm can be described as follows. Input parameters consist of the compressed sensing matrix A and the measurement vector y. Output vector $\hat{x}$ is initialized with zeros and the residual is initialized with values of the vector y. Then the algorithm tries to find the column of matrix A which is most correlated to the residuum r. The index of the most correlated column is then added to the support. Then the estimate of K-sparse signal representation x is updated. In the end, the residual of the measurements is updated. If the stopping criteria are met, then the K-sparse representation of signal x is obtained. Otherwise, these steps are repeated. OMP is a simple algorithm to obtain l_1 minimization, but its simplicity is connected to one weak point of this algorithm. If the wrong index is added to the support during the computation of this algorithm, it cannot be removed later. The following algorithm was invented as an attempt to overcome this weak point.

The second algorithm is the Compressive sampling

matching pursuit (CoSaMP) [20]. The formal transcription of CoSaMP is displayed below.

Compressive sampling matching pursuit

Input: Compressed sensing matrix A, measurement vector y, sparsity level k
Initial values: $\hat{x}_0 = 0, r = y, T_0 = \emptyset, i = 0$
Iteration step: Repeat until the stopping criterion is met
$i := i + 1$
$T_i \leftarrow supp(\hat{x}_{i-1}) \cup supp\ H_{2k}(A^T r)$ // Update the support
$\hat{x}_i | T_i \leftarrow H_k(A_{T_i}^{\dagger} y)$ // Update the estimate of the signal
$r \leftarrow y - A\hat{x}_i$ // Update the residual
Output: $\hat{x}_i$

This algorithm is similar to the OMP algorithm. The main difference is that CoSaMP requires an additional input which is the expected sparsity level k. This input is then used to enlarge the support at each iteration not by one but by at least k new entries. Other steps are equivalent to the OMP.

The last algorithm explained is called Iterative Hard Thresholding (IHT) [21]. This algorithm is more directive in updating the K-sparse signal representation x than the OMP. The algorithm can be described in the following way. Similar to CoSaMP, IHT also assumes the sparsity level k is known. In the beginning, K-sparse signal representation x is initialized with zeros and gradually updated by the following formula (7).

$$\hat{x}_i = H_k(\hat{x}_{i-1} + A^T(y - A\hat{x}_{i-1})) \quad (7)$$

The formal transcription of IHT is shown below.

Iterative Hard Thresholding

Input: Compressed sensing matrix A, measurement vector y, sparsity level k
Initial values: $\hat{x}_0 = 0, i = 0$
Iteration step: Repeat until the stopping criterion is met
$i := i + 1$
$\hat{x}_i = H_k(\hat{x}_{i-1} + A^T(y - A\hat{x}_{i-1}))$ // Update the estimate of the signal
Output: $\hat{x}_i$

IV. Future works and conclusion

At the time of writing the article, one method of performing the CS on ECG signals was published [9]. Since the use of CS for medicine signals became very popular, future works could be focused on this area. It is possible to continue researching new methods for ECG signal CS but also for other medical signals like blood pressure, etc.… For this purpose, it was created and submitted a Multiwavelet toolbox [22], which could be used to obtain a new basis for the sparse signal reconstruction.

CS as a new method of acquiring signals by using sub-Shannon sampling frequency became popular in recent years. Every month discoveries are being made about this topic and its applications. Throughout this article, the basic theory about signal sparsity and compressed sensing was explained. An example of performing CS on a periodic signal was shown as well as the current applications of CS in the signal processing field. In the end, l_1 minimization and greedy algorithms were outlined as a way of performing sparse signal recovery.

Acknowledgment

The work is a part of the project supported by the Science Grant Agency of the Slovak Republic (No. 1/0413/22).

References

[1] A. C. Gilbert, P. Indyk, M. Iwen, and L. Schmidt, "Recent Developments in the Sparse Fourier Transform: A compressed Fourier transform for big data," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 91–100, Sep. 2014, doi: 10.1109/MSP.2014.2329131.

[2] M. Mozammel, H. Chowdhury, and A. Khatun, "Image Compression Using Discrete Wavelet Transform," *Int. J. Comput. Sci. Issues*, vol. 9, ervenec 2012.

[3] G. Strang, "Wavelet transforms versus Fourier transforms," *Bull. Am. Math. Soc.*, vol. 28, no. 2, pp. 288–305, 1993, doi: 10.1090/S0273-0979-1993-00390-2.

[4] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006, doi: 10.1109/TIT.2006.871582.

[5] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005, doi: 10.1109/TIT.2005.858979.

[6] E. J. Candes and T. Tao, "Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006, doi: 10.1109/TIT.2006.885507.

[7] W. Shi, F. Jiang, S. Liu, and D. Zhao, "Image Compressed Sensing Using Convolutional Neural Network," *IEEE Trans. Image Process.*, vol. 29, pp. 375–388, 2020, doi: 10.1109/TIP.2019.2928136.

[8] J. C. Ye, "Compressed sensing MRI: a review from signal processing perspective," *BMC Biomed. Eng.*, vol. 1, no. 1, p. 8, březen 2019, doi: 10.1186/s42490-019-0006-z.

[9] J. Šaliga, I. Andráš, P. Dolinský, L. Michaeli, O. Kováč, and J. Kromka, "ECG compressed sensing method with high compression ratio and dynamic model reconstruction," *Measurement*, vol. 183, p. 109803, Oct. 2021, doi: 10.1016/j.measurement.2021.109803.

[10] J. Yang, T. Jin, C. Xiao, and X. Huang, "Compressed Sensing Radar Imaging: Fundamentals, Challenges, and Advances," *Sensors*, vol. 19, no. 14, Art. no. 14, Jan. 2019, doi: 10.3390/s19143100.

[11] H. Boche, *Compressed sensing and its applications*. New York, NY: Springer Science+Business Media, 2015.

[12] G. Chen and D. Needell, "Compressed sensing and dictionary learning," in *Finite Frame Theory: A Complete Introduction to Overcompleteness*, Providence, 2016, vol. 73, pp. 201–241. doi: 10.1090/psapm/073/00633.

[13] S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge: Cambridge University Press, 2019. doi: 10.1017/9781108380690.

[14] Y. C. Eldar and G. Kutyniok, Eds., *Compressed sensing: theory and applications*. Cambridge ; New York: Cambridge University Press, 2012.

[15] J. Šaliga, O. Kováč, and I. Andráš, "Analog-to-Information Conversion with Random Interval Integration," *Sensors*, vol. 21, no. 10, Art. no. 10, Jan. 2021, doi: 10.3390/s21103543.

[16] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, 2013th edition. New York: Birkhäuser, 2013.

[17] Q. Lyu, Z. Lin, Y. She, and C. Zhang, "A comparison of typical ℓp minimization algorithms," *Neurocomputing*, vol. 119, pp. 413–424, 2013, doi: 10.1016/j.neucom.2013.03.017.

[18] mahdi Khosravy, Ed., *Compressive sensing in healthcare*. Waltham: Elsevier, 2020.

[19] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993, doi: 10.1109/78.258082.

[20] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, květen 2009, doi: 10.1016/j.acha.2008.07.002.

[21] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 265–274, listopad 2009, doi: 10.1016/j.acha.2009.04.002.

[22] J. Kromka, O. Kováč, and J. Šaliga, "Multiwavelet toolbox for MATLAB," accepted for the conference RADIOELEKTRONIKA 2022.

CNN Approach for Parkinson's Disease Detection from Voice recordings on the ItalianPVS dataset

[1]*Máté HIREŠ (3rd year),*
Supervisor: [2]*Peter DROTÁR*

[1,2]Dept. of Computers and Informatics, FEI TU of Košice, Slovak Republic

[1]mate.hires@tuke.sk, [2]peter.drotar@tuke.sk

***Abstract*—This work provides a deep learning-based approach for Parkinson's disease (PD) detection from recordings of sustained vowels. We analyze six voice-specific augmentation techniques to improve the model accuracy further. To evaluate the proposed approach, we utilized the Xception model on the vowel subset of the Italian Parkinson's voice and speech dataset. Our model was able to achieve accuracies beyond 90% without considering any augmentations; we were able to achieve 100% accuracy in all cases while using the augmentations.**

***Keywords*—convolutional neural network, Parkinson's disease, spectrogram, augmentation, pathological voice**

I. Introduction

Parkinson's disease is the second most common neurodegenerative disorder, which affects about 1% of the population above 60 years. People with PD can develop both motor- and non-motor dysfunctions [1]. Originally, PD was described as a movement disorder, which can cause rigidity, bradykinesia, or tremor [2]. Lately, the diagnosis process was revised, taking to account the non-motor disorders as well [3], which has been given more attention in the past years [4]. It was found out that people with PD may exhibit various speech disorders, which are connected to bradykinesia and rigidity. The speech of these patients is soft, monotonous, often breathy with variable volume [5].

Human speech can be analyzed in various aspects, such as phonation, articulation, prosody, linguistics, which can all be affected by PD [6]. However, in this work, we only analyze the phonatory aspects of the human voice, which can be measured in sustained phonation of vowels [7].

The computerized detection of PD got investigated in the past years. This process does not require trained personnel to perform the diagnosis; it's relatively fast and cheaper than the traditional diagnosis methods. Most often, the automatic detection of PD includes the analysis of handwriting [8], gait [9] and speech [10]. Gómez-García et al. [11], [12] proposed a review on automatic analysis of speech disorders, including dysarthria, which is caused by PD.

II. Dataset

We utilized the Italian Parkinson's voice and speech (ItalianPVS) database in this work. It contains recordings from 22 subjects from the healthy control group (HC), with no reported speech disorder, and 28 people with Parkinson's disease. All the samples were recorded at 16 kHz frequency. One recording session contains the following tasks: (1) 2 readings of a phonemically balanced text; (2) execution of the syllables /pa/ and /ta/; (3) 2 repetitions of the sustained vowels /a/, /e/, /i/, /o/, /u/; (4) reading of phonemically balanced words and phrases. The detailed description of the dataset can be found in [13].

III. Experiments and Results

A. Convolutional neural networks

Convolutional neural network (CNN) is a deep neural network with a high number of hidden layers between the input and output layer. CNNs use convolution operation in place of matrix multiplication [14]. The goal of the convolutional layers is to learn certain patterns in the grid-like input data, such as images. This is done by a feature extracting process, which is done automatically, no manual feature selection is needed, like in case of the traditional machine learning approaches [15]. The more layers in the CNN, the more complex features can be extracted from the data.

B. Data pre-processing

We present the audio samples in image format by transforming them into spectrograms. The audio-to-image conversion is done by first applying the short-time Fourier transform (STFT) to the original audio data. Using the STFT, the time and frequency components of the data are both preserved. The converted data is then displayed as log-frequency power spectrograms.

***Data augmentation**:* We consider six voice-specific augmentations in this study: (1) a band-pass filter is applied to filter out the frequencies outside the interval [500, 1500]; (2) a random noise is added to the audio, so samples from other environments can also be processed; (3) the frequency components are randomly shifted upwards; (4-5) the original recordings are sped-up and slowed-down by a random rate to ensure, the model does not fit to the speed of the samples; (6) the audio signals are randomly moved to the right by a random amount, so the model does not fit to the time localization of the samples.

C. Experimental setup

We utilized the Xception model [17] to evaluate the performance of our proposed approach. This model was pretrained on the ImagNet dataset [18]. With this knowledge, we further train the model on the target ItalianPVS dataset while updating the weights of all the layers. In addition, the

TABLE I
ACHIEVED RESULTS ON THE ITALIANPVS DATASET WITH AND WITHOUT APPLYING AUGMENTATIONS

TASK	WITHOUT AUGMENTATIONS			WITH AUGMENTATIONS			COMPARISON (ACC)
	ACC (%)	SP (%)	SE (%)	ACC (%)	SP (%)	SE (%)	Tripathi et al. [16]
Vowel /a/	95.45 ± 13.64	98.33	92.00	100.00	100.00	100.00	76.00
Vowel /e/	97.50 ± 5.34	95.00	100.00	100.00	100.00	100.00	76.40
Vowel /i/	95.45 ± 6.10	96.67	94.00	100.00	100.00	100.00	72.00
Vowel /o/	92.50 ± 16.01	85.00	100.00	100.00	100.00	100.00	72.40
Vowel /u/	95.83 ± 12.50	91.67	100.00	100.00	100.00	100.00	72.00

classification block of the Xception model is replaced by a block of three Dense layers with 128 neurons and a one neuron Dense layer with a Sigmoid activation since we have a binary classification problem. For model optimization, we use the mini-batch stochastic gradient descent (SGD) algorithm with a 0.0005 learning rate, which was chosen experimentally. We use 10-fold cross-validation for model validation.

We artificially extend the training data by augmenting the original audio data to prevent the model overfitting. We apply the augmentations mentioned in Section III-B before converting the audio data into image format. Since the default input size of the Xception model is 299×299, we also resize the spectrograms to match this size.

D. Results

In this study, we only consider the vowel subset of the ItalianPVS dataset. The experiments were run separately for each vowel.

Table I contains the results achieved on each vowel subset. We evaluate the performance in terms of accuracy (ACC), specificity (SP), and sensitivity (SE).

As the results reflect, the choice of the data representation was appropriate enough to achieve accuracies beyond 90%. By further applying the proposed augmentation techniques, we achieved 100% accuracy in all five cases.

Tripathi et al. [16] evaluated a 1D-CNN model for the detection of PD on the vowel subset of the ItalianPVS dataset. Other works did not consider the vowel subset of the ItalianPVS dataset. The comparison of the achieved results on the vowel subset of the given dataset is presented in Table I.

Since only Tripathi et al. published their results on the ItalianPVS dataset's vowel subset, no objective comparison can be made.

IV. CONCLUSION

We have demonstrated a deep CNN-based approach for PD detection in this work using log-frequency spectrograms as input images. We further improved the method by using voice-specific data augmentations. This approach applied on the vowel subset of the ItalianPVS dataset enabled to boost the prediction accuracy to 100%. Nonetheless, the ItalianPVS database is a relatively small and unbalanced set; therefore, the model cannot be generalized, and we can not expect the approach to work in every situation. Furthermore, there is a need for a massive amount of training data to evaluate the model accurately. Testing the approach on other datasets would also provide more objective results.

Future research may validate the proposed approach on other datasets. Another possible research would be to consider also the words and sentences subset, respectively.

ACKNOWLEDGMENT

This work was supported by the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic and the Slovak Academy of Sciences under contract VEGA 1/0327/20 and by the Slovak Research and Development Agency under contract No. APVV-16-0211.

REFERENCES

[1] O.-B. Tysnes and A. Storstein, "Epidemiology of Parkinson's disease," *Journal of neural transmission*, vol. 124, no. 8, pp. 901–905, 2017.
[2] D. J. Gelb, E. Oliver, and S. Gilman, "Diagnostic criteria for Parkinson disease," *Archives of neurology*, vol. 56, no. 1, pp. 33–39, 1999.
[3] R. B. Postuma, D. Berg, M. Stern, W. Poewe, C. W. Olanow, W. Oertel, J. Obeso, K. Marek, I. Litvan, A. E. Lang *et al.*, "Mds clinical diagnostic criteria for Parkinson's disease," *Movement disorders*, vol. 30, no. 12, pp. 1591–1601, 2015.
[4] P. J. Garcia-Ruiz, K. R. Chaudhuri, and P. Martinez-Martin, "Non-motor symptoms of Parkinson's disease a review… from the past," *Journal of the neurological sciences*, vol. 338, no. 1-2, pp. 30–33, 2014.
[5] M. Trail, C. Fox, L. O. Ramig, S. Sapir, J. Howard, and E. C. Lai, "Speech treatment for Parkinson's disease," *NeuroRehabilitation*, vol. 20, no. 3, pp. 205–221, 2005.
[6] S. Skodda, W. Visser, and U. Schlegel, "Short-and long-term dopaminergic effects on dysarthria in early Parkinson's disease," *Journal of Neural Transmission*, vol. 117, no. 2, pp. 197–205, 2010.
[7] A. M. Goberman, "Correlation between acoustic speech characteristics and non-speech motor performance in Parkinson disease," *Medical science monitor*, vol. 11, no. 3, pp. CR109–CR116, 2005.
[8] M. Gazda, M. Hireš, and P. Drotár, "Multiple-fine-tuned convolutional neural networks for Parkinson's disease diagnosis from offline handwriting," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–12, 2021.
[9] P. Ghaderyan and S. M. Ghoreshi Beyrami, "Neurodegenerative diseases detection using distance metrics and sparse coding: A new perspective on gait symmetric features," *Computers in Biology and Medicine*, vol. 120, p. 103736, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482520301189
[10] M. Hireš, M. Gazda, P. Drotár, N. D. Pah, M. A. Motin, and D. K. Kumar, "Convolutional neural network ensemble for Parkinson's disease detection from voice recordings," *Computers in biology and medicine*, p. 105021, 2021.
[11] J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. part i: Review of concepts and an insight to the state of the art," *Biomedical Signal Processing and Control*, vol. 51, pp. 181–199, 2019.
[12] ——, "On the design of automatic voice condition analysis systems. part ii: Review of speaker recognition techniques and study on the effects of different variability factors," *Biomedical Signal Processing and Control*, vol. 48, pp. 128–143, 2019.
[13] G. Dimauro, V. Di Nicola, V. Bevilacqua, D. Caivano, and F. Girardi, "Assessment of speech intelligibility in Parkinson's disease using a speech-to-text system," *IEEE Access*, vol. 5, pp. 22 199–22 208, 2017.
[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.
[15] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.
[16] A. Tripathi and S. K. Kopparapu, "CNN based Parkinson's disease assessment using empirical mode decomposition," in *CIKM*, 2020.
[17] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.
[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

Real world use of the elastomagnetic effect

[1]*Šimon GANS (1st year)*
Supervisor: [2]Ján MOLNÁR

[1,2]Dept. of Industrial and Theoretical Electrical Engineering, FEI TU of Košice, Slovak Republic

[1]simon.gans@tuke.sk, [2]jan.molnar@tuke.sk

***Abstract*—This paper serves as a brief introduction into the research area of magnetoelastic sensors. These sensors are based on the Villari effect which is introduced in chapter II. Real world uses of these sensors are introduced in chapter III by describing such sensors in the field of medical research and in vivo monitoring of tissue and artificial bone structures as well as measuring torque, compressive and tensile forces. Directions of further sensor development are estimated in the last chapter.**

***Keywords*—biomedical sensors, force sensing, magnetoelastic senor, magnetostriction, torsional forces, Villari effect**

I. Introduction

Sensors are one of the most important parts of the structures of today's world. They are present in nearly every scientific field that humans take part in in addition to everyday life. Such a broad range of applications puts a lot of conditions that sensors must fulfill to be suitable for that particular use.

There are multiple physical effects that can be exploited for various types of sensors. Some don't require an energy source for their operation, some do. Some can remotely monitor the measured quantities, and some don't. There is a vast majority of characteristics that a sensor must satisfy. When it comes to magnetoelastic sensors they benefit from being able to quantify the measured value from afar and to operate under harsh environmental conditions which makes these sensors ideal for some highly specific use cases.

II. Theoretical basis of the magnetoelastic effect

The effect of the change of length of a ferromagnetic substance when it is exposed to an external magnetic field has been first observed by James Joule in 1842. He did the experimental work with a system of levers that were attached to one end of an iron rod while the other end was fixed. When he magnetized the iron rod, he observed that the dimensions of the rod were growing in the direction of the applied magnetic field. The effect was most prominent at small fields. The more he magnetized the rod, the lesser increase of the length he observed. This is also visualized in the figure (Fig. 1). The x-axis represents the applied magnetic field denoted as H and the y-axis represents the relative lengthening of the rod[1].

This effect is now known as magnetostriction, and it is present in all ferromagnetic materials to some extent. Usually the effect observed is quite small and the change in dimensions is in the range of micrometers per meter. For example the magnetostriction coefficient in the 1-1-1 crystallographic direction of iron (in the direction of the body diagonal of the crystal lattice cube) is around 20 micrometers per meter[1].

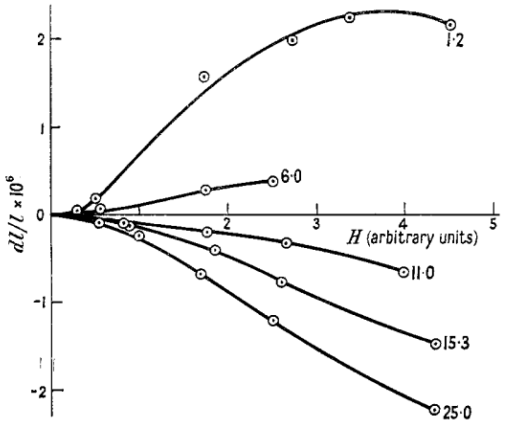


Fig. 1: The original measurements of the magnetostriction of iron done by Joule in 1842. The numbers near the graphs represent the applied mechanical tension on the iron rod in the direction of magnetization. It is clear that it plays a big role in the lengthening of the rod[1].

The figure shown below (Fig. 2) illustrates a simplified model of a ferromagnetic material in which magnetostriction emerges. Without the presence of an external magnetic field H, the magnetic domains in which a magnetization vector **M** exists are randomly oriented so that the net magnetization of the material is 0. However when an external magnetic field **H** is present, the domains try to orient themselves in such a way, that their magnetization vector **M** will be as parallel as possible to the external field **H**. This results in the change of dimensions of the material and when the vectors **M** and **H** are fully parallel this is known as the saturation magnetization[2].

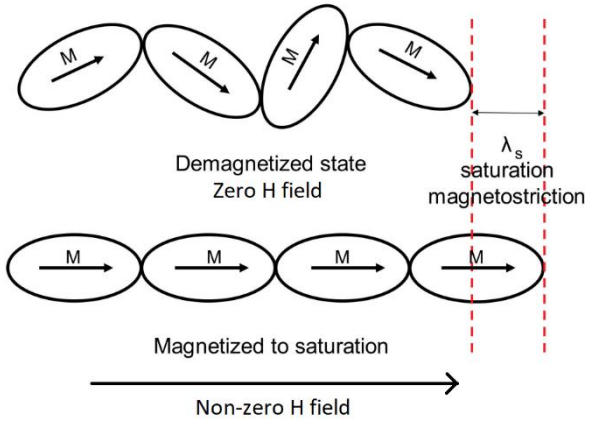


Fig. 2: Simplified explanation that illustrates the emergence of magnetostriction in a ferromagnetic material when an external magnetic field is present[2].

Magnetostriction is also known as the Joule effect. There are many magnetostrictive effects (Wiedemann effect, Mateucci effect) and they all have their inverse effects. In the case of the Joule effect it is the Villari effect which has been observed for the first time by Emilio Villari in 1865. It is defined by the following thermodynamic relation (1), where l is the lengthening of the sample, H is the magnetic field intensity and σ is the stress exerted to the normal of the surface of the material. The Villari effect can be interpreted as the change in magnetization due to external stresses [1].

$$\frac{1}{l}\left(\frac{dl}{dH}\right)_{\sigma} = \left(\frac{dI}{d\sigma}\right)_{H} \quad (1)$$

III. Quantitative relationship of the magnetoelastic effect

The Villari (magnetoelastic) and the Joule (magnetostrictive) effect are both magnetomechanical effects because they both show that the magnetic and mechanical properties of ferromagnetic materials are closely related. The mechanical part is expressed by a presence of external stress acting on the material. The magnetic part is expressed by the material's permeability which is the most widely used magnetic characteristic. The Villari effect can be formulated by the change of permeability due to stress by (2), where λ_{ms} is the magnetostriction coefficient at magnetic saturation, B_S is the magnetic flux density at magnetic saturation, μ is the magnetic permeability of the material when no mechanical stress is acting on it and σ is the mechanical stress[3].

$$\Delta\mu = \frac{2\lambda_{ms}}{B_S^2}\mu^2\sigma \quad (2)$$

Formula (2) was derived by only considering the magnetoelastic and magnetic energy of a ferromagnetic material by putting an equality sign between them. This neglects other energy types, but it is still a pretty good representation of the effect of stress on a ferromagnetic material[4].

Throughout the years there have been many different models of the Villari effect based on different models of ferromagnetic materials like the Jiles-Atherton model ([5]) and the piezomagnetic constant ([6]). Both methods yield good results in their specific use cases.

IV. The real-world use of the magnetoelastic effect

A. The use of magnetoelastic sensors in healthcare

Magnetoelastic sensors are not only used for measuring applied mechanical forces. They can also detect stress, strain, torque, they can continuously measure various chemical indexes like the concentration of ammonia and carbon dioxide, pH, glucose concentrations and the measurement of parameters like the degradation rate and force distribution in artificial bones. Strictly speaking they are prominent in the healthcare system. The main reason for that is that they can be used as wireless passive sensors which means that they can be inserted into living tissue without the need of wired connections to the sensor and without any requirements for batteries or other power sources for their operation[7].

Sensors of this type in this area are mostly composed of amorphous ferromagnetic ribbons. Amorphous ferromagnetic alloys are a special kind of materials which are produced by rapid heat quenching via the melt spin extrusion technique. This creates a nonperiodic amorphous crystal lattice in their internal structure. A typical commercial material used for these sensors is Metglas®2826 MB. The basic characteristics of such materials is that they are soft from a magnetic perspective, they have a high magnetoelastic coupling factor, low hysteresis, low coercivity, high permeability and due to their means of production they are also cheap[7].

As it was mentioned when a magnetic field is acting on a ferromagnetic material it undergoes structural changes that also change the dimensions of the material. When a material is excited via a periodic signal the material will vibrate accordingly. Constructive acoustic waves at the mechanical resonant frequency of the material will introduce a state in which the maximum vibration amplitude is present which in turn will give rise to the maximum permeability of the material at that frequency. Due to internal changes at this state, the flux density variation exhibits a local maximum. This is usually sensed by a pick-up coil in which a corresponding voltage signal is induced[7].

The mechanical resonant frequency of an object which has a rectangular shape is computed via the following formula (3), where l is the length of the sensor material, E is the Young's modulus of the material, ρ is the density of the sensor, v is the Poisson's ratio of the material and the index n denotes that there are multiple resonant frequencies which are integer multiples of the base harmonic frequency which is the frequency when n is equal to one[7].

$$f_n = \frac{n}{2l}\sqrt{\frac{E}{\rho(1-v)^2}} \quad n = 1, 2, 3 \ldots \quad (3)$$

The resonance frequency of the material can be detected by multiple methods. One of them is monitoring the inductance of the pick-up coil if the magnetoelastic material is used for the core of the coil. The relation is shown by formula (4), where L is the inductance of the coil, μ is the permeability of the core material, N is the number of turns of the coil, l is the length of the solenoid and S is the area enclosed by one turn of the coil. The inductance of the solenoid is linearly dependent on the permeability of the core, which is then evaluated[7].

$$L = \frac{\mu S N^2}{l} \quad (4)$$

Another common technique for evaluating the change in permeability is to excite the material with a known waveform and then to analyze the voltage induced in the sensing coil in the frequency domain. The frequency component with the largest amplitude will correspond to the resonant frequency of the magnetic material[7].

A technique is also used in which the excitation coil is operated with a signal with a known waveform which produces a specific voltage waveform in the pick-up coil. When force is applied to the sensor, the variation of the magnetic flux density **B** creates a variation in the pick-up coil's voltage waveform which is proportional to the applied force[7].

Sensors that are designated for sensing gases (like CO_2) are coated with special materials that absorb these substances. The absorption is typically a fast process which is also reversible, and it depends on the concentration of these compounds in the environment. When the sensor coating absorbs some of the gases the distribution of mass changes, which in turn changes the mechanical resonant frequency of the sensor. When the mass increase can be considered small compared to the sensor mass, the shift of the mechanical resonant frequency can be expressed by (5), where Δm is the absorbed mass of the measured substance, M is the mass of the sensor and f is the resonant frequency of the sensor without any additional mass. This means that the amount of substance the coating has absorbed will manifest itself in a shift of the resonant frequency[8].

$$\Delta f = -\frac{f}{2}\frac{\Delta m}{M} \quad (5)$$

Magnetoelastic sensors can be used to measure the viscosity of the environment like for example the state of blood coagulation in the veins. Blood coagulation is a complicated process that depends on many factors, but at one point of the process the viscosity of the blood changes. Because the resonant frequency doesn't only depend on the geometry of the sensor itself but also on the damping forces exerted on the sensor by the environment, the shift in the resonant frequency can be expressed by equation (6), where ρ_s and ρ_1 are the densities of the magnetoelastic material and the liquid it is submerged in (in this case blood) respectively, η is the viscosity of the blood, d is the sensor thickness and f_0 is the resonant frequency of the sensor in vacuum[9].

$$\Delta f = -\frac{\sqrt{\pi f_0}}{2\pi\rho_s d}\sqrt{\eta\rho_1} \quad (6)$$

In addition to that they can be also used for monitoring the state of bone fracture fixation. By analyzing the forces that a healthy tibia bone of a sheep experiences researchers found that a specific combination of tensile and compressive forces is present in bone when walking, which is more closely depicted in the figure below with sensor placement for the measurement of these forces (Fig. 3)[10].

From these examples it is obvious that magnetoelastic sensors have a strong position in the field of healthcare. However because magnetostrictive metals mainly consist of iron, nickel, and cobalt they have poisonous effects on live cells, which is a reason why they need to be coated in a biocompatible substance like TiO_2 or some form of biocompatible silicone before their use for in vivo measurements. The researchers also must consider the heat generated by the human body which affects the sensor performance. The human body also to some extent attenuates the external magnetic fields which the sensor experiences which also must be considered[10].

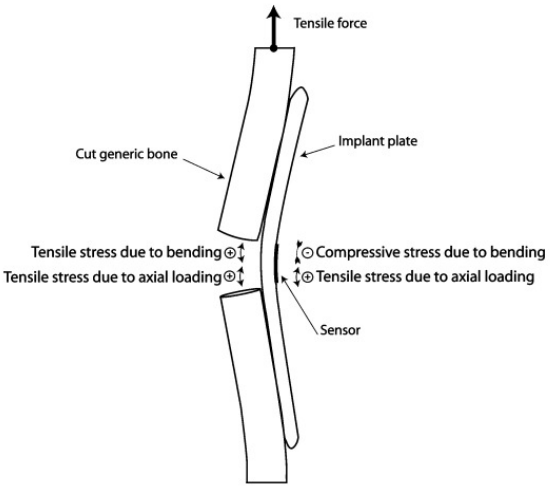


Fig. 3: A magnetoelastic sensor placed in the middle of an implant plate. The strain exerted on the plate gives rise to a proportional strain on the sensor, which changes its magnetization[10].

B. The use of magnetoelastic sensors in the car industry

Sensing torque like for example the torque exerted on a drive shaft by a car engine is utmost important in the design of an automatic transmission. When the torque applied to the wheels is too high or low the transmission should switch gears to make the torque lower or higher respectively. In most applications of such nature the shaft is permanently magnetized, and the presence of torque changes the anisotropy energy that rotates the magnetic domains inside the material which manifest itself by a slight rotation of the magnetic field (Fig. 5). This rotation is linear until a specific magnetic yield point is reached after which the dependence of the rotation on the torque is nonlinear. The figure below (Fig. 4) shows that the sensor is very linear[11].

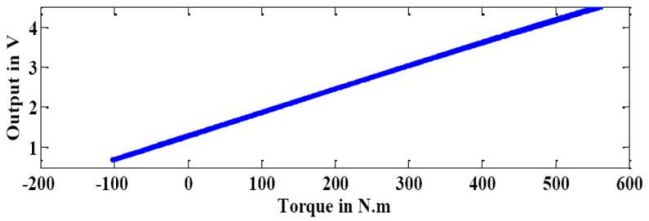


Fig. 4: The output voltage as a function of the applied torque on a torque sensor described in [11]

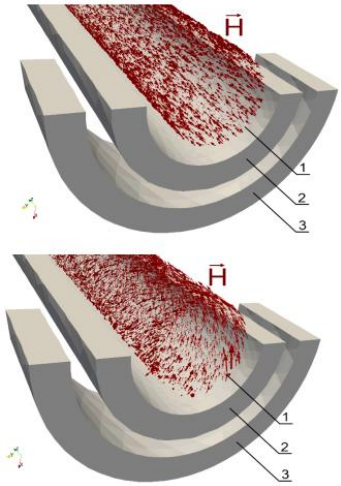


Fig. 5: Visualization of the magnetic field strength distribution inside a ferromagnetic rod manufactured into a driving shaft. The upper image is solved for 0 Nm torque and the lower one for 1.3 Nm torque. A change of the rotation of the magnetic field can be observed. 1-shaft (transparent) 2-magnetizing coil 3-sensing coil. The sensor is more closely described in [12].

C. The use of magnetoelastic sensors for force measurements

One of the most important aspects of magnetoelastic force sensors is their shape. A suitable shape can be defined as one in which by mechanical and magnetic excitation a homogenous mechanic and a homogenous magnetic field exists. In the picture below (Fig. 6) is a magnetoelastic force sensor in which such a homogenous mechanic field does not exist. At this time there are no mathematical analytical models which could describe the depicted sensor operation. Numerical finite element method (FEM) simulations must be used to quantify the sensor characteristics[13].

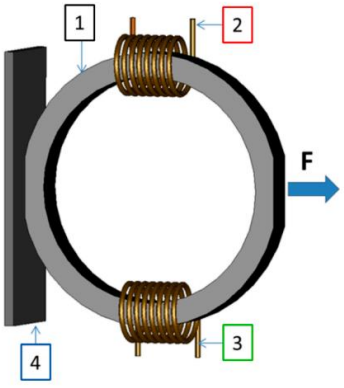


Fig. 6: A ring shaped magnetoelastic force sensor described in [13]. 1-Magnetoelastic core, 2-Excitation winding, 3-Sensing winding, 4-Base plane.

Another common design is the so called “Pressductor” type sensor (Fig. 7). It is characterized by its specific shape which usually consists of multiple single thin magnetic sheets. This type of magnetoelastic force sensor is in production since 1954. Since then continuous development was done especially in the field of materials used. New magnetoelastic sensors are made of mostly amorphous materials of various compositions. The output signal expressed as a function of the applied force is also highly linear (Fig. 8). The linearity is also highly dependent on the drive signal characteristics[14].

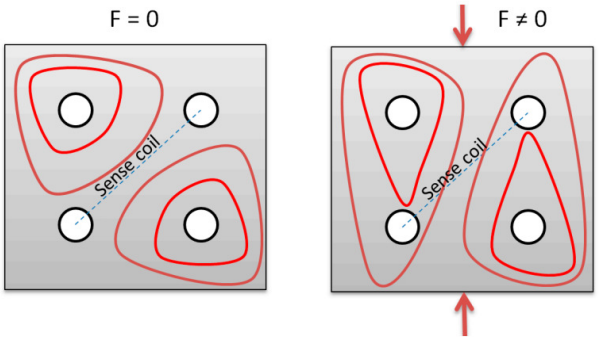


Fig. 7: Typical pressductor shape. When no force is applied no magnetic flux density lines (red curves) intersect the diagonal coil. When a force is present the stress induced anisotropy changes the path of the magnetic flux in the material that then intersects the sensing coil which generates an output voltage[14].

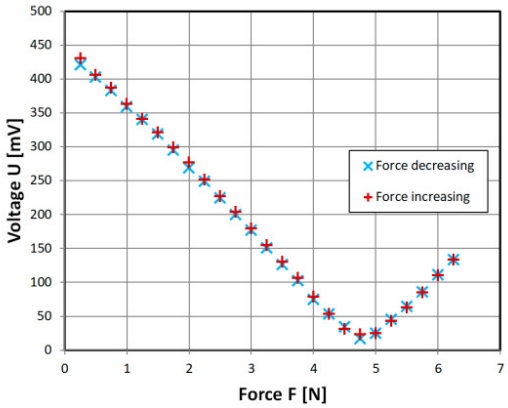


Fig. 8: The output voltage dependence on the applied stress of a pressductor type force sensor. Because the points of decreasing and increasing forces are nearly identical, the sensor has just a small hysteresis error[14].

A common sensor type is shown in the picture below (Fig. 9). It consists of a magnetostrictive rod which has its open ends connected to a ferromagnetic frame that closes the magnetic circuit into a loop. The shape of the sensor makes it quite easy to mathematically describe its operation, but in [15] it is shown that even such geometrically basic sensors yield complicated formulas[15].

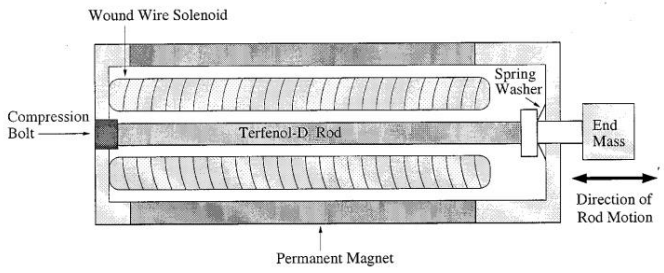


Fig. 9: The cross-section of a magnetostrictive force transducer with a Terfenol-D magnetostrictive rod. Notice that the material is excited via permanent magnets which constrict this sensor to measure dynamic forces only[15].

Magnetoelastic force sensors, especially the pressductor type sensor, is most widely used in the metal industry. It has been used since its development mainly in rolling mills in factories which work with enormous forces. These forces are applied on metals to flatten them into thin electrical sheets. It is one of the only types of sensors which can be used in such an environment because of their extreme mechanical strength, resilience to external magnetic fields and output signal characteristics. There are other areas of application of such sensors, but in them they aren’t nearly as dominant in the market as they are in the metal industry. However, continuous research is being made with respect to their metrological characteristics which in the long run will allow them to be used in multiple areas too like for example in heavy machinery where big forces are present[17][16].

In the present-day strain gauges are more widely used. They usually consist of 4 strain sensitive resistive elements placed on different places of different mechanical structures which detect strain upon their bending. They have a linear characteristic, but their output signal is very weak which is the main advantage of magnetoelastic sensors compared to strain gauges[18].

V. Next directions of the magnetoelastic sensor development

Continual research is being done on both the theory of ferromagnetic materials and the theory of the Villari effect. Even today, ferromagnetic materials are usually modeled via the Jiles-Atherton model of ferromagnetic materials or with some of its variations. Some work has been done in the temperature dependence of ferromagnetic material characteristics by developing formulas that express the parameters of the model as functions of temperature. Also a frequency dependency of the model properties has been investigated. The model of ferromagnetic materials is crucial when developing a force sensor since the structural changes in ferromagnetic materials due to stress must be accurately described[19].

One of the most important parts of the Jiles-Atherton model is the anhysteretic magnetization model of the material. Anhysteretic curves can be measured experimentally but only with great experimental difficulty (even though some easier methods have been lately developed[20]), which is the reason why the curves are usually extracted by optimizing the model parameters first from which then the anhysteretic curve is obtained. Special functions are used to model the anhysteretic magnetization like for example the Langevin function. But it remains unknown whether there exists a general model for all kinds of magnetic materials and a lot more work needs to be done in this area[20].

As it was mentioned, the geometry of the sensor plays a big role in the sensor performance when subjected to mechanical stress. A general formula for an arbitrary sensor shape is unlikely to be developed. The only feasible methods of describing sensor performance are experimental work, which tends to be tedious since magnetoelastic sensors are usually metallic objects that must be manufactured into a specific shape, or finite element method simulations that must be done to sufficient accuracy. Coupling different physics (heat flow in solids, magnetic fields, structural mechanics) into one simulation model introduces a big computation cost, so advances in the development of new and more powerful computers and the optimizations of the finite element method solvers could aid in the better and especially faster development process of magnetoelastic sensing devices[21].

There also has been done a great deal of work in the development of new magnetic materials with better and more desirable magnetic and mechanical characteristics. Recent research has been mainly focused on materials with an amorphous internal structure. There seems to be no currently visible boundary that would stop the development of high-permeability, low electrical conductivity, and high mechanical strength materials[14].

From the current standpoint there are also many other parameters of a physical magnetoelastic sensor that can be optimized to enhance its metrological characteristics which means that the doors in this area of research are wide open for new ideas.

VI. Conclusion

In this paper a brief review of the current state of knowledge in the field of magnetoelastic sensors has been introduced. Some major drawbacks in the development process of new magnetoelastic sensors have been put forth and some improvement paths have been proposed.

Acknowledgment

The authors thank for the financial support from the GRANT FEI 2022 project for the grant specified by the index FEI-2022-82 which helped in the research.

References

[1] Lee, E. W., Magnetostriction and Magnetomechanical Effects, 1955, Rep. Prog. Phys. 18 184, IOP Publishing. Available on the internet: Magnetostriction and Magnetomechanical Effects - IOPscience

[2] Biénkowski, A., Szewczyk R., Magnetostrictive Properties of Mn0.70Zn0.24Fe2.06O4 Ferrite, MDPI materials, 2018, 11(10), 1894. Available on the internet: https://www.mdpi.com/1996-1944/11/10/1894?type=check_update&version=1

[3] Tomčíková, I., Modeling of magnetic field distribution in magnetoelastic force sensor, ReseachGate, 2018, Available on the internet: https://www.researchgate.net/publication/266292425_MODELING_OF_MAGNETIC_FIELD_DISTRIBUTION_IN_MAGNETOELASTIC_FORCE_SENSOR

[4] Hajko, V., Potocký L., Zentko A., Magnetizačné procesy, Alfa, 1. Vydanie, 1982, 63-119-82, strany 78 – 85.

[5] Dapino M, J., Smith R.C., Calking F.T, Flatau A.B., A Coupled Magnetomechanical Model for Magnetostricive Transducers and its Application to Villari-effect Sensors, first published 2012, Journal of Intelligent Materials Systems and Structures, Volume: 13, issue: 11, pages: 737-747 Available on the internet at: https://journals.sagepub.com/doi/abs/10.1177/1045389X02013011005

[6] Deng Z., Nonlinear Modeling and Characterization of the Villari Effect and Model-guided Development of Mangetostrictive Energy Harvesters and Dampers, Dissertation Thesis, The Ohio State University, 2015. Available on the internet at: https://etd.ohiolink.edu/apexprod/rws_etd/send_file/send?accession=osu1437607426&disposition=inline

[7] Ren, L, Yu K., Tan Y., Applications and Advances of Magnetoelastic Sensors in Biomediacl Engineering: A Review, MDPI materials, 2019, 12(7): 1135. Available on the internet at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6479581/

[8] Cai Q. Y., Cammers-Goodwin A., Grimes C. A., A wireless, remote query magnetoelastic CO_2 sensor, Journal of Environmental monitoring, January 2001, 2(6):556-60, Available on the internet: https://www.researchgate.net/publication/12036921_A_Wireless_Remote_Query_Magnetoelastic_CO2_Sensor

[9] Puckett, L G., Barrett G., Kouzoudis D., Grimes C., Bachas L. G., Monitoring blood coagulation with magnetoelastic sensors, Elsevier biosensors and bioelectronics, May 2003, 18(5-6):675-81. Available on the internet: https://pubmed.ncbi.nlm.nih.gov/12706578/#:~:text=The%20magnetoelastic%20sensors%20emit%20magnetic,of%20a%20soft%20fibrin%20clot.

[10] Oess N. P., Weisse B., Nelson B. J., Magnetoelastic Strain Sensor for Optimized Assessment of Bone Fracture Fixation, IEEE Sensors Journal, 2009, Vol. 9, No. 8. Available on the internet at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5155915

[11] Pietron G. M., Fujii Y., Kucharski J., Yanakiev D., Kapas N., Hermann S., Development of Magneto-Elastic Torque Sensor for Autoatic Transmission Applications, SAE International Journal of Passenger Cars – Mechanical Sytems, 2013, 6(2), pages 529-534. Available on the internet at: https://www.sae.org/publications/technical-papers/content/2013-01-0301/

[12] Ostaszewska-Lizewsla A., Nowicki M., Szewczyk R., Malinen M., A FEM-Based Optimization Method for Driving Frequency of Contactless Magnetoelastic Torque Sensors in Steel Shafts, 2021, MDPI materials. Available on the internet at: https://www.mdpi.com/1996-1944/14/17/4996

[13] Lizewska A. O., Szewczyk R., Raback P., Malinen M., Modelling the Characteristics of Ring-Shaped Magnetoelastic Force Sensor in Mohri's Configuration , MDPI sensors, 2020, 14(17), 4996. Available on the internet at: https://www.mdpi.com/1424-8220/20/1/266

[14] Nowicki, M., Tensductor – Amorphous Alloy Based Magnetoelastic Tensile Force Sensor, 2018, 18(12), 4420, MDPI sensors. Available on the internet at : https://www.mdpi.com/1424-8220/18/12/4420

[15] Dapino, M- J., Smith R. C., Flatau A. B., Structural Magnetic Strain Model for Magnetostrictive Transducers, IEEE Transactions on magnetics, Vol.36, No.3, 2000. Available on the internet at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=846217

[16] Otosson Martin, Fifty years of Pressductor, 2004, ABB Review 3/2004. Available on the internet at: https://library.e.abb.com/public/dca7e44a17500667c1256f09004caa48/45-49%203M879.pdf

[17] Bienkowski A., Szewczyk R., Salach J., industrial Application of Magnetoelastic Force and Torque Sensors, 2010, Acta Physica Polonica A, No. 5, Vol. 118. Available on the internet at: http://przyrbwn.icm.edu.pl/APP/PDF/118/a118z5p120.pdf

[18] Kamble, V. A., Shide V. D., Kittur J. K., Overview of Load Cells, Journal of Mechanical and Mechanics Engineering, 2020, Volume 6, Issue 3. Available on the internet at: https://www.researchgate.net/publication/355370308_Overview_of_Load_Cells

[19] Walker J. M., Measurement, and modeling of the anhysteretic magnetization of magnetic cores for temperature and frequency dependent effects, 2007, Dissertation thesis. Available on the internet at: https://digitalcommons.usf.edu/etd/2401/

[20] Nowicki M., Anhysteretic Magnetization Measurement Methods for Soft Magnetic Materials, MDPI materials, 11(10), 2021, 2018. Available on the internet at: https://www.mdpi.com/1996-1944/11/10/2021

[21] Belahcen A., Magnetoelastic Coupling in Rotating Electrical Machines, IEEE Transactions on magnetics, Vol. 41, No. 5, 2005. Available on the internet at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1430925

Overview of semantic segmentation applications in medical imaging

[1]*Ing. Dávid Jozef Hreško,*
Supervisor: [2]*doc. Ing. Peter Drotár PhD.*

[1,2]Department of Computers and Informatics, FEI TU of Košice, Slovak Republic

[1]david.jozef.hresko@tuke.sk, [2]peter.drotar@tuke.sk

***Abstract*—In recent years, deep learning techniques were one of the most groundbreaking innovations used in automated medical image analysis. These computer-aided diagnostic systems equipped with deep learning models proved to be accurate enough to complement human judgement in real world medical scenarios. Currently, there are various applications of deep learning models, but semantic segmentation is one of the most promising area of ongoing research to precisely recognize possible regions of interests for medical diagnosis and treatment. It also had significant impact on daily medical tasks, such as radiological diagnostics, computer-aided detection and surgery simulation. The underlying article provides a brief overview of state-of-the-art deep learning architectures used primarily in semantic segmentation tasks, concrete use-cases of semantic segmentation in medical scenarios, together with short description of publicly available medical datasets used as benchmarks to validate provided solutions. Moreover, the article focuses on the previous experiments and obtained results, which can be used as potential research directions in medical image semantic segmentation.**

***Keywords*—convolutional neural networks, deep learning, healthcare, medical imaging, semantic segmentation**

I. Introduction

In the area of medical imaging, image segmentation represents the task, where the main goal of medical specialists is to precisely identify concrete areas of various anatomical or pathological structures, defined by biomedical images composed of series of elementary blocks known as pixels (2D domain) or voxels (3D domain). This task is primarily required in applications, such as radiotherapy, computer-aided diagnosis and investigation of multiple severe diseases. To be more concrete, these tasks typically include brain structures segmentation in MRI scans [1], [2], lungs segmentation in chest X-Ray [3], abdominal segmentation in CT scans [4], [5], [6], retinal vessel segmentation [7], breast mass segmentation [8], etc. This wide variety of mentioned use cases, results in fact that medical segmentation is currently active area of scientific research and thus provides opportunities for further improvements.

One of the most promising results of recent years was brought by deep learning architectures, which have achieved impressive improvements and set the new standard for computer-aided medical segmentation. The turning point came in 2015, when papers describing fully convolutional networks (FCN) [9] and U-Net [10] architecture were published and thus provided a backbone structures for later designed models.

Additionally, the number of publicly available medical datasets used for models training and validation grows rapidly due to increased attention around deep learning architectures, which helps researchers to further improve existing architectures and reach performance comparable to human expertise.

To sum up, the main contributions of this paper are:

- We provide structured brief of the most recent state-of-art deep learning models used for medical image segmentation to identify the key concepts behind these architectures.
- We analyze actual applications of medical image segmentation, where one can easily obtain an overall picture of the problems and methodologies in this research area.
- We collect list of commonly used medical datasets for training and validation of deep learning architectures in medical imaging.

II. State-of-art-architectures

This section describes the key concepts behind the state-of-art deep neural network architectures and provides brief overview of their applications in the area of medical imaging.

A. nnU-Net

Based on the latest statistics, domain of medical image segmentation is currently dominated by deep convolutional neural networks (CNNs). However, authors of nnU-Net [11] hypotetize, that some of the published architectural modifications and general improvements are overfitted to specific problems or are compared to sub-optimal reimplementations of the state-of-the-art algorithms. Based on these assumptions they proposed the nnU-Net (no-new-Net) framework to overcome mentioned bottlenecks.

The key concept of the nnU-Net relies on a set of three basic U-Net modules, which are similar to originally proposed version. To be more specific, nnU-Net contains:

- 2D U-Net - In the context of 3D medical image segmentation, it appears to be as sub-optimal solution, because information along the z-axis cannot be considered, but training speed is more favorable over 3D based architectures.
- 3D U-Net - This configuration seems to be ideal, but in the reality limitations caused by available GPU capacity, allows to train this architecture only on image patches.
- U-Net Cascade - To address limitations of a 3D U-Net, authors additionally propose a cascaded model, where a 3D U-Net is trained first on downsampled images. After that, the obtained results are upsampled to the original

size and passed to a 3D U-Net, which is trained on patches at full resolution.

These modules are trained from the scratch on same configurations and require five-fold cross validation during validation stage. To train neural networks, authors use combination of dice and cross entropy loss function together with Adam optimizer with gradual decreasing of learning rate, where loss function is defined as:

$$\mathcal{L}_{total} = -\frac{2}{|K|}\sum_{k\in K}\frac{\sum_{i\in I} u_i^k v_i^k}{\sum_{i\in I} u_i^k + \sum_{i\in I} v_i^k} + \sum_{i=1}^{N_c} y_i \log(\hat{y}_i) \quad (1)$$

Additionally, authors also focused on steps where much of the models performance can be gained or possibly lost. These steps include preprocessing (augmentation, normalization and sampling), inference mechanism (patch-based strategy and ensembling) and potential post-processing in automated manner.

Authors also proved, that proposed method performs competitively on 7 various datasets, achieving the highest mean dice scores for all classes (except class 1 in the BraTS dataset) at the time of manuscript submission.

Currently, the nnU-Net is subject of further research and is actively modified to provide even better results on various datasets. For example, one of the modifications provided by authors of [12], was to combine cropped input images with mix-up technique to overcome baseline nnU-Net in kidney and kidney tumor segmentation task.

B. ResNet

The regular models based on a U-Shaped structure had certain shortcomings. To address these issues, researchers proposed various enhanced architectures, where one of them was Residual Neural Network (ResNet) [13]. It was initially developed for image classification task however, further research proved that combination of residual blocks with U-Shaped models can be also beneficial for image segmentation task. The key concept of ResNet is utilization of residual block to enhance feature representation in deep learning networks with increased number of hidden layers without falling into the problem of vanishing gradient as it was common in regular U-Shaped architectures.

In [14] the authors proposed U-Net with additional residual blocks. The paper investigated the impact of both long and short skip connections on FCN for biomedical image segmentation, paying specific attention to parameter updates at each layer of the network during training process. As the result, they observed that residual connections speed up the convergence of training process and also showed that a very deep architecture with a small number of parameters can reach near-state-of-the-art performance on EM ISBI 2012 dataset without any further post-processing.

C. DenseNet

The authors of [15] proposed another novel deep learning architecture referred as Dense Convolutional Network (DenseNet), which takes away the traditional concept of previously mentioned ResNet architecture and improved network response by advancing a more comprehensive approach. They proposed an architecture, which utilizes simple connectivity pattern to ensure maximum information and gradient flow between layers in the network. Specifically, DenseNet preserve the feed-forward nature, where each layer obtains additional inputs from all preceding layers and passes on its own features to all subsequent layers. The key difference between ResNet and DenseNet architecture is that, DenseNet does not combine features through summation before they are passed into a concrete layer instead, it combines features by concatenation.

A novel hybrid densely connected U-Net abbreviated as H-DenseUNet was proposed in [16] to address the issues connected with insufficient leverage of spatial information in 2D segmentation models and high computational costs required by 3D segmentation models. The proposed architecture consists of a 2D DenseUNet for efficiently extracting intra-slice features and a 3D counterpart for hierarchically aggregating volumetric contexts. The learning process was formulated in an end-to-end manner, where the intra-slice and inter-slice features were jointly optimized through a hybrid fusion layer. H-DenseUNet was evaluated on the data set of the MICCAI 2017 Liver Tumor Segmentation Challenge and 3DIRCADb data set, where it achieved competitive performance and outperformed other state-of-the-arts.

III. Applications of semantic segmentation

In this section we describe latest successful applications of deep learning concepts related to medical image segmentation, where we primarily focus on abdominal part of the human body.

A. Kidneys and tumor segmentation

Kidney cancer is one of the most common genitourinary cancers in adults around the world with the highest mortality rate [12]. Fortunately, many kidney tumors can be discovered in early stage, which is main precondition for appropriate treatment. In the context of computer-aided therapy, an accurate segmentation of kidney and tumor mass from CT scans is the key for exploring the relationship between the tumor and its surroundings, which aids doctors to provide more accurate decisions. However, the manual segmentation is highly time-consuming, because a radiologist needs to mark specific regions in multiple slices just for one patient. Example representation of slice with segmented regions of interests can be seen in the 1.

In recent years, many researchers put their effort towards the automatic segmentation to address mentioned issues of manual segmentation. In particular, novel deep learning techniques have been studied and presented as state-of-art solutions on on medical segmentation competitions, like KiTS19 or KiTS21.

Besides the regular usage of previously mentioned nnU-Net, the authors of [17] enhance this architecture by coarse-to-fine method to achieve the first place and state-of-art segmentation results on KiTS21 challenge. Firstly, they used nnU-Net to obtain the coarse segmentation and crop the kidneys region-of-interest (ROI) as the kidneys were always contained in the original CT image. This step was important, because they were able to further work with smaller image while still retaining the necessary areas to segment. The fine segmentation of kidneys was then obtained from the cropped CT scan by a another nnU-Net model. The last step was to segment tumor and cyst, which was achieved by usage of another two nnU-Net models. The final segmentation was then obtained as combination of all previous segmentation masks.

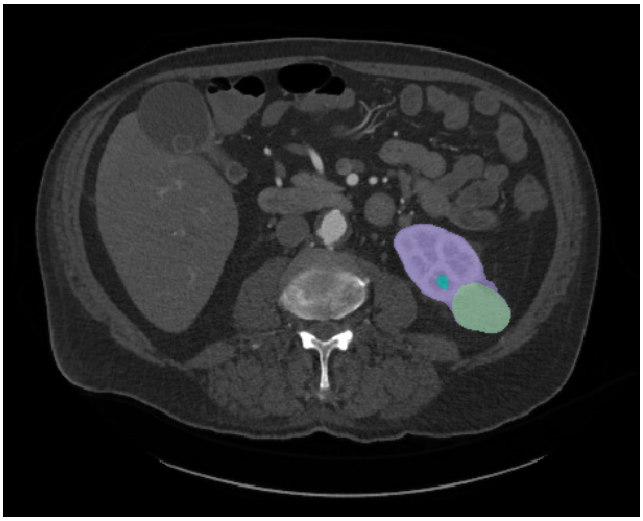


Fig. 1. Axial slice of CT scan with segmented regions of kidney (purple), kidney tumor (green) and cyst mass (blue)

Based on the results from the validation set, which contained 60 cases, they achieved average Dice score 0.9099. Their Dice score for kidney, kidney masses, kidney tumor were respectively 0.9752, 0.8851, and 0.8693.

B. Liver and tumor segmentation

Due to its detoxification function, the liver is one of the most important organs in the human body. Medical experts used to analyze CT scans or magnetic resonance images (MRI) to look for livers anomalies, which are important indicators for various disease diagnosis. It is common case, that tumors of the abdominal part of the body, such as breast or pancreas cancer spread metastases to the liver. From a wider perspective liver cancer is the second most common cause of cancer death and is the sixth most frequent cancer [18].

From deep learning perspective, a fully-automated segmentation of liver and its lesion remains still an open problem, mainly because of inconsistencies between existing datasets, which affects overall models performance. In order to judge the state-of-the-art architectures and compare different methods in a fair way, LiTS challenge was organized in 2017.

Among many of submitted solutions to LiTS challenge, we observed that, the best solutions relied on deep learning architectures. Concretely, the most popular choice was cascaded U-Net, which was very similar to the winning architecture of KiTS challenge. One of the major modifications was integration of residual connections, similar to the ones mentioned in ResNet model. Additionally, to overcome high class imbalance presented in LiTS dataset authors proposed usage of weighted cross entropy as a loss function.

C. Pancreas segmentation

Pancreatic cancer is one of the most lethal malignant tumors, which can be characterized by delayed diagnosis with difficult treatment. Moreover, studies says, that the overall five-year survival rate of patients is less than 9%. Based on these characteristics, we can state that, accurate segmentation of pancreatic cancer plays the key role in early diagnosis and proper treatment. However, the manual segmentation of pancreatic cancer is challenging and time consuming, mainly because of blurred boundaries and small size of pancreatic cancer. Thus, it is necessary to develop automated segmentation algorithms for pancreatic cancer to address this issues [5].

One of the proposed method, which even took place in the top three at Medical Segmentation Decathlon (MSD), was utilization of a fully-supervised uncertainty-aware multi-view co-training strategy. Generalization and robustness of the model was achieved by ensemble method, including three separate 2D pre-trained models, where each model was trained on a different view of image (coronal, saggital and axial). To be more specific, they utilized pre-trained weights of mentioned 2D models to asymmetric kernels in 3D networks. By this way, they were able to gain more spatial information with less computational effort. To obtain even better results they additionally applied a combination of augmentation strategies, such as affine transformations, image flipping and random cropping.

Even this segmentation model was not the absolute winner of MSD challenge, it obtained competitive results to the solution based on nnU-Net architecture. Based on the official challenge leaderboard, a difference between median of the Dice Similarity Coefficient (DSC) was only 0.02 points.

IV. Datasets for medical image segmentation

In this section we provide short overview of abdominal datasets used to train and benchmark mentioned models.

A. KiTS

KiTS (Kidney Tumor Segmentation) dataset [12] includes 300 publicly available CT scans with kidney, kidney tumor and cyst annotations. The patients in this dataset underwent partial or radical nephrectomy for suspected renal malignancy between 2010 and 2020 at either an M Health Fairview or Cleveland Clinic medical center. The kidney and tumor annotations were provided by medical students and laypeople (people who received no training other than a brief sheet with instructions) under the guidance of radiologists and urologic cancer surgeons.

B. LiTS

LiTS (Liver Tumor Segmentation) dataset [19] contains 201 CT cases with liver and liver tumor annotations available for training and testing purposes. The cases are collected from seven medical centers around the world and the gathered screenings have a variety of primary cancers, including hepatocellular carcinoma, as well as metastatic liver disease derived from colorectal, breast, and lung primary cancers. Annotations of the liver and tumors were performed by experienced radiologists from every clinic site and manually blind reviewed by another group of independent radiologists.

C. MSD

MSD (Medical Segmentation Decathlon) pancreas dataset [20] consists of 421 portal-venous phase CT scans with pancreas and tumor annotations, which were manually labeled by radiologists. The data was acquired by Memorial Sloan Kettering Cancer Center located in New York, US. The patients in this dataset underwent resection of pancreatic masses, including intraductal mucinous neoplasms, pancreatic neuroendocrine tumors, or pancreatic ductal adenocarcinoma.

V. Conclusion

In this paper we provide short overview of current state-of-art deep learning architectures used in the area of medical image segmentation, so one can easily observe common patterns and methodologies, which can be subjects for further research and possible improvements. We also analyze applications of these models in segmentation task, specifically we focus on abdominal part of human body. Selected applications and deep learning architectures are tightly coupled with training and benchmarking process, which includes various medical datasets. These datasets, which were also used on various medical segmentation challenges, are also shortly described in this paper.

From an overall perspective, this paper can be used as starting point for future research of medical image segmentation in abdominal part of human body.

Acknowledgment

The responsibility of the research content is on primary author and does not represent any view from funding authorities.

References

[1] L. Dora, S. Agrawal, R. Panda, and A. Abraham, "State-of-the-art methods for brain tissue segmentation," *IEEE Reviews in Biomedical Engineering*, vol. 10, 2017.

[2] C. Han, L. Rundo, R. Araki, Y. Nagano, Y. Furukawa, G. Mauri, H. Nakayama, and H. Hayashi, "Combining noise-to-image and image-to-image gans: Brain mr image augmentation for tumor detection," *IEEE Access*, vol. 7, 2019.

[3] F. Munawar, S. Azmat, T. Iqbal, C. Gronlund, and H. Ali, "Segmentation of lungs in chest x-ray image using generative adversarial networks," *IEEE Access*, vol. 8, 2020.

[4] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P. A. Heng, "H-denseunet:hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, 2018.

[5] X. Chen, Z. Chen, J. Li, Y. D. Zhang, X. Lin, and X. Qian, "Model-driven deep learning method for pancreatic cancer segmentation based on spiral-transformation," *IEEE Transactions on Medical Imaging*, vol. 41, 2021.

[6] J. Ma, Y. Zhang, S. Gu, C. Zhu, C. Ge, Y. Zhang, X. An, C. Wang, Q. Wang, X. Liu, S. Cao, Q. Zhang, S. Liu, Y. Wang, Y. Li, J. He, and X. Yang, "Abdomenct-1k: Is abdominal organ segmentation a solved problem?" 2021.

[7] X. Guo, C. Chen, Y. Lu, K. Meng, H. Chen, K. Zhou, Z. Wang, and R. Xiao, "Retinal vessel segmentation combined with generative adversarial networks and dense u-net," *IEEE Access*, vol. 8, 2020.

[8] H. Li, D. Chen, W. H. Nailon, M. E. Davies, and D. Laurenson, "Dual convolutional neural networks for breast mass segmentation and diagnosis in mammography," *IEEE Transactions on Medical Imaging*, vol. 41, no. 1, 2021.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, p. 3431–3440.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Springer*, 2015.

[11] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, and K. H. Maier-Hein, "nnu-net: Self-adapting framework for u-net-based medical image segmentation," *Nature methods*, vol. 18, 2020.

[12] M. Gazda, P. Bugata, J. Gazda, D. Hubacek, D. J. Hresko, and P. Drotar, "Mixup augmentation for kidney and kidney tumor segmentation," 2021.

[13] M. Z. Khan, M. K. Gajendran, Y. Lee, and M. A. Khan, "Deep neural architectures for medical image semantic segmentation: Review," *IEEE Access*, vol. 9, 2021.

[14] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections inbiomedical image segmentation," *Springer*, 2016.

[15] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[16] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P. A. Heng, "H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE Transactions on Medical Imaging*, vol. 37, pp. 2663 – 2674, 2018.

[17] Z. Zhao, H. Chen, and L. Wang, "A coarse-to-fine framework for the 2021 kidney and kidney tumor segmentation challenge," 2021.

[18] P. B. et al., "The liver tumor segmentation benchmark (lits)," 2019.

[19] J. Ma, Y. Zhang, S. Gu, C. Zhu, C. Ge, Y. Zhang, X. An, C. Wang, Q. Wang, X. Liu, S. Cao, Q. Zhang, S. Liu, Y. Wang, Y. Li, J. He, and X. Yang, "Abdomenct-1k: Is abdominal organ segmentation a solved problem?" 2021.

[20] Y. Zhang, J. Wu, Y. Liu, Y. Chen, W. Chen, E. X. Wu, C. Li, and X. Tang, "A deep learning framework for pancreas segmentation with multi-atlas registration and 3d level-set," *Medical Image Analysis*, vol. 68, 2021.

Particulate Matter and the Methods of Its Measurement: An Overview

[1]*Simona KIREŠOVÁ (1st year)*
Supervisor: [2]Milan GUZAN

[1,2]Dept. of Theoretical and Industrial Electrical Engineering, FEI TU of Košice, Slovak Republic

[1]simona.kiresova@tuke.sk, [2]milan.guzan@tuke.sk

***Abstract*—This paper is an overview of the research on particulate matter (PM), its sources, composition, negative impact on human health, methods and instruments of measurement, as well as a summary of air quality with respect to PM in Slovak republic in recent years. Future development of measuring PM is proposed, namely integrating measuring other quantities along with PM and finding if there is any correlation between them.**

***Keywords*—air pollution, air quality, fine particles, measurement, particulate matter, sensors, ultrafine particles.**

I. Introduction

Air quality is a measure of how clean or how polluted the air is. Measuring air quality is important, as the exposure to air pollutants can cause a myriad of health issues. Chemical substances released from the sources of pollution are subject to atmospheric dispersion, horizontal and vertical transfer, and chemical transformations. Some of them settle on the earth's surface and penetrate the surface and groundwater. From there they can be released back into the air. In any part of the cycle, substances can enter the chemical reactions. Among the substances which are classified as air pollutants are benzo(a)pyrene (BaP), ozone (O_3), heavy metals (such as lead, cadmium, nickel, arsenic, and recently mercury has been added to this category as well), benzene (C_6H_6), sulfur dioxide (SO_2), nitrogen oxides (NO_X) and carbon monoxide (CO) and particulate matter (PM) [1], which is the air pollutant that will be the subject of this paper. Due to the impact of particulate matter on human health, it is important to measure it. It is true that in the recent years, the air quality in Slovak republic has improved significantly compared to the historical precedent. However, despite the fact that the concentrations of air pollutants have decreased, the situation is still far from ideal, and the air quality does not reach a level that would not affect the quality of human health and the environment [2].

II. Particulate Matter

A. Composition of Particulate Matter

Particulate matter is a term that describes a mixture of solid particles and liquid droplets (aerosols) of varying size (Fig. 1) and composition. Particulate matter with aerodynamic diameter smaller than 10 µm (PM10) can be classified into three categories – coarse particles (2,5 – 10 µm), fine particles (smaller than 2,5 µm, or PM2.5) and ultrafine particles (smaller than 0,1 µm, or PM0.1) [3]. Most fine particles follow a core–shell model (Fig. 2), where a source-specific core is covered

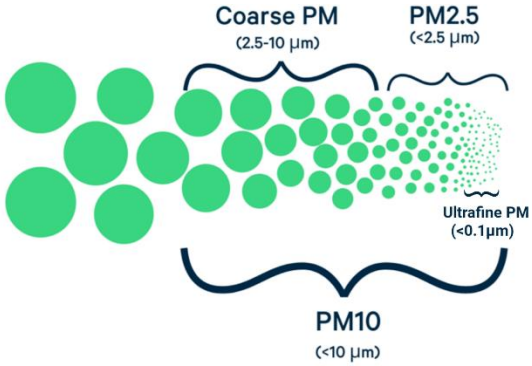


Fig. 1. Particulate matter sizes [4].

with a coating dominating the mass. This coating is of similar composition for most particles. The constituents that made up the shell include sulfate, nitrate, ammonia, and organics: hydrocarbon-like organic aerosol (HOA) or oxidized organic aerosol (OOA). Among the cores of fine particles are crustal cores, biomass cores, cores made up from sea salt, diesel and spark cores (which consist of organic aerosol, black carbons, and metals), as well as cores which arise from new-particle formation (their composition consists of sulfate, ammonium and LVOC/ELVOC – low and extremely low volatility organic compounds). “Core” and “shell” might sometimes mix freely. As for shell, it may include more than one phase (e.g., organic and inorganic) [3][5].

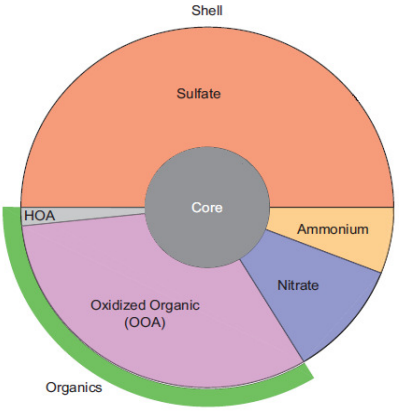


Fig. 2. Core-shell model of fine particles [3].

B. Sources of Particulate Matter

The sources of PM can be either natural or anthropogenic. Among the anthropogenic sources of PM is the residential combustion. Even though residential wood combustion is labelled as a renewable fuel, the combustion of biomass and

coal is currently a source of PM. This is the case especially in winter when residential heating with biomass and coal is more intense [3]. According to [6], between Europe and North America, central Europe has the highest proportion of outdoor PM2.5. This can be traced to residential heating with solid fuels (21% in 2010). In urban areas, among the major sources of PM is road transport [3]. Study [7] found that the emissions from road traffic contributed between 24% and 62% of the hourly average roadside PM10 concentrations in London. As for Slovakia, road transport is the dominant source of air pollution in Bratislava. Most cars in Bratislava cross the D1 motorway bypass from the port bridge towards Žilina and the D2 motorway bypass behind the Lafranconi bridge to Austria and Hungary [1]. Another source of PM are emissions from energy and manufacturing industries. However, effect of industries on the concentration of PM in urban areas is less pronounced than that of road transport [3], with the exception are certain cities close to industrial environments, such as Košice, Slovakia. In Košice-Šaca, there is an industrial complex focused on iron metallurgy, steel, and coke production, which is the dominant industrial source of air pollution. Other industrial sources include cement plants. A secondary source of air pollution in Košice is road traffic [1]. Agricultural, metal processing, construction and mining activities also present a notable contribution to PM10. As for natural sources of particulate matter, this category includes windblown dust (which can be local and desert dust), sea salt aerosols, volcano eruptions and wildfires [3].

C. Health Effects of Particulate Matter

Particulate matter affects negatively primarily the respiratory and cardiovascular systems in children, adults, as well as the elderly (although children and the elderly are more vulnerable to the effects of PM). Among the difficulties caused by high concentrations of PM are decreased lung function, increased respiratory symptoms (e.g., coughing, difficulty breathing, etc.), irregular heartbeat or non-fatal heart attacks. The impact of particulate matter on human health varies with size. Their toxicity increases with smaller size. The smaller the particles, the deeper they penetrate the respiratory system, the stronger their negative effects are. While coarse particles settle in the upper respiratory system (they usually land in the airways), particles with the diameter smaller than 1 µm can reach lung alveoli. Compared to fine particles, ultrafine particles cause more pulmonary infection. Exposure to ultrafine particles induces cough and worsens asthma and is also linked to diabetes and cancer [8] – [14]. There is evidence that being exposed to particulate matter leads to the increased morbidity and mortality, including premature death of individuals with lung or heart problems [15] – [18]. According to the current estimates, the inhalation of fine particles (PM2.5) causes between 2 to 3 million deaths per year [3]. It was estimated in [19] that exposure to PM2.5 in 2017 caused from 3 100 to 6 200 premature deaths in Slovakia. This means that on average, almost 10% of all natural deaths in Slovakia were premature due to exposure to PM2.5 with one premature death reducing life expectancy by 11 years on average. The concentration of indoor and outdoor particulate matter, weather and climate conditions also have an impact on the increased number of confirmed COVID-19 cases and deaths, as was found in study [20] conducted in Milan, Italy, where the elevated values of PM10 and PM2.5 occur frequently. However, it should be noted that while PM concentrations affect new cases of COVID-19, the measures taken to prevent the spread of COVID-19 also affect PM concentrations (the measures result in reduced mobility, which means that less traffic leads to lower concentration of PM) [21].

D. Concentration of Particulate Matter in Slovak Republic

In the years 2010 – 2020, a significantly declining trend in concentration of fine particles (PM2.5) was recorded, which had a positive impact on human health and the quality of life in Slovak republic. Fig. 3 compares the annual average of PM2.5 concentrations and the Average Exposure Indicator for PM2.5 against the EU limit value of the PM2.5 average, which is 20 µg/cm^3 [22][23]. All air quality limit values for PM2.5 and PM10 are shown in Tab. 1, though it is important to note that up until 1st January 2020, the limit value for annual PM2.5 average used to be 25 µg/cm^3 [1]. Evaluation of air quality with respect to hourly averages of PM10 and PM2.5 is shown in Tab. 2. [24]

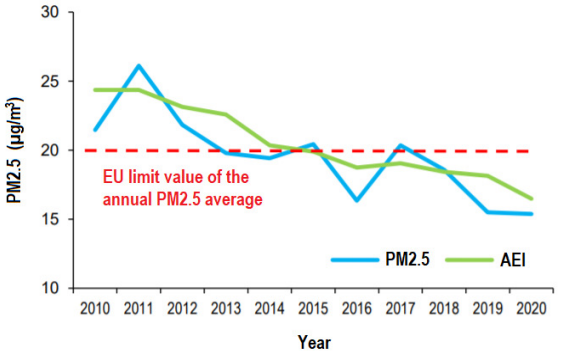


Fig. 3. Annual PM2.5 average (blue), Average Exposure Indicator (ERI), which is a three-year moving average of the annual PM2.5 averages (green) and EU limit value of the annual PM2.5 average (red) [22].

TABLE I. LIMIT VALUES FOR PM10 AND PM2.5 [1].

PM category	Averaging interval	Limit value (µg/m^3)
PM10	24 hours	50
PM10	1 year	40
PM2.5	1 year	20

TABLE II. AIR QUALITY WITH RESPECT TO HOURLY AVERAGES OF PM10 AND PM2.5 [24].

Air Quality	Hourly average of PM10 (µg/m^3)	Hourly average of PM2.5 (µg/m^3)
Very Good	0 - 20	0 – 14
Good	>20 – 40	>14 – 25
Worsened	>40 – 100	>25 – 70
Bad	>100 – 180	>70 – 140
Very Bad	>180	>140

III. MEASURING PARTICULATE MATTER

Particulate matter measurement can be divided into four categories: particle mass, number, surface area and size distribution measurements. Particle mass measurements are generally used when estimating the impact of PM on human health. On the other hand, particle number and size distribution measurements of PM are preferred when investigating the

impact of PM on the climate [25][27]. Different instruments used for measuring PM are shown in Fig. 4, classified [27] by the metric by which PM is measured. Tab. 3 compares those devices.

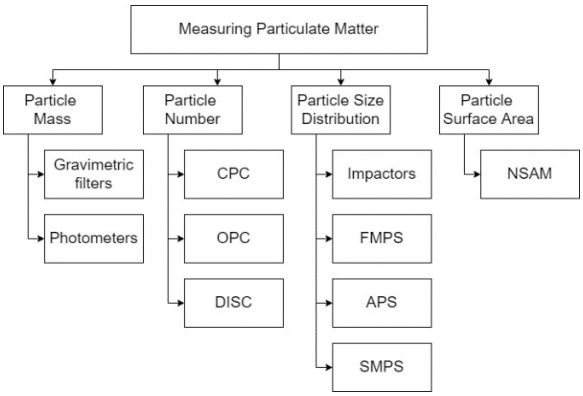


Fig. 4. Instruments used for measuring particulate matter, classified by the metric by which PM is measured, namely particle mass, number, size distribution, and surface area.

A. Measuring Particle Mass

The most common metric to measure particulate matter is by measuring its particle mass. It is a mass of particles per unit of volume (usually measured in μg/cm^3) [27]. The most commonly measured categories of particle mass are PM10 (mass of particles with diameter smaller than 10 μm) and PM2.5 (mass of particles with diameter smaller than 2.5 μm, or fine particles), although there are sensors which can offer other measurements, e.g., PM4 (mass of particles with diameter smaller than 4 μm) or PM1 (mass of particles with diameter smaller than 1 μm) [37]. There are different methods of measurement for particle mass, namely gravimetric, optical and microbalance methods.

Gravimetric methods are based on calculating particle mass from the difference in weight between the samples taken before and after a sampling period. The samples of particulate matter are collected on the filter paper. The sampling period is 15 minutes or more, which means that the identification of fast changes in PM is not possible for this method. However, the samples collected in the filter can be chemically analyzed [25] – [28].

Optical methods of measuring particle mass are based on the way in which particles interact with light. When light beam hits the particle, it is either scattered or absorbed. Measuring light intensity and angle of the scattered beam are the basis of optical measurements of particle mass. Light scattering, light absorption, and light extinction (which is the sum of scattering and absorption) are the three main principles of optical measuring. Based on the principle of light scattering, photometers measure the intensity of scattered light in one or more directions. Photometers collect real time measurements with a frequency of 1 Hz [25] – [27], [29].

B. Measuring Particle Number

Particle number is a metric used for measuring indoor PM, though it is increasingly used in studies about ambient PM as well. Originally particle number was used for measuring and characterizing emissions from vehicles. Particle number is given per a unit of volume (usually particles/cm^3). Most particles making up particle number are smaller particles, specifically 0.5 – 18 μm or smaller (PM0.5), which is why this metric is often measured in studies with focus on ultrafine particles [27].

Optical Particle Counters (OPC), the most used instrument, use a light source (usually a diode laser) to light a sample of particles. A photodetector measures the scattered light. The intensity of the flash allows the particles to be counted and measured [25][26]. The working principle of OPC is shown in Fig. 5.

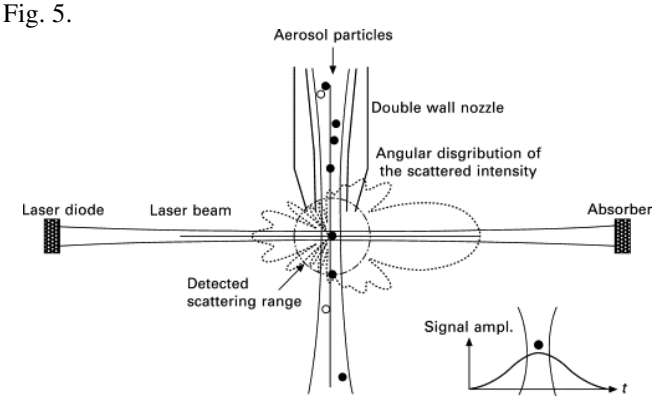


Fig. 5. The working principle of OPC - the particle crossing the illumination zone produces a photoelectric impulse, which relates to the size of the particle being measured [30].

Condensation Particle Counters (CPCs) work similarly to OPCs (working principle is illustrated in Fig. 6). The difference is that they can count particles which are much smaller than those that are counted by OPCs. CPCs condense a solvent (butanol, isopropyl alcohol, or more recently water) onto the surface of particles. This causes the particles to grow to a size where they can pass through the laser beam of CPC and be counted [27][31].

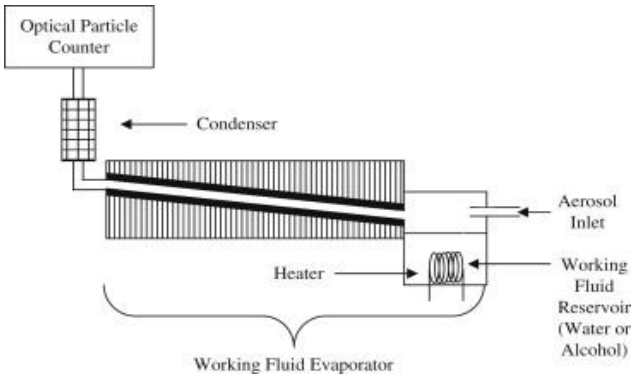


Fig. 6. The working principle of CPC a fluid is condensed on particles that are too small to be detected in order for the particles to grow to a size which is detectable by CPC [31].

The Diffusion Size Classifier (DiSC) can estimate particle number. PM is charged in a unipolar diffusion charger. Then it passes through two electrometer stages – the first stage or "diffusion stage," consists of a stack of stainless-steel screens connected to sensitive electrometers, and the second stage is a HEPA filter connected to an electrometer. Deposition of particles in each of these areas generates a current, from which it is possible to calculate particle number and the average diameter of the particles. One advantage of the DiSC is that it is small, portable, and cheaper than CPCs, which makes it suitable for implementing into embedded systems. The trade-off is that there are certain circumstances in which the DiSC can perform poorly in terms of accuracy, e.g., larger particles can carry more charge which leads to overcounting. [27][32].

TABLE III. AIR QUALITY WITH RESPECT TO HOURLY AVERAGES OF PM10 AND PM2.5 [27].

Measured quantity	Measuring device	Sampling period	Size range	Detection limits	Advantages	Disadvantages
Particle mass	Gravimetric filters	15 min	>150 nm	>10 µg/m^3	Cheap. Simple to deploy indoors.	Processing is time consuming.
	Photometers	1 s	10 nm - 10 µm	0,001 - 200 mg/m^3	Portable, reliable, accurate, relatively cheap.	PM is measured indirectly.
Particle number	CPC	0.5 - 3 s	>15 nm	<1.10^4 - 1.10^6 particles/cm^3	Able to measure rapid atmospheric processes in real time.	Mixing CPC is more complicated to accurately measure sample flow.
	OPC	1 s	0.3 - 20 µm	<1.10^4 particles/cm^3	Lower cost and more portable than CPC.	Unable to measure particles smaller than 0.3 µm.
	DISC	1 s	10 - 700 nm	<5.10^4 - 1.10^6 particles/cm^3	Portable, reliable, robust, and lightweight.	Less accurate than CPS. PM is measured indirectly.
Particle size distribution	Impactors	-	1 - 10 µm	-	Can determine specific chemical characteristics of aerosols.	Unable to measure ultrafine PM. Processing is time consuming.
	FMPS	1 s	5 - 560 nm	-	Much higher time resolution than an SMPS.	Lower resolution of size distribution than SMPS.
	APS	10 s	0.5 - 20 µm	1000 particles/cm^3	Used as a supplement equipment unable to measure the larger sizes.	Unable to measure ultrafine PM.
	SMPS	1 min	2.5 nm - 1 µm	<1 - 1.10^7 particles/cm^3	Provides the highest resolution size distribution of particles.	Much longer sampling period than FMPS.
Particle surface area	NSAM	1 s	20 – 400 nm	<10000 µm^2/cm^3	Portable.	Highly sensitive to environmental factors.

C. Measuring Particle Size Distribution

The size of PM the impact on human health. The smaller the particles are, the deeper they can penetrate into the respiratory tract, the more negatively they affect health [8] – [14]. Therefore, measuring particle size becomes relevant when we want to assess health effects of PM.

Impactors measure particle size distribution. The working principle of impactors is gravimetry with multiple stages. The type of impactor which are used most often are Cascade Impactors, where larger particles are collected. Particles smaller than the cutoff diameter flow through and are collected in the next stage and so on, until the smallest particles are removed in the after-filter. As is the case with measuring particle mass using gravimetric filters, impactors are not suited for measuring the changes in particle size distribution in real time [26][33].

Scanning Mobility Particle Sizer (SMPS) is a spectrometer composed of a particle loader, a classification column, and a series of detectors. Particles which enter the system are first neutralized, after which they enter a Differential Mobility Analyzer (DMA). There the particles are classified according to electrical mobility, after which CPC determined particle concentration at the given size. It is used for measuring the particle size distribution from vehicle emissions and biomass combustion because of its high resolution. It is however slower than Fast Mobility Particle Sizer (FMPS). Minimum scanning time for SMPS is 1 minute [26][27][34].

FMPS uses a technique similar to the one used by SMPS. The difference between them is that FMPS spectrometer uses multiple electrometers for detecting the particles instead of CPC (as is the case for SMPS). This results in measurements within one second, which is suitable when measuring dynamic changes in the particle size distribution [26][35].

Aerodynamic Particle Sizer (APS) is a spectrometer. Its working principle is based on the acceleration of the flow of the PM sample through an accelerating orifice. The size of the particle determines the rate by which it accelerates (i.e., slower acceleration is attributed to the larger particles due to their larger inertia). At the exit, the particles cross two laser beams and scatter their light on an avalanche photodetector. After that, APS converts the light pulse to electric pulse [25] – [27].

D. Measuring Particle Surface Area

Particle surface can be measured by Diffusion Chargers, specifically Nanoparticle Surface Area Monitors (NSAM). The charge transferred from ions to the particles is then measured by the electrometer, through the recording of electrical current. NSAM is the best instrument for measuring the surface of ultrafine particles. It can measure lung-deposited particle surface area concentration, based on lung particle deposition models [25] – [27], [36].

IV. FUTURE DEVELOPMENT

The reason why measuring particulate matter is important was laid out in Chapter II of this paper, which summarized the impact of PM on human health. Measuring and monitoring PM and finding the sources of air pollution is the first step to reducing air pollution and improving the air quality.

Internet of Things (IoT) and wireless sensor technologies (WSN) are both suited for the development of air quality monitoring systems. Integrating multiple sensors within those systems, which can measure multiple physical quantities, has great importance. Those measured quantities can be other air pollutants or meteorological factors like temperature, humidity, and pressure. When measuring indoor air quality, it is relevant to measure, e.g., the levels of carbon dioxide. With implementation of technologies like Wi-Fi, or Bluetooth, it is possible to send the measured data to a web server and display their value via an application or a website. A warning system can be implemented for when the levels of air pollution exceed the limit values. Sensors can be interfaced with development boards like Arduino, ESP8266, ESP32, Raspberry Pi and more. One of the PM sensors which is suitable for interfacing with these development boards is sensor SPS30 [37]. The sensor is based on the principle of laser diffraction. It detects the intensity and angle of a laser beam after it has been struck by a particle which causes the laser beam to scatter. An algorithm then determines mass concentration, number concentration and typical size of the particles. With measuring more physical quantities open more options for future development. Another possibility is to perform a thorough data analysis and determine if there is significant correlation between PM and other physical quantities (e.g., other air pollutants or CO_2 or

meteorological factors like temperature, humidity, pressure) and find out if certain meteorological conditions lessen or exacerbate the effects of PM on human health.

V. Conclusion

This paper serves as an overview of the information about particulate matter, namely the composition of PM, its sources and most importantly, the impact particulate matter has on human health as well as the methods and instruments of measurement. Future development regarding measuring particulate matter is outlined, with the possibility of integrating the measuring of particulate matter along with measuring other physical quantities and finding if there is any significant correlation between them. It is suitable to use IoT or WSN technologies for the development of the measurement systems.

Acknowledgements

The paper has been prepared under the support of Grant FEI-2022-82.

References

[1] M. Kremier et al., "Report on Air Quality in Slovak Republic in 2020", Slovak Hydrometeorological Institute, Air Quality Department, Oct. 2021. Available: <https://www.shmu.sk/File/oko/rocenky/2020_Sprava_o_KO_v_SR_v2.pdf>. (In Slovak).

[2] Z. Jonáček, M. Zemko, J. Szemesová, L. Zetochová, "Quality Improvements of Air Emission Accounts and Extension of Provided Time-Series Focusing on Households Heating", Meteorological Journal, vol. 23, no. 1, pp. 41-46, Jul. 2020. (In Slovak).

[3] R. M. Harrison et al., "Airborne particulate matter: sources, atmospheric processes and health", Royal Society of Chemistry, UK, 2016, ISBN: 978-1-78262-491-2.

[4] D. Smith, (2022, Feb 3). *A Guide to Understanding Particulate Matter (PM)* [Online], Feb. 2020. Available: <https://learn.kaiterra.com/en/air-academy/particulate-matter-pm>.

[5] F. Dominici, Y. Wang, A. W. Correia, M. Ezzati, C. A. Pope III, D. W. Dockerye, "Chemical Composition of Fine Particulate Matter and Life Expectancy In 95 US Counties Between 2002 and 2007", Epidemology, vol. 26, no. 4, pp. 556-564, Jul. 2015.

[6] Z. Chafe et al., "Residential heating with wood and coal: health impacts and policy options in Europe and North America", World Health Organization, 2015. Available: <https://www.euro.who.int/en/publications/abstracts/residential-heating-with-wood-and-coal-health-impacts-and-policy-options-in-europe-and-north-america>.

[7] A. Suleiman, M. R. Tight, A. D. Quinn, "Assessment and prediction of the impact of road transport on ambient concentrations of particulate matter PM10", Transportation Research Part D: Transport and Environment, vol. 49, pp. 301-312, Dec. 2016.

[8] C. Mühlfeld, B. Rothen-Rutishauser, F. Blank, D. Vanhecke, M. Ochs, P. Gehr, "Interactions of nanoparticles with pulmonary structures and cellular responses", The American Journal of Physiology-Lung Cellular and Molecular Physiology, vol. 294, no. 5, L817-L829, May 2008.

[9] M. Kampa, E. Castanas, "Human health effects of air pollution", Environmental Pollution, vol. 151, no. 2, pp. 362-367, Jan. 2008.

[10] D. E. Schraufnagel, "The health effects of ultrafine particles", Experimental & Molecular Medicine, vol. 52, pp. 311-317, Mar. 2020.

[11] S. Y. Kyung, S. H. Jeong, "Particulate-Matter Related Respiratory Diseases", Tuberculosis and Respiratory Diseases, vol. 83, no. 2, pp. 116-121, Mar. 2020.

[12] E. J. Jo et al., "Effects of particulate matter on respiratory disease and the impact of meteorologival factors in Busan, Korea", Respiratory Medicine, vol. 124, pp. 79-87, Feb. 2017.

[13] T. Götschi, J. Heinrich, J. Sunyer, N. Künzli, "Long-term effects of ambient air pollution on lung function: A review", Epidemiology, vol. 19, no. 5, pp. 690-701, Sep. 2008.

[14] R. D. Brook, "Cardiovascular effects of air pollution", Clin Sci (Lond), vol. 115, no. 6, pp. 175-187, Aug. 2008.

[15] C. A. Pope, M. Ezzati, D. W. Dockery, "Fine-Particulate Air Pollution and Life Expectancy in the United States", New England Journal of Medicine, vol. 360, no. 4, pp. 376–386, Jan. 2009.

[16] U. Gehring et al., "Air Pollution Exposure and Lung Function in Children: The ESCAPE Project", Environmental Health Perspectives, vol. 121, no. 11-12, pp. 1357-1364, Jan. 2013.

[17] Y. Zhang, M. He, S. Wu, Y. Zhu, S. Wang, M. Shima, K. Tamura, and L. Ma, "Short-Term Effects of Fine Particulate Matter and Temperature on Lung Function among Healthy College Students in Wuhan, China," International Journal of Environmental Research and Public Health, vol. 12, no. 7, pp. 7777–7793, Jul. 2015.

[18] T. Omori, G. Fujimoto, I. Yoshimura, H. Nitta, M. Ono, "Effects of Particulate Matter on Daily Mortality in 13 Japanese Cities", Journal of Epidemiology, vol. 13, no. 6, pp. 314-322, Nov. 2007.

[19] D. Štefánik, J. Matejovičová, "Mean Exposure to PM2.5 Particles in Slovakia in 2017 and Number of Premature Deaths", Meteorological Journal, vol. 23, no. 1, pp. 31-40, Jul. 2020. (In Slovak).

[20] M. A. Zoran, R. S. Savastru, D. M. Savastru, M. N. Tautan, "Assessing the relationship between surface levels of PM2.5 and PM10 particulate matter impact on COVID-19 in Milan, Italy", Science of The Total Environment, vol. 738, Oct. 2020.

[21] J. Beňo, D. Štefánik, "Impact of COVID-19 Protective Measures on Concentrations of Pollutants in Slovakia - Analysis of The First Month", Meteorological Journal, vol. 23, no. 1, pp. 5-14, Jul. 2020. (In Slovak).

[22] M. Mladý, K. Pukančíková, V. Mináriková, B. Paveleková, "Air and Health in Slovakia within the Years 2010-2020", Meteorological Journal, vol. 24, no. 2, pp. 105-116, Dec. 2021. (In Slovak).

[23] J. Matejovičová et al, (2022, Feb 4). *Brief Overview of the Development of Air Pollution by PM10 Particles in the Years 2005 - 2020 with a Focus on the Assessment of Zones and Agglomerations by Exceeding the Limit Value for PM10 in 2019* [Online]. Available: <https://shmu.sk/File/Infridgement_podklady_fin4opr.pdf>. (In Slovak).

[24] Slovak Hydrometeorological institute. (2022, Feb 4). *Hourly concentrations of air pollutants* [Online]. Available: <https://.shmu.sk/sk/?page=1&id=oko_imis>. (In Slovak).

[25] R. Ravi Krishna., S. M. Shiva Nagendra, M. Saraswati, Diya, "Urban Air Quality Monitoring, Modelling and Human Exposure Assessment", Springer, Singapore, 2020, ISBN: 978-981-15-5510-7.

[26] S. S. Amaral, J. A. De Carvalho Jr., M. A. Martins Costa, C. Pinheiro, "An Overview of Particulate Matter Measurement Instruments", Atmosphere, vol. 6, no. 9, pp. 1327-1345, Sep. 2015.

[27] S. D. Lowther, K. C, Jones, X. Wang, J. D. Whyatt, O. Wild, D. Booker, "Particulate Matter Measurement Indoors: A Review of Metrics, Sensors, Needs, and Applications", Environmental Science & Technology, vol. 53, no. 20, pp. 11644-11656, Oct. 2019.

[28] S. O'Connor et al., "Gravimetric Analysis of Particulate Matter using Air Samplers Housing Internal Filtration Capsules", Gefahrst Reinhalt Luft, vol. 74, no. 10, pp. 403-410, Oct. 2014.

[29] T. Lanki, S. Alm, J. Ruuskanen, N. A. H. Jassen, M. Jantunen, J. Pekkanen, "Photometrically measured continuous personal PM2.5 exposure: Levels and correlation to a gravimetric method", Journal of Exposure Science & Environmental Epidemiology, vol. 12, pp. 172–178, May 2002.

[30] B. Denker, E. Shklovsky, Eds., "Handbook of solid-state lasers: Materials, systems and applications", Cambridge, England, Woodhead Publishing, 2013, ISBN: 978-0-85709-272-4.

[31] R. Kohli, K.L. Mittal, Eds. "Developments in Surface Contamination and Cleaning", Norwich, CT, William Andrew Publishing, 2011, ISBN: 978-1-4377-7883-0.

[32] M. Fierz,C. Houle,P. Steigmeier, H. Burtscher, "Design, Calibration, and Field Performance of a Miniature Diffusion Size Classifier", Aerosol Science and Technology, vol. 45, no. 1, pp. 1-10, Jun. 2011.

[33] L. Bogrese et al., "Assessment of Integrated Aerosol Sampling Techniques in Indoor, Confined and Outdoor Environments Characterized by Specific Emission Sources", Applied Sciences, vol. 11, no. 10, May 2021.

[34] T. Moreno et al., "Using miniaturised scanning mobility particle sizers to observe size distribution patterns of quasi-ultrafine aerosols inhaled during city commuting", Environmental Research, vol. 191, Dec. 2020.

[35] Y. Lin, L. Pham, X. Wang, R. Bahreini, H. S. Jung, "Evaluation of Fast Mobility Particle Sizer (FMPS) for Ambient Aerosol Measurement", Aerosol and Air Quality Research, vol. 21, no. 4, Jan. 2021.

[36] L. Ntziachristos, A. Polidori, H. Phuleria, M. D. Geller, C. Sioutas, "Application of a Diffusion Charger for the Measurement of Particle Surface Concentration in Different Environments", Aerosol Science and Technology, vol. 41, no. 6, pp. 571-580, May 2007.

[37] Sensirion, "Automotive Grade Humidity and Temperature Sensor", SHT3xA-DIS datasheet Dec. 2019. Available: <https://sensirion.com/media/documents/68C84E66/61641CAE/Sensirion_Humidity_Sensors_SHT3xA_Datasheet.pdf>

Aspects of usability in clinical decision support systems

[1]*Oliver LOHAJ (1st year)*
Supervisor: [2]Ján PARALIČ

[1,2]Dept. of Cybernetics and Artificial Intelligence, FEEI TU of Košice, Slovak Republic

[1]oliver.lohaj@tuke.sk, [2]jan.paralic@tuke.sk

***Abstract*—This research paper deals with usability aspects that can be examined in clinical decision support systems based on data analytic models. In the analytical part, we first describe the evolution of usability, then we define it according to dominant perspective and according to international standards (ISO 9241-11). We analyze current state of the art discussing selected studies that deal with different views of usability. As a result of our analysis, we propose particular extensions of the Van Welie model, which reflect some missing usability aspects of the original model, specifically important for CDSS based on data analytic models.**

***Keywords*—decision support systems, data analytic models, medical domain, usability, user interface**

I. Introduction

This research paper is focused on one of many aspects that are very important in developing responsive and usable decision support systems. Usability as a term was firstly used more than 40 years ago, when it replaced the term user friendly [1]. In another study [2], authors talk about dominant perspective that defines usability as quality of use. Quality of use emerges another aspect, that needs to be covered. The quality of use perspective hypothesizes that usability of a product varies on who is using the product, how and for what purpose it is being used. The view of interaction with product that needs to be considered when specifying usability is known as the context of use [3]. According to International Standardized Organization (ISO) and the norm ISO 9241-11, usability is defined as "*the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use*" [4]. This paper is organized as follows. The state of the art is presented in the next section II. while also the research gap is identified. Section III. introduces also the Van Welie's layered model of usability and its suggested extension based on identified research gap. At the end of research paper, important facts from this work are summarized in concluding section IV.

II. Analysis of current state of Knowledge

As mentioned previously, usability is one of very important aspects of decision support systems to be assessed. According to Caicedo [5], usability is very important, if a good quality of developed product needs to be ensured. This applies to terms of interaction. Usability also affects how users feel about the final software product, which applies to the user experience part. The main difference between usability and user experience is that usability deals with how the user interacts with a product, user experience deals about user feels about mentioned interaction. It is an emotional and subjective assessment of how users perceive it and interact with it, whether they are satisfied and glad to use it, or frustrated and demotivated.

In the study from Zytek. A. et.al. [6] authors assess many challenges of usability from more qualitative fields. Authors also ran two different formal user studies. According to them, machine learning, data science and data visualization professionals have provided multitude of algorithms and tools to augment ML (Machine Learning) predictions and address many usability challenges. These algorithms are referred to as ML augmentation tools and, if the domain is chosen correctly, they have great power to improve usability of ML models for decision making. Examples of tools like mentioned include visualization (graphs), local and global explanations [7], cost-benefit analysis, performance metrics and information about historic usage and results of the ML model.

Authors of this study focus on child screening services. According to them [6], research aimed at augmenting ML predictions often focuses on audience of data experts [9] or domain experts in fields such as medicine, because these are more technical and also more data-driven fields. For example, Zhang. et.al. [10] developed a framework for helping data scientists and ML experts interpret and debug ML models. For concrete assessment of usability challenges, authors investigated the need for additional auxiliary information alongside ML predictions through a literature review. Many factors can influence the usability of ML models. The set of factors, found in Table 1, has been identified that decreases the usability in ML models. According to authors of study [6], the main challenges stem from user's different points of view. Not understanding where a model's prediction is coming from is making it difficult for human decision-maker to trust the model *(TR)* and to handle any miscommunication between human and model's output *(DIS)*. Other challenges are coming from the lack of information about real effects of a decision, that would be made according do DSS (decision support system). Model prediction alone usually doesn't explicitly indicate the expected result of decision *(CON),* suggest accountability *(ACC)* or provide some ethical assurances *(ETH).* Lastly, many challenges can arise when the output of model is not direct suggestion of decision, but for the end user, an auxiliary information – output may be confusing *(CT)* or entirely irrelevant *(UT).* In the following Table 1, these 7 challenges of usability are presented, together with mitigating tools.

TABLE 1
USABILITY CHALLENGES ACCORDING TO [6]

Usability Challenge	Code	Mitigating tools
Lack of trust	TR	Local/global explanation, performance metrics, historical predictions, and results
Difficulty reconciling human-ML disagreements	DIS	Local explanations
Unclear consequences of actions	CON	Cost-benefit analysis, historical predictions, and results
Lack of Accountability or protection from it	ACC	Local explanations, performance metrics
Ethical concerns	ETH	Local/global explanations, ML fairness metrics, historical predictions, and results
Confusing/unclear prediction target	CT	Cost-benefit analysis, further analysis of prediction target impact
Unhelpful prediction target	UT	Retraining model with new prediction target

Authors of mentioned study [6] focused on 3 main questions, while important for this work are the first two. First question was about what ML usability challenges exist in the domain of child welfare screenings. Four out of seven usability challenges from Table 1 were identified – Lack of Trust *(TR),* Difficulty reconciling Human-ML disagreements *(Dis),* Unclear prediction target *(CT),* and Concerns about ethics *(ETH).* Second question is about what interfaces can be helpful in mitigating mentioned ML usability challenges. According to oral feedback from user study, local factor contributions are most useful in mentioned domain.

One of the main reasons why study [6] was deemed important is, that authors performed two user studies, on two different user groups. First study involved 12 data and/or social scientist. Authors decided to run first study with non-experts, so usability problems can be fixed immediately and iteration on UI/UX elements of the tool can be done. Second study engaged 13 collaborating experts (child welfare screeners). Important part of the user study is, that participants were first shown video how to use tested tool. Next, they were shown 7 case descriptions with model and interface of examined tool with simulated data. Case descriptions were real paragraph-form narratives provided by concerned parties during past referrals [8]. Names were changed by a representative of the county for de-identification. Questionnaire consisted of 3 kinds of questions and answers: 1. Five-point Likert-scale style questions, 2. Multiple choice questions and the last 3. Free response questions.

According to results, Case-Specific Details interface was deemed the most useful (91.8% experts, 90.7% non-experts). Non-experts also were more likely to report listening to the model without considering the added information. In the section of information presentation, there were three main problems. First problem is according to both, experts, and non-experts, that too many factors were shown, and it was confusing. Second is confusion caused by correlated factors The last one is confusion caused by Boolean terminology.

In the next study collective of authors Tsopra, R. et.al. [11] focus more on interfaces of an application for prescription of antibiotics. It is significant to understand why systems like this are important. General practitioners and other professionals are not always able to keep up with the rapid changing of medical knowledge. If they are unaware of any changes, incorrect decision can be made, resulting in inappropriate treatment. This way, it has consequences on both, patients, and the healthcare system [12]. However, many physicians do not use CDSS (Clinical decision support system) because of barriers it brings [13].

Authors of this study analyze four types of how to present the decision process for physicians and doctors in general. First one to be mentioned is a *Decision tree diagram*. These systems basically consist of static presentation with node and link diagrams. They can be useful because they provide physicians an overview of the whole process. However, decision trees tend to get very branched, so they require large amount of available display and also, the deciding information is limited to be displayed as a simple text at every decision node [14].

Second one is *Hypertext linking* that is used in many CDSS [15]. Textual links correspond to the nodes of the decision process (or to decision tree mentioned previously). Doctors should move through the hypertext links manually by clicking on them. However, this approach is subject to several limitations, e.g., the number of clicks depends on the depth of the hierarchy and may be large, there is no overview of the decision process, physicians may not remember where they are in the decision process (the 'lost in hyperspace' phenomenon [16]) potentially leading to navigation errors.

The third one includes *Data entries and checkboxes* [17]. Doctors input data (as free text or boxes to check) and press navigational buttons to go through the decision process. Presentations of this type preserve the autonomy of the physician, unlike previously mentioned *Hypertext linking.* However, it takes time to input the necessary data, all fields must be completed even if only a fraction of the data is required for the decision process, and there is no overview of the branching.

The last one used is *Expand/contract interface* [18]. The user has a global hierarchical view of the decision process because the information content is concealed within individual nodes. Users may gradually reduce the search field by dynamic filtration, this involves clicking on the node, so that items in the level immediately below appear and adapting navigation across the hierarchy according to the result obtained after each mouse click. This approach also facilitates reading because elements are read from top to bottom. However, *expand/contract* interfaces have also their own limitations, such as the number of clicks depends on the depth of the hierarchy, there is no complete view of the overall branching, the user must have previous knowledge about navigation of the system.

Important part of this study [11] are mentioned techniques for user interface according to usability principles. Some of these techniques relate to CDSS content: the use of concise and unambiguous language, based on a consistent terminology, so it would promote simplicity, consistency, efficient interactions, and effective use of language; the presentation of explanations and justifications to increase physician confidence; the provision of advice, suggestions and alternatives, rather than orders, to increase compliance and to respect the autonomy of the physician [19].

Various techniques have been proposed for CDSS display, such as reduction of the number of screens to facilitate navigation and to promote efficient interactions; use of appropriate font sizes, acceptable contrast between text and background and meaningful colors to improve readability [20]; organization of information to facilitate on-screen searches; display of important information in more prominent positions to ensure that it is seen [21]; use of tables, graphs, buttons,

scroll bars and iconic languages to ensure that the density of information is appropriate. Space-filling approaches may also help to maximize the amount of information that can be displayed in the available display space.

Authors of the study [11] decided to test two types of user interface on two groups of doctors. First tested UI (user interface) was the *Expand/contract Interface*, that was mentioned in previous section. The second interface was designed according to mentioned usability principles and it is called *At-a-glance Interface.* For usability evaluation in the user study, one important scale was used – *System Usability Scale* or *SUS.* The System Usability Scale (SUS) is a widely used standardized questionnaire for the assessment of perceived usability [22]. In its standard (most often used) form, the SUS has 10 five-point items with alternating positive and negative tone. Every one of these questions has a five-point answer (also called Likert scale), ranging from Strongly disagree up to Strongly agree, that user needs to answer. Then these equations are used for calculating *Result,* while Score (Q_x) is the evaluation from Likert Scale.

$$Score_1 = \sum \text{Score}\ (Q_{1,3,5,7,9}) - 1$$
$$Score_2 = \sum 5\text{- Score}\ (Q_{2,4,6,8,10})$$
$$Result = (S_1 + S_2) \times 2.5$$

According to [22], if the resulting score is above 80.3, app is very usable. Getting a score around 68 is OK but could be improved and score of 51 or lower is bad and usability of system should be the priority in fixing.

The results from user study [11] show us, that SUS score was significantly higher for the 'at-a-glance' interface than for the 'expand/contract' interface (score 76 vs 62) over the entire study period. In period 1, physicians evaluated the *expand/contract* interface in group 1 and the *at-a-glance* interface in group 2. There was no significant difference between the two interfaces in period 1 (score 72 vs 70). In period 2, physicians evaluated the other interface —that is, the *at-a-glance* interface in group 1 and the *expand/contract* interface in group 2. A highly significant difference was detected in period 2: the SUS score was 51 for the *expand/contract* interface and 81 for the *at-a-glance* interface. According to the SUS [22] perceived usability was OK for the *expand/contract* interface with score of 51 and excellent for the *at-a-glance* interface in period 2 with score of 81.

Authors Hardenbol A.X., et.al. in the study [23] evaluate the usability aspects of medication related CDSS in the outpatient setting. For this purpose, authors analyzed relevant articles. The query authors used was composed of four separate parts: Articles were relevant, if the first part had keywords relating to electronic prescription and DSS, the second part to medication related keywords, the third part to the inpatient setting and the fourth part to usability. The four parts were combined using "and" statements resulting in the final query. Altogether 3970 articles were searched and only 22 of them were classified into three usability aspects: Effectiveness, Efficiency and Satisfaction. Other articles were excluded because after thorough read, the lacked focus on usability/CDSS.

Using Van Welie et al.'s layered usability model [24], authors categorized usability aspects in terms of usage indicators and means. Evidence could mainly be found for Effectiveness and showed high rates of positive results in reducing medication errors. The effects of Efficiency and Satisfaction of clinical decision support systems regarding medication prescription remain understudied [23]. A total of 19 out of the 31 combinations (61.3%) could be categorized in category *Effectiveness*, 6 (19.4%) in *Efficiency* and 6 (19.4%) in *Satisfaction*. Of which, 22 (71.0%) studies found a positive effect, 1 (3.2%) study found no effect, 6 (19.4%) studies found a negative effect and 2 (6.5%) studies found mixed effects (meaning both positive and negative results were found). As we can see, Van Welie et.al. layered model of usability does not cover every aspect of usability e.g., in CDSS. That is the main reason for improving this model and adding new aspects to it.

III. Van Welie et.al., Layered Model of Usability and its Proposed Extension

For purpose of evaluating the applicability of clinical decision support systems (CDSS), the Van Welie usability model according to the study [23] was used, which divides the three usability aspects into three sublayers (Fig. 1). The first of them contains the main three aspects of applicability, e.g., efficiency, effectiveness, and satisfaction defined according to ISO 25000. The next level contains several indicators, which are indicators of the level of usability that can be observed in practice when users work with decision support system. Each of these indicators contributes to higher-level abstract aspects. For example, low error rates contribute to better efficiency and good performance speeds indicate good effectiveness. The third layer provides how usage indicators can be measured. These include e.g., consistency, availability of rollback operations, warnings, presence of feedback and adaptability. This layer is specified by factors that affect the concept to which it belongs. For example, consistency positively affects learning ability and warnings can reduce errors.

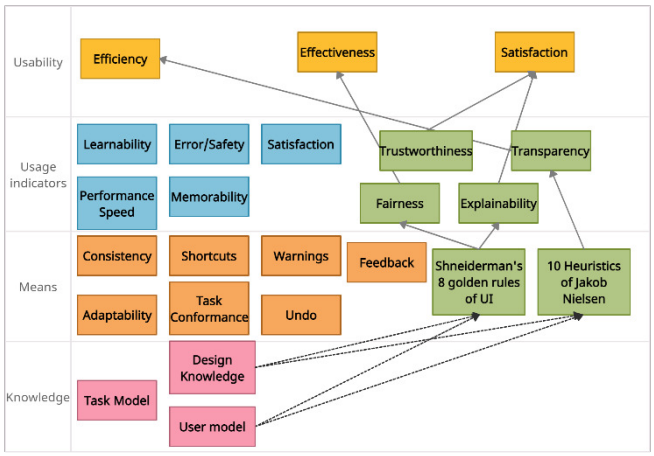


Fig. 1 Van Welie et.al. model of usability and its proposed extension [24]

Van Welie model of usability was introduced in 2001 and it is missing more aspects that need to be evaluated while working on a high-level usability product, in particular a decision support system based on data analytic models. Good example would be the quality of the solution. Van Welie model does not look if proposed solution is of high quality, it just looks on performance. This, and other missing aspects were identified as an opportunity for research.

First layer remained unchanged. The second layer was supplemented by 4 elements from study [24]: trustworthiness, explainability, transparency and fairness (green boxes). These 4 added elements represent another important usage indicators that need to be monitored while working on good usability of CDSS. Trustworthiness is very important, because doctors need to trust the system to apply its proposed decision. Explainability is another important part. According to results in user study from article [11], users want to see the explanation behind every decision system made. Transparency and fairness have to be involved in extended model, because users need to

see every part of the decision that is being made and the decision has to be always the same and fair – consistency is the key.

After careful consideration, Jakob Nielsen's 10 heuristics for UI design [27] were chosen to be added to the Means section. They include visibility of system status, match between system and the real world, user control and freedom, consistency and standards, error prevention, recognition rather than recall, flexibility and efficiency of use, aesthetic and minimalist design, recognition, diagnose and recover from errors, and help and documentation. These heuristics are added because they are directly related to the previous layer. For example, one of heuristics is consistency and standards, that is important for transparency and fairness in previous layer. Another heuristic is flexibility and efficiency of use – that goes perfectly with performance speed in usage indicators section. We also added 8 golden rules by Ben Shneiderman [26], consisting of: strive for consistency, seeking universal usability, offering feedback, designing dialogues to yield closure, preventing errors, permitting easy reversal of actions, keeping users in control and reducing short-term memory load. Shneiderman's golden rules are tied with Nielsen's heuristics, e.g., in consistency, dialogues to yield closure and keeping users in control. As for the last level, it remained unchanged, as the designer of the application must be able to handle correctly and especially know the three means mentioned - User, Knowledge of design and Task model. All this can be seen in Fig.1. For cleaner graphics, only relations between added aspects (green boxes) were drawn.

IV. Conclusion

In a conclusion, usability is a very broad term. We can find usability and its aspects everywhere in our ordinary lives. Usability is the key for successful application also in the case of CDSS. There are many rules to be followed and designers can get lost easily. That is why one universal model should help them. Van Welie et.al. did a great job with providing a layered model of usability but from its original introduction in 2001 it lacks some aspects to be considered in creating a high-level usability application, when it is based on data analytic models. That was identified as a research opportunity and after consideration of many aspects of usability, the most important for CDSS were added as seen in Figure 1.

Acknowledgment

This work was partially supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and Academy of Science of the Slovak Republic under grant no. 1/0685/21 and The Slovak Research and Development Agency grant under grant no. APVV-17-0550.

References

[1] N. Bevan, et.al., "What is usability," in Proc. 4 Int. Conf. On HCI, Stuttgart: 1991.

[2] N. McNamara and J. Kirakowski, "Defining usability: quality of use or quality of experience?," IPCC 2005. Proceedings. International Professional Communication Conference, 2005., 2005, pp. 200-204

[3] N. Bevan & M. MacLeod, "Usability measurement in context," Behavior and Information Technology, vol. 13, pp. 132-145, 1994

[4] ISO 9241, Ergonomic Requirements for Office Work with Visual Display Terminals: Part 11: Guidance on Usability, 1998.

[5] D. Güiza Caicedo, 2017. The real importance of usability and user experience. [online] Medium design community.

[6] A. Zytek, D. Liu, et.al., "Sibyl: Understanding and Addressing the Usability Challenges of Machine Learning In High-Stakes Decision Making," in IEEE Transactions on Visualization and Computer Graphics, vol. 28, no. 1, pp. 1161-1171, Jan. 2022

[7] F. Doshi-Velez and B. Kim, Towards A Rigorous Science of Interpretable Machine Learning, Mar. 2017.

[8] W. Kenton, "How Cost-Benefit Analysis Process Is Performed", Investopedia, 2021.

[9] H. Strobelt, S. Gehrmann, et.al., "LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks" in InfoVis, Phoenix, Arizona, USA:IEEE, vol. 24, pp. 667-676, Oct. 2017.

[10] J. Zhang, Y. Wang, et.al., "Manifold: A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models", IEEE Transactions on Visualization and Computer Graphics, vol. 25, no. 1, pp. 364-373, Jan. 2019.

[11] R. Tsopra, J-P. Jais, et.al., Comparison of two kinds of interface, based on guided navigation or usability principles, for improving the adoption of computerized decision support systems, Journal of the American Medical Informatics Association, Volume 21, Issue e1, February 2014

[12] D. PG. Marwick, Appropriate vs. inappropriate antimicrobial therapy. Clin Microbiol Infect 2008;14(Suppl 3):15–21.

[13] M.A. Robertson, J. Newby et al.: Computerized clinical decision support for prescribing: provision does not guarantee uptake. J Am Med Inform Assoc2010;17:25–33.

[14] B. Johnson, B. Shneiderman: Tree-Maps: a space-filling approach to the visualization of hierarchical information structures. IEEE Computer Society Press, 1991:284–91.

[15] CB. Litvin, et.al., Use of an electronic health record clinical decision support tool to improve antibiotic prescribing for acute respiratory infections: the ABX-TRIP study. J Gen Intern Med2013;28:810–16.

[16] Y.L. Theng, H. Thimbleby, et.al. Lost in hyperspace: psychological problem or bad design? Asia-Pacific Computer-Human Interaction 1996:387–96.

[17] L.J. Hoeksema, et al. Accuracy of a computerized clinical decision-support system for asthma assessment and management. J Am Med Inform Assoc2011;18:243–50.

[18] S.J. Johnson: An object-oriented taxonomy of medical data presentations. J Am Med Inform Assoc2000;7:1–20.

[19] J. Horsky, et al. . Interface design principles for usable decision support: a targeted review of best practices for clinical prescribing interventions. J Biomed Inform2012;45:1202–16.

[20] J.L. Belden, R. Grayson, J. Barnes: Defining and Testing EMR Usability: Principles and Proposed Methods of EMR Usability Evaluation and Rating. Healthcare Information and Management Systems Society (HIMSS); 2009.

[21] R. Jaspers, Khajouei MWM . The impact of CPOE medication systems' design aspects on usability, workflow and medication orders: Methods Inf Med2010;49:3–19.

[22] J. R. Lewis (2018) The System Usability Scale: Past, Present, and Future, International Journal of Human–Computer Interaction

[23] A.X. Hardenbol, et al. "Usability Aspects of Medication-Related Decision Support Systems in the Outpatient Setting: A Systematic Literature Review." Health Informatics Journal, Mar. 2020, pp. 72–87,

[24] M. Welie, et.al,. (2001). Breaking down Usability. Proceedings of Interact.

[25] C.M. Cutillo, K.R. Sharma, L. Foschini, et al. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. npj Digit. Med. 3, 47 (2020).

[26] N. Aottiwerch and U. Kokaew, "Design computer-assisted learning in an online Augmented Reality environment based on Shneiderman's eight Golden Rules," 2017 14th (JCSSE), 201

[27] A. Lodhi, "Usability Heuristics as an assessment parameter: For performing Usability Testing," 2010 2nd International Conference on Software Technology and Engineering, 2010, pp. V2-256-V2-259, doi: 10.1109/ICSTE.2010.5608809.

Link prediction in knowledge graphs

[1]*Jakub Ivan VANKO (1st year)*
Supervisor: [2]*Peter BEDNÁR*

[1,2]Dept. of Cybernetics and Artificial Intelligence, FEEI TU of Košice, Slovak Republic

[1]jakub.ivan.vanko@tuke.sk, [2]peter.bednar@tuke.sk

***Abstract*—In this paper, we have researched the current literature for detailed description of knowledge graphs and its representations. We focused mainly on graph incompleteness problem, where we discussed the definition and usage of link prediction. In the second part, we researched current methods used for link prediction problem, and described them, focused mainly on knowledge graph embedding models. We also described data we will use in our future research provided by a company we collaborate with, Ayanza Inc.**

***Keywords*—embedding, knowledge graph, link prediction**

I. Introduction

Knowledge Graphs offer a wide range of uses in industry and in academia, which has stimulated significant research on large-scale information extraction from a variety of sources. For many applications, Knowledge Graphs have shown to be an effective technique to connect diverse data sources and model underlying relationships. They are currently widely used in a variety of disciplines, ranging from question answering to information retrieval and content-based recommendation systems, due to their capacity to describe structured, complicated material in a machine-readable manner (they are important to any semantic web project) [1]. However, their biggest problem is their incompleteness. Task for correcting this incompleteness is commonly known as *Knowledge Graph Completion* or *Knowledge Graph Augmentation* [1]. One of the latest approaches for this task, is called *Link Prediction*. Link prediction has become a more active topic of research in recent years, thanks to the proliferation of machine learning and deep learning techniques. Several strategies based on various representation techniques have been presented to perform the link-prediction challenge. In the following chapters, we will describe important concepts for this article, and analyze recent methods and groups of methods for link prediction problem.

II. Knowledge Graphs

Despite its current popularity in research, term knowledge graph does not constitute a new technology, since graph-based representation of knowledge has been researched for decades. However, in 2012, when Google introduced the *Knowledge Graph* as a semantic enhancement of Google's search function that does not match strings but enables searching for "things" (in other words, real-world objects), the popularity of the term began to grow [2]. Large number of knowledge graphs have been created and applied to many real-world applications in recent years, such as FreeBase [3], DBpedia [4], YAGO [5] and NELL [6].

Ristoski and Paulheim described that knowledge graphs are used to map data from multiple sources and create connections between entities in each subject matter [7]. It can be considered as a type of network which forms the basis of associations between entities in a network (between real-life objects). Despite the facts, that terms *Knowledge Graph* and *Knowledge Base* (often used as synonym for *Ontology*) are often used interchangeably, these terms are not the same. Ehrlinger and Wolfram described size as very important characteristic of knowledge graphs and indicated, it can be described as very large ontology [2]. However, knowledge graphs provide more features than ontologies, whereupon are superior. The difference could be interpreted as a matter of quantity, or matter of extended requirements [2]. By that assumption, they described knowledge graph as a knowledge-based system, that contains both a knowledge base and a reasoning engine, which is used to generate new knowledge (architecture is described in Figure 1) [2].

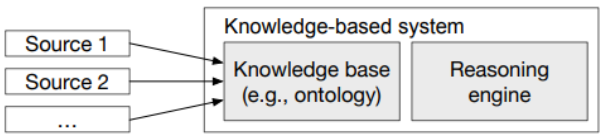


Fig. 1. *Architecture of a knowledge graph [2]*

Based on the architectures, Ehrlinger and Wolfram defined knowledge graphs as follows:

"A knowledge graph acquires and integrates information into an ontology and applies a reasoner to derive new knowledge."[2]

As mentioned before, knowledge graphs model information in form of objects and connections between them. Specifically, it can be described as directed labeled graph consisting of nodes, edges and labels [9]. Nodes are objects, that can represent anything from real world. Edge connects pair of nodes; it represents a relationship between them. Labels describe the meaning of relationship. Formally, we have set of nodes N, labels L; knowledge graph is subset of cross product $N \times L \times N$, where each member of this set is called *triple* [9]. In some data models, the triples are represented in as *SPO – subject, predicate, object* – where subject and object are entities (nodes) and predicate is the relation between them [8].

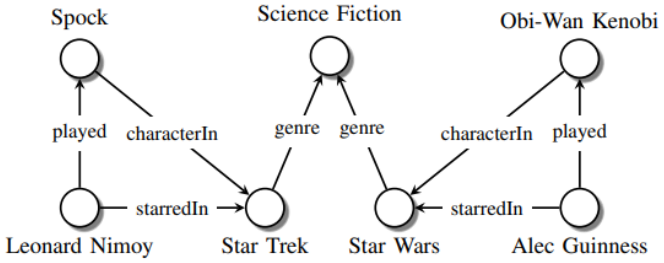


Fig. 2. *Sample knowledge graph* [8]

In Fig. 2, a sample knowledge graph is presented describing relationships between actors and their roles in movies. These relationships can also be expressed with following triples:

TABLE I
TRIPLE REPRESENTATION OF SAMPLE KNOWLEDGE GRAPH

Subject	Predicate	Object
(Leonard Nimoy,	*starredIn,*	*Star Trek)*
(Leonard Nimoy,	*played,*	*Spock)*
(Spock,	*characterIn,*	*Star Trek)*
(Star Trek,	*genre,*	*Science Fiction)*
(Alec Guinness,	*starredIn,*	*Star Wars)*
(Alec Guinness,	*played,*	*Obi-Wan Kenobi)*
(Obi-Wan Kenobi,	*characterIn,*	*Star Wars)*
(Star Wars,	*genre,*	*Science Fiction)*

The direction indicates whether entities are subjects, or objects. Knowledge graphs also can provide type hierarchies – Leonard Nimoy is an actor, who is a person – and type constraints – person can marry only other person, not a thing [8].

Existing triples represent true relationships, facts, but there are different paradigms for interpretation of non-existing triples [8]:

1. *Closed world assumption* indicates, that what is not known to be true, must be false. If relation is not contained in our knowledge graph, it doesn't exist. In our sample knowledge graph, it would mean that Leonard Nimoy never played in Star Wars, because such a relationship does not exist.
2. *Open world assumption:* if triple doesn't exist in our knowledge graph, then it's interpreted as unknown. The relationship can either be true, or false. If there is no *starredIn* connection between Leonard Nimoy and Star Wars, that did not necessarily mean, he didn't star in the movie. The interpretation in this case is, that it's unknown if he starred in Star Wars.

III. LINK PREDICTION IN KNOWLEDGE GRAPHS

Knowledge Graphs are only as powerful as their applications, no matter how carefully curated and high quality they are. They have a wide range of applications in both research and industry. Many applications of a directed graph form may be found in computer science, such as data flow graphs, binary decision diagrams, state charts, and so on [9]. Question answering [10], recommendation systems [11], and information retrieval [12] are also an important areas where they have been used.

On the other hand, knowledge graphs have some flaws. Usually, they are incomplete, due to its difficulty to incorporate all the human concepts, and because real-world data are dynamic, and are always evolving [13]. As a result, a great deal of effort has gone into developing an effective method for completing the Knowledge graphs. This method is also formulated as *Link Prediction* [14]. Link prediction has become a more active topic of research in recent years, thanks to availability of machine learning and deep learning techniques and rising popularity of knowledge graphs in recent years. The aim of link prediction is to use the available facts in a KG to predict missing ones. Fact is a triple $\langle h, r, t \rangle$, where h is the *head* (subject), r is the *relation* (predicate) and t is the *tail* (object) of the fact. By guessing the correct entity, we can complete the problem of tail prediction $\langle h, r, ? \rangle$, or head prediction $\langle ?, r, t \rangle$.

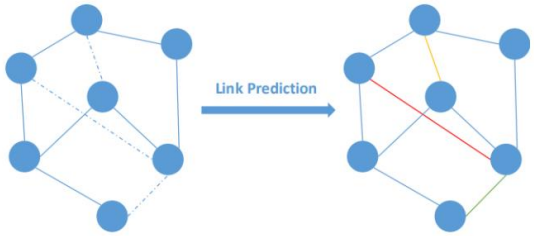


Fig. 3. *An example of link prediction* [13]

There are various methods used for the link prediction problem, including decomposition-based methods, path-based methods and embedding-based methods [15]. In decomposition-based models, entity–relation triples are encoded as tensors. They are based on possible semantic information and may have a large number of parameters. Such models may be inefficient and have poor scalability. Path-based models analyze a path from point a to point b via a series of edges. The earliest models include random walks and the path-ranking algorithm. Embedding-based methods translate a knowledge graph into a low-dimensional vector space while keeping its underlying semantics [13]. Many link prediction models nowadays utilize original elements from knowledge graphs to learn low-dimensional representations known as *Knowledge graph embeddings*, which are then used to infer new facts [1].

IV. EMBEDDING MODELS FOR LINK PREDICTION

Wang et. al performed an investigation on most popular knowledge graph embedding models in last few years. They divided into three categories - translation-distance-based models, semantic information-based models, and neural network-based models [13].

A. Translation-distance-based models

These models are additive, and they usually use distance-based functions to define the scoring function for link prediction. Wang et. al classify these models into three subcategories – TransE and its extensions, Gaussian embedding and others.

The first category includes mentioned *TransE* model and its extensions, such as *TransH* and *TransR* (and many more) [13]. The primary idea behind TransE [16] model is to get the sum of the head and related vectors as close to the tail vector as possible. It is a well-known and simple model, that regards a relation as a translation from a head entity to tail entity. TransE assumes $h + r \approx t$, when *(h, r, t)* is a golden (positive) triple, indicating that t should be nearest neighbor of $h+r$ [16]. TransE was the first translation-based embedding model, however it struggles with multirelational graphs. Its simple translation process, as well as the lack of a

discriminating policy for all types of relationships, limit it. As mentioned above, TransE had many extensions in recent years. One of these extensions, *TransH* [17] models the relation *r* on a hyperplane w_r as the normal vector. It projects entity embeddings into relation hyperplanes, allowing them to play different roles in different relations. The other extension we mentioned was *TransR* [17]. This model separates entities and relations into two spaces: entity space and multiple relation spaces, which are connected by relation-specific matrices.

The second category describes Gaussian embedding models. It considers entity and relation uncertainties and model them as random variables. We know two Gaussian embedding models – *KG2E* and *TransG* [13]. The first mentioned model regards entities and relations as random vectors derived from multivariate Gaussian distributions, and the distance between the two random vectors is used to score a triple [18]. The TransG models entities with Gaussian distributions, using a mixture of those distributions to obtain multiple semantics [13]. The uncertainty of the entities and relations are taken into account in Gaussian embedding models, although this results in a complex model.

Other translation-distance-based models worth mentioning are *RotatE* and *HAKE* [13]. *RotatE* [19] generates graph embeddings which can model and infer various relation patterns such as symmetry/antisymmetry, inversion, and composition. Each relation in the RotatE model is defined as a rotation in the complex vector space from the source entity to the target entity. *HAKE* [20] uses polar coordinates to represent the semantic hierarchy of entities, in which concentric circles can naturally reflect the hierarchy. It also considers whether the entities are on the same level of the hierarchy, which is divided into two parts: the modulus and the phase, which are used to differentiate the two sorts of entities. The modulus information is represented by the radial coordinate, whereas the phase information is represented by the angular coordinate.

B. Semantic information-based models

These types of models typically use similarity-based functions to define scoring functions for traditional semantic-based models or introduce additional information to mine more knowledge for recently proposed models [13]. The traditional semantic information-based models measure the plausibility of a triple by matching the latent semantics of entities and relation embeddings. Traditional semantic-based models focus solely on the information contained within the triple and do not combine any additional information. These models also suffer from higher computational complexity. Example of traditional semantic information-based models are *DistMult* [21] and *Complex* [22]. Recently proposed models combined various additional information (path information, order information, concepts, entity attributes, entity types and so on) to improve performance to mine deeper semantic information at the bottoms of graphs. Some notable models like *RESCAL* and *PTransE* also belongs to this category [17]. When applied to relational data, RESCAL [17] can take advantage of a collective learning effect. It has demonstrated excellent performance in variety of canonical relational learning tasks (such as link prediction). It also has an extension called *TRESCAL* [17], which tries to encode rules into RESCAL, but focuses only on one rule (a relation's arguments should be entities of specific types.). *PTransE* [23] is a translation-based model that extends TransE by introducing contextual information via multiple-step relation paths. PTransE uses TransE's scoring function, but for the multiple-step relation path, it adds another score item to the triple's score.

C. Neural network-based models

Knowledge graph embedding's requirements cannot be met with traditional distance-based and semantic-matching-based approaches. Due to popularity of neural networks in many fields, it was also introduced into knowledge graph embeddings for link prediction problems. Neural link prediction models consist of an encoding component and scoring component. The encoding component uses the input triple to map entities to their distributed embedding representation, and the scoring component score those two entity embeddings [24]. Interesting examples of network-based models are *ConvE and HypER* [13].

The first model, ConvE [24], uses 2D convolutions over embeddings for link prediction. It's the most basic multi-layer convolutional architecture for link prediction, consisting of single convolution layer, a projection layer to the embedding dimension, and an inner product layer. The main characteristic of this model is that 2D shaped embeddings define the score. ConvE additionally includes a 1-N scoring program that takes a head-entity–relation pair and matches all the tail entities at the same time, saving time in the evaluation process. In the architecture, the entity and relation embeddings are reshaped and concentrated as first. The matrix, that we get is then used as input to a convolutional layer. After that, we vectorize the resulting feature map tensor, project it into a k-dimensional space and match with all candidate object embeddings.

HypER [25] uses hypernetwork architecture to generate convolutional filter weights for each relation, and it's based on ConvE. There are three main differences between these two models. First, after resizing and concatenating the entity and relation embeddings, ConvE constructs convolution operators using 2D filters, while HypER uses 1D relation-specific filters, to handle entity embeddings, which can simplify the interaction between entities and relational embeddings. The second difference is in interaction between entities and relations. In ConvE, the interaction is affected by how they are reshaped and concatenated before being fed into the convolutional layers, while HypER uses a convolution operator with a set of relation-specific filters produced by hypernetwork from relation embeddings for head-entity embeddings (the hypernetwork is a fully connected layer). The third difference is in projecting and vectorizing the feature maps into a k-dimensional space. ConvE uses a linear parameterized transformation, while HypER uses a weight matrix to which the ReLU activation function is applied.

V. PROVIDED DATA

In our research, we collaborate with company Ayanza, Inc. [26]. Ayanza helps teams achieve their dreams by committing to a vision they co-create together with better collaboration and excellence in execution and building healthy relationships. They developed an application where teams can design their workflows, organize tasks transparently and schedule their meeting cycle in the most useful way. In Ayanza, they created an application, where you can do it all in

one collaborative environment with newsfeed, many reaction possibilities, document connection and task management functions.

Ayanza provides us the data from beta version of their application. It's a network created from documents in their database, containing 99 047 nodes, from which we extracted 10 metanodes and 16 metaedges. Each metanode with its count is described below:

- *Organization* (1061) – describes the organization space in Ayanza
- *Workspace* (8471) – every team in organization can have their own workspace (or individual private workspace)
- *Widget* (17 054) – a document, where users can share their visions, activities, recent work
- *Activity* (69) – recurrent activities related to widgets (daily standups etc.)
- *Comment* (1061) – users can add comments on widgets
- *Reaction* (655) – similar to social networks, you can leave likes on widgets and comments
- *User* (185) – a person using Ayanza
- *Invitation* (152) – user invitation
- *Notification* (70 905) – notifications about new activities, news, mentions…
- *Group (611)* – user permission groups

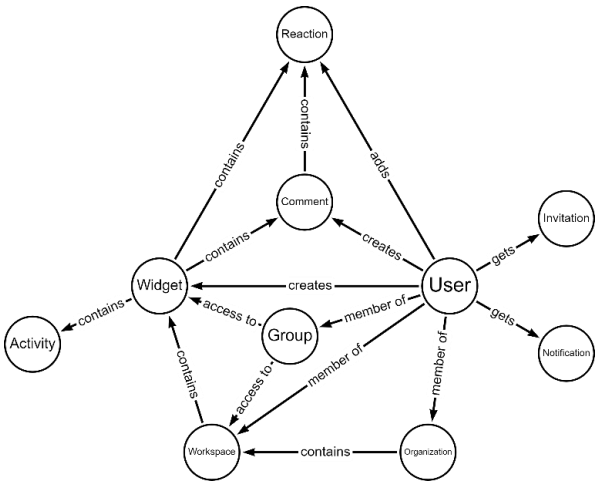


Fig. 4. *Metagraph diagram illustrates the connectivity between different types of nodes and edges in the network.*

With Ayanza data, we will focus mostly on using link prediction models to find different ways to display personalized content in the newsfeed for each user, with emphasis on work efficiency and achieving the goal.

VI. Conclusion And future direction

Knowledge graphs are getting more popular in recent years and their use is rising every day. In this article, we defined what knowledge graphs are, showed some examples and talked about link prediction problems. We described the solutions of this problem with various current methods, where we mostly focused on embedding models. These models made great progress in recent years and are very effective.

This research of link prediction methods inspired us to focus mostly on knowledge graph embedding models, which we will research even further, so we can apply them on our data in further experiments. Besides the experiments on Ayanza data, we would like to focus our research also on medical data, where are very interesting tasks related to link prediction in knowledge graphs.

References

[1] Rossi, A., Barbosa, D., Firmani, D., Matinata, A. and Merialdo, P., 2021. Knowledge Graph Embedding for Link Prediction. ACM Transactions on Knowledge Discovery from Data, 15(2), pp.1-49.
[2] Ehrlinger, Lisa & Wöß, Wolfram, 2016. Towards a Definition of Knowledge Graphs.
[3] K. Bollacker, C. Evans, P. Paritosh, T. Sturge and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge", Proc. ACM SIGMOD Int. Conf. Manage. Data, pp. 1247-1250, 2008.
[4] J. Lehmann et al., "DBpedia: A large-scale multilingual knowledge base extracted from Wikipedia", Semantic Web J., vol. 6, no. 2, 2015.
[5] F. M. Suchanek, G. Kasneci and G. Weikum, "YAGO: A core of semantic knowledge", Proc. 16th Int. Conf. World Wide Web, pp. 697-706, 2007.
[6] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr. and T. M. Mitchell, "Toward an architecture for never-ending language learning", Proc. 24th AAAI Conf. Artif. Intell., pp. 1306-1313, 2010.
[7] Ristoski, P. and Paulheim, H. (2016) 'Semantic Web in data mining and knowledge discovery: A comprehensive survey, Journal of Web Semantics', Journal of Web Semantics, 4(2), pp.1–22.
[8] Nickel, M., Murphy, K., Tresp, V. and Gabrilovich, E. A review of relational machine learning for knowledge graphs (2015). Proceedings of the IEEE, 104(1).
[9] Web.stanford.edu. 2021. What is a Knowledge Graph?. [online] Available at: <https://web.stanford.edu/class/cs520/2020/notes/What_is_a_Knowledge_Graph.html>.
[10] Wang, R.; Wang, M.; Liu, J.; Chen, W.; Cochez, M.; Decker, S. Leveraging Knowledge Graph Embeddings for Natural Language Question Answering. In Proceedings of the DASFAA 2019, Chiang Mai, Thailand, 22–25 April 2019; pp. 659–675.
[11] Musto, C.; Basile, P.; Semeraro, G. Embedding Knowledge Graphs for Semantics-aware Recommendations based on DBpedia. In Proceedings of the UMAP 2019, Larnaca, Cyprus, 9–12 June 2019; pp. 27–31.
[12] Wang, Q.; Mao, Z.; Wang, B.; Guo, L. Knowledge Graph Embedding: A Survey of Approaches and Applications.
[13] Wang, M., Qiu, L. and Wang, X., 2021. A Survey on Knowledge Graph Embeddings for Link Prediction. Symmetry, 13(3), p.485.
[14] Siddhant, A. A Survey on Graph Neural Networks for Knowledge Graph Completion. arXiv 2020, arXiv:2007.12374.
[15] Ma, J., Qiao, Y., Hu, G., Wang, Y., Zhang, C., Huang, Y., Sangaiah, A., Wu, H., Zhang, H. and Ren, K., 2019. ELPKG: A High-Accuracy Link Prediction Approach for Knowledge Graph Completion. Symmetry, 11(9), p.1096.
[16] Bordes, A., Usuneir, N., Garcia-Duran, A., Weston, j. and Yakhnenko, O., 2013. Translating Embeddings for Modeling Multi-relational Data.
[17] Lin, H., Liu, Y., Wang, W., Yue, Y. and Lin, Z., 2017. Learning Entity and Relation Embeddings for Knowledge Resolution.
[18] He, S., Liu, K., Ji, G. and Zhao, J., 2015. Learning to Represent Knowledge Graphs with Gaussian Embedding. Proceedings of the 24th ACM International on Conference on Information and Knowledge Management.
[19] Sun, Z., Deng, Z., Nie, J. and Tang, J., 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space.
[20] Zhang, Z., Cai, J., Zhang, Y. and Wang, J., 2019. Learning Hierarchy-Aware Knowledge Graph Embeddings for Link Prediction.
[21] Yang, B., Yih, W., He, X., Gao, J. and Deng, L., 2014. Embedding Entities and Relations for Learning and Inference in Knowledge Bases.
[22] Trouillon, T., Welbl, J., Riedel, S., Gaussier, É. and Bouchard, G., 2016. Complex Embeddings for Simple Link Prediction.
[23] Lin, Y., Liu, Z., Luan, H., Sun, M., Rao, S. and Liu, S., 2015. Modeling Relation Paths for Representation Learning of Knowledge Bases.
[24] Dettmers, T., Minervini, P., Stenetorp, P. and Riedel, S., 2017. Convolutional 2D Knowledge Graph Embeddings.
[25] Balažević, I., Allen, C. and Hospedales, T., 2019. Hypernetwork Knowledge Graph Embeddings. Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions, pp.553-565.
[26] Ayanza.com. 2021. Ayanza. [online] Available at: <https://ayanza.com/>.

Intelligent rehabilitation platform for upper limb rehabilitation

[1]*Stanislav HUSÁR (1st year),*
Supervisor: [2]*Marek BUNDZEL*

[1,2]Department of Cybernetics and Artificial Intelligence, FEI TU of Košice, Slovak Republic

[1]stanislav.husar@tuke.sk, [2]marek.bundzel@tuke.sk

***Abstract*—This paper describes the development of a rehabilitation platform based on the Rehapiano device. The Rehapiano device is designed for upper limb diagnostics and rehabilitation purposes, focusing on Parkinson's disease. We describe sensoric systems for intelligent rehabilitation in subject part. Described are devices based on computer vision, force transducer and inertial measurements. We describe individual components of proposed rehabilitation platform. These components include minicomputer connected to the Rehapiano device, physician's interface and central backend server. Individual components are in various development stages ranging from concept to functional prototype.**

***Keywords*—intelligent rehabilitation, load cell, data acquisition, service oriented architecture**

I. Introduction

Many people worldwide are suffering from motoric impairments caused by injuries or neurodegenerative diseases. Most effective way to regulate and possibly heal motoric impairments is extensive rehabilitation. Traditionally, rehabilitation consists of a huge amounts of repetitive movements. Patients get bored before making any significant improvement, and rehabilitation eventually fails. Advances in technology enabled development of rehabilitation games. While basic principle of repetitive movements remains, for the patient rehabilitation becomes entertaining.

From the physician's point of view, traditional assessment of impairments requires constant attention. Any small inattention results in inaccurate diagnosis. During rehabilitation game session, all collected data can be saved. Physician may then inspect this data as long as required, and as many times as required. This eases physician's work and results in more accurate diagnosis.

There is a huge market of sensing devices that may be used for rehabilitation purposes. There are high quality sensors with very accurate measurements, but they are often too expensive. Another complication for some of the sensors is calibration requirement. Sensors must be fixed to resulting position and calibrated by trained personnel. Recent research proved, that consumer-grade equipment may be successfully used for rehabilitation purposes [1] [2] [3] [4] [5] [6] [7]. These devices are targeting prices low enough for patients. Any required calibration is performed during manufacturing in a way, that it remains valid in any location.

Problem description part of this paper contains summary of selected types of sensoric devices. Solution design part contains description of various parts of the whole rehabilitation platform. Their development status ranges from functional prototype to draft of required functionality.

Fig. 1. AR based rehabilitation game screenshot source:([8])

II. Problem description

Most of intelligent rehabilitation sensory systems are based on one of these principles:

- Computer Vision
- Inertial measurements
- Force measurements

A. Computer Vision based

Computer vision based rehabilitation devices employ a set of cameras. Stereovision RGB cameras, or a time-of-flight camera provides nearly full body motion tracking. All parts of body with direct line of sight are tracked. This enables use of virtual reality within rehabilitation game. Devices that have at least one RGB camera, additionally enable use of augmented reality - artificial graphics are rendered to camera footage. Professional-grade camera systems often require in-place calibration performed by trained personnel. This greatly limits their usage. Fortunately, consumer-grade devices do not have this requirement, and are often suitable for intelligent rehabilitation.

Article [8] describes example of computer vision based rehabilitation solution. Authors are using Microsoft Kinect ™sensor. This device utilizes another approach to motion tracking - infrared pattern projector and infrared camera. A pattern with known geometry is projected to scene. Pattern geometry distortion as observed with camera allows depth

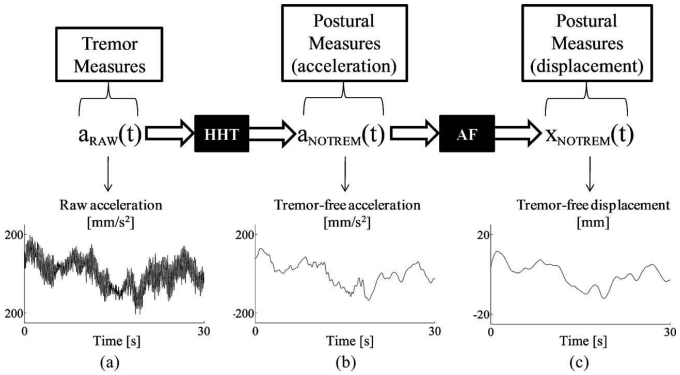


Fig. 2. Signal transformations (source: [12]) HHT: Hilbert-Huang transform AF: anthropometric filter

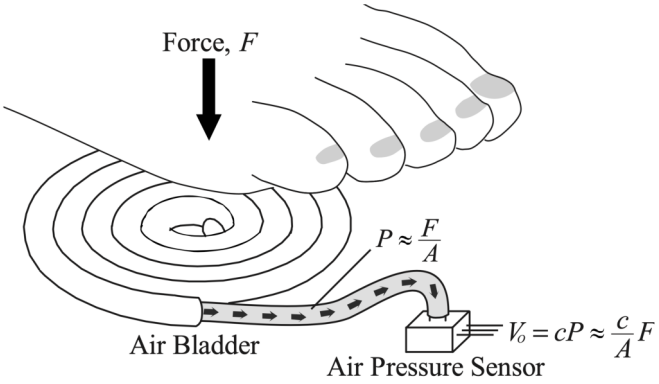


Fig. 3. Novel force sensing sensor (source: [15])

perception. Rehabilitation game is based on augmented reality. Example screenshot provided by authors is on figure 1. Authors proposed solution with a local database as storage for data generated during rehabilitation session.

B. Inertial measurements based

Inertial measurement based rehabilitation devices employ inertial measurement units (accelerometer, gyroscope, magnetometer). This provides limited motion tracking availability of one part of body per inertial measurement unit. However, they do not require direct line of sight. These devices enable limited use of virtual reality, and do not allow use of augmented reality. Calibration of inertial measurement units is valid worldwide. Moreover, these devices do not loss their tracking availability during any movement, so they can be used during route to another place(To work, from work, etc.). Some commonly used devices (for example smartphone, smart watch) already contain inertial measurement devices. With proper software, they can be used for intelligent rehabilitation purposes [9] [10].

After proper preprocessing, inertial measurements may be classified using traditional machine learning methods [11]. For preprocessing, signal transformations may be used [12]. Feature selection using similarity networks is also possible [13]. Inertial measurements may also be fused with additional sensors [14].

C. Force measurements based

Force measurement based rehabilitation devices employ force transducer sensors (for example load cell). This does not provide feasible motion tracking capability. On the other hand, this allows precise measurement of forces and centers of forces exerted by muscles. These devices enable very limited use of virtual reality, and do not allow use of augmented reality. Similar to inertial measurements, calibration of force transducers is valid worldwide, and is not invalidated by any movement.

Article [15] proposes a novel force sensing system embedded to shoes (picture 3. Human ground force is transduced to air pressure and sampled using pressure sensor. Force is measured on four important locations, this adds capability to measure center of ground force. These shoes may be worn during common everyday activities. Article [16] deals with assessment of force sensing device attached to calf. Article [17] proposes a device for measurement of postural balance. A

Fig. 4. Mechanical part

total of four load cells (left foot front, left foot rear, right foot front, right foot rear) are used to measure center of pressure. Study [18] uses multiple sensors including force measurement to assess compensatory backwards stepping after waist pull.

III. Solution Design

This paper focuses on proposal of intelligent rehabilitation platform for upper limb rehabilitation.

Whole solution may be divided into these parts:

- Mechanical device design
- Data acquisition electronics
- Rehabilitation game computer
- Central backend server
- Frontends

A. Mechanical device design

Rehapiano device (figure 4) consists of a total of ten load cells (one for each finger), mounted on a carrier part made from plastic. In front of load cells are splints for patient's forearm. In place between load cells is enclosure, that houses the electronics.

B. Data acquisition electronics

Electronics system of Rehapiano is shown on figure 5. Load cells convert physical force exerted by fingers to electronic signal. Attached to load cells are amplifier subboards. These amplify very weak signals from load cells, which otherwise

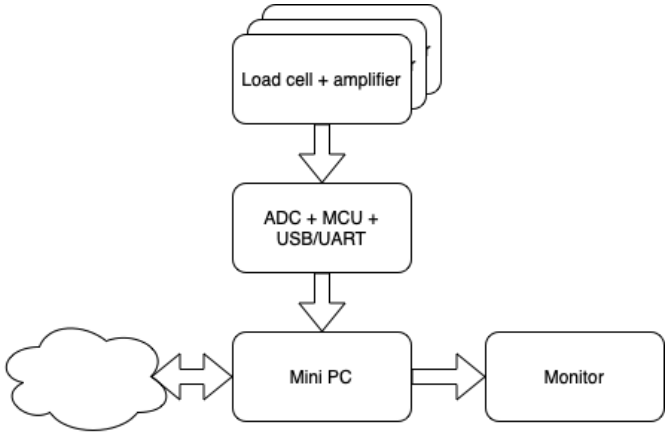


Fig. 5. Electronics

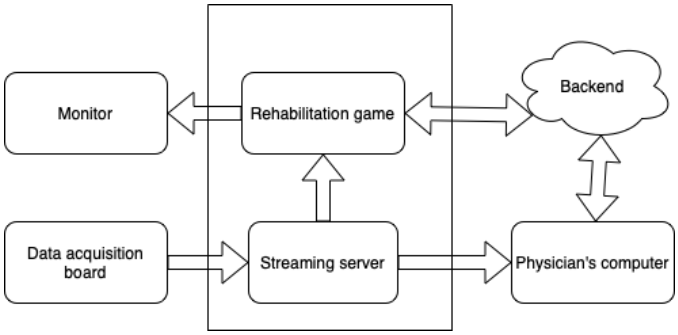


Fig. 6. Rehabilitation game computer

cannot be processed. There is an analog/digital converter, microcontroller, and USB to serial port converter on mainboard. A/D converter samples electronic signals and converts them to numerical representation. Microcontroller is responsible for initializing A/D converter, decoding incoming data, and encoding them for transmission to computer. USB to serial port converter takes encoded data from microcontroller, and transmits them to computer via USB interface.

C. Rehabilitation game computer

Detailed view at software running on computer is shown on figure 6. Streaming server is responsible for reading data incoming from USB interface, and streaming them to any connected clients over local network. Rehabilitation game is currently in design phase. It will receive data stream from streaming server, and display in-game graphics on monitor. Rehabilitation game will be controlled remotely from physician's computer. All controls, data collecting and processing and reports will be controlled by backend server.

D. Central backend server

Backend server is currently in design phase. Its architecture will contain database and API microservices. Database will contain all collected data (for AI research purposes) and commands to be sent between Rehapiano computer and physician's computer. Services provided by backend will utilize microservices architecture. Each microservice will handle a single well defined part of whole functionality. All internal and external communications will utilize REST API interface.

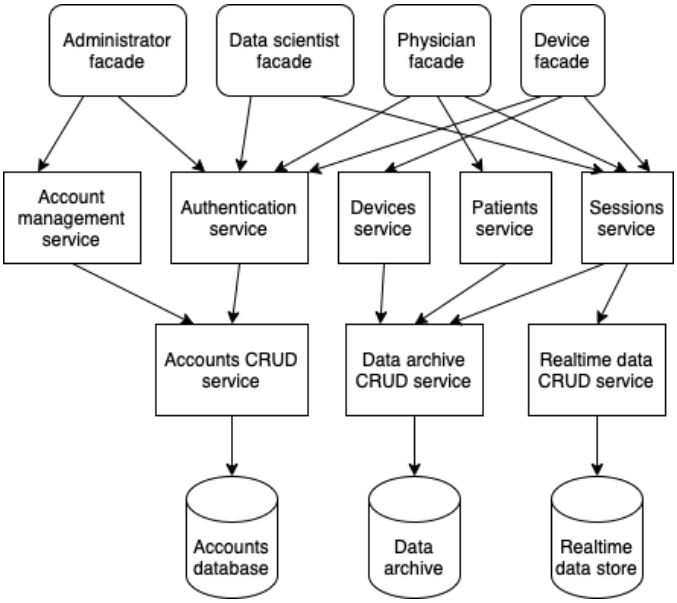


Fig. 7. Backend services

E. Frontends

Frontend development is currently postponed, until backend design will be finalized. Proposed solution will use web interface.

Physician frontend will be used to specify all rehabilitation sessions. After completing a session, physician will be able to view all results. Primary data source to diagnose the patient will be reports from backend. Proposed is also a feature allowing physician to view realtime data from Rehapiano. Data scientist frontend will be used to retrieve data for machine learning purposes. All data will be anonymized or statistically aggregated. Administrator frontend will be used to create accounts and assign permissions to individual clients. These clients will include data scientists, Rehapiano devices and physicians which use them.

IV. Future work

We are planning complete rework of mechanical design. Current state is functional prototype, but with poor aesthetics. Grip of the fingers only allows measurement in one direction. We are considering other arrangements that will allow measurement in both directions. Additionally, ergonomics may be enhanced.

Data acquisition electronics may be considered final for a maximum of 10 load cells. We are considering adding more load cells to additional locations. This would require redesign of electronics. Amplifier subboards are functional with good performance. However it will be required to increase number of A/D converters on mainboard. At the same better A/D converter may be used, enhancing sampling rate or other parameters. This may lead to requirement to replace microcontroller.

We have designed a measurement protocol. This protocol splits rehabilitation session into multiple executions of a few types of challenges. These include assessment of reaction speed, maximum exerted force and fine motoric skills. Each type of challenge is coupled to a scene inside rehabilitation game. Individual scenes are under development.

Central backend server is under development. Draft of individual microservices and their interactions is on figure

7. At the bottom end of architecture is database layer, containing all data used in the whole solution. Above it is layer of CRUD services, decoupling microservices from database type and implementation. Next is bussiness layer consisting of microservices handling each type of data. This includes authentication service, responsible for creating authentication tokens. Other services will use authentication tokens to validate client permissions. At the top end is layer of facades simplifying access for individual types of clients.

- Administrator facade : Serves as API for administrator frontend. This will allow management of accounts and their permissions.
- Data scientics facade: Serves as API for data scientist frontend. This will allow retrieval of data collected during rehabilitation sessions.
- Physician facade: Serves as API for physician's facade. This will allow management of rehabilitation sessions and retrieval of their results.
- Device facade: Serves as API for rehabilitation game.

Complete solution will also require frontends. These will be implemented as a single web portal. All clients will connect using their web browser. We will start development after finalizing backend design.

Calibration of the Rehapiano device has two parts. Offset calibration (known as TARE) will be executed on each boot of rehabilitation game computer. This will reject any load cell offset, ensuring that zero value corresponds to no load condition. Gain calibration will be performed once after finalizing mechanical and hardware part. A weight of known value will be attached to load cells, recording corresponding raw value. From this we can compute gain multiplier value, that will convert raw values to kilograms. This value will be inserted to streaming server configuration.

Data acquisition validation may be divided into two parts. First is assessment in time domain, validating latency of the whole system. Computer-controlled device will exert a momentary force to load cells, monitoring incoming data to compute latency. This latency will essentially define a minimum value for reaction time. Second is assessment in frequency domain, validating response of the whole system. Computer-controlled device will exert a periodic signal with certain frequencies (one at time), collecting incoming data. After preprocessing with fourier transform and peak detection, we can conclude range of frequencies the device is able to detect, and their deviation from ground truth value.

After finalizing design of the whole solution, we can start gathering data. These may be used in machine learning with two basic targets:

- Diagnosis classification : Classify if patient is healthy, or what could be his/her diagnosis.
- Intelligent rehabilitation: AI design of rehabilitation session scenario, aiming for most effective rehabilitation.

References

[1] Y.-J. Chang, S.-F. Chen, and J.-D. Huang, "A kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities," *Research in developmental disabilities*, vol. 32, no. 6, pp. 2566–2570, 2011.

[2] R. A. Clark, R. McGough, and K. Paterson, "Reliability of an inexpensive and portable dynamic weight bearing asymmetry assessment system incorporating dual nintendo wii balance boards," *Gait & posture*, vol. 34, no. 2, pp. 288–291, 2011.

[3] G. Saposnik, R. Teasell, M. Mamdani, J. Hall, W. McIlroy, D. Cheung, K. E. Thorpe, L. G. Cohen, and M. Bayley, "Effectiveness of virtual reality using wii gaming technology in stroke rehabilitation: a pilot randomized clinical trial and proof of principle," *Stroke*, vol. 41, no. 7, pp. 1477–1484, 2010.

[4] J.-F. Esculier, J. Vaudrin, P. Beriault, K. Gagnon, and L. E. Tremblay, "Home-based balance training programme using wii fit with balance board for parkinsons's disease: a pilot study." *Journal of Rehabilitation Medicine*, vol. 44, no. 2, pp. 144–150, 2012.

[5] J.-A. Lozano-Quilis, H. Gil-Gomez, J.-A. Gil-Gomez, S. Albiol-Perez, G. Palacios, H. M. Fardoum, and A. S. Mashat, "Virtual reality system for multiple sclerosis rehabilitation using kinect," in *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*. IEEE, 2013, pp. 366–369.

[6] J. Synnott, L. Chen, C. D. Nugent, and G. Moore, "Wiipd—objective home assessment of parkinson's disease using the nintendo wii remote," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 6, pp. 1304–1312, 2012.

[7] B. Lange, C.-Y. Chang, E. Suma, B. Newman, A. S. Rizzo, and M. Bolas, "Development and evaluation of low cost game-based balance rehabilitation tool using the microsoft kinect sensor," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 1831–1834.

[8] G. Palacios-Navarro, I. García-Magariño, and P. Ramos-Lorente, "A kinect-based system for lower limb rehabilitation in parkinson's disease patients: a pilot study," *Journal of medical systems*, vol. 39, no. 9, pp. 1–10, 2015.

[9] S. Barrantes, A. J. Sánchez Egea, H. A. González Rojas, M. J. Martí, Y. Compta, F. Valldeoriola, E. Simo Mezquita, E. Tolosa, and J. Valls-Solè, "Differential diagnosis between parkinson's disease and essential tremor using the smartphone's accelerometer," *PloS one*, vol. 12, no. 8, p. e0183843, 2017.

[10] R. LeMoyne, T. Mastroianni, M. Cozza, C. Coroian, and W. Grundfest, "Implementation of an iphone for characterizing parkinson's disease tremor through a wireless accelerometer application," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 2010, pp. 4954–4958.

[11] C. Ahlrichs, A. Samà, M. Lawo, J. Cabestany, D. Rodríguez-Martín, C. Pérez-López, D. Sweeney, L. R. Quinlan, G. Ò. Laighin, T. Counihan *et al.*, "Detecting freezing of gait with a tri-axial accelerometer in parkinson's disease patients," *Medical & biological engineering & computing*, vol. 54, no. 1, pp. 223–233, 2016.

[12] L. Palmerini, L. Rocchi, S. Mellone, F. Valzania, and L. Chiari, "Feature selection for accelerometer-based posture analysis in parkinson's disease," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 3, pp. 481–490, 2011.

[13] E. Rastegari, S. Azizian, and H. Ali, "Machine learning and similarity network approaches to support automatic classification of parkinson's diseases using accelerometer-based gait analysis," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[14] N. H. Ghassemi, F. Marxreiter, C. F. Pasluosta, P. Kugler, J. Schlachetzki, A. Schramm, B. M. Eskofier, and J. Klucken, "Combined accelerometer and emg analysis to differentiate essential tremor from parkinson's disease," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 672–675.

[15] J. Bae, K. Kong, N. Byl, and M. Tomizuka, "A mobile gait monitoring system for abnormal gait diagnosis and rehabilitation: a pilot study for parkinson disease patients," *Journal of biomechanical engineering*, vol. 133, no. 4, 2011.

[16] D. Giansanti, G. Maccioni, and S. Morelli, "An experience of health technology assessment in new models of care for subjects with parkinson's disease by means of a new wearable device," *TELEMEDICINE and e-HEALTH*, vol. 14, no. 5, pp. 467–472, 2008.

[17] J.-H. Park, S. Youm, Y. Jeon, and S.-H. Park, "Development of a balance analysis system for early diagnosis of parkinson's disease," *International Journal of Industrial Ergonomics*, vol. 48, pp. 139–148, 2015.

[18] M. A. McVey, S. Amundsen, A. Barnds, K. E. Lyons, R. Pahwa, J. D. Mahnken, and C. W. Luchies, "The effect of moderate parkinson's disease on compensatory backwards stepping," *Gait & posture*, vol. 38, no. 4, pp. 800–805, 2013.

Autonomous Configuration Change of Network Devices Based on Network Flow Analysis

[1]*Martin HAVRILLA (2nd year),*
Supervisor: [2]*Martin CHOVANEC*

[1,2]Dept. of computers and informatics, FEI TU of Košice, Slovak Republic

[1]martin.havrilla@tuke.sk, [2]martin.chovanec@tuke.sk

***Abstract*—This work deals with the autonomous configuration change of network devices based on the analysis of network flows. The work first deals with the methods of collecting network flows and proposes a basic architecture for the collection and analysis of network flows based on machine learning. Later, the work is focused on the option of changing the configuration of network devices based on the database of instructions.**

***Keywords*—Network flows, Computer learning, Streamlining, Autonomous configuration**

I. Introduction

Currently, more and more emphasis is being placed on increasing efficiency in almost every area of computer technology. The same is true in the field of computer networks. Our World in Data reports a significant year-on-year increase in the use of Internet services, which may result in higher computer network operating costs. For this reason, too, a large number of companies and researchers are looking at ways to streamline and secure network operations.

The main ideas of this work include the creation of an autonomous mechanism for changing the configuration of network devices based on the information stored in the instruction database. By solving this problem, it is possible to streamline the process of configuring a large number of network devices and thus eliminate possible routines. The work deals with the discovery of an anomaly in network traffic in the form of a DDoS attack and the subsequent change of the network device configuration in order to eliminate this anomaly.

The design of the whole solution consists of several parts such as collection of network flows, storage of network flows, search for similarity of measured data with samples and interpretation of results, translation of results into configuration of network devices.

II. Collection and Analysis of Network Flows

In the first step, it was necessary to collect network flows by a suitable method. In the case of the mechanism for collecting network flows, experiments were performed with protocols such as NetFlow, NFStream, sFlow. The NetFlow protocol was chosen for further use due to its wide range of implementation options.

A. Data Collection

The first part of the mechanism is the sensor, or also the probe, a component that covers the capture of the network and subsequently captures data about its operation[1]. Among the main information that needed to be obtained during the flow collection was source and destination IP adress, port and frequency. Another part of the mechanism is the collector. This is most often software that expects data from the sensor and writes that data to memory[2].

This network flow collection mechanism can be implemented as follows:

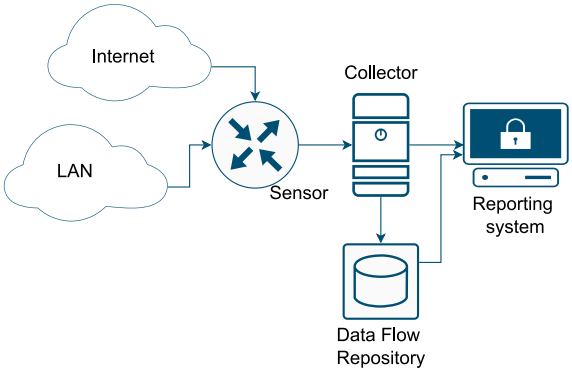


Fig. 1. Basic architecture for data flow collection

Fig. 1 shows the basic architecture for collecting network flows[3]. The measurement was performed in the time interval from 18.1.2022 to 20.1.2022, while 133990250 records with a total capacity of 100.2GB were obtained.

B. Data Processing

The next step that has been mastered is the analysis of network flows based on machine learning. For these needs, a decision regression tree algorithm was chosen from among the possible machine learning methods and algorithms[4]. The basic idea was to initially create a network flow sample to resemble an anomaly in network traffic. In this case, it was a DDoS attack which is shown in Fig. 2. This form was chosen because DDoS is easily separable from standard network traffic[5]. The mentioned sample served as a tool for developing decision regression trees so that they could detect a similar DDoS attack in a sample of network flows. The solution itself was implemented in the ELK stack environment and its SIEM interface[6].

The result of the analysis was the identification of the suspicious part of the network flow and the recording of the necessary information about the suspicious network flow so

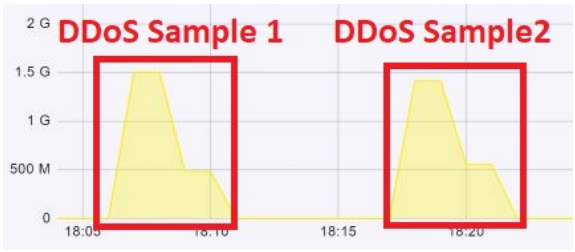


Fig. 2. Measured DDoS attack

that it was possible to subsequently transform the results of the analysis into a configuration change. Among the basic attributes that are preserved after the analysis and discovery of the DDoS attack is the IP address where the attack came from, which IP was attacked and what service was the target (Port).

Fig. 3. Measured attack in real network traffic

In 3 the attack is shown in real network traffic. By comparing the attack in real operation via the SIEM interface, the similarity of 84.37% was derived with the model attack. To increase success, it is necessary to collect a larger number of samples of anomalies, for the needs of machine learning.

III. Autonomous configuration change mechanism

The second part of this work was the creation of an autonomous mechanism to change the configuration of network devices based on the initiative from the analysis of network flows. The technical solution consisted of the following architecture shown in Fig. 4.

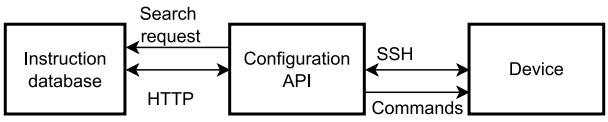


Fig. 4. Architecture block diagram for reconfiguring network components

The first block is the instruction database. This database contains instructions for reconfiguring network components. The instructions contain basic attributes such as the IP address to be disabled in the case of DDoS attacks, on which device this operation is to be performed, or the service to be disabled.

TABLE I: Selection from instruction database parameters

Action	ACL_ID	PermitDeny	IP_adr_space
add	103	Deny	147.232.144.144
remove	102	Deny	any

The mechanism is designed to load the configuration index in the ElasticSearch database in the Python scripting language using the *„Python Elasticsearch Client“* library. The library is designed to use the REST API of the ElasticSearch database.

The second block is the API which covers the configuration of network components. The "Netmiko" API was chosen from among the available solutions such as Paramico, NAPALM, etc. This option was created because Netmiko creates communication based on the standard SSH protocol. One of the other reasons for using this platform was the fact that it sends commands to network devices directly in their language.

The result of this section was to create a mechanism that can change the configuration of network components based on instructions from the database. Thus, in the case of a DDoS attack, a record was stored in the instruction database with the attacker's IP address, the target of the attack and the service.

This command was interpreted to a network device via the Netmiko API. An ACL record subsequently appeared on this device.

access-list 103 deny ip host 147.232.207.186 host 147.232.144.144

IV. Conclusion and future work

The aim of the work was research with the task of creating an autonomous mechanism that can create a more efficient configuration of network devices based on the nature of the network so that higher efficiency and security of network operation is achieved. Among the available protocols for data collection, the NetFlow protocol was chosen due to implementation possibilities on devices from different manufacturers and also because of its low computational requirements.

Based on the results of the analyses, a database of instructions was created which contained the basic parameters for changing the configuration of network devices.

The third goal was to create a mechanism for autonomous reconfiguration of network components based on the instruction database. The Python and Netmiko API environment was chosen for this purpose.

From the point of view of the work performed, methods for the collection of network flows were chosen, and the architecture of the collection and storage of network flows was proposed. A survey was also carried out of possible mechanisms for changing the configuration of network devices and the subsequent implementation of autonomy.

The main objective for future work is to increase the accuracy of the network flow analysis mechanism using machine learning techniques. Secondly it is necessary to extend the options of autonomous configuration with additional functionalities.

References

[1] C. Wagner, J. François, T. Engel *et al.*, "Machine learning approach for ip-flow record anomaly detection," in *International Conference on Research in Networking*. Springer, 2011, pp. 28–39.

[2] J. Bakker, B. Ng, W. K. Seah, and A. Pekar, "Traffic classification with machine learning in a live network," in *2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, 2019, pp. 488–493.

[3] P. Asrodia and H. Patel, "Network traffic analysis using packet sniffer," *International journal of engineering research and applications*, vol. 2, no. 3, pp. 854–856, 2012.

[4] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Routledge, 2017.

[5] R. Zhong and G. Yue, "Ddos detection system based on data mining," in *Proceedings of the 2nd International Symposium on Networking and Network Security, Jinggangshan, China*, vol. 2. Citeseer, 2010, p. 062.

[6] E. Anumol, "Use of machine learning algorithms with siem for attack prediction," in *Intelligent Computing, Communication and Devices*. Springer, 2015, pp. 231–235.

Edge AI - intelligent computing and sensing

[1]*Alexander BRECKO (1st year),*
Supervisor: [2]Iveta ZOLOTOVÁ, Consultant: [3]Erik KAJÁTI

[1,2,3]Dept. of Cybernetics and Artificial Intelligence, FEEI TU of Košice, Slovak Republic

[1]alexander.brecko@tuke.sk, [2]iveta.zolotova@tuke.sk, [3]erik.kajati@tuke.sk

***Abstract*—This article discusses the various devices used at the edge of the network and in IoT solutions. Sending all the data to the cloud for processing is not always possible with the growing amount of data generated everywhere around us. In such cases, the data should be processed at the network's edge. This article describes the individual edge devices and their hardware parameters. The primary goal is a comparison of individual edge devices in data pre-processing tasks. These can include image recognition and the creation of simple object detection models. The next step will be testing various artificial intelligence machine learning methods on individual devices and using this knowledge in the medical environment.**

***Keywords*—Artificial Intelligence, Computer Vision, Edge AI, Edge Computing**

I. Introduction

With the arrival of the Internet of Things (IoT) ideas, the amount of data that needs to be transferred over the network has increased significantly. Usually, this data is sent directly to the cloud [1]. However, cloud computing resources are often in remote data centres, away from end-users. It causes high latency, the need to ensure data security, and of course, increase the demands on network, since large amounts of data are transferred over the network [2]. Based on this, an edge computing architecture has been created that brings benefits. Edge computing aims to provide computing power for processing and storing data without moving data to data centres. Edge computing combines with cloud computing to overcome the specific limitations of different computing paradigms and offer more efficient services [3] [4].

Devices located at the edge of the network should have specific computing power according to their task to ensure efficient use of hardware resources. Edge devices can offer a variety of functions and can also have special hardware requirements. Each type of network edge solution requires a different approach, especially different hardware. High-performance computers and servers are not required for intelligent home management, and the performance of single-chip microcomputers will not be enough for image recognition [5]. In my research, I would like to focus on edge devices used at the edge of the network, which collect data from sensors and actuators and then pre-process the data. The primary goal will be to compare individual intelligent edge devices in pre-processing data such as data from cameras. This document provides an overview of edge devices from microcontrollers to common computers as we know them today. I also want to focus on federated learning on these edge devices and focus on its advantages and disadvantages in my next years of doctoral study.

II. Review of edge devices

Edge devices lie between the data source and the cloud. Depending on the architecture, edge devices may have different purposes, such as aggregation of data, pre-processing or providing computation power that would otherwise depend on the cloud [6]. The general architecture of the cloud, fog, edge network as shown in the figure 1. Edge facilities are beginning to be used regularly. We can most often encounter edge devices in, for example, autonomous vehicles, healthcare systems, security systems and various other industries [7].

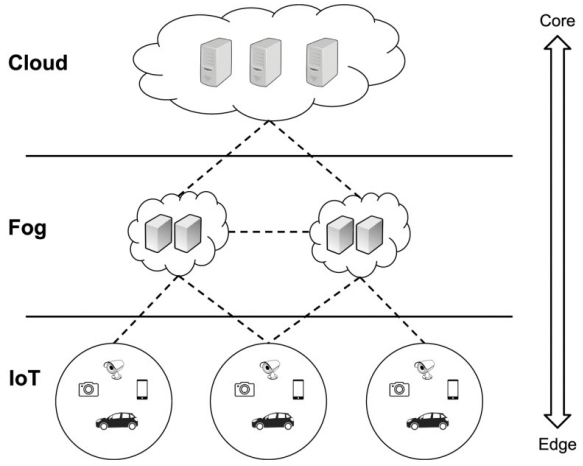


Fig. 1. Architecture of cloud, fog, edge computing for IoT [8]

This section introduces the most used devices at the edge of the network. Edge devices can be divided as follows:

- Microcontroller Unit (MCU)
- Single Board Computer (SBC)
- Mini Personal Computer (mini PC)
- Personal Computer and server

Each of these devices has its advantages and disadvantages. They differ mainly in the work for which they are intended. This section will describe the individual devices and their hardware parameters. We will also look at works that have been devoted to these devices. FPGAs are also nowadays being used as edge devices and will be the subject of further research.

A. Microcontroller Unit

This section will focus on MCUs, their performance, consumption and the price. As technology advances, manufacturers can integrate memory, input/output pins, serial ports, analog-to-digital converters, input timers, and other peripherals into a MCU. An MCU is a small, self-contained computer that resides on a single integrated circuit or microchip. MCUs

contain a CPU, RAM, ROM, and similar peripherals to perform (simple) tasks independently. MCUs are limited in performance and functionality but can perform simple functions independently. A MCU primary focus is accessing and interacting with other hardware. It has only a less capable resources in terms of CPU and RAM. However, it provides very low power consumption and the availability of real-time, fine-grained processing of connected hardwares' signals. Moreover, an MCU does not have a full, general purpose operating system but is very specialized for its use case. [9].

We meet daily with the use of MCUs. The MCUs are used to control functions in household appliances such as washing machines, vacuum cleaners, office appliances such as printers and scanners, motor vehicles, medical devices or Industry 5.0 for process control, and many other devices and systems. Their advantage is the possibility of long-term battery power and low price. Other advantages of development boards include easier development of new technologies thanks to the possibility of simple work (firmware upload, communication, peripherals) with a single-chip MCU. The disadvantages are low computing power, and small memory [10] [11].

One of the best known and frequently used MCUs in IoT solutions is ESP32 [12]. The advantage of this MCU is its very low power consumption, which allows this device to be powered from batteries for a long time. Microcontrollers are not often used for more complex tasks such as machine learning tasks. In general, this article [13] analyze the possibility of using ESP32 with a built-in camera for machine learning algorithms. The ability to apply simple machine learning algorithms (logistic regression) and the impact of Pseudo-Static RAM (PSRAM) implementation on performance were tested. The research focused on the computation demands of image processing. However, some logistic regression calculations had to be moved to the cloud because ESP32 is not powerful enough. This MCU was used to differentiate two handwritten letters on the grayscale pictures. In this case, ESP32-CAM produced by AI-Thinker was used. The advantage of this device is that it has 4MB PSRAM, which allows working with simple pictures.

The second MCU, which is used very often, is Arduino Uno rev.3 [14] with ATmega328P chip. This MCU makes it possible to create simple programs for reading data from sensors and actuators. The disadvantage of this device is that it has a small operating memory and does not handle more complicated calculations. It has an operating voltage of 5 volts.

One of the leading MCU manufacturers is the Raspberry Pi Foundation. The Raspberry Pi Foundation is a British charity and company founded in 2009 to promote the study of basic computer science in schools and is responsible for developing the Raspberry Pi microcontrollers. [15]

From the family of Raspberry Pi [14], Raspberry Pi Pico is the first high performance, low-cost microcontroller board built with the new RP2040 chip. An thorough comparison of the Arduino Uno and Raspberry Pi Pico microcontrollers is described in [14], which discusses the individual speeds and calculations of microcontrollers with a change in the operating frequency of the processor. The study results show that the speed of the Raspberry Pi Pico board is higher than the speed of the Arduino Uno board but less stable when the processor is overclocked.

Main MCUs hardware parameters are summarized in TABLE I. Devices are very similar in power consumption. Due to the operating frequencies of the processor and the size of the RAM memory, the ESP32 and RPi Pico seem to be more powerful and cost less in this comparison. There are several other similar MCUs not currently listed here.

TABLE I
MCUs COMPARISON

	ESP32	**Arduino Uno rev.3**	**Raspberry Pi Pico**
CPU	Dual-Core LX6	ATmega328P	Dual-Core Cortex-M0
Frequency	240MHz	16MHz	133MHz
RAM	520KB	2KB	264KB
ROM	4MB	32KB	2MB
Communication	Wi-Fi, Bluethooth	-	-
GPIO pins	32	20	30
Consumption	1.5W	1.5W	2W
Price	10€	25€	6€

B. Single Board Computer

This section describes an SBC, it describes the differences between SBCs and MCUs, their advantages and disadvantages, and summarizes the many things that can be done with SBCs. The SBC is a complete computer built on a single circuit board, with a microprocessor, memory, input/output pins and other features required of a functional computer. A large number of manufacturers use ARM architecture. We can also find ARM architecture in today's mobile devices, tablets and laptops. SBCs are easy to manufacture and come to market quickly compared to personal computers or laptops. They have lower weight, compact dimensions, are more reliable and more efficient than multi-board computers [16]. They can be used for real-time data processing, for running simple machine learning algorithms or for simple image processing from a cameras. The advantages are small size, low power consumption, affordable price, and large number of available operating systems for various deployments [17]. SBCs are very popular devices that are used at the edge of the network, due to the fact that they have sufficient power to process data without the need to send it to the cloud [18].

There are several differences between SBC and MCU. The biggest difference is mainly the size of RAM, ROM and computing power. Another difference is the operating system. SBCs such as the Raspberry Pi 4 have a Linux-based operating system and can perform multiple tasks in parallel, while MCUs usually perform a single task in a loop. On SBC, it is possible to do more complex tasks than in the case of MCUs due to their higher computing power and more RAM. Tasks may involve more complex calculations, localization, image and video processing, or the use of artificial intelligence and machine learning.

The Raspberry Pi Foundation not only manufactures MCUs but also SBCs. The most powerful and the latest from the Raspberry Pi family is the Raspberry Pi 4 model B [19]. This model is more suitable for demanding computing tasks because it contains a more powerful processor and RAM. The advantage is still low consumption and compact dimensions.

On the other hand, the disadvantage of this device is its weaker GPU, which can cause problems with processing images. [20]

Another SBC manufacturer is NVIDIA with the NVIDIA Jetson family [21]. The smallest board in this series is the NVIDIA Jetson Nano board, similar in size to the Raspberry Pi 4. The advantage of this device is that it contains a graphics card with CUDA cores, which makes this device more suitable for AI. These two devices were also compared in [22], where their performance in classifying 2D images were tested. The TensorFlow library was used for image classification. However, RPi 4 has a newer, faster and better processor than NVIDIA Jetson Nano. On the other hand, Jetson Nano has a better graphics card for artificial intelligence and working with images. The study results show that Jetson Nano is faster in this case than the mentioned RPi 4 in the classification of images.

The third main manufacturer that makes SBCs is Google with the Google Coral board [23]. This board is also focused on artificial intelligence calculations at the edge of the network. The Google Coral Board works well with the Edge Tensor Processing Unit (Edge TPU). This coprocessor is suitable for learning a pre-trained computer vision model. Edge TPU can perform 4 Tera-operations per second (TOPS), while his consumption is 2W. The Edge TPU can also be connected to other edge devices such as RPi model 3b+ or RPi model 4 via USB. The disadvantage of machine learning models is that they are not fully optimized for edge devices, and not all devices always support artificial intelligence models and prediction algorithms. Due to the growing performance of edge devices, the models that can be performed on these devices are gradually being optimized. [24] This SBCs hardware parameters are summarized in TABLE II.

TABLE II
SBCs COMPARISON

	Raspberry Pi 4	**Nvidia Jetson Nano**	**Google Coral Board**
CPU	Quad-Core Cortex-A72	Quad-Core Cortex-A57	Quad-Core Cortex-A53
Frequency	1.5GHz	1.42GHz	1.5GHz
RAM	8GB LPDDR4	4GB LPDDR4	4GB LPDDR4
GPU	Broadcom Video Core VI	NVIDIA Maxwell 128 CUDA cores	GC7000 Lite + Google Edge TPU coprocessor
ROM	Micro-SD card	16GB eMMC	16GB eMMC
Network	10/100/1000 + Wi-Fi	10/100/1000	10/100/1000 + Wi-Fi
GPIO pins	40	40	40
Consumption	2.5W-7W	5W-10W	10W-15W
Performance	64 GFLOPS (FP16)	472 GFLOPS (FP16)	4 TOPS (int8)
Price	60€	130€	120€

C. Mini Personal Computer

Computer sets with compact dimensions are suitable for running and training simpler artificial intelligence methods, analysis of high data flow in real-time, and easier image processing from multiple cameras. The advantage is the computing power, small dimensions and lower purchase price. Although these computers offer relatively high computing power, they often have a cooling problems, because every part of these components are in small box without good cooling system. Some parameters such as CPU, RAM or hard drive can be replaced with newer and more powerful ones in some cases. However, most of these computers have graphics cards that are not too powerful. In this case, it is possible to use the Google Edge TPU and connect it to a computer for more computing power [25]. Mini PC hardware parameters can be as follows:

- CPU up to eight cores and frequency up to 3.5GHz, 64bit architecture,
- RAM memory up to 64GB,
- internal memory up to 4TB,
- GPU up to 3.6TFLOPS with 6GB of memory
- power consumption usually 50W-200W.

The work [26] describes the consumption of processors based on the x86/64 architecture and the ARM architecture. As a result, it can be seen that mini PCs have slightly higher power consumption but also achieve higher performance than computers with ARM architecture than those used in SBC.

D. Personal Computer and server

A network edge server or computer system with high computing power is suitable for complex artificial intelligence methods, such as training deep neural networks or training and validating machine learning models. They also offer performance sufficient for complex multi-camera image processing methods, such as object detection and location. The advantage of servers is their high computing power, extensive use and fast internal memory with high capacity. Edge server hardware parameters can be as follows:

- multiple CPUs up to 56 computing cores and frequency up to 5GHz, 64 bit architecture,
- RAM memory up to 1.5TB,
- internal memory up to 100TB,
- GPU up to 56TFLOPS with 64GB of memory.

The disadvantages are larger dimensions, energy consumption and higher purchase price. The consumption of these computers is an order of magnitude higher than that of the mini PC. Work [27] provides an overview of the latest and most powerful graphics cards used in today's computers and servers. These GPUs are mainly used for complex calculations when training the most demanding neural networks. It works with chip manufacturers such as Intel, AMD, NVIDIA and Google, which are most often used in training deep neural networks for image recognition and speech recognition. It provides an overview of the performance and consumption of individual devices.

III. EDGE COMPUTING

A large number of mentioned devices can be used at the edge of the network. We obtained a project granted by TUKE Grant for young scientists. In this project, we would like to look at SBC boards and edge AI accelerators suitable for network's edge. We plan to purchase additional edge AI devices from this project. Subsequently, we would like to examine the possibilities of artificial intelligence at the edge of the network and the associated methods used.

Our goal will be to focus mainly on the affordable devices that have been described in section II-B. Their high performance, low power consumption, and acceptable price suit them to perform various tasks at the edge of the network. The aim will be to find out which of these devices will be most suitable for processing data in video and images. Artificial intelligence techniques such as deep neural networks will be used. At the same time, it would be possible to test calculations on several devices simultaneously. Because only simple and small models can be trained on AI devices, it would be appropriate to test federated learning, which takes place on several devices simultaneously. We would like to examine these devices' performance, power consumption, and speed when training new models at the edge of the network without the need to send sensitive data to the cloud. This work will be a continuation of our efforts to improve intelligent gateways, highlighted in [28].

IV. Future Work

Peripheral devices finding used in many industries. The use of these devices can be, for example, in transport and healthcare to follow Industry 5.0 principles. The devices can process the image from the video camera, provide data collection and data processing from sensors such as LIDAR or Time-of-Flight (ToF) sensors. In healthcare, these devices can be used to collect and pre-process data from devices such as magnetic resonance imaging (MRI), computed tomography (CT) and others. The devices have enough power to train simple models. As part of our work, we would look at the possibilities of federal learning on the edge of devices. Federated learning has recently become increasingly popular, especially in the medical sector, because it is unnecessary to have data stored in data centres during these training models. These are new possibilities for training models in order to protect data. In this case, all data is stored on the client-side. Given that it will be video image and image processing, it will also be necessary to explore the possibilities of computer vision at the edge of the network.

This publications is a result of cooperation between Technical University of Kosice and Siemens Healthineers, within whitch Intelligent Health Lab - Siemens Healthineers Space was established.

Acknowledgment

This publication was supported by the grant Acceleration of Artificial Intelligence at the Edge of the Network (07/TUKE/2022).

References

[1] M. A. Zamora-Izquierdo, J. Santa, J. A. Martínez, V. Martínez, and A. F. Skarmeta, "Smart farming iot platform based on edge and cloud computing," *Biosystems Engineering*, vol. 177, pp. 4–17, 1 2019.

[2] M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. Netto, and R. Buyya, "Big data computing and clouds: Trends and future directions," *Journal of Parallel and Distributed Computing*, vol. 79-80, pp. 3–15, 2015, special Issue on Scalable Systems for Big Data Management and Analytics.

[3] P. Corcoran and S. K. Datta, "Mobile-edge computing and the internet of things for consumers: Extending cloud computing and services to the edge of the network," *IEEE Consumer Electronics Magazine*, vol. 5, pp. 73–74, 10 2016.

[4] M. Fazio, R. Ranjan, M. Girolami, J. Taheri, S. Dustdar, and M. Villari, "A note on the convergence of iot, edge, and cloud computing in smart cities," *IEEE Cloud Computing*, vol. 5, pp. 22–24, 9 2018.

[5] P. Papcun, E. Kajáti, and J. Vaščák, "Smart/intelligent EDGE – hardware parameters of network edge devices," *ATP Journal*, vol. 6, pp. 42–43, Jun 2019.

[6] N. Hassan, S. Gillani, E. Ahmed, I. Yaqoob, and M. Imran, "The role of edge computing in internet of things," *IEEE Communications Magazine*, vol. 56, pp. 110–115, 11 2018.

[7] I. I. at work. (2020) Real-life use cases for edge computing. [Online]. Available: https://bit.ly/3GNfC25

[8] M. De Donno, K. Tange, and N. Dragoni, "Foundations and evolution of modern computing paradigms: Cloud, iot, edge, and fog," *IEEE Access*, vol. 7, pp. 150 936–150 948, 2019.

[9] M. K. Parai, B. Das, and G. Das, "An overview of microcontroller unit: from proper selection to specific application," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 2, no. 6, pp. 145–147, 2013.

[10] M. Babiuch, P. Foltynek, and P. Smutny, "Using the esp32 microcontroller for data processing," *2019 20th International Carpathian Control Conference (ICCC)*, pp. 1–6, 5 2019.

[11] A. Maier, A. Sharp, and Y. Vagapov, "Comparative analysis and practical implementation of the esp32 microcontroller module for the internet of things," *2017 Internet Technologies and Applications (ITA)*, pp. 143–148, 9 2017.

[12] I. Allafi and T. Iqbal, "Design and implementation of a low cost web server using esp32 for real-time photovoltaic system monitoring," in *2017 IEEE Electrical Power and Energy Conference (EPEC)*, 2017, pp. 1–5.

[13] K. Dokic, "Microcontrollers on the edge – is esp32 with camera ready for machine learning?" in *Image and Signal Processing*, A. El Moataz, D. Mammass, A. Mansouri, and F. Nouboud, Eds. Cham: Springer International Publishing, 2020, pp. 213–220.

[14] M. Thothadri, "An analysis on clock speeds in raspberry pi pico and arduino uno microcontrollers," *American Journal of Engineering and Technology Management*, vol. 6, no. 3, pp. 41–46, 2021.

[15] R. P. Foundation. (2022) About us. [Online]. Available: https://www.raspberrypi.org/about/

[16] N. Alee, M. Rahman, and R. B. Ahmad, "Performance comparison of single board computer: A case study of kernel on arm architecture," *2011 6th International Conference on Computer Science & Education (ICCSE)*, pp. 521–524, 2011.

[17] P. J. Basford, S. J. Johnston, C. S. Perkins, T. Garnock-Jones, F. P. Tso, D. Pezaros, R. D. Mullins, E. Yoneki, J. Singer, and S. J. Cox, "Performance analysis of single board computer clusters," *Future Generation Computer Systems*, vol. 102, pp. 278–291, 1 2020.

[18] L. Pomsar, E. Kajati, and I. Zolotova, "Deep learning powered class attendance system based on edge computing," in *2020 18th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, 2020, pp. 538–543.

[19] R. Chancharoen and K. Maneeratana, "Introducing raspberry pi and its peripherals to a mechatronics course under covid-19 disruption," in *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 2020, pp. 883–888.

[20] A. Brecko, F. Burda, P. Papcun, and E. Kajáti, "Applicability of opc ua and rest in edge computing," sAMI Conference 2022.

[21] L. Pomšár, A. Brecko, and I. Zolotová, "Brief overview of edge ai accelerators for energy-constrained edge," sAMI Conference 2022.

[22] A. A. Süzen, B. Duman, and B. Şen, "Benchmark analysis of jetson tx2, jetson nano and raspberry pi using deep-cnn," in *2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, June 2020, pp. 1–5.

[23] Google. (2020) Dev board. [Online]. Available: https://coral.ai/products/dev-board/

[24] M. Antonini, T. H. Vu, C. Min, A. Montanari, A. Mathur, and F. Kawsar, "Resource characterisation of personal-scale sensing models on edge accelerators," in *Proceedings of the First International Workshop on Challenges in Artificial Intelligence and Machine Learning for Internet of Things*, ser. AIChallengeIoT'19. New York, NY, USA: Association for Computing Machinery, 2019, p. 49–55. [Online]. Available: https://doi.org/10.1145/3363347.3363363

[25] A. C. Inc. (2020) Mini pc. [Online]. Available: https://www.asus.com/bt/Displays-Desktops/Mini-PCs/PN-PB-series/Mini-PC-PN61T/

[26] P. Dukan, A. Kovari, and J. Katona, "Low consumption and high performance intel, amd and arm based mini pcs," in *2014 IEEE 15th International Symposium on Computational Intelligence and Informatics (CINTI)*, 2014, pp. 127–131.

[27] Y. Wang, Q. Wang, S. Shi, X. He, Z. Tang, K. Zhao, and X. Chu, "Benchmarking the performance and energy efficiency of ai accelerators for ai training," in *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*, 2020, pp. 744–751.

[28] M. Kvetko, J. Mocnej, L. Pomsar, and I. Zolotová, "Raspberry pi and windows 10 powered intelligent modular gateway for decentralized iot environments," *Acta Electrotechnica et Informatica*, vol. 20, pp. 44–50, 06 2020.

Simplified inverse fuzzy model for an induction motor drive control

[1]*Marek FEDOR (2nd year)*
Supervisor: [2]Daniela PERDUKOVÁ

[1,2] Department of Electrical Engineering and Mechatronics, Technical University of Kosice, Slovak Republic

[1]marek.fedor.4@student.tuke.sk, [2]daniela.predukova@tuke.sk

***Abstract*— Even after the adoption of some simplifying assumptions, the three-phase induction motor presents a complex nonlinear dynamic system of the 5th order with multiple inputs. Modern methods of artificial intelligence (fuzzy approach, neural networks) make it possible to find its much simpler and practically usable models, based on the description of its important qualitative properties. The presented article deals with the construction of a simplified inverse fuzzy model of an induction motor using the estimation of critical parameters of the Kloss relation, which will be used for its control (note: this will be presented in the final paper). To build such a model, it is necessary to perform several standard measurements only – step responses on the electrical motor, and no a priori knowledge of motor parameters is required. The obtained simple fuzzy model of the induction motor (IM) is then used for linearization of its torque loop and a design of superior control loops (speed, position). This also eliminates the need for transformations of motor quantities, as is usual in vector control.**

***Keywords*— nonlinear systems control, induction motor, Kloss formula, fuzzy inverse model**

I. INTRODUCTION

Nowadays, the control of linear systems is elaborated in detail, and, in practice, it is possible to design it analytically for every real system. However, this does not apply in the field of nonlinear systems control, where existing control methods are limited to certain selected groups of nonlinear systems due to the invalidity of the superposition principle. This means, there are not any universal analytical methods to design their control. Fuzzy systems present a successful approach to control of many complex nonlinear systems. The fuzzy logic control (FLC) can be considered as one of the most successful applications of the fuzzy systems, which in many cases has become an alternative to the use of conventional control techniques in various areas including: power systems [1-2], mechanical-robotic systems [3-5], automotive systems [6-7], electrical motors [8-9], power electronic systems [10] and other areas. The control based on the inverse fuzzy model is also known in the field of FLC [11]. This method of the control has many modifications, depending on the specific application [12], while the quality of the fuzzy model of the controlled system is also very important. In the field of electric drives, a typical representative of a strongly nonlinear system is a three-phase induction motor, the analytical description of which, under certain simplifying assumptions, consists of a system of five nonlinear differential equations. This paper shows a methodology of a nonlinear system control design using its linearization by the method of inverse fuzzy model and, based on a general derivation, its application to simple drive control with an induction motor is presented.

II. MATHEMATICAL MODEL OF AN INDUCTION MOTOR

Several different types of mathematical models of an IM are published in the literature, depending on which motor quantities are chosen as state variables of the model and on the reference system of the selected variables. In general, we can say that from the point of view of control the IM presents a system of five first-order nonlinear differential equations, which cannot be solved analytically. If the IM stator current and the rotor flux are chosen as system state variables, then its mathematical model can be described in state space by the following set of equations [16], expressed in the matrix form:

$$\begin{bmatrix} \frac{\mathrm{d}i_{1x}}{\mathrm{d}t} \\ \frac{\mathrm{d}i_{1y}}{\mathrm{d}t} \\ \frac{\mathrm{d}\psi_{2x}}{\mathrm{d}t} \\ \frac{\mathrm{d}\psi_{2y}}{\mathrm{d}t} \end{bmatrix} = \begin{bmatrix} -\omega_0 & \omega_1 & -K_{12}\omega_g & -K_{12}\omega_m n_p \\ -\omega_1 & \omega_0 & K_{12}\omega_m n_p & -K_{12}\omega_g \\ M\omega_g & 0 & -\omega_g & \omega_2 \\ 0 & M\omega_g & -\omega_2 & \omega_g \end{bmatrix} \begin{bmatrix} i_{1x} \\ i_{1y} \\ \psi_{2x} \\ \psi_{2y} \end{bmatrix} + \begin{bmatrix} K_{11} & 0 \\ 0 & K_{11} \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_{1x} \\ u_{1y} \end{bmatrix} \quad (1)$$

$$n_p \frac{M}{L_2}\left(\psi_{2x} i_{1y} - \psi_{2y} i_{1x}\right) - T_L = J\frac{d\omega_m}{dt} \quad (2)$$

In this model, the motor variables (like the input stator voltage vector $\mathbf{U}_1$, stator current vector $\mathbf{i}_1$ and rotor flux vector $\boldsymbol{\psi}_2$) are expressed by their components in a rectangular coordinate system {x, y}, which relatively to the stator rotates at the angular speed of the stator field ω_1. The parameters in (1), (2) can be determined directly from the motor parameters according to the following relations:

$$K_{11} = \frac{3}{2}\left(L_{s1} + \frac{L_{s2}L_h}{L_{s2} + L_h}\right)^{-1} \quad (3)$$

$$K_{12} = -\frac{3}{2}\left(L_{s1} + L_{s2} + \frac{L_{s1}L_{s2}}{L_h}\right)^{-1} \quad (4)$$

$$\omega_0 = K_{11}\left[R_1 + \left(\frac{M}{L_2}\right)^2 R_2\right] \quad (5)$$

$$M = \frac{2}{3}L_h \quad (6)$$

$$\omega_g = \frac{R_2}{L_2} \quad (7)$$

$$L_2 = \frac{2}{3}(L_{s2} + L_h) \quad (8)$$

The dynamics of the motor variables after its connection to the stator voltage $U_1 = 220$ V with the frequency $\omega_1 = 314$ rad/s is shown in Fig. 1. This relatively complex IM model can be simply replaced by a nonlinear dynamic system of the 1st order, while the nonlinearity of the IM can be described using the so-called Kloss relationship for the motor static torque with the nonlinearity:

$$T = \frac{2\,T_{max}}{\frac{s}{s_{max}} + \frac{s_{max}}{s}} \quad (9)$$

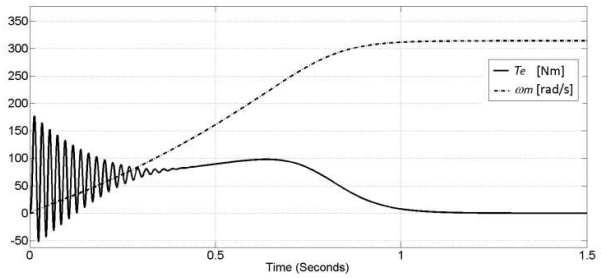


Fig. 1. Time course of the torque and angular speed when connected the IM to $U_1 = 220$ V, $\omega_1 = 314$ rad/s

where:

$$X_\delta = \omega_1 L_\delta \quad (10)$$

For the slip s and maximum slip s_{max} one applies:

$$s = \frac{\omega_1 - \omega_m}{\omega_1} \quad (11)$$

$$s_{max} = \pm\frac{R_2'}{\sqrt{R_1^2 + X_\delta^2}} \quad (12)$$

Note: calculation of t_r will be presented in the final paper.

III. Linearization of the Nonlinear Torque Characteristic of an Asynchronous Motor

The nonlinearity of the IM torque characteristic is approximately given by the Kloss relation (9), which corresponds to the simplified model IM in Fig. 2. Linearization of a static function is generally possible by a product of its inverse function. According to this relation, the torque of the IM is a function of the difference between the angular speed of the stator rotating magnetic field and the mechanical angular speed of the rotor ω_2, where:

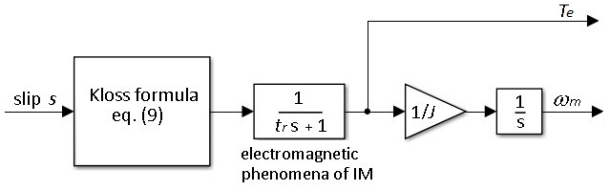
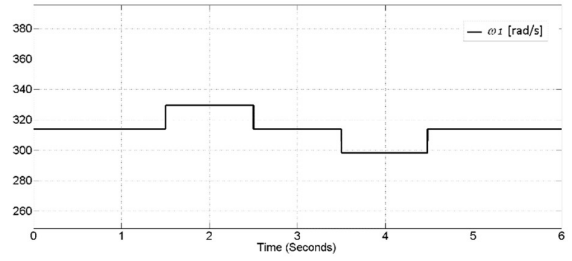
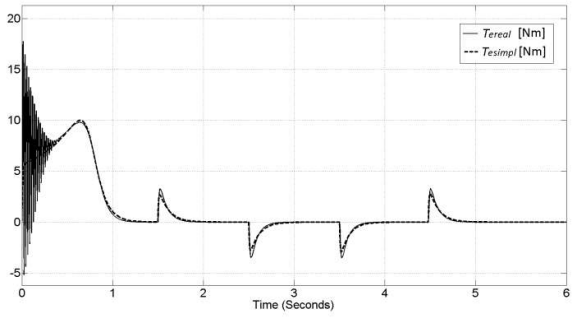


Fig. 2. Block diagram of an induction motor simplified model



a)



b)

Fig. 3. a) Stator frequency steps ω_1; b) Comparison of the torque dynamics of a complete (T_{mreal}) and of a simplified (T_{msimpl}) model of the IM

$$\omega_2 = \omega_1 - \omega_m = s\,\omega_1 \quad (13)$$

$$\omega_{2max} = s_{max}\,\omega_1 \quad (14)$$

$$T = \frac{2M_{max}\,s_{max}\,\omega_1\,\omega_2}{s_{max}^2\,\omega_1^2 + \omega_2^2} \quad (15)$$

$$T = \frac{2M_{max}\,\omega_{2max}\,\omega_2}{\omega_{2max}^2 + \omega_2^2} \quad (16)$$

Graphically, this nonlinear function is shown in Fig. 4a and its inverse function is in Fig. 4b. It is clear from Fig. 4b that the inversion of this particular nonlinearity is not unambiguous, i.e., for example: for the motor torque $T_e = 7.12$ Nm the slip angular speed can have two values: $\omega_2 = 41$ rad/s and $\omega_2 = 41$ rad/s.

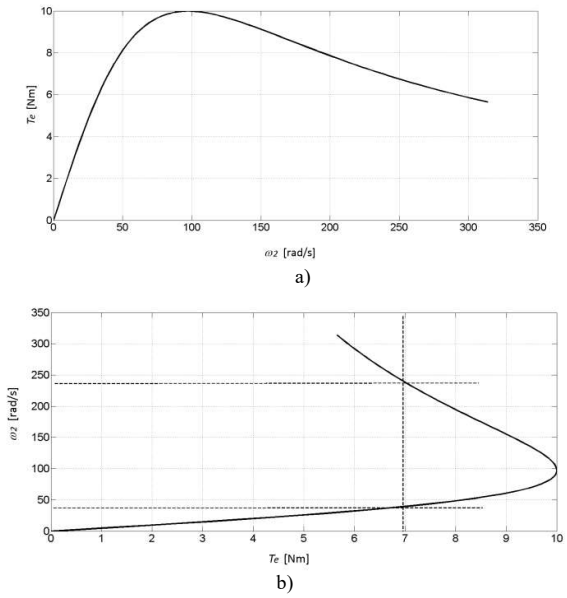


Fig. 4. Static torque characteristic of theinduction motor: a) original function, b) its inversion

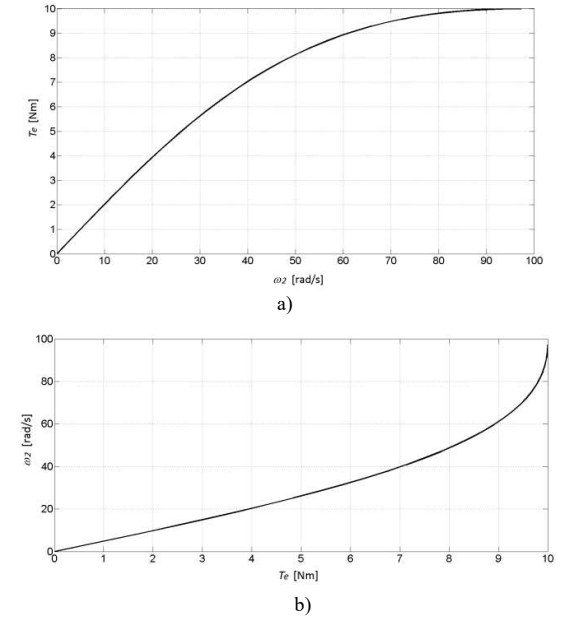


Fig. 5. Modified static torque characteristic IM: a) original function, b) its inversion

When linearizing this nonlinear function, therefore it will be necessary to work only in an unambiguous part of the nonlinearity. For this reason, the range of the quantity ω_2 is limited to the interval <0, ω_{2max}>. The nonlinearity adjusted in this way is shown in Fig. 5a and its inversion is in Fig. 5b.

IV. Fuzzy Linearization of IM Torque Characteristicthe equations are as follows

Today the fuzzy systems are considered as a suitable tool for the approximation of nonlinear functions. Chapter III presented the procedure for linearization of the nonlinear characteristic of the IM for the selected operating point.

To replace the inverse nonlinearity of the IM torque characteristic, a fuzzy system of the Sugeno type with three rules was used, which approximates the nonlinearity of IM (Fig. 6).

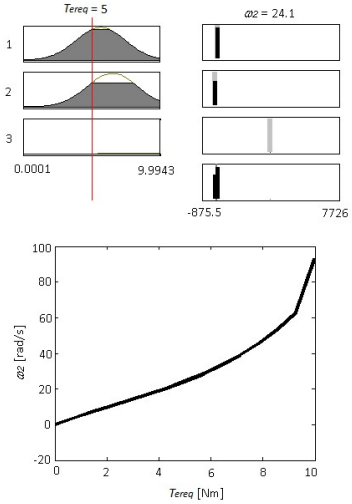


Fig. 6. Aproximation of the inverse nonlinearities of the IM using the fuzzy system

The dynamics of the torque control loop IM based on the fuzzy system is practically the same as the dynamics shown in Fig. 5b.

V. CONCLUSION

The presented paper shows an approach to linearization of an IM drive using fuzzy inversion of its fundamental nonlinearity expressed analytically by the Kloss relation. The use of fuzzy approximation of this nonlinearity also eliminates the need for its analytical knowledge and analytical inversion. Such a linearization of the torque loop of the IM drive in principle does not require any transformations and allows the design of superior control structures (speed or position control) on the basis of known standard methods of linear control.

The use of an inverse fuzzy system of a specific IM also allows its simple application to multi-parameter input values and the adjustment of the nonlinearity of the drive for various operating states and variable motor parameters.

Acknowledgment

This work was supported by the Slovak Research and Development Agency under the Contract no. APVV-19-0210. I also wish to thank Mrs. Professor Daniela Perduková and Mr. Professor Pavol Fedor for their guidance so far.

References

[1] N. K. Kumar and V. I. Gandhi, “Implementation of fuzzy logic controller in power system applications,” J. of Intelligent & Fuzzy Systems, vol. 36, no. 5, pp. 4115-4126, 2019. DOI: 10.3233/JIFS-169971.

[2] M. Gaurkar and Ch. Gowder, “Aapplication of fuzzy logic for power system,” IJARIIE, vol. 3, no. 3, 2017. ISSN(O)-2395-43.

[3] J. Bačík J., F. Ďurovský, P. Fedor, and D. Perduková, “Autonomous flying with quadrocopter using fuzzy control and ArUco markers,” Intelligent Service Robotics, vol. 10, no. 3, 2017, pp. 185-194. ISSN 1861-2776. DOI: 10.1007/s11370-017-0219-8.

[4] P. Fedor and D. Perduková, “Model based fuzzy control applied to a real nonlinear mechanical system,” Iranian Journal of Science and

Technology, Trans. of Mechanical Eng., vol. 40, no. 2, pp. 113-124, 2016. ISSN 2228-6187. DOI 10.1007/S40997-016-0005-9.

[5] Ch. H. Chen, Ch. Ch. Wang, Y. T. Wang, and P. T. Wang, "Fuzzy logic controller design for intelligent robots," Mathematical problems in Eng., vol. 2017, Article ID 8984713, 2017. https://doi.org/10.1155/2017/8984713.

[6] V. Ivanov, "A Review of fuzzy methods in automotive engineering applications," European Transportation Research Review, pp. 7-29, 2015.

[7] E. Uzunsoy, "A brief review on fuzzy logic used in vehicle dynamics control," Journal of Innovative Science and Eng. (JISE), vol. 2, no. 1, 2018, pp. 1-7.

[8] L. Qin, Y. Jiang, X. X. Tan, L. Y. Li, and J. Y. Zhao, "Application research of fuzzy control for AC motor on the PLC platform," Advances in Civil Eng., PTS 1-6, vol. 255-260, pp. 2136-2140, 2011.

[9] M. Strefezza, H. Kobayashi, K. Fujikawa, and Y. Dote, "Fuzzy robust-control for AC drive system," J. Systems Eng., vol. 4, no. 2, pp. 87-96, 1994.

[10] I. J. Balaguer, L. Qin, Y. Shuitao, U. Supatti and Z.P. Fang, "Control for grid-connected and intentional islanding operations of distributed power generation," IEEE Transactions on Industrial Electronics, vol. 58, no. 1, pp. 147-157, 2011, ISSN: 0278-0046.

[11] T. Kumbasar, I. Eksin, M. Guzelkaya and E. Yesil, "An inverse controller design method for interval type-2 fuzzy models," Soft Computing - A Fusion of Foundations, Methodologies and Applications, Vol. 21, no.10, pp 2665–2686, 2017. https://doi.org/10.1007/s00500-015-1966-0.

[12] D. Perduková and P. Fedor, „A model-based fuzzy control of an induction motor," Advances in Electrical and Electronic Engineering, VSB-Technical University of Ostrava, vol. 12, no. 5, pp. 635-641, 2015. ISSN 1336-1376. DOI: 10.15598/aeee.v12i5.1229.

Detection of Malware Samples Using Machine Learning Algorithms

[1]*Jakub PALŠA (3rd year),*
Supervisor: [2]*Liberios VOKOROKOS*

[1,2]Dept. of Computer and Informatics, FEI TU of Košice, Slovak Republic
[1]jakub.palsa@tuke.sk, [2]liberios.vokorokos@tuke.sk

***Abstract*—The goal is to perform a dynamic analysis of programs designed for the Windows operating system and then process the results of the analysis into a dataset. Afterwards, we analyze, implement and test methods of decision tree, random forest, support vectors and naive Bayes. Their ability to distinguish between a malware and a benign sample will be verified and their success rate evaluated using the metric of classification accuracy.**

***Keywords*—machine learning, dynamic analysis, malware**

I. Introduction

With the advent of artificial intelligence and its integration into many spheres of human life, the field of computer security also became interested in its use. Therefore, artificial intelligence practices are currently being incorporated into computer security tools in an effort to detect new forms of malware that security analysts have not yet encountered [1]. Machine learning is an area of artificial intelligence that allows the machine to be trained on available, known forms of malicious code in order to subsequently detect its new variations [2].

It is the detection of malware using machine learning models that is the focus of this work. The aim is to perform a dynamic analysis of the collected program samples, the results of which will be processed in the form of a data set. The data set is used in the implementation of various classification models. These will then be evaluated and their success in recognizing malicious code compared to each other.

II. Dynamic Analysis

The principles of dynamic analysis are used after static analysis did not provide enough information to confirm or refute the dangers of the analyzed software, or static analysis could not be performed due to the use of packaging or obfuscation methods by the malware's author [3].

A. Acquiring the test samples

Malicious samples were obtained from the repository *www.virusshare.com*. The package of September 15, 2015 was downloaded. The work focuses on the analysis of malware that attacks devices with the Windows operating system, because this system is the target for the largest amount of malware created and therefore the date of September 15, 2015 is potentially the most appropriate for two main reasons:

1) The most common version of the Windows operating system in 2015 was Windows 7 with a 62.31% market share [4].
2) The version of Windows 10 was released on July 29, 2015. From this it can be concluded that at that time a large amount of malware was uploaded to *VirusShare.com* by users of Windows 7, the target system for this work.

Benign samples used in this work are made up of executable programs. They were obtained from *www.portablefreeware.com* and *www.portableapps.com*. For our needs, **3000 malware samples** were selected in an executable format with the extension *.exe*. Benign samples are represented by **838 executable programs**, see the TABLE I.

TABLE I
Number of Samples Prepared for Analysis

sample type	number of samples
malware	3000
benign	838
total	3838

B. Performing analysis

The first step was to configure the **Cuckoo Sandbox** [5] environment. The option to create a dump of the operating memory after the analysis of the sample was turned off because it would quickly fill the workstation's storage space.

The analysis will be performed in the background and it will not be possible to monitor the behavior of the virtual system on the screen. This saved system resources. For the same reason, the tool that takes screenshots of the virtual system has been deactivated. The ability to simulate mouse movements in a virtual system during analysis has been enabled to create a more trusted impression of the real system in which the malicious code was running.

III. Dataset Creation

Processing of the output reports of the analysis consisted of the following steps:

1) Write the names of all called Windows API functions The function names will form the attributes of the individual samples.
2) Creating a dataset structure. The number of analyzed samples indicates the number of rows in the dataset.
3) The first transition was needed to determine the number of samples and the total number of their attributes to create the structure of the dataset. The second time the

script adds each function call found in the report to the appropriate column.
4) A binary identifier has been added as the first attribute of the sample to indicate whether it is a malware (1) or a benign sample (0).
5) Save the dataset to a comma-separated values file in csv format.

The resulting dataset consists of **3765 samples** with **299 unique attributes**, see the TABLE II.

TABLE II
STRUCTURE OF THE FINAL DATASET

sample class	number of samples	number of attributes
malware	2937	299
benign	828	299
malware + benign	3675	299

IV. CREATING MACHINE LEARNING MODELS

In this work, classifiers based on four different methods were analyzed and implemented, namely the **Decision Tree (DTC), Random Forest (RFC), Support Vector Machine (SVC)** and the **Naive Bayes method (GNB)**.

The dataset was divided into training and testing sets. The training set is used to train the classifier, while the test set is used to test it and determine its success in classifying on unknown data. In this work, all classifiers were trained on **75%** of the input data set (2756 samples) and tested on the remaining **25%** (919 samples). The same ratio of classes in the training and test set was ensured.

The training and testing of the classifier consisted of the following seven steps:

1) Division of the dataset into training and testing sets.
2) Data scaling (not needed for Decision Tree and Random Forest Classifiers).
3) Choosing the classifier hyperparameter combination
4) Classifier training on training data.
5) Prediction by classifiers on test data.
6) Evaluation of classification by evaluation metric.
7) Saving prediction results to a file for each classifier.

The sequence of steps 1 to 7 was repeated 30 times to ensure both the diversity of the training and test set and also the verification of the performance of the classifier with the same combination of its hyperparameters, see the TABLE III.

TABLE III
HYPERPARAMETERS AND THEIR VALUES USED IN MODEL TRAINING

MODEL	HYPERPARAMETER	HYPERPARAMETER VALUE
DTC RFC	criterion	gini; entropy
	splitter	best; random
	min_samples_split	2; 3; 4
	min_samples_leaf	1; 2; 3; 4
	max_features	auto; sqrt; log2; None
	class_weight	balanced; None
SVC	kernel	linear
	C	0.001; 0.01; 0.1; 1; 1.5; 2; 5
	tol	0.001; 0.01; 0.1; 1; 1.5; 2; 5
	class_weight	balanced; None
GNB	priors	None
	var_smoothing	1e-12; 1e-9; 1e-6; 0.001; 0.1; 0.0; 0.1; 1; 2; 5

V. RESULTS

The **Decision Tree Classifier** achieved high classification accuracy values above **94%** with standard deviations below **1%**. None of the tested combinations of hyperparameters achieved a significantly better result compared to the others.

The best classification results were achieved by the **Random Forest Classifier**. The highest value of classification accuracy it achieved was **95.95%** with a standard deviation of only **0.58%**. In addition to the highest classification result, its effectiveness is also proven by the lowest values of the standard deviations. **Support Vector Classifier** did not achieve significantly different results based on the kernel function used. The highest result was achieved by the use of a linear function, namely **92.38% ± 0.83%**.

The use of the **Naive Bayes method** in the classification of malware and benign samples by the Gausian Naive Bayes classifier did not achieve an acceptable result. The highest value obtained was **59.53%** with a standard deviation of **1.76%**. You can see the results in the TABLE IV.

TABLE IV
COMPARISON OF THE BEST RESULTS OF INDIVIDUAL CLASSIFIERS

classifier	accuracy (%)
RFC	95,95 ± 0,58
DTC	94,53 ± 0,74
SVC_linear	92,38 ± 0,83
SVC_poly	92,17 ± 0,79
SVC_rbf	91,93 ± 0,84
GNB	59,53 ± 1,76

VI. CONCLUSION

In this work, classifiers based on four different methods were analyzed and implemented, namely the decision tree, random forest, support vector machine and the Naive Bayes method.

When evaluating the results of the classifiers, it was confirmed that the Decision Tree Classifier, despite its simplicity, can effectively predict malware and healthy samples. It has also been confirmed that Random Forest Classifier can, by aggregating the results of many Decision Tree Classifiers, obtain an even higher score than the Decision Tree Classifier itself.

In this work, we pointed out part of the research, which in this case deals with a comparison of some basic methods of machine learning. Of course, the research will continue to do so in the field of evaluation and comparison of other machine learning methods. The success of the best method will be used to build antivirus software.

REFERENCES

[1] R. Veeramani and N. Rai, "Windows api based malware detection and framework analysis," in *International conference on networks and cyber security*, vol. 25, 2012.
[2] S. Madeh Piryonesi and T. E. El-Diraby, "Using machine learning to examine impact of type of performance indicator on flexible pavement deterioration modeling," *Journal of Infrastructure Systems*, vol. 27, no. 2, p. 04021005, 2021.
[3] M. Sikorski and A. Honig, *Practical malware analysis: the hands-on guide to dissecting malicious software*. no starch press, 2012.
[4] H. Zhang, "Exploring conditions for the optimality of naive bayes," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 19, no. 02, pp. 183–198, 2005.
[5] C. Guarnieri, A. Tanasi, J. Bremer, and M. Schloesser, "Cuckoo sandbox-automated malware analysis," *URL: https://cuckoosandbox. org*, 2021.

Utilization of Wide Area Monitoring System for power system control in real time measurement

[1]*Vladimír KOHAN (3rd year)*
Supervisor: [2]Michal KOLCUN

[1,2]Dept. of Electric Power Engineering, FEI TU of Košice, Slovak Republic

[1]vladimir.kohan@tuke.sk, [2]michal.kolcun@tuke.sk

***Abstract*— The article deals with modern systems in the field of monitoring, which are related to improving the quality of control and operation of power systems, distribution gird and power plants. The points briefly describe the various areas in which WAMS systems based on the so-called PMU Units - Phasor Measurement Unit.**

***Keywords*— power system monitoring, power system control, WAMS, PMU, synchrophasor, transient phenomena.**

I. INTRODUCTION

Monitoring systems based on synchro-phasor measurements allow dispatch centers to better understand the limitations and behavior of their electrical systems. Synchro-phasors can communicate with a SCADA / EMS system to improve e.g., quality of system state estimation (estimation). Although highly sophisticated automatic stability control algorithms are not yet commercially available as standard products, the operator has at least clear information (indication) on how far the system is from a "crash" and that it must respond in a timely manner to a given disturbances. [1], [3], [8].

Advanced application of WAMS technology will allow the load distribution of the system to be directly measurable, allowing maximum safe load transfer. Impedance of branches, busbars of the system, branches, etc.

Broadband monitoring with active management brings more opportunities to further improve the operation of electricity systems.

II. VOLTAGE STABILITY

From the PMU we know the values of voltages from the phasors - from them we can determine below/overvoltage's, asymmetries... i.e., voltage instability / stability. For better evaluation, we can graphically process the voltage profiles of individual phases.

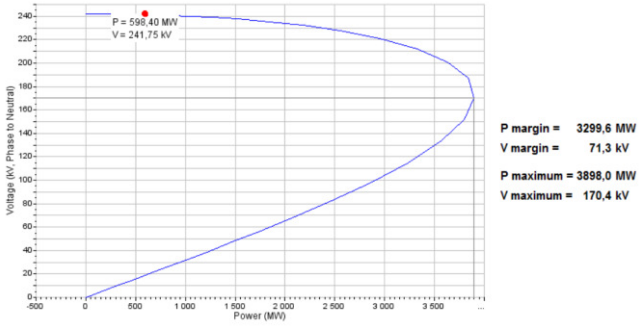


Fig. 2 Visual view of voltage stability on lines

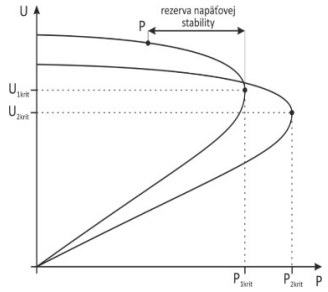


Fig. 1 Curve P-V, $P=f(U)$

III. FREQUENCY INSTABILITY

Measured values of frequency from the PMU where we can determine +/- Hz and evaluate frequency instability. In the current display, we can point out frequency instability, while in the evaluation system we can set specific limits to indicate critical values.

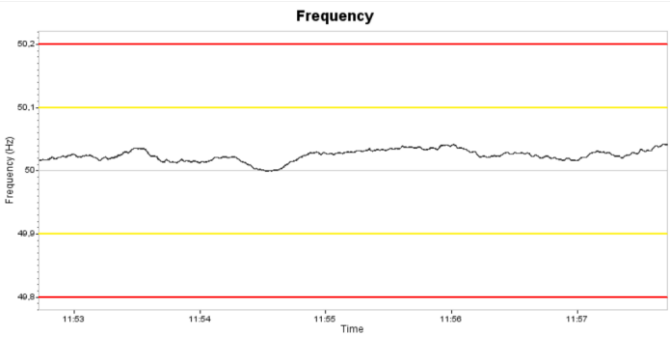


Fig. 3 Frequency monitoring with limit (yellow) and alarm (red)

IV. OSCILLATION DAMPING CONTROL

From the measurements, evaluate the most accurate values of P, Q and draw a graphical (best visual view) representation of whether the abnormal condition is dampening or not (or at what pace in the time domain). Possibility to measure on both border and national lines (e.g., at power plant outlets). The measurement takes place online at a given point in time with the oscillation damping control set (in the opposite effect, an alarm is activated). [4], [7], [14]

V. ISLAND DETECTION

Through the measured phasors (on the border lines one of the WAMS will not prove the operation of: (e.g.: none or

deformed V, I phasors) - lines with voltage but without current (no power flows through any border lines = island operation). There will be zero transmission (P, Q = 0) Domestic production will be equal to domestic consumption PMUs on border lines will not register any current.

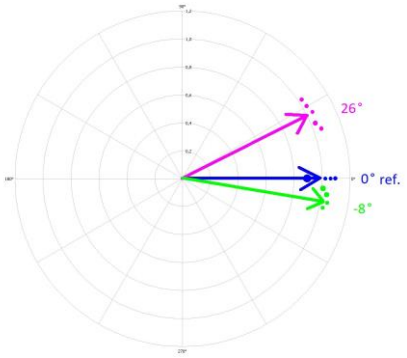


Fig. 4 Polar chart - blue-reference station, green-synchronous with reference and pink-out-of-synchronous area otherwise island operation

VI. SYSTEM ESTIMATION

Evaluation of the measurement from the PMU (location on border and significant lines) while the operation in the natural steady state speaks of the normal state of the system. Any other change from the nominal or setpoint value (f, P, Q, U, I) may lead to an abnormal system operating condition. Each parameter is monitored online and the measurement is evaluated at a given point in time. From the estimated consumption and production values, we can partially determine the power flow, with WAMS helping to refine the forecast using online actual values. Knowledge of the actual parameters of the line enables more precise setting of protections and is the basis for the correct function of calculations such as condition estimation, loss calculation and network models. The possibility of analyzing the time dependence of line parameters, e.g., in connection with weather, profile change, season, etc. [5]

VII. LINE STABILITY (STATIC STABILITY) AND LINE AMPACITY AND THERMAL LOAD

Evaluation of line capacity (measurement of line load while maintaining retention power in case of undesired events). By setting the limit values from the measurement, we can use WAMS to signal limit / critical values that will begin to draw attention to the instability of the operation of individual lines. Based on the equations [10]: P=f(ϑ) and P=((3*U_1f*U_2f)/X)*sin(ϑ).

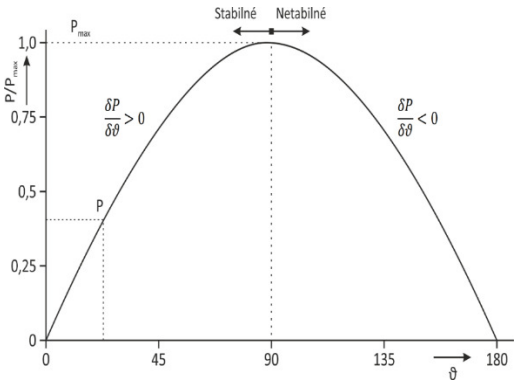


Fig. 5 Curve Pδ

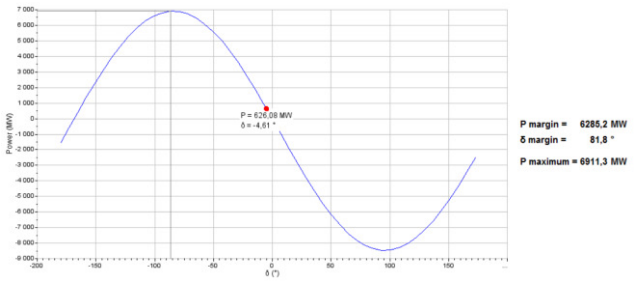


Fig. 6 A real-time visual view of an oveheadline in the Pδ plane

By measuring the currents taking into account the maximum capacities of the lines, we can calculate the thermal stress of the lines while the (approximate) temperature of the air around the line must be known. When overheating, the line will be signaled by the line overheating alarm. Factors influencing the final conductor temperature: Ambient temperature, max. wind speed, min. wind speed, max. sunlight, conductor temperature, max. temperature, line current, current limit for range, highest load for lowest limit. [6], [9]

VIII. SWING DETECTION AND ROTOR ANGLE DYNAMIC STABILITY

WAMS located at the power plant outlets provide the most accurate measurements (P, Q) from which generator oscillations can be detected. From the point of view of e.g., on border lines, WAMS can provide outputs suitable for detecting power fluctuations in both its own and the neighboring operating system. In this way, dispatching can detect power fluctuations and eliminate this undesirable event as quickly as possible. By measuring the power of the generator and the required voltages and subsequent conversion to tgδ using a selected method such as: the method of surfaces (we start using the differential equation of oscillation) determine the stability of the rotor rotor angle. This can prevent an unwanted generator failure from synchronism and the impact of CCT - critical clearing time. [11], [12]

CONCLUSION

WAMS research is a very extensive and evolving topic. However, over time, PMUs have become part of common practice and are increasingly being implemented. However, the topic of WAMS is still not definitively assessed and is worth further research in the reliability, security and quality of electricity supply, transmission and production. It can be stated that real-time measurements will play an important role in any sector focused on the field of electricity and will become an essential part of every measurement.

REFERENCES

[1] LI, Y. – YANG, D. – LIU, F. – CAO, Y. – REHTANZ, CH. 2016. Interconnected Power Systems; Wide-Area Dynamic Monitoring and Control Applications, Berlin Heidelberg: Springer-Verlag, 2016. 98-112 p. ISBN 978-3-662-48625-6.

[2] MONTI, A. – MUSCAS, C. – PONCI, F. 2016. Phasor Measurement Units and Wide Area Monitoring Systems; From the Sensors to the System, London: Academic Press is an imprint of Elsevier, 2016. 2-111 p. ISBN: 978-0-12-804569-5.

[3] MA, J. 2018. Power system; Wide-Area Stability Analysis and Control, Beijing-China: Wiley-Science Press Beijing, 2018. 23-154 p. ISBN 9781119304876.

[4] PHADKE, A. G. – THORP, J. S. 2017. Synchronized Phasor Measurements and Their Applications, Second Edition, USA: Springer International Publishing, 2017. 14-243 p. ISBN 978-3-319-50582-4.
[5] MESSINA, A. R. 2015. Wide-Area Monitoring of Interconnected Power Systems, London: The Institution of Engineering and Technology, United Kingdom, 2015. 2-98 p. ISBN 978-1-84919-853-0.
[6] VACCARO, A. – ZOBAA, A. F. 2016. Wide Area Monitoring, Protection and Control Systems; The enabler for Smarter Grids, London: The Institution of Engineering and Technology, United Kingdom, 2016. 3-86 p. ISBN 978-1-84919-830-1.
[7] BEVRANI, H. – WATANABE, M. – MITANI, Y. 2014. Power System Monitoring And Control, New Jersey: Published by John Wiley & Sons, Inc, Hoboken-IEEE Press, 2014. 31-185 p. ISBN 978-1-118-45069-7.
[8] SANTOS, L. F. – ANTONOVA, G. – LARSSON, M. – SERGIO, F. 2015. The Use of Synchrophasors for Wide Area Monitoring of Electrical Power Grids. [online]. Semantic Scholar: 1-15 p. [cit. 2020.09.21] Available on inernet: https://library.e.abb.com/public/9357d370a88948e3adcd1fdc1a216741/The%20Use%20of%20Synchrophasors%20for%20Wide%20Area%20Monitoring%20of%20Electrical%20Power%20Grids.pdf.
[9] BABNIK, T. – GÖRNER,K. – MAHKOVEC, B. 2014. Wide Area Monitorig System. [online]. Berlin Heidelberg: Springer-Verlag, 1-18 p. [cit. 2020.07.10] Available on inernet: https://www.kth.se/social/files/5829b140f276540cc86d5d1a/9%20WAMS.pdf.
[10] ABB. 2012. Wide Area Monitoring Systems; Portfolio, applications and experiences. [online]. ABB Group. 1-35 p. [cit. 2020.10.27] Available on inernet: https://library.e.abb.com/public/94fab39c67b4ac00c125784f00293520/1KHL501042%20PSGuard%20WAMS%20Overview%202012-04.pdf.
[11] ABB. 2015. Jornadas Técnicas Peru 2015 Protection, monitoring and control solutions using Wide Area Monitoring Systems. [online]. ABB Group. 1-29 p. [cit. 2020.11.07] Available on inernet: https://new.abb.com/docs/librariesprovider78/documentos-peru/presentaciones-primeras-jornadas-tecnicas-abb-peru/ps/wide-area-network-technology-and-synchrophasors.pdf?sfvrsn=2.
[12] ENTSO-E. 2021. System Separation in the Continental Europe Synchronous Area on 8 January 2021 – update. [online]. ENTSO-E. [cit. 2021.01.16] Available on inernet: https://www.entsoe.eu/news/2021/01/15/system-separation-in-the-continental-europe-synchronous-area-on-8-january-2021-update/
[13] ENTSO-E. 2021. System split registered in the synchronous area of Continental Europe – Incident now resolved. [online]. ENTSO-E [cit. 2021.01.16] Available on inernet: https://www.entsoe.eu/news/2021/01/08/system-split-registered-in-the-synchronous-area-of-continental-europe-incident-now-resolved/
[14] VANFRETTI, L – BAUDETTE, M – DOMÍNGUEZ-GARCÍA, J-L – ALMAS, M. S – WHITE, A – GJERDE, J. O. 2016. A Phasor Measurement Unit Based Fast Real-time Oscillation Detection Application for Monitoring Wind-farm-to-grid Sub–synchronous Dynamics, Stockholm: Electric Power Components and Systems, 2016. 123-134 p. ISSN: 1532-5016.

Control of energy storage system for electric midibus

[1]*Dávid BODNÁR (1st year)*
Supervisor: [2]František ĎUROVSKÝ

[1,2]Dept. of Electrical Engineering and Mechatronics, FEI TU of Košice, Slovak Republic

[1]david.bodnar@tuke.sk, [2]frantisek.durovsky@tuke.sk

***Abstract*—This paper deals with design and management of an energy storage system for electric midibus. The energy storage system (ESS) contains a battery and a supercapacitor. The role of supercapacitor is to reduce the load on the battery during acceleration and regenerative braking and thus prolong its life. An energy management system has been set up between the battery, supercapacitor, and the electric motor, which is based on the kinetic and potential energy of the vehicle. The battery model also includes an aging model. The functionality of ESS was verified by simulation in the MATLAB program. The simulation results are analyzed in detail, including an evaluation of the supercapacitor's impact on battery life, internal storage losses, and vehicle energy consumption. Finally, possible improvements and the next direction of the work are discussed.**

***Keywords*—Battery model, supercapacitor, QSS toolbox, energy management, energy storage system, battery aging.**

I. INTRODUCTION

The European Union's efforts to decarbonize the economy are putting pressure on the massive use of electric vehicles in transport. Traction batteries are still a critical component of electric vehicles (EVs). Current battery technologies do not provide such energy density, safety, and battery life that electric cars can fully compete with vehicles with internal combustion engines. Electric cars are currently suitable for operation in cities where they are not required to travel long distances, and they can advantageously use the kinetic energy of the vehicle during braking due to their ability to recover. Night charging is battery-friendly and at the same time helps to balance the consumption of electricity from the public network. An interesting possibility is their ability to accumulate (supply) part of the electricity from the public network, and thus help to compensate fluctuations in energy supply and consumption in the public distribution network. Despite these possibilities, electric vehicles have not yet found enough customers among ordinary users, who prefer long-range vehicles and convenient use outside cities. The situation is different in the area of public transport. Public transport vehicles travel on permanent routes with frequent starts and braking. In addition, they drive in an area where the emphasis is on minimal or no emissions. The deployment of electric buses is supported by states and local governments, so a wider deployment of electric vehicles can be expected in this area.

II. CURRENT STATUS OF THE ACTUAL KNOWLEDGE

The development of battery technologies is moving in several directions. The main challenge is to change the technology of the battery itself in order to increase its energy density, the ability to fast charge and discharge without damaging the battery and increase its safety. In addition, there are other ways to extend battery life. High charging and discharging currents are not desirable for battery lifetime. However, during its operation, EV requires increased power during acceleration as well as the ability to absorb energy quickly during braking. Such driving cycles are typical, especially for urban traffic.
To protect the battery from high currents, but at the same sustain possibility to accelerate rapidly, a device capable of receiving and providing a large amount of energy in a short time is needed. One solution is to add a supercapacitor (SC) to the traction battery. The SC is characterized by high specific power, which means that it can receive and supply high currents. On the other hand, it has low specific energy, so it does not have much effect on extending the range of the vehicle. Ensuring a sufficient range of the vehicle remains a responsibility of the battery. As a result, it is advantageous for the high power peaks to be covered by the SC and the long-term load by the battery. The SC has a significantly lower internal resistance than the average battery, which will reduce the overall internal losses of the power storage and increase its efficiency. However, the use of a battery-SC combination also has its disadvantages. The SC voltage varies significantly based on the state of charge of the SC, which requires the use of an additional DC/DC converter. The SC would be connected to the DC bus via the DC/DC converter. In addition, the inclusion of the SC and another converter in the vehicle's traction chain increases the cost and weight of the vehicle. The size of the batteries and the SC must therefore be optimized for the application.

A. Topologies of the energy storage system

There are several possible battery and SC connection topologies [1]. Improvements in efficiency and performance are achieved when the battery and SC are connected in parallel without the use of a DC/DC converter, but the SC potential remains largely unused. By adding a DC/DC converter, the possibility of operating the SC in a wide range of voltages is achieved and a larger percentage of the SC capacity can be used [2]. The use of a bidirectional DC/DC converter for SC seems

to be a good compromise between achieving the desired effect and the cost of the system. Due to these advantages, the mentioned solution is currently the most frequently used topology [3]. One of the proposed possibilities is to use a pair of complementary transistors instead of a DC/DC converter [4]. Either the battery or the SC would be in operation. This would improve the efficiency of the ESS, but the DC bus voltage would fluctuate significantly with respect to the SC voltage, which would require a more complex motor-drive control algorithm.

B. Energy management

The energy flows between the battery, SC, and motor are controlled by energy management. Many different approaches have been proposed, which have been based on optimization techniques, fuzzy logic, neural networks, etc. One way is to use the minimization function [5]. This design covers a wide range of EV powers, but the SC is sometimes discharged to zero and relatively heavily loaded. The energy management algorithm is well described in [6] and is based on the control of the state of charge of the SC with respect to the kinetic energy of the vehicle. A good option is to use fuzzy logic [7]. In this design, however, discharging prevailed over charging, so there were situations where the SC was significantly discharged even at low speeds. Another possibility is that the SC would perform the function of a high-pass filter, where the low-frequency load would cover the battery and the high-frequency load SC [8].

C. Battery models

In order to verify the correctness of the design by simulation, it is necessary to have an accurate battery model. There are several different battery models. An example is a battery model from MATLAB/Simulink. It is a block "Equivalent Circuit Battery" compiled according to [9] or "battery" from the Simscape library [10], [11]. With the "Equivalent Circuit Battery", the battery voltage is determined using table values, which are a function of the state of charge and the battery temperature. The advantage of this model is that the user can set the battery parameters through the discharge characteristics of the battery. Another model named "battery" from the Simscape library is a bit more complicated, the discharge characteristic is described by equations. The thermal model [12], [13], and the battery aging model [14] are also included. This block is more suitable for the abovementioned ESS because of the possibility to simulate battery aging. In addition, a supercapacitor model is included in [15] and can be used in the simulation.

III. MATHEMATICAL MODEL OF MIDIBUS

The mathematical model of the vehicle can be based on the quasi-static modeling method or on the dynamic modeling method. The quasi-static method was used in [17], the vehicle model with the ESS was created in the QSS Toolbox [16]. QSS TB provides the ability to easily build a model of different vehicle propulsion systems and estimate its consumption for different driving cycles, while does not require the high computing power of the computer. QSS TB uses a quasi-static modeling method, in which only mechanical transients are considered. The main idea of the QSS TB is to reverse the usual cause-and-effect relationships of dynamic systems. Instead of calculating speeds from given forces, the QSS TB calculates accelerations and determines traction resistances of the vehicle according to predetermined speeds (at discrete times). [16]

In the simulation, the entire driving cycle is divided into short periods of time also called a step size. The step size has to be chosen so that the value of traction resistances does not change significantly between two adjacent steps. Based on the prescribed speed profile, the acceleration of the vehicle, and the traveled distance, are calculated. Based on them and on the parameters of individual components of the traction chain, the required motor power and energy are calculated.

The model of the whole powertrain consists of several subsystems (Fig. 1). The "Driving cycle" block allows you to select one of the commonly used driving cycles, but also offers the possibility to load your own driving cycle data measured on a real route. The block also allows loading data about altitude and charging time. In the next block named "Vehicle", the traction resistances are calculated according to the vehicle parameters. The "Single Transmission" block represents a single-speed transmission model. In the "Electric Motor" block, the required electrical input of the motor is calculated from the angular speed and torque of the motor. Motor efficiency is determined here based on the efficiency map. The calculated power must be provided by the ESS. This power is divided between the battery and the SC by the energy management system.

A. Energy management system

The energy management system in [6] works on the principle of controlling the state of charge of the SC with respect to the kinetic energy of the vehicle. This allows high utilization of the energy received during regenerative braking and provides energy for high accelerations at any time. The goal is to discharge the SC in such a way that there is anytime

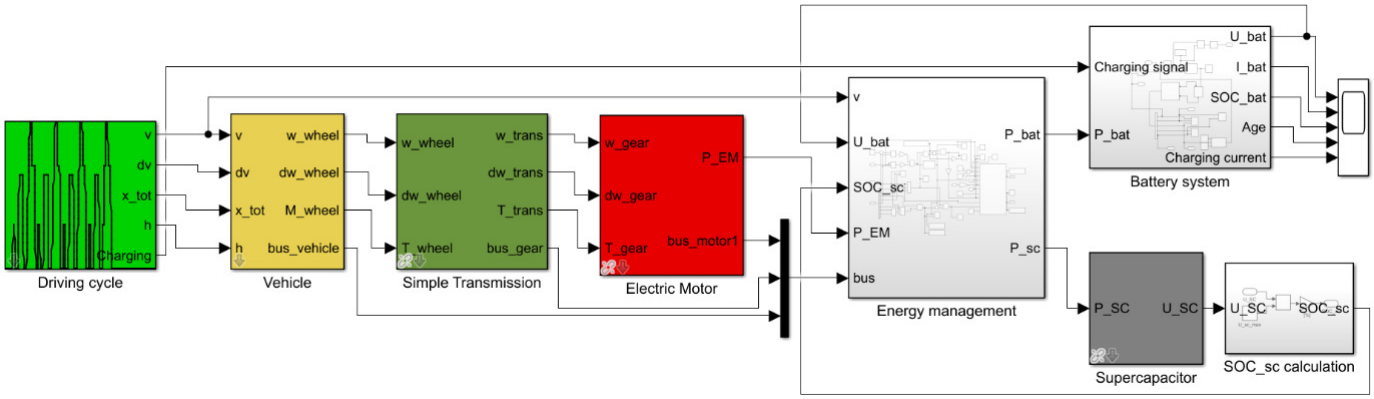


Fig 1 Model of midibus built by using QSS Toolbox and model of the ESS with energy management.

enough space left to absorb the energy from the regenerative braking. The SC is discharged by lower power, so the SC is fully charged when braking at zero speed. This requirement prevents excessive discharge, and therefore it is possible to start a new driving cycle even at high acceleration requirements. The "Energy management" unit contains relatively complicated calculations and an algorithm, which is described in detail in [6]. As a result, in the case of a long-term load, energy is taken from the battery, and in the case of a short-term peak load, the SC also cooperates, while only the SC works in the case of recuperation braking. An example of power distribution in the ECE driving cycle is shown in fig. 2.

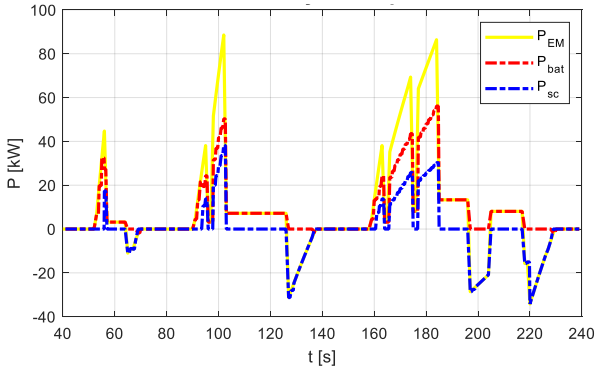


Fig 2 Power distribution between battery and SC

B. Energy storage system

The model of ESS consists of two main parts, the battery model from Simscape and the supercapacitor from the QSS toolbox. Additional blocks were added to them in [17], which are used to control the charging and discharging of the battery, to determine the state of charge of the SC, to calculate the range of the vehicle, and to calculate internal losses. More details are given in [17] and [18].

IV. Simulation of Midibus Model

The aim of the simulations in [17] was to find out how significant benefits is the SC capable to provide. The change in battery aging, energy consumption, internal losses, vehicle range, and other factors were monitored. Two driving cycles, ECE and NEDC, were used in the simulation, with the maximum speed at NEDC being limited to 90 km/h. In the first simulation, the ESS contained only the battery. Recuperative and non-recuperative driving was simulated, so it was possible to calculate the amount of energy that can be potentially saved by regenerative braking. In other simulations, the ESS was a combination of the battery and the SC. The total capacity of the SC was composed of several supercapacitor modules connected in parallel. The number of SC modules can be changed with respect to the driving cycle. The simulations also monitored the influence of the maximum current flow from the battery.

A. Recuperative and non-recuperative simulations

First, the ECE and NEDC cycles were simulated when powered only by the battery, the SC was not used. In fig. 3 are current flows for the ECE cycle with and without recuperation. The results of energy consumption showed that without recuperation there was a consumption of 24.85 kWh and with recuperation 18 kWh. By recuperation, savings of up to 27.5% can be achieved in this driving cycle. During the NEDC driving cycle, the consumption reached 59.56 kWh without recuperation and 47.3 kWh with recuperation. In this case, the savings were 20.57% of energy.

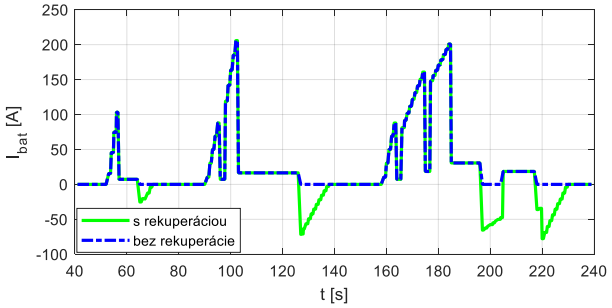


Fig 3 Current flow supplied by the battery only with (green) and without (blue) regenerative braking for ECE driving cycle

B. Simulations of battery aging

Subsequently, simulations were performed using SC. Battery aging, internal losses, vehicle range, and more were monitored. Simulations showed that the current flowing through the battery was significantly lower when using SC, which corresponds to a slowing down of battery aging. The average current flowing through the battery was almost twice as small. Acceleration peaks are significantly reduced and peaks during regenerative braking are even completely removed (Fig. 4). The range of the vehicle was extended only slightly. However, the efficiency of the ESS components has a huge influence on the range. More accurate battery and DC-DC converter models are required for more accurate vehicle range simulations.

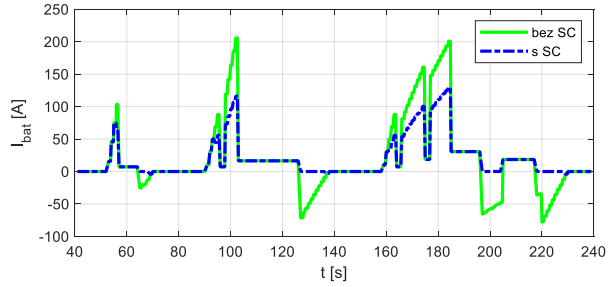


Fig 4 Battery current flow with (blue) and without (green) the SC for ECE driving cycle

Another way to solve energy management could be recovering small currents to the battery and large currents to the SC. The advantage is that a smaller total SC capacity could be used. It could be also a way to improve the overall efficiency of the ESS.

V. Possibilities for Model and Simulations Improvement

The paper presents the possibility of using a combined energy storage system for an electric midibus. The ESS contains a battery and a supercapacitor and is equipped with an energy management system that controls the distribution of energy between the two components of the ESS. The analysis was performed by simulation in the Matlab/Simulink, and the battery model also provides battery aging. The simulations have confirmed that the currents flowing through the battery can be significantly reduced by using a supercapacitor. This mainly applies to peak battery currents during the acceleration and braking of the vehicle. Although the extension of the vehicle's range did not increase significantly with the use of the supercapacitor, the aging of the battery slowed down, which

was about half as slow according to the simulation results. As a result, the battery lasts longer in operation without any need for replacements. The energy management of the ESS was based on the use of the vehicle's kinetic energy and ensured that the supercapacitor was fully charged and ready for the vehicle's subsequent acceleration. The management was supplemented by an algorithm that was also able to take into account the potential energy of the vehicle when driving on irregular terrain. However, for high altitude changes, the SC must be oversized, so power supplies from the battery to overcome elevation seem to be a better option.

Based on the above simulations, it can be stated that the use of SC in the source chain of the vehicle can significantly reduce battery stress. The aging data were obtained only on the basis of the aging model of the "Battery" block in Simscape, which is not sufficient. The ideal way to determine the degree of aging of the battery is to physically discharge it on a battery cell tester, which would determine the aging rate at different currents. The results would later be used to modify the battery model and optimize the sizing of the battery-SC assembly. The optimal sizing is important due to the high initial costs of the SC and DC-DC converter. Another drawback of the ESS is higher weight, which could cause a lowering of the range. The ESS also requires more space in the vehicle and is more complex than a single battery ESS.

The internal losses of the ESS were also discussed. Although the SC has a lower internal resistance than a battery cell, the overall internal resistance of the SC module is usually just slightly lower than the battery pack internal resistance. With high current flow through the ESS, the internal losses increase significantly. Hence, the overall efficiency of the ESS could be studied and optimized in the future. Optimized efficiency could not just prolong battery lifetime, but also improve the range of the vehicle. The efficiency of the DC-DC converter also significantly affects the overall efficiency of the ESS. Modern DC-DC converters reach efficiency up to 98%, but it also depends on the transferred power and actual voltage of the SC. Therefore, the optimization of the operation point of the DC-DC converter is crucial for ESS efficiency. A new control algorithm could improve efficiency and the range of the vehicle.

Besides that, different battery technologies could be studied in this type of energy storage system. The battery pack can be specified on high energy density, so the battery pack is lighter. However, such a pack has a lower power density and higher internal resistance. High power demands are supplied by SC, so it is possible to lower the weight of the vehicle and to preserve the good driving characteristics of the vehicle. However, the benefits of this design are questionable because of the higher internal resistance of the battery and the unknown aging of different battery technologies. It requires more battery testing and modeling.

The most significant benefit of combined energy storage is the extension of battery life. By maintaining sufficient capacity for a longer period of time, the battery does not need to be replaced and recycled. The energy requirements for battery recycling with current technologies are significantly higher than with the production of a new battery. Therefore, the possibility of longer operating life of the original battery in a vehicle is a significant contribution to environmental protection.

ACKNOWLEDGMENT

This work was supported by Slovak Research and Development Agency on the basis of Contract no. APVV-15-0750.

This work was supported by Slovak Research and Development Agency on the basis of Contract no. APVV-18-0436.

REFERENCES

[1] A. Khaligh, L. Zhihao, “Battery, ultracapacitor, fuel cell, and hybrid energy storage systems for electric, hybrid electric, fuel cell, and plug-in hybrid electric vehicles: state of the art”. IEEE Transactions on Vehicular Technology, 2010; 59(6): pp. 2806–2814.

[2] L. Gao, R. A. Dougal, and S. Liu, “Power enhancement of an actively controlled battery/ultracapacitor hybrid,” IEEE Transaction on Power Electronics., vol. 20, no. 1, pp. 236–243, Jan. 2005.

[3] L. Kouchachvili, W. Yaïci, E. Entchev, “Hybrid battery/supercapacitor energy storage system for the electric vehicles”, J. Power Sources 374 (2018) pp. 237–248.

[4] L. Shuai, K. A. Corzine, M. A. Ferdowsi, “A new battery/ultracapacitor energy storage system design and its motor drive integration for hybrid electric vehicles”, IEEE Transactions on Vehicular Technology, 2007; 56(4): pp. 1516–1523.

[5] M.-E. Choi, S.-W. Seo, “Robust energy management of a battery/supercapacitor hybrid energy storage system in an electric vehicle”, in: Electric Vehicle Conference (IEVC), 2012 IEEE International, March 2012, pp. 1-5.

[6] J. Armenta, C. Núñez, N. Visairo, I. Lázaro, “An advanced energy management system for controlling the ultracapacitor discharge and improving the electric vehicle range”, J. Power Sources 284 (2015) pp. 452–458.

[7] L. Rosario, P.-K. Luk, “Applying management methodology to electric vehicles with multiple energy storage systems”, in: International Conference on Machine Learning and Cybernetics, 2007, vol. 7, Aug 2007, pp. 4223 -4230.

[8] E. Schaltz, A. Khaligh, P. O. Rasmussen, “Influence of battery/ultracapacitor energy-storage sizing on battery lifetime in a fuel cell hybrid electric vehicle”. IEEE Transactions on Vehicular Technology, 2009; 58(8), pp.3882–3891.

[9] R. Jackey, "A Simple, Effective Lead-Acid Battery Modeling Process for Electrical System Component Selection." SAE Technical Paper 2007-01-0778. doi:10.4271/2007-01-0778, 2007.

[10] O. Tremblay, L. A. Dessaint, and A. I. Dekkiche, “A generic battery model for the dynamic simulation of hybrid electric vehicles,” in Proc. IEEE Vehicle. Power Propulsion Conf. (VPPC 2007), pp. 284–289.

[11] Tremblay, O., L.A. Dessaint, "Experimental Validation of a Battery Dynamic Model for EV Applications." World Electric Vehicle Journal. Vol. 3, May 13–16, 2009.

[12] Saw, L.H., K. Somasundaram, Y. Ye, and A.A.O. Tay, “Electro-thermal analysis of Lithium Iron Phosphate battery for electric vehicles.” Journal of Power Sources. Vol. 249, pp. 231–238.

[13] Zhu, C., X. Li, L. Song, and L. Xiang, “Development of a theoretically based thermal model for lithium ion battery pack.” Journal of Power Sources. Vol. 223, pp. 155–164.

[14] Omar N., M. A. Monem, Y. Firouz, J. Salminen, J. Smekens, O. Hegazy, H. Gaulous, G. Mulder, P. Van den Bossche, T. Coosemans, and J. Van Mierlo. “Lithium iron phosphate based battery — Assessment of the aging parameters and development of cycle life model.” Applied Energy, Vol. 113, January 2014, pp. 1575–1585.

[15] Guzzella L., Onder C. H.: Introduction to Modeling and Control of Internal Combustion Engine Systems, Springer Verlag, Berlin, 2004.

[16] QSS toolbox. ETH Zürich – Swiss Federal Institute of Technology. https://idsc.ethz.ch/downloads.

[17] D. Bodnár, Predikcia spotreby elektrického midibusu. Diplomová práca. FEI TU Košice, 2021.

[18] Riadenie zásobníka energie elektrického midibusu / Dávid Bodnár, František Ďurovský - 2021. In: Electrical Engineering and Informatics 12 : Proceedings of the Faculty of Electrical Engineering and Informatics of the Technical University of Košice. - Košice (Slovensko) : Technická univerzita v Košiciach s. 399-405 [CD-ROM]. - ISBN 978-80-553-3940-5

A Position Controller with Low Speed Area based on Switching of Non-linear Functions

[1]*Pavol SMOLEŇ (4th year)*
Supervisor: [2]František ĎUROVSKÝ

[1]BWG, k.s. (member of Redex Group), Prešov, Slovak Republic
[2]Dept. of Electrical Engineering and Mechatronics, FEI TU of Košice, Slovak Republic

[1]pali.smolen@gmail.com, [2]frantisek.durovsky@tuke.sk

***Abstract* – The paper presents a position controller based on the switching between algebraic non-linear functions with low speed area. It is suitable for positioning of steel or aluminum slabs on the exit section of furnaces, feeding units, transport systems - coil cars. The control system with the implemented controller is robust, ideal for drive systems with variable moment of inertia of the load.**

The paper describes the principle of the controller, simulation results and practical implementation - feeding unit of zinc ingots at continuous galvanizing strip line. Designed positioning controller is used on many applications in industry and worked stable.

***Keywords* – low speed area, non-linear position controller, robust controller, transport systems**

I. Introduction

Auxiliary devices play an important role in industry, especially position orientated actuators. Typical example are transports system, which move products from one part of production line to another (transport of coils at entry and exit part of continuous strip processing line). They are characterized by different weight (different type of material, width), which makes the system variable from point of view moment of inertia.

There are several solutions for this problem, which are based on the application of different control methods, like the shaping of the open-loop transfer function in the Nyquist diagram [1] or a nominal characteristic trajectory following (NCTF) control [2]. These methods are very sophisticated, complicated and could cost a lot of time during commissioning to reach satisfactory results.

Industrial control systems offer various types of controllers in the software library with self-tuning options and predefined parameters. However, in case of positioning systems with a variable parameter range (moment of inertia), this solution does not give satisfactory system dynamics and it is complicated to adjust parameters of controller for high variation of system parameters. Therefore, special solutions like our proposal are preferred in industrial applications.

II. Principle of Positioning Controller

As shown in Fig. 1, non-linear positioning controller is implemented in the cascade structure of the drive. Output Y from the controller is velocity setpoint for drive. Actual position can be taken from motor encoder (appropriate calculation must be done) or from external sensor as laser device. Using the relations of trapezoidal motion profile between acceleration (deceleration), velocity and position [3] and [4], low speed way area x_s is calculated (1) and used in one of the non-linear functions (2).

$$x_s = \frac{1}{2} a_{cc} \left(\frac{v_{sa}}{a_{cc}} \right)^2 + x_{sa}\,, \tag{1}$$

where a_{cc} is acceleration (deceleration), v_{sa} is velocity in low speed area and x_{sa} is defined as low speed area. Position controller is built by three functions Y_1, Y_2 and Y_3:

$$\begin{aligned} Y_1 &= \sqrt{a_{cc} f_k \left(x_{set} - x_{act} \right)} \\ Y_2 &= \left(\sqrt{x_a f_k K_p} \right) \left[\left(x_{set} - x_{act} \right) / x_a \right] \\ Y_3 &= \sqrt{a_{cc} f_k \left[\left(x_{set} - x_{act} \right) - x_s \right]} \end{aligned} \tag{2}$$

The key variable of the final profile is a control deviation (x_{set} - x_{act}). Depends of deviation, gain of controller is changing what lead to change the curves of functions Y_1, Y_2 and Y_3.

Acceleration a_{cc} affects functions Y_1 and Y_3, K_p is gain and affects function Y_2, x_{set} is desired position (position setpoint), x_{act} is actual position (from encoder or external sensor), f_k is factor influences all three functions, x_a is area close to

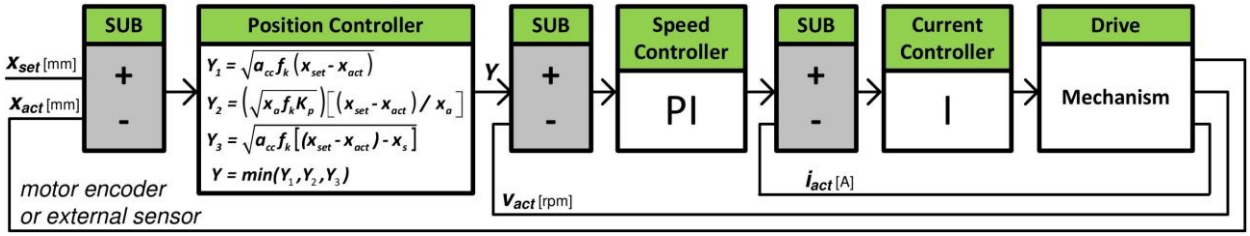


Fig. 1 Block diagram

desired position (switching point between Y_1 and Y_2).

Parameters a_{cc}, K_p, x_a, must be appropriate adjusted to achieve desired behavior of the controlled system. Parameter f_k is constant value 7,2 for the velocity unit *m/min*. As described above, a_{cc} affects function Y_1 which ensures a rough approach to desired position and K_p affects function Y_2 and ensures fine-positioning. Parameter x_a is constant value, defined as area close to desired position. Detailed analysis how these parameters influence the system is done by simulations in [5]. The output of position controller is minimum of functions Y_1, Y_2 and Y_3 represented as velocity speed setpoint.

$$Y = min(Y_1, Y_2, Y_3) \tag{3}$$

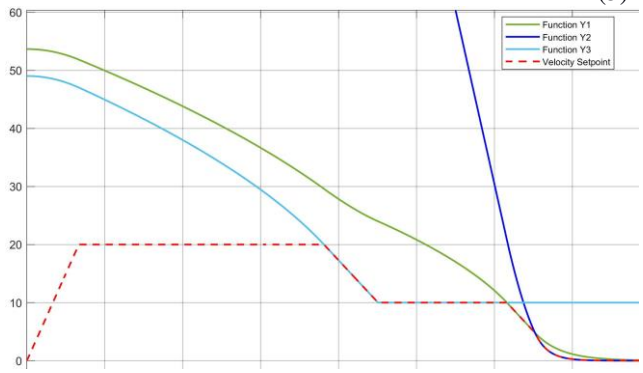


Fig.2. Non-linear functions with speed velocity profile

Ramp function and limiter are used to achieve trapezoidal velocity profile as shown on Fig. 2 (red dashed line). Before commissioning is possible to find approximate parameters by simulation in PLC program or using MATLAB Simulink and on commissioning with real mechanics do fine-tuning of controlled system.

III. Practical Implementation

Designed position controller has rich application potential. One of application where is successfully used is zinc ingot feeder unit at galvanic strip processing line.

The zinc ingot feeder unit is used to put zinc ingots into the zinc bath and essentially consists of a cart on which sinking device is located. The cart (traverse drive) is on rails and moves between two positions. The distance is measured by a laser measuring system. The cart is equipped with a lifting mechanism for tilting the sinking device. All drives are equipped by designed position controller. The cart moves empty or loaded by ingots varying weight (from 1,5 to 2 tons), so the load inertia is variable. Even during feeding ingot to zinc bath, ingot is melted, so weight of whole unit is changing.

Fig. 3 shows the data of cart movement (traverse drive) from loading position (3500 *mm*) to bath position (1000 *mm*). Feeder is loaded by ingot. Maximum velocity of drive is 15 *m/min*; a_{cc} = 0,1; K_p = 0,1; v_{sa} = 4 *m/min*; x_{sa} = 100 *mm* and x_a = 10 *mm*.

During deceleration velocity setpoint follows function Y_3 until the deviation position (x_{set} - x_{act}) is less than 100 *mm* (x_{sa}). After that is velocity setpoint limited to 4 *m/min* (v_{sa}). In the area closer to final position, velocity setpoint follows function Y_1 and then (switching point x_a) follows Y_2 until the desired position is reached. The center of gravity of feeding unit is high, therefore actual torque is rippling at the low speed and influences the positioning.

IV. Conclusion

In this paper, the positioning controller based on switching of non-linear functions with low speed area is presented. Adjusting of parameters can be done by simulation and fine-tuning during commissioning.

Positioning controller is robust and suitable to use on simple applications, where precise positioning is required and its performance does not depend on the variation of moment of inertia. It is ideal solution for transport systems like coil cars or feeding units. The controller is successfully applied on many applications in industry.

This work was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic under the project VEGA 1/0493/19.

References

[1] A. Karimi, M. Kunze, R. Longchamp, "Robust controller design by linear programming with application to a double-axis positioning system", Control Engineering Practice 15 (2007) 197–208

[2] Shin-Homg, Chong Kaiji Sato, "Practical and Robust Control for Precision Positioning Systems", Proceedings of the 2011 IEEE International Conference on Mechatronics, 2011, Istanbul, Turkey

[3] Dr. H. Gürocak, "Industrial Motion Control: Motor Selection, Drives, Controller Tuning, Applications", Washington State University, Vancouver, USA, 2016, ISBN: 9781118350812

[4] Khan Academy, "Kinematic formulas and projectile motion", Accessed February 02.2021, https://www.khanacademy.org

[5] D. Magura, V. Fedák, K. Kyslan, "A Simple Position Controller with Swiching of Non-linear Functions for Positioning of the Web End on Uncoiler", Elektrotechnické listy, ISSN 2453-8981, 2016

Fig. 3 Velocity, Position and Torque profiles during positioning of Ingot Feeder Unit (traverse drive)

Modelling photovoltaic system power output based on historical Meteorological data

[1]*Dávid MARTINKO (5th year)*
Supervisor: [2]Michal Kolcun

[1,2]Dept. of Electric Power Engineering, FEI TU of Košice, Slovak Republic

[1]david.martinko@student.tuke.sk, [2]michal.kolcun@tuke.sk

***Abstract*— In this study, we tested several models to evaluate annual photovoltaic (PV) system power output compared to a real existing PV generator in Hainburg, Austria. Calculations are based on a time series historical meteorological dataset from Solcast. The PV generator power output can be defined as the relationship between power output based on standard test conditions and actual weather conditions.**

***Keywords*— Photovoltaic systems, Historical meteorological data, Power output profile modelling, Losses, Solcast**

I. INTRODUCTION

One of the essential topics of our research calculations of photovoltaic systems and usage of electric vehicles on the load profile of residential LV grid is modelling power generation of photovoltaic systems, respectively photovoltaic panels or cells. Photovoltaic cell models are usually based on the equivalent electrical circuit as either four or five-parameter models consists of current and resistance. This technique can be categorized as a direct forecasting method. Since we do not have data for these variables, it is necessary to use indirect forecasting methods based on solar irradiance and weather data.

II. METHODOLOGY

The photovoltaic systems power output can be described as a relationship between weather conditions and rated power output based on standard test conditions (STC) [1]:

$$P_{PV(t)} = \eta\, P_{PV\,ref}\, \frac{G}{G_{ref}} \left(1 + \gamma \left(T_C - T_{ref}\right)\right)$$

where $P_{PV(t)}$ is power output of photovoltaic system in time interval, η is photovoltaics system efficiency, respectively the summary of different losses, G is the solar irradiance (W/m^2) received on module plane, γ is temperature correction for maximum power and T_c is panel temperature. Subscript "ref" refers to panel standard testing conditions (STC) as G=1000 W/m^2, T= 25° C. P_{PVref} is maximum panel power output at reference conditions.

It is unrealistic to think that all the incident radiation on the module will be converted into energy produced. In real PV systems, we must also include losses such as temperature losses, mismatch losses, cable losses, shadows, snow, dirt and dust in the calculations.

Table I shows range of typical values of the different losses in PV system, expressed as percentage of annual power [2].

TABLE I
TYPICAL RANGE OF ENERGY LOSSES IN A GENERATOR OF PV SYSTEM

Type of losses	Typical percentages of annual energy	
	Minimum (%)	Maximum (%)
Thermal	5	15
Angular and spectral	0.5	7
Tolerance and degradation	2	5
Shading	0	5
Dirt and Dust	0.5	4.5
Mismatch	2	4
Low irradiance	0.5	3
Ohmic cable	0.5	1.5

Thermal Losses

Part of solar radiation is transformed into electric energy by solar cells. Another part is converted into heat energy and since solar cells are semi-conductor, they are sensitive to temperature. They are multiple models to predict cell temperature:

A. Standard approach

Cell temperature T_c is a function of irradiation and ambient temperature and does not include cooling effect from wind. The NOCT Standard formula is represented:

$$T_C = T_a + \frac{I}{I_{NOCT}}\left(T_{NOCT} - T_{a,NOCT}\right)$$

T_a is ambient temperature, I is in-plate irradiance on cell. Subscript "NOCT" refers to nominal operating cell temperature on irradiance I_{NOCT}=800 W/m^2, ambient temperature T_{aNOCT}= 20° C and wind speed 1 m/s. Temperature TNOCT is determined on cell technology, with typical value in the range from 41 to 45° C.

B. Koehl

The Koehl simple model is a function of ambient temperature T_a, irradiation and wind speed near cell v_w :

$$T_C = T_a + \frac{I}{U_0 + U_1 v_w}$$

The constants U_0 and U_1 are specified for exact cell technology by Koehl, e.g. Monocrystalline silicon (m-Si) has U_0=30.02 and U_1=6.28.

Meteorological data usually contain only wind speed measured 10 meters above the ground. For the transformation to wind speed close to the module we used [3]:

$$v_w = 0.68\, v_f - 0.5$$

where v_f is wind speed in 10 meters above the ground.

C. Mattei

The Mattei model [4] consider moreover solar cell properties, such as temperature coefficient for maximum power temperature β, cell efficiency η, transmittance of the cover system τ and absorption coefficient of the cells α.

$$T_C = \frac{U_{PV} T_a + I\,[\tau.\alpha - \eta_{STC}(1 - \beta_{STC} T_{STC})]}{U_{PV} + \beta_{STC}\,\eta_{STC}\,I}$$

The relation $\tau.\alpha$ is given as a predefined value of 0.81. The heat exchange coefficient for the total surface of the module U_{PV} is expressed with two different functions:

$$U_{PV} = 26.6 + 2.3\,v_w \text{ (called Mattei 1)}$$
$$U_{PV} = 24.1 + 2.9\,v_w \text{ (called Mattei 2)}$$

D. Skoplaki

The Skoplaki model [5] is an advanced model with three variants, one considers in addition the wind direction.

$$T_C = T_a + \frac{I}{I_{NOCT}}\left(T_{NOCT} - T_{a,NOCT}\right)\frac{h_{wNOCT}}{h_{w(v)}}\left[1 - \frac{\eta_{STC}}{\tau.\alpha}\left(1 - \beta_{STC} T_{STC}\right)\right]$$

The wind convection coefficient h_w is linear function of wind speed in wind velocity close to PV module v_w or 10 meters above ground v_f:

$$h_w = 8.91 + 2\,v_f \text{ (called Skoplaki 1)}$$
$$h_w = 5.7 + 2.8\,v_w \text{ (called Skoplaki 2)}$$

Skoplaki 3 parametrization [6] for hw is based on wind direction, perpendicular (±45°) and parallel (±45°).

$$h_w = 8.3 + 2.2\,v_w \text{ (perpendicular } \pm 45°)$$
$$h_w = 6.5 + 3.3\,v_w \text{ (parallel } \pm 45°)$$

E. Kurtz

The Kurtz formula [7] calculate with ambient temperature T_a, solar irradiation I and wind speed close to module v_w, but does not distinguish solar cell technology:

$$T_C = T_a + I.e^{-3.473 - 0.0594.v_w}$$

Cable Losses

The cable or ohmic losses normally for an installation should be less than 2% and this proportion shouldn't rise over time. According to simulation studies [8], it has been shown that cable losses are 1.7%, 0.6% and 0.2% for the cables of 1.5, 4 and 10 mm^2 cross-sections.

Mismatch Losses

The differences in PV modules, partial shading and arrangement of the connection of PV arrays are commonly referred to as mismatch losses. The largest losses are in series with a single string.

Low irradiance Losses

The efficiency of the inverter decreases rapidly when the radiation level or PV panels power output decreases to zero. It has a logarithmic dependence, e.g. from 85% efficiency at 5% PV power to 98% efficiency at 100% PV power [9].

III. Conclusion

In Table II are annual power output calculated under different approaches for evaluate PV module [10] temperature and real data measured by inverter. Calculations are based on 30 minutes time-series historical meteorological dataset from Solcast [11]. Dataset includes air temperature, wind speed and direction, humidity, cloud opacity and several types of solar irradiance calculated for specific GPS coordinated, tilt and azimuth according to the actual parameters of our existing PV system.

TABLE II
ANNUAL POWER OUTPUT CALCULATED UNDER DIFFERENT APPROACHES

Calculation model	P (MWh)
Real data measured by inventer	4.796
Standard approach	4.731
Koehl	4.765
Sklopaki 1 / 2 / 3	4.807 / 4.763 / 4.78
Mattei 1/ 2	4.784 / 4.779
Kurtz	4.745

Losses that could not be inferred from the meteorological dataset such as thermal, low irradiance and ohmic losses, average values from Table I were used.

In Fig. 1 are real and calculated PV output profiles in different weather conditions with detail view.

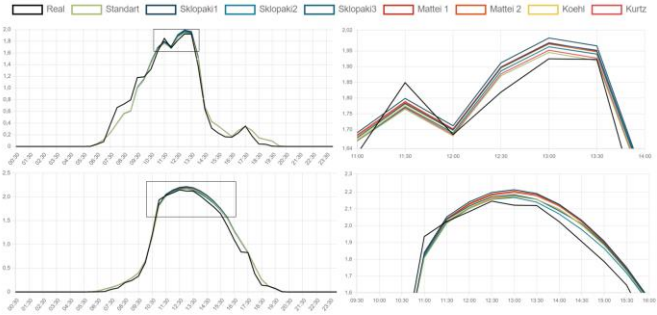


Fig. 1. Examples real and calculated PV output profiles (detail view on right)

IV. Next Steps

The next step is the usage of these models to calculate the impact of PV on the load profile of residential low voltage power grid at different penetrations of PV systems based on the parameters of real systems.

References

[1] J.I. Rosell, M. Ibanez, "Modelling power output in photovoltaic modules for outdoor operating conditions", Energy Conversion and Management 47, 2006, p. 2424–2430

[2] C. Rus-Casas, J. D. Aguilar, P. Rodrigo, F. Almonacid, P. J. Pérez-Higueras, "Classification of methods for annual energy harvesting calculations of photovoltaic generators", 2013, Energy Conversion and Management

[3] D. L. Loveday, A. H. Taki, "Convective heat transfer coefficients at a plane surface on a full-scale building facade." Int J Heat Mass Trans, 1996, p. 1729 – 1742.

[4] M. Mattei, G. Notton, G. Cristofari, M. Muselli, P. Poggi, "Calculation of the polycrystalline PV module temperature using a simple method of energy balance.", Renew Energ 2006, p. 553-567.

[5] E. Skoplaki, A.G. Boudouvis, J. Palyvos. "A simple correlation for the operating temperature of photovoltaic modules ofarbitrary mounting.", Sol Energ Mat Sol C 2008, p. 1393-1402.

[6] M. Koehl, M. Heck, S. Wiesmeier, J. Wirth, "Modeling of the nominal operating cell temperature based on outdoor weathering." Sol Energ Mat Sol C, 2011, p. 1638-1646.

[7] S. Kurtz, K. Whitfield, D. Miller, J. Joyce, J. Wohlgemuth, M. Kempe, et al. "Evaluation of high-temperature exposure of rackmounted photovoltaic modules." IEEE Photovoltaic Specialists Conference, 2009

[8] S. Ekici, M. A. Kopru, "Investigation of PV System Cable Losses", International journal of renewable energy research, 2017, Vol.7, No.2

[9] "Fronius Symo 5-0-3M", Fronius [Online], Available at https://www.fronius.com/en/solar-energy/installers-partners/technical-data/all-products/inverters/fronius-symo/fronius-symo-5-0-3-m

[10] "Rec TwinPeak 2 SERIES Datasheet", Rec [Online], Available at https://www.renugen.co.uk/content/Special_Offers/pdf/rec_twinpeak_2_renugen.co.uk.pdf

[11] "Solcast API Toolkit", Solcast [Online], Available at https://solcast.com/solar-data-api/api-toolkit/

Distribution of multiple deep convolution neural networks for support system of BLUE protocol ultrasound examination

[1]*Maroš HLIBOKÝ (1st year),*
Supervisor: [2]*Marek BUNDZEL*

[1,2]Dept. of Cybernetic and Artificial Intelligence, FEI TU of Košice, Slovak Republic

[1]maros.hliboky@tuke.sk, [2]marek.bundzel@tuke.sk

***Abstract*—This paper describes a proposed solution to support the diagnosis of examination of the lungs and their surroundings using an ultrasonograph. The following text describes the problem of the formation of convolutional deep neural networks for the classification and segmentation of ultraonographic examination records for the localization of pathological phenomena. In this paper, we describe modeling lung sliding detection, A and B line detection, and lung consolidation. In the second part of the article, we describe the process of deploying models for inference in the pipeline, which will process and evaluate newly created ultrasonographic records.**

***Keywords*—ultrasound, BLUE protocol, deep convolution neural network, ML Ops**

I. Introduction

Today, digitization and automation are becoming an exception in every area of society and medical medicine is no exception. In this area, more and more emphasis is placed on the accuracy of the determination of digests in the shortest possible time. Thanks to these conditions, doctors are able to establish a correct diagnosis for the patient in sufficient time to prevent permanent consequences in some cases and death. The medical sector has long sought support in artificial intelligence. However, this does not mean that artificial intelligence will replace doctors. These are supportive and recommended systems that should determine the direction of the diagnosis to the doctor and check it and point out things you did not notice.

II. Chest ultrasound examination

Ultrasound was used first in 1983, on occasion in François Fraisse's Intensive Care Unit (ICU) in 1985–1989, using machine devoted to cardiac assessment, in actual fact suitable for whole body and lung assessment [1]. At this time, although an old idea, ultrasound was not routine in the ICUs and had neglected lungs as vital organ [2]. For a long time, doctors did not believe that the ultrasound of the lung would bring helpful results in the diagnosis of the disease. Critically ill patients frequently need thoracic imaging due to the constant evolution of their clinical conditions. Today we know that ultrasound achieves higher examination accuracy with many other benefits [3].

CT scans are still the gold standard imaging tool for thoracic evaluation, but moving patients outside of the ICU is complicated and possibly dangerous. CT scans of the chest expose patients to high levels of radiation. For many diagnostic applications in the Intensive Care Unit, bedside chest X-ray (CXR) is still regarded the gold standard at ICU. However, this imaging technique has many methodological flaws and is frequently inaccurate [4]. It is necessary to consider that these types of devices require harmful equipment. Multiple radiologic imaging exams are linked to a higher risk of radiation-induced cancer over time [5]. For many typical uses in the ICU, bedside ultraonsography has become indispensable. Lung ultrasound (LUS) in particular has been shown to be superior to CXR as a screening tool for the diagnosis of certain lung conditions in critically ill patients [6]. As a consequence, in some case scenarios, LUS may be considered a viable alternative to CXR. Also, the use of bedside LUS could lead to reduced medical costs, as ultrasound scanners are relatively low-cost regarding maintenance and high durability compared to other imaging modalities [7].

A. Principle of ultrasound

Ultrasound is a sound with such a high frequency that one cannot hear it. It propagates through the environment like a wave of alternating thickening and dilution of molecules. The wave source is a piezoelectric crystal in an ultrasonic probe.

Different tissues have different echogenicity - the ability to reflect ultrasound waves (Ultrasound in different environments and tissues). The ultrasound strikes the tissue interfaces as they pass through the body, which are places where one tissue is adjacent to another, such as the kidney cortex and the surrounding fat.

At the tissue interface, part of the wave bounces off and the rest passes on to the next tissue interface so that part bounces off again and part passes through, etc. The different the echogenicities of the adjacent tissues, the greater the reflection of the wave. Due to this principle of operation, ultrasound examination had no support in practice because ultrasound energy was rapidly dissipated through the air, ultrasound imaging is not useful for evaluating the pulmonary parenchyma [8]. The presence of air creates a significant acoustic discrepancy with the surrounding tissues, resulting in complete reflection of the ultrasound beam and preventing direct imaging of the pulmonary parenchyma [9].

B. Ultrasound lung signs

The pleura, visible as a hyperechoic horizontal line in a normally aerated lung, is the only detectable structure. It is debatable whether this line is an artifact caused by a reflection phenomenon at the interface between alveolar air and the soft tissues of the thoracic wall, or if it is an image of the real pleura. The pleural line moves synchronously with respiration [10]

There are also some hyperechoic, horizontal lines that emerge at regular intervals from the pleural line: the A-lines (Figure 1). When combined with lung sliding, these reverberation artifacts indicate normal or excessive air content in the alveolar spaces.

Fig. 1. View of the A line on the USG examination record.

When the air content of the lung decreases and lung density increases due to the presence of exudate, transudate, collagen, blood, and other substances in the lung, the acoustic mismatch between the lung and the surrounding tissues decreases, and the ultrasound beam can be partially reflected at deeper zones and repeatedly. This phenomenon produces B-lines, which are vertical reverberation artefacts (Figure 2). B-lines belong to the family of the comet-tail artifacts, well known in the setting of abdominal ultrasound [11].

Multiple B-lines are considered a sonographic sign of lung interstitial syndrome, and their number increases as air content decreases and lung density increases [12]. When the air content is reduced further, as in lung consolidations, the acoustic window on the lung becomes completely open, and the lung can be directly visualized as a solid parenchyma, similar to the liver or spleen. Lung consolidation can occur as a result of an infectious process, an infarction caused by pulmonary embolism, cancer localization and metastasis, compression or obstructive atelectasis, or a contusion in thoracic trauma. Additional sonographic signs, such as the quality of the deep margins [13], the presence of air or fluid bronchogram [14], or the vascular pattern within the consolidation, may aid in determining the aetiology of the consolidation.

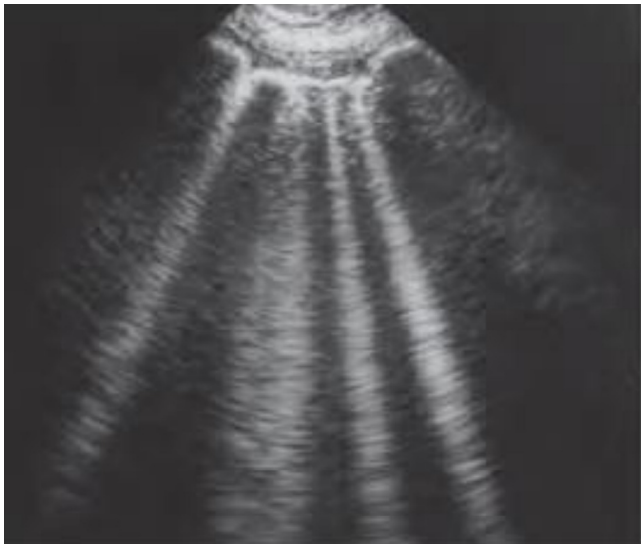


Fig. 2. View of the B line on the USG examination record.

C. BLUE protocol

All these symptoms were written in the so-called BLUE protocol, which represents as a decision tree (Figure 3) in which there are partial pathological findings in individual nodes, while the final nodes speak of the diagnoses themselves. The BLUE-protocol is a fast protocol (<3 minutes), which allows diagnosis of acute respiratory failure. It includes a venous analysis done in appropriate cases. Pulmonary edema, pulmonary embolism, pneumonia, chronic obstructive pulmonary disease, asthma, and pneumothorax yield specific profiles [15]. Pulmonary edema, e.g., yields anterior lung rockets associated with lung sliding, making the “B-profile". (see at the picture)

III. Deep learning methods

In the individual nodes of the BLUE protocol, the expert in our case must find the examining painter and describe the pathological symptom on the video from the ultrasonographic device. In both cases, the correct diagnosis must be made immediately, and without the possibility of making a wrong judgment because the diagnosis is determined, a treatment is set that can cost the patient life.

We therefore decided to propose a solution that will help the doctor in determining the final diganosis. As mentioned, the BLUE protocol consists of different decision points in different steps. When designing the solution, we decided that we would not cover such a complex problem by implementing one algorithm, so we divided this problem into smaller subtasks which are individual prepossessing of USG videos, detection of individual pathological phenomena in the form of separate models. not less important tasks include the evaluation of the results itself and the representation of the final diagnosis with visualization..

A. Data prepair

The data that we provide from doctors or the data that we found for the development of models are represented by video recordings. Models of deep neural networks often work with input in the form of a single frame. Before we deal with the models themselves for individual phenomena, it is necessary to process the video into a series of images. We have multiplied

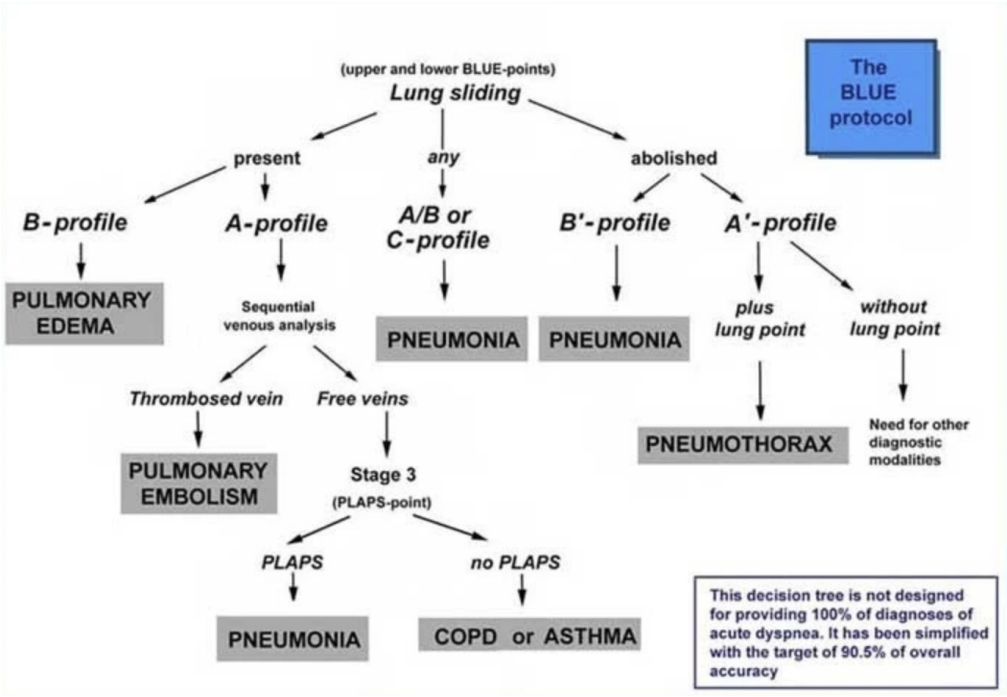


Fig. 3. The BLUE-protocol decision tree [15]

the preprocessing process for each model and it is included at the beginning of the training and inference pipeline. The images are generated and stored in a central repository from where I will be able to draw the models themselves with only the necessary pre-prepared frames. Metadata on the basic models will be attached to the images, and I will be able to filter the necessary images, for example with some models for optimal calculation of each 5 frames from the video. Metadata allows you to quickly filter data before querying it from the data repository, which saves the necessary time for inference.

B. Pathology sings

In the first level of the BLUE protocol, we get a partial stalk in which it is necessary to detect lung sliding. Here it is necessary in the first step to locate the area of the lungs, ribs and pleura. To do this, we use a series of semantic-segmentation models for binary segmentation that locate the required areas on the ultrasonographic data. Using localized parts of the thoracic region, we can best determine the regional incision to generate an M-mode image. Subsequent data enter the classification model ResnNet 18 which evaluates the presence or absence of lung sliding [16]. A special feature of lung sliding is that the symptom occurs and changes over time. The presence of the sing cannot be determined on the basis of a single image, but it is necessary to use a series of frames from the provided ultrasonographic record. In our next research we will focus on the modification of the solution by implementing deep neural networks on the principle of LSTM (Long short-term memory) [17].

In the next level of the BLUE protocol, it is necessary to determine whether there are A lines and B lines on the input ultrasonographic record. Unlike lung sliding, this task is pure classification. The A and B lines do not change during the patient's breathing, they do not disappear during inhalation and exhalation, nor do they change their shape. That is why it sets the space to use single frame classification here. The individual frames from the video examination will be processed by the model, while each frame will be assigned a separately assigned probability of the occurrence of a pathological phenomenon of the A or B line. By subsequently aggregating the results of the selected images from the examination video, we obtain the probability of the occurrence of a pathological phenomenon for the examination itself. In current experiments we work with models for binary classification which in practice means the development of two separate models for A and B lines.

Another pathological phenomenon that we will deal with is lung consolidation. There are several ways to approach the detection of this phenomenon. One of the flammable solutions will be the use of the deep neural network Inception V3, which uses inception blocks which brings several benefits [18]. Lung consolidation is a symptom that evolves and changes over time [19], so it is necessary to consider monitoring the princely patterns on the continuous section of the examination record. The next thought is to apply LSTM as a final classifier. As a symptom extractor, a flood-categorized classification model of the Inception V3 architecture would be used, which would omit the classification part of the network. In other words, entering the LSTM model would represent the extracted feature maps from the Inception V3 model.

C. Explainable AI

The solution we propose consists exclusively of models of deep neural networks, often called black boxes. Unlike the approach of decision storms in neural networks, for example, we cannot explicitly describe the rules based on the model that determined its decision [20]. Our solution in the initial stages of the solution should only be a support tool for doctors, but we still want to show the doctors symptoms and give at least a partial explanation of how our models converged to make a diagnosis of repositive detection of pathological sing.

In the case of the A and B lines, we provide the doctor's percentage probability of having a symptom for the entire

ultrasound record of the patient's examination. Our idea is to implement a mechanism for visualizing the area of interest in the input image, which means for the model the greatest information in its final classification decision. In this way we can prove to the doctor the highlighted areas of the input image that contributed the most to the decision. One possible implementation would be to use the Grad-CAM architecture [21]. Grad-CAM uses the gradient information flowing into the last convolution layer of the CNN to understand the importance of each neuron for a decision of interest [22]. We can use this approach in different levels of the BLUE protocol to clarify the decisions of individual models.

IV. Machine Learning Operations

We want to transfer our solution to production use, so we must meet the requirements of stability, availability and sustainability [23]. The more we try to consider these questions due to the application of applications in the medical environment.

A. Security of data

The focus on data security comes first in today's IT industry. In the case of medical data, high emphasis is placed on privacy data. A typical solution to allow access to medical data while maintaining security is to delete all personal data through anonymization. Unfortunately, this is not always a viable solution, as patient data are often unstructured, which in many cases requires manual action. In addition, it sometimes requires the removal or modification of properties that may provide relevant information for the machine learning model, e.g. age, locality, ethnicity. An alternative solution to data anonymisation is to use homomorphic encryption (HE): a special type of encryption that allows operations to be performed directly on encrypted data without decryption, and thus without access to real data. In this case, the machine learning approach consists of training on encrypted data and generating encrypted results, thus securing input and output data [24].

Another important step is the establishment of a central repository where doctors upload record ultrasonographic records for the development of our models. Due to GDPR, we should not have used publicly available cloud services as a data repository. In agreement with the University Hospital in Martin, we have used the Microsfot sharepoint interface as a central data warehouse, which is connected to the local data repository of the Technical University in Košice. All partially processed data, intermediate calculations as well as final results and trained models will be stored here under local administration.

B. Model distribution

This complex solution involves several researchers who often use different approaches to problem solving, such as choosing a framework for creating deep neural models or a library of initial data preprocessing. In many cases, it is essential to ensure the current type of operating system and the correct version of the code library. We will pack individual cleanup tasks, which represent separate models with their necessary prerequisites, into Docker Image. This approach allows us to separate the environment by virtualization, which allows us to configure the operating system and deliver the right types and versions of libraries exactly according to the requirements of different models on a single physical server without worrying about collisions or configuration consistency.

Some steps require running on CPU (Central Processing Unit) architectures such as the video slicer. In the case of deep neural network models, it is possible to infer models on the CPU, GPU (Graphics Processing Unit) or (Tensor Processing Unit). Accessing the code wrapper to a separate entity allows us to deploy the inference model in seconds on the CPU, GPU, or TPU architecture without having to change and reconfigure the guest operating system. Another advantage of the approach of virtualization containers is the possibility of replication of individual instances of the models in the event of increased demand for inference of the model for the same pathological symptom.

The whole solution consists of certain subtasks. The first necessary step is to paint the individual models. Subsequently, to ensure their inference in their correct order as some parts process the excerpts from the previous parts. All necessary features are offered by the KubeFlow solution, which in practice can be used in various scientific fields from universities and research institutions to medical applications. [25], [26]. The tool allows to integrate, for example, the process of parameter tuning or the process of retelling models in a production environment with an evaluated A / B test [27].

V. Conclusion

We can divide our final solution into 2 main areas. In the first part we will focus on the development of modules for the detection of pathological phenomena. We will evaluate the probabilities of the findings of the BLUE protocol, which will give us a final diagnosis. The second part is related to the deployment of artificial intelligence in the medical environment. It will be necessary to design and create pipilens for inference of deep neural network models. Last but not least, the design was for the doctors themselves, where I will be able to record videos from ultrasonographic examinations, which will also allow you to view the results and the proposed diagnosis.

References

[1] F. Jardin, J.-C. Farcot, L. Boisante, N. Curien, A. Margairaz, and J.-P. Bourdarias, "Influence of positive end-expiratory pressure on left ventricular performance," *New England Journal of Medicine*, vol. 304, no. 7, pp. 387–392, 1981.

[2] B. S. Slasky, D. Auerbach, and M. L. Skolnick, "Value of portable real-time ultrasound in the icu." *Critical Care Medicine*, vol. 11, no. 3, pp. 160–164, 1983.

[3] Y. Amatya, J. Rupp, F. M. Russell, J. Saunders, B. Bales, and D. R. House, "Diagnostic use of lung ultrasound compared to chest radiograph for suspected pneumonia in a resource-limited setting," *International journal of emergency medicine*, vol. 11, no. 1, pp. 1–5, 2018.

[4] D. Lichtenstein, I. Goldstein, E. Mourgeon, P. Cluzel, P. Grenier, and J.-J. Rouby, "Comparative diagnostic performances of auscultation, chest radiography, and lung ultrasonography in acute respiratory distress syndrome," *The Journal of the American Society of Anesthesiologists*, vol. 100, no. 1, pp. 9–15, 2004.

[5] L. Gargani and E. Picano, "The risk of cumulative radiation exposure in chest imaging and the advantage of bedside ultrasound," *Critical ultrasound journal*, vol. 7, no. 1, pp. 1–4, 2015.

[6] C. Arbelot, M. Blaivas, B. Bouhemad, R. Copetti, A. Dean, S. Dulchavsky, M. Elbarbary, L. Gargani, R. Hoppmann, A. W. Kirkpatrick *et al.*, "Ultrasound performs better than radiographs," *Thorax*, vol. 66, no. 9, pp. 828–829, 2011.

[7] L. Zieleskiewicz, A. Cornesse, E. Hammad, M. Haddam, C. Brun, C. Vigne, B. Meyssignac, A. Remacle, K. Chaumoitre, F. Antonini *et al.*, "Implementation of lung ultrasound in polyvalent intensive care unit: impact on irradiation and medical cost," *Anaesthesia Critical Care & Pain Medicine*, vol. 34, no. 1, pp. 41–44, 2015.

[8] D. Kasper, A. Fauci, S. Hauser, D. Longo, J. Jameson, and J. Loscalzo, *Harrison's principles of internal medicine, 19e*. Mcgraw-hill New York, NY, USA:, 2015, vol. 1, no. 2.
[9] G. Volpicelli, "Lung sonography," *Journal of Ultrasound in Medicine*, vol. 32, no. 1, pp. 165–171, 2013.
[10] G. Volpicelli, M. Elbarbary, M. Blaivas, D. A. Lichtenstein, G. Mathis, A. W. Kirkpatrick, L. Melniker, L. Gargani, V. E. Noble, G. Via *et al.*, "International evidence-based recommendations for point-of-care lung ultrasound," *Intensive care medicine*, vol. 38, no. 4, pp. 577–591, 2012.
[11] M. Ziskin, D. Thickman, N. Goldenberg, M. Lapayowker, and J. Becker, "The comet tail artifact." *Journal of Ultrasound in Medicine*, vol. 1, no. 1, pp. 1–7, 1982.
[12] G. Baldi, L. Gargani, A. Abramo, L. D'Errico, D. Caramella, E. Picano, F. Giunta, and F. Forfori, "Lung water assessment by lung ultrasonography in intensive care: a pilot study," *Intensive care medicine*, vol. 39, no. 1, pp. 74–84, 2013.
[13] A. Reissig and C. Kroegel, "Transthoracic ultrasound of lung and pleura in the diagnosis of pulmonary embolism: a novel non-invasive bedside approach," *Respiration*, vol. 70, no. 5, pp. 441–452, 2003.
[14] D. Lichtenstein, G. Mezière, and J. Seitz, "The dynamic air bronchogram: a lung ultrasound sign of alveolar consolidation ruling out atelectasis," *Chest*, vol. 135, no. 6, pp. 1421–1425, 2009.
[15] D. A. Lichtenstein, "Lung ultrasound in the critically ill," *Annals of intensive care*, vol. 4, no. 1, pp. 1–12, 2014.
[16] M. Jaščur, M. Bundzel, M. Malík, A. Dzian, N. Ferenčík, and F. Babič, "Detecting the absence of lung sliding in lung ultrasounds using deep learning," *Applied Sciences*, vol. 11, no. 15, p. 6976, 2021.
[17] S. Kulhare, X. Zheng, C. Mehanian, C. Gregory, M. Zhu, K. Gregory, H. Xie, J. M. Jones, and B. K. Wilson, "Ultrasound-based detection of lung abnormalities using single shot detection convolutional neural networks," in *POCUS/BIVPCS/CuRIOUS/CPM@MICCAI*, 2018.
[18] S. Kulhare, X. Zheng, C. Mehanian, C. Gregory, M. Zhu, K. Gregory, H. Xie, J. McAndrew Jones, and B. Wilson, "Ultrasound-based detection of lung abnormalities using single shot detection convolutional neural networks," in *Simulation, image processing, and ultrasound systems for assisted diagnosis and navigation*. Springer, 2018, pp. 65–73.
[19] J.-x. Yang, M. Zhang, Z.-h. Liu, L. Ba, J.-x. Gan, and S.-w. Xu, "Detection of lung atelectasis/consolidation by ultrasound in multiple trauma patients with mechanical ventilation," *Critical Ultrasound Journal*, vol. 1, no. 1, pp. 13–16, 2009.
[20] C. Rudin and J. Radin, "Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition," *Harvard Data Science Review*, vol. 1, no. 2, 11 2019, https://hdsr.mitpress.mit.edu/pub/f9kuryi8. [Online]. Available: https://hdsr.mitpress.mit.edu/pub/f9kuryi8
[21] R. Arntfield, D. Wu, J. Tschirhart, B. VanBerlo, A. Ford, J. Ho, J. McCauley, B. Wu, J. Deglint, R. Chaudhary *et al.*, "Automation of lung ultrasound interpretation via deep learning for the classification of normal versus abnormal lung parenchyma: A multicenter study," *Diagnostics*, vol. 11, no. 11, p. 2049, 2021.
[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
[23] J. Bogatinovski, S. Nedelkoski, A. Acker, F. Schmidt, T. Wittkopp, S. Becker, J. Cardoso, and O. Kao, "Artificial intelligence for it operations (aiops) workshop white paper," *arXiv preprint arXiv:2101.06054*, 2021.
[24] A. Vizitiu, C. I. Niţă, A. Puiu, C. Suciu, and L. M. Itu, "Towards privacy-preserving deep learning based medical imaging applications," in *2019 IEEE international symposium on medical measurements and applications (MeMeA)*. IEEE, 2019, pp. 1–6.
[25] D. Y. Yuan and T. Wildish, "Bioinformatics application with kubeflow for batch processing in clouds," in *International Conference on High Performance Computing*. Springer, 2020, pp. 355–367.
[26] D. Golubovic and R. Rocha, "Training and serving ml workloads with kubeflow at cern," in *EPJ Web of Conferences*, vol. 251. EDP Sciences, 2021, p. 02067.
[27] C. Xu, G. Lv, J. Du, L. Chen, Y. Huang, and W. Zhou, "Kubeflow-based automatic data processing service for data center of state grid scenario," in *2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*. IEEE, 2021, pp. 924–930.

Advancements in the design of a framework for evaluation of cosmic ray trajectories

[1]*Daniel GECÁŠEK (3rd year),*
Supervisor: [2]*Ján GENČI, Consultant:* [3] *Pavol BOBÍK*

[1,2]Dept. of Computers and Informatics, FEI TU of Košice, Slovak Republic
[3]Institute of Experimental Physics, Slovak academy of sciences, Slovak Republic

[1]daniel.gecasek@tuke.sk, [2]jan.genci@tuke.sk, [3]bobik@saske.sk

***Abstract*—This contribution describes the progress done on the COR (CutOff Rigidity) system made since the last. The main three points are the presentation of the COR system to the physics community, quantitative comparison of COR and existing systems, and examination of existing standard data formats. Based on results, the COR system produces results within acceptable accuracy while still being comparatively easy to use in relation to existing command-line simulation software. Potential benefits of the usage of standard data formats are also discussed.**

***Keywords*—data analysis, scientific computing, data formats, disk operations**

I. Introduction

In the world of today's physics research, numerical simulations of physical processes via computer software are often employed. This puts a strain on physicists where they not only need to be proficient in their domain but also be proficient in the software engineering domain. The aim of the COR system and the main contribution of my work is to help minimize this factor - simplify the simulation of cosmic ray trajectories in the Earth's magnetosphere, provide a basic analysis of results and tools for advanced analysis.

II. State of the art

In previous publication [1], a state of the art analysis was done. Since then, a new system, Cutoff2050 rigidity visualizer[1], described in manuscript [2], in domain of cosmic ray trajectory evaluation was introduced. To evaluate trajectories the system offers the following magnetic field models: dipole model or a combination of IGRF and models of external fields T89 [3], T96 [4] and TS02 [5], [6]. Temporarily, it is not possible to log in as of February 2022.

Based on the cited manuscript and our previous experiences it offers very thorough visualizations for the evaluation of a single trajectory of cosmic ray particles. The most innovative aspect is a graph of the length of time that the particle took traveling the evaluated trajectory. Based on this, we can easily determine whether the trajectory is trapped in the magnetosphere.

However, the Cutoff2050 rigidity visualizer does not offer facilities for mass simulation of many cosmic ray trajectories. Also, the source code of this simulation software is not publicly available.

[1]Located at https://tools.izmiran.ru/cutoff2050/

III. Solved tasks

Since last years article, the following goals were fully or partly fulfilled:

- Determination of suitability of existing standard data formats
- Simulation of historical cut-off rigidities

The aim that was not fulfilled is the examination of existing magnetic field models based on their performance and results. The aim that was not planned but was fulfilled is the comparison of our trajectory tracing code with other existing codes. This section contains a brief overview of fulfilled plans.

A. Presentation of COR system to the community

We presented the results of our work since the last SCYR conference manuscript at ICRC online conference [7]. Our contribution got invited to a discussion session #22 "Atmospheric effects of CR" where the need for such a system was expressed by the physics community. In the article, the capabilities of the system and a result directly from the system are presented.

The simulation described contains an evaluation of vertical cut-off rigidities on a grid covering the whole surface of the Earth with a resolution of 10° in latitude and 5° in longitude in the period from year 1 CE to the year 1901 CE. We use computed rigidities to evaluate cosmic ray intensities at the top of the atmosphere in two limit cases of solar activity taken from [8]. In the result part, we show that changes in cosmic ray intensities were significantly larger than changes in the Sun's activity. This can have a significant impact on the generation of carbon isotope ^{14}C and thus impact dating of archeological artifacts using methods of radiocarbon dating first described in [9].

B. Quantitative comparison of COR characteristics and characteristics of existing systems

In our paper [10], currently in a revision process, planned to be published in Advances in Space Research, we are presenting the COR system to a wider audience and comparing numerical results to other existing software.

In the article, we present the preexisting features but also an extension for simulation data analysis that uses Jupyter Notebooks. We also present the new simulation code that was rewritten from Fortran to C. This refactorization improved its

readability via using modern programming practices and thus should improve its maintainability.

The quantitative comparison was done by the analysis of the results from the COR system and two other authors' models, specifically, data published in [11] and [12]. We compared relative and absolute differences of vertical cut-off rigidities. In table I, median and average of relative differences between published results and the fraction of compared area where relative differences are smaller than 5%. Area of Earth where the relative difference in effective cut-off rigidity between [11] and COR is smaller than 5% is 80.24%. In second comparison between [12] and COR it is 90.61%. This shows a general agreement between the models examined and confirms an acceptable accuracy of the COR code.

TABLE I
TABLE OF MEDIAN AND AVERAGE OF RELATIVE DIFFERENCES BETWEEN PUBLISHED RESULTS (SMART SHEA [11], GERONTIDOU [12], AND COR) AND THE FRACTION OF COMPARED AREA WHERE RELATIVE DIFFERENCES ARE SMALLER THAN 5%.

	Median	Average	$A_{rd} < 5\%$
Smart Shea / Gerontidou	1.40%	7.63%	90.61%
Gerontidou / COR	1.73%	7.87%	88.62%
Smart Shea / COR	2.76%	7.10%	80.24%

C. Examination of existing standard data formats

With our research team, we are also preparing an article that aims to show the performance gain of using specialized data formats for the storage of scientific data compared to ad-hoc ASCII solutions. The data formats we compared in the article are: CDF [13], NetCDF [14] and HDF [15] and a baseline of ad-hoc ASCII formats stored in directories.

Reading and writing speed is hardware dependent so the exact results may not apply for all devices but we expect some sort of improvement on all hardware caused by the minimization of the number of system calls, specifically for file opening and reading/writing. The measurements were done on a conventional SSD but also on a RAM disk[2], temporary storage using system memory instead of a disk [16], to measure and minimize the impact of disk operations. Depending on the type of operation, the RAM disk was predictably faster. The most significant reductions were 54.45% reduction in reading times and 91.17% reduction for writing times.

The preliminary results from SSD measurements show that dedicated data formats may be up to 5 times faster for speed of writing, up to 2.6 times faster for reading speed in the case of a few big files. In the case of a lot of (over 120 000) small (few kilo-bytes) files, dedicated data formats have 336 times faster writing speed and 1888 times faster reading speed.

Size occupied on disk should be constant as the file size is the same on all supported devices, which is guaranteed by specific standards of used data formats. The preliminary results show that binary formats without compression use 69% of the capacity compared to using ASCII solutions, 34% with lossless compression, and 8% of capacity with lossy compression.

IV. CONCLUSION AND FUTURE RESEARCH

In this paper, we described the progress made since last year. We showed that the created system for automation, post processing, and interactive data analysis of cosmic ray trajectories has the potential to be useful for scientists working in its domain. An example of this is the planned detailed examination of the impact of historical cosmic ray intensities in the last two millennia on the accuracy of radiocarbon dating methods.

In the article concerned with quantitative comparison, we showed that the results of the simulations in our system are comparable to the results of other models. We also highlighted the advantage of the automated approach where a request to run over 60000 simulations can be achieved by filing input fields of one web form.

Based on our examination of standard data formats, it would be beneficial for the COR system to adopt them. Adoption of one of these formats should consist of reimplementation of program segments implementing file operations in all the modules forming the processing pipeline.

[2]We used tmpfs to create a RAM disk.

V. ACKNOWLEDGMENT

This work was partially supported by the TUKE Space Forum ESA PECS project.

REFERENCES

[1] D. Gecášek, "Design of web platform for radiation models automation: Technology review," in *SCYR 2020: Nonconference Proceedings of Young Researchers*. Faculty of Electrical Engineering and Informatics Technical University of Košice, April 2020, pp. 51–54.
[2] S. Belov, E. Zobnin, V. Yanke *et al.*, "Cutoff rigidity and particle trajectories online calculator," in *NMDB@ Home 2020: Proceedings of the 1st virtual symposium on cosmic ray studies with neutron detectors*, 2021, pp. 197–203.
[3] N. A. Tsyganenko, "A magnetospheric magnetic field model with a warped tail current sheet," *Planetary and Space Science*, vol. 37, no. 1, pp. 5–20, 1989.
[4] ——, "Effects of the solar wind conditions in the global magnetospheric configurations as deduced from data-based field models," in *International conference on substorms*, vol. 389, 1996, p. 181.
[5] N. Tsyganenko, "A model of the near magnetosphere with a dawn-dusk asymmetry 1. mathematical structure," *Journal of Geophysical Research: Space Physics*, vol. 107, no. A8, pp. SMP–12, 2002.
[6] ——, "A model of the near magnetosphere with a dawn-dusk asymmetry 2. parameterization and fitting to observations," *Journal of Geophysical Research: Space Physics*, vol. 107, no. A8, pp. SMP–10, 2002.
[7] D. Gecášek, P. Bobík, and J. Genči, "Transparency of magnetosphere for cosmic rays in last two millennia," in *Proceedings of the 37th International Cosmic Ray Conference*, July 2021.
[8] I. G. Usoskin, A. Gil, G. A. Kovaltsov, A. L. Mishev, and V. V. Mikhailov, "Heliospheric modulation of cosmic rays during the neutron monitor era: Calibration using pamela data for 2006–2010," *Journal of Geophysical Research: Space Physics*, vol. 122, no. 4, pp. 3875–3887, 2017.
[9] W. F. Libby, "Atmospheric helium three and radiocarbon from cosmic radiation," *Physical Review*, vol. 69, no. 11-12, p. 671, 1946.
[10] D. Gecášek, P. Bobík, J. Genči, J. Villim, and M. Vaško, "Cor system: a tool to evaluate cosmic ray trajectories in the earth's magnetosphere," May 2022, in review process.
[11] D. Smart and M. Shea, "Vertical geomagnetic cutoff rigidities for epoch 2015," in *36th International Cosmic Ray Conference (ICRC2019)*, vol. 36, 2019, p. 1154.
[12] M. Gerontidou, N. Katzourakis, H. Mavromichalaki, V. Yanke, and E. Eroshenko, "World grid of cosmic ray vertical cut-off rigidity for the last decade," *Advances in Space Research*, vol. 67, no. 7, pp. 2231–2240, 2021.
[13] *CDF User's Guide*, 11 2020.
[14] R. Rew and G. Davis, "Netcdf: an interface for scientific data access," *IEEE computer graphics and applications*, vol. 10, no. 4, pp. 76–82, 1990.
[15] *HDF5 User's Guide*, 6 2019.
[16] T. Wickberg and C. Carothers, "The ramdisk storage accelerator: a method of accelerating i/o performance on hpc systems using ramdisks," in *Proceedings of the 2nd International Workshop on Runtime and Operating Systems for Supercomputers*, 2012, pp. 1–8.

Properties enhancement of dielectric fluids for power transformers

[1]*Miloš ŠÁRPATAKY (2nd year)*
Supervisor: [2]Juraj KURIMSKÝ

[1,2]Dept. of Electric Power Engineering, FEI TU of Košice, Slovak Republic

[1]milos.sarpataky@tuke.sk, [2]juraj.kurimsky@tuke.sk

***Abstract*— Current research in the field of insulating fluids focuses on biodegradable oils that are in line with sustainable development and improvement of the properties of these oils using nanoparticles. The main goal of actual research is to create more efficient and environmentally friendly oil for high voltage applications. The combination of biodegradable oil (synthetic ester or natural ester) and nanoparticles may fill these requirements and become one of the possible replacements for currently used insulating oils.**

***Keywords*—nanofluids, nanoparticles, biodegradable oils**

I. Introduction

In recent decades, there has been an increase in the number of activities and projects aimed at innovating current technologies and materials. This is caused by the demand for sustainable industrial development, which is also evident in the power engineering sector [1]. This important part of the industry is under great pressure to improve technology to ensure the supply of electricity, considering all strict regulations regarding environmental protection, quality, and reliability of electricity supply [2]. Much attention is paid to electricity generation and consumption, but many other parts of the sector, such as electricity transmission and distribution, are key to meeting the goals and needs of sustainable development [1]. The main goal in the field of insulating liquids is the application of biodegradable oils in as many numbers as possible in power transformers and research of new materials that will improve the properties and prolong the life of researched and currently used insulating oils. The most hopeful alternative for high voltage applications are synthetic esters (SE) and natural esters (NE) with biodegradability over 80 % (NE over 97 %) [2]. Mentioned research on improvement of insulating fluids are focused on nanofluids. Nanofluids are colloidal solutions formed by adding nanoparticles to a base fluid [2].

II. Nanofluids

Nanofluids are mainly created by two step method represented in Fig. 1. Very important quality indicator of nanofluid is stability which is evaluated according to the distribution of nanoparticles in the fluid and the elimination of clusters. This is achieved by ultrasonication, magnetic and mechanic stirring and the addition of dispersants [3], [4].

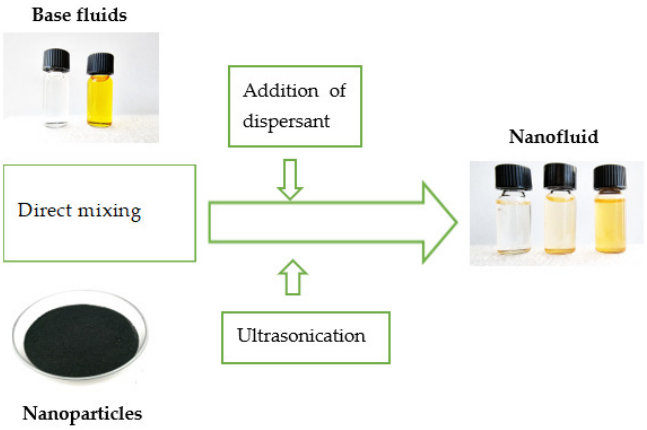


Fig. 1. Simple illustration of two-step method of preparing nanofluids.

Nanofluid research aims to confirm improvements in properties such as thermal conductivity, viscosity, safety, and dielectric properties by using different types of nanoparticles.
Review [5] describes the dielectric parameters of nanofluids divided into individual sections according to the nanoparticles used in different scientific papers published after year 2018. The most frequently used nanoparticles for research according to this review are Al_2O_3, TiO_2, Fe_3O_4 and SiO_2. As for the next research Fe_3O_4 and C_{60} nanoparticles will be used. Nanofluids made of these nanoparticles and their dielectric properties were described in mentioned review.
Fe_3O_4 nanoparticles are conductive type of nanoparticles mainly used for enhancement of dielectric properties [3]. As for the AC breakdown voltage (AC-BDV) enhancement moved from 9.41 % to 47.78 % that can be seen in TABLE I. Khaled et al. [6] compared AC-BDV of nanofluids made of SE with Al_2O_3, Fe_3O_4 and SiO_2 nanoparticles. Fe_3O_4 nanoparticles of diameter 50 nm enhanced the AC-BDV up to 47.78 % at concentration 0.4 g/l that is the highest enhancement among all samples. Average AC-BDV enhanced its value from 60.03 kV (pure SE sample) to 88.81 kV (concentration 0.4 g/l) that is outstanding improvement of dielectric strength. Other concentrations of Fe_3O_4 nanoparticles had worse results. Concentration 0.3 g/l of nanoparticles in SE enhanced AC-BDV by 17.83 % but samples with lower concentration (0.05 g/l and 0.2 g/l) exhibited decreased AC-BDV by 6.06 % and 0.05 %, respectively. These facts confirms that when researching nanofluids, an increased attention is necessary to investigate the optimal concentration of nanoparticles in concrete insulation oil.

TABLE I
Fe_3O_4 AC-BDV OVERVIEW TABLE [5]

Base fluid	Preparation of nanofluid	Size of NP (nm)	Optimal concentration	Highest enhancement	Reference
SE	Two-step; magnetic stirring, ultrasonication	50	0.4 g/l	47.78 %	[6]
NE	Two-step; mechanic stirring, ultrasonication	10-20	0.2 kg/m³	15.1 %	[12]
MO	Two-step; ultrasonication	10	0.2 g/l	9.41 %	[13]
NE	Two-step; magnetic stirring, ultrasonication	50-100	0.004 w/w%.	33.4 %	[14]
SE	Two-step; magnetic stirring, ultrasonication	50-100	0.0022 w/w%	30.7 %	[14]
NE	Two-step; magnetic stirring, ultrasonication	10-20	0.2 kg/m³	16 %	[15]
NE	Two-step; ultrasonication	-	0.012 w/w%	17 %	[16]

DC breakdown voltage (DC-BDV) enhancement moved around 10 % in [7] and [8]. DC-BDV were mainly enhanced at lower concentrations specifically in NE enhancement 10.56 % was achieved at nanoparticles concentration 0.2 g/l and this result surpass other examined combinations (Fe_3O_4, Al_2O_3 and SiO_2 with NE). Remaining concentrations (0.05 g/l, 0.3 g/l and 0.4 g/l) decreased value of DC-BDV from 0.62 % to 13.84 %. A similar experiment [8] with the different base fluid combined SE MIDEL 7131 with Fe_3O_4 nanoparticles with an average diameter of 50 nm with the highest enhancement 9.8 % at the lowest concentration 0.05 g/l. Higher concentrations decreased values of DC-BDV by 7.89 %, 1.05 % and 2.03 % at concentrations 0.2 g/l, 0.3 g/l and 0.4 g/l respectively.

Lightning breakdown voltage (LI-BDV) enhancement of SE and NE with Fe_3O_4 nanoparticles were compared in [9]. Enhancement of nanofluid with SE decreased with increasing concentration from 25.57 % (0.05 g/l) to 6.37 % (0.4 g/l). NE nanofluid did not achieve such results when enhancement moved from 7.51 % at concentration 0.2 g/l to -8.29 % (decrement) at concentration 0.4 g/l. It can be concluded that SE according to this research is more appropriate option for application in nanofluid than NE.

Fullerene (C_{60}) nanoparticles are insulating type of nanoparticles. In [10] fullerene with a diameter of approximately 0.7 nm mixed with NE before and after 164 h of accelerated ageing at temperature 150°C showed decrement around 10 % in AC-BDV before ageing but after accelerated ageing enhancement reached value 23 % in comparison to aged pure insulating oil. Dissipation factor of nanofluid was higher before but also after accelerated ageing that suggest worsened dielectric properties. As for thermophysical properties, the viscosity of the nanofluid remains almost unchanged even after ageing so the nanofluid is suitable as a cooling medium as well as a base fluid. In [11] there is comparison of AC-BDV of NE and MO nanofluids with fullerene. Enhancement of NE nanofluid moved from around 2 % to around 8 % and MO nanofluid at concentration 200 mg/l reached enhancement around 21 % that made according to this research MO more suitable for this mixture.

III. RESEARCH DIRECTION

Next step of research will be examination of SE and NE (MIDEL 7131 and MIDEL eN 1215) with magnetite (Fe_3O_4) and fullerene (C_{60}) nanoparticles. Research will be aimed at dielectric parameters before and after accelerated ageing. Options for dielectric properties examination are breakdown voltage (AC, DC, lightning), dissipation factor, partial discharges, and volume resistivity. Accelerated ageing will be performed at temperature 130°C for 30 days to simulate the aging of the transformer insulation for almost 27 years as described in [17].

ACKNOWLEDGMENT

This research was funded by the Ministry of Education, Youth and Sports within the project VEGA 2/0011/20 and 1/0154/21 and the Slovak Agency for Research and Development based on contracts no. APVV-15-0438, APVV-17-0372, and APVV-18-0160.

REFERENCES

[1] CHU, S. - MAJUMDAR, A. Opportunities and challenges for a sustainable energy future. In *Nature* . 2012. Vol. 488, no. 7411, s. 294–303. .
[2] RAFIQ, M. et al. The impacts of nanotechnology on the improvement of liquid insulation of transformers: Emerging trends and challenges. In *Journal of Molecular Liquids* . 2020. Vol. 302, s. 112482. .
[3] SAJID, M.U. - ALI, H.M. Thermal conductivity of hybrid nanofluids: A critical review. In *International Journal of Heat and Mass Transfer* . 2018. Vol. 126, s. 211–234. .
[4] HWANG, Y. et al. Production and dispersion stability of nanoparticles in nanofluids. In *Powder Technology* . 2008. Vol. 186, no. 2, s. 145–153. .
[5] ŠÁRPATAKY, M. et al. Dielectric Fluids for Power Transformers with Special Emphasis on Biodegradable Nanofluids. In *Nanomaterials* . 2021. Vol. 11, no. 11, s. 2885. .
[6] KHALED, U. - BEROUAL, A. AC dielectric strength of synthetic ester-based Fe3O4, Al2O3 and SiO2 nanofluids — conformity with normal and weibull distributions. In *IEEE Transactions on Dielectrics and Electrical Insulation* . 2019. Vol. 26, no. 2, s. 625–633. .
[7] KHALED, U. - BEROUAL, A. DC breakdown voltage of natural ester oil-based Fe3O4, Al2O3, and SiO2 nanofluids. In *Alexandria Engineering Journal* . 2020. Vol. 59, no. 6, s. 4611–4620. .
[8] BEROUAL, A. et al. DC Breakdown Voltage of Synthetic Ester Liquid-Based Nanofluids. In *IEEE Access* . 2020. Vol. 8, s. 125797–125805. .
[9] BEROUAL, A. - KHALED, U. Statistical Investigation of Lightning Impulse Breakdown Voltage of Natural and Synthetic Ester Oils-Based Fe3O4, Al2O3 and SiO2 Nanofluids. In *IEEE Access* . 2020. Vol. 8, s. 112615–112623. .
[10] SZCZEŚNIAK, D. - PRZYBYLEK, P. Oxidation Stability of Natural Ester Modified by Means of Fullerene Nanoparticles. In *Energies* . 2021. Vol. 14, no. 2, s. 490. .
[11] HUANG, Z. et al. Significantly enhanced electrical performances of eco-friendly dielectric liquids for harsh conditions with fullerene. In *Nanomaterials* . 2019. Vol. 9, no. 7. .
[12] MÉNDEZ, C. et al. Effect of Magnetic and Non-magnetic Nanoparticles on Insulation and Cooling Behaviour of a Natural Ester for Power Transformers. In *2020 International Symposium on Electrical Insulating Materials (ISEIM)* . 2020. s. 111–114. .
[13] PRIMO, V.A. et al. AC breakdown voltage of Fe3O4 based nanodielectric fluids. Part 1: Analysis of dry fluids. In *IEEE Transactions on Dielectrics and Electrical Insulation* . 2020. Vol. 27, no. 2, s. 352–359. .
[14] HUSSAIN, M.R. et al. Dielectric Performance of Magneto-Nanofluids for Advancing Oil-Immersed Power Transformer. In *IEEE Access* . 2020. s. 1–1. .

[15] OLMO, C. et al. Maghemite Nanofluid Based on Natural Ester: Cooling and Insulation Properties Assessment. In *IEEE Access* . 2019. Vol. 7, s. 145851–145860. .
[16] CHARALAMPAKOS et al. Dielectric Insulation Characteristics of Natural Ester Fluid Modified by Colloidal Iron Oxide Ions and Silica Nanoparticles. In *Energies* . 2019. Vol. 12, no. 17, s. 3259. .
[17] RAFIQ, M. et al. Effect of Al2O3 nanorods on dielectric strength of aged transformer oil/paper insulation system. In *Journal of Molecular Liquids* . 2019. Vol. 284, s. 700–708. .

Measurement of leakage current in wet and polluted conditions using different sensing electrodes

[1]*Ľuboš ŠÁRPATAKY (2nd year)*
Supervisor: [2]Bystrík DOLNÍK

[1,2]Dept. of Electric Power Engineering, FEI TU of Košice, Slovak Republic

[1]lubos.sarpataky@tuke.sk, [2]bystrik.dolnik@tuke.sk

***Abstract*— Research and diagnostics of insulators are evolving in different directions, but recent trends emphasize measurement and diagnostics that can be implemented online. Thanks to the modernization of technology, it is possible to collect and analyze a large amount of data. Therefore, the research focuses on the online measurement, where data collection is continuous and the system can evaluate the measured quantities immediately, which can prevent the failure of insulators and thus increase the reliability of transmission and distribution of electricity. One of the biggest problems with insulators is the susceptibility to flashover. The chance of flashover increases with increasing humidity and pollution level. Research is focused on various sensing electrodes which can be used to measure quantities to determine the severity of pollution at different humidity.**

***Keywords*—sensing electrodes, pollution, humidity, leakage current, glass insulator**

I. Introduction

Many research institutes are involved in the development of insulation and a lot of money is spent on it. Insulators, which serve to support and separate electrical conductors without allowing current through themselves, are one of the most important insulating elements, especially in the transmission and distribution of electricity on high-voltage lines.

The operating conditions in which the insulators operate vary based on the design of the insulators (mechanical stress), the environment, and the voltage level at which they are installed (electrical stress). The insulators are constantly exposed to a combination of these stresses, especially electrical stress in combination with an unsuitable environment or weather. In addition to electrical stress, frost or snow increases and suddenly changes the mechanical stress of the insulator. These stresses cause degradation and aging on the insulation systems. Insulators will only operate reliably throughout their planned service life if their design and use take into account all relevant stresses [1], [2].

The mechanism of aging by electrical stress in electrical insulation systems is a very complex phenomenon. Precise methods have yet not been defined to describe accelerated aging to determine insulation life as well as analytical methods for thermal aging. However, there are well-developed working practices based on some empirical knowledge. It is important to realize that the negative effects of electrical aging are strongly dependent on the nature and type of electrical insulation system. The size of the electrodes, the thickness of the insulation, the degree of stress, the maximum intensity and frequency of the applied voltage, and the degree of probability of failure are factors and variables that affect the magnitude of the electrical stress. Electrical stress is caused by AC, DC, or impulse voltage [3]-[6].

Nowadays a lot of researchers use leakage current to analyze the risk of flashover and changes in parameters that describe the probability of insulators failure mainly when the insulator is polluted. A conductive layer of pollution on insulators increase leakage current and this parameter can predict failure and helps to understand different environmental conditions in which insulator can be placed. Also, lots of research shows that humidity rapidly increases the leakage current as well as the probability of flashover. Besides the basic measurement of leakage current which is very perspective and can be measured online, new research is focused on harmonic analysis. Leakage current magnitude seems to indicate the influence of relative humidity and harmonic compounds indicate the distribution of pollutants at the insulator's surface. Different indexes were created to predict the danger of flashover via harmonic compounds [7]-[15].

II. Measurement of leakage current

Measurement of leakage current was performed in a high-voltage laboratory at the Technical University in Košice. Glass insulators were used to measure leakage current at different relative humidity (from 40 % to 90 %) and under different pollution levels. Two types of sensing electrodes were used.

A. Materials and methods

The first measuring electrodes were formed by a conductive paste applied to the surface of the glass insulator. The applied conductive paste was fired in a muffle furnace. The distance between the measuring electrodes of the circular shape was 4 cm. Glass insulator with measuring electrodes is in Fig. 1.

The second type of measuring electrodes was formed with copper conducting tape paste on the insulator's surface. The copper electrodes were slightly elevated on the surface because of the thickness of the tape.

Measurement conditions were equal for both electrodes. Insulators with different electrodes were placed in a closed chamber. Relative humidity was increased stepwise from 40 % to 90 %. Measurement was performed on insulators with a light pollution layer (L1) and with a heavy pollution layer (L4). The

conductivity of light pollution solution is 28 µS/cm and conductivity of heavy pollution is 286 µS/cm.

Fig. 1 Electrodes made of conductive paste applied to the surface of the glass insulator.

Fig. 2 Copper tape electrodes on a glass insulator during measurement.

The measurement results of the leakage current on glass insulators under various environmental conditions, shown in Fig. 3 and Fig. 4, were performed using a function/arbitrary waveform generator Agilent 33210A (G) and an oscilloscope Agilent DSO7104A (DSO). To prevent electromagnetic interference from the supply network, two different frequencies of the measuring voltage were used. The first frequency close to the industrial frequency is 113 Hz, the second frequency several times greater than the industrial frequency is 1000 Hz. The resulting leakage current (i_L) value was calculated as the average of 128 periods of measuring voltage.

B. Measurement results

In Fig. 3 and Fig. 4, the dependence of leakage current on relative humidity for different pollution levels with different sensing electrodes is depicted.

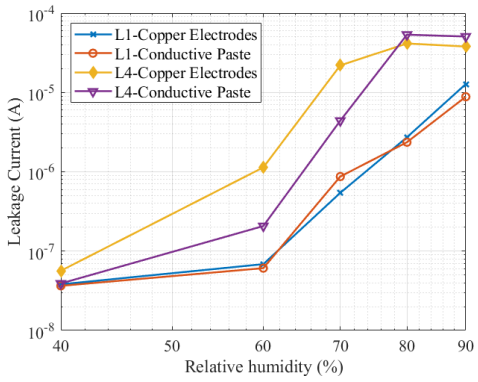


Fig. 3 Measurement of leakage current at frequency 113 Hz.

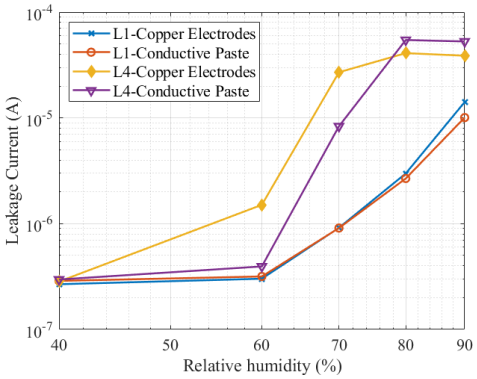


Fig. 4 Measurement of leakage current at frequency 1 kHz.

Dependencies of leakage current on relative humidity show that contamination of the glass insulator surface and relative humidity significantly affect leakage current over the surface of the glass insulator. The increase of leakage current between relative humidity 60 % and 70 % is the most pronounced. The increase is significant with increasing pollution level too. For heavily polluted insulator, leakage current increases more rapidly.

The difference between the individual electrodes is evident when measuring with heavy contamination. Light pollution measurements showed almost identical results. Copper electrodes show more than one-third higher leakage current at relative humidity 60 % and 70 %. At relative humidity from 80 % to 90 % the difference was minimal.

III. Conclusion and Future Research

Research of insulators, especially of contaminated insulators, is still in progress. Problems associated with flashover through the surface of the insulator are complex due to the large number of factors that can affect this phenomenon. Research mainly focuses on contamination as the most severe factor. My research focuses on the identification of the pollution level at different relative humidity. The result of the leakage current measurement confirms a rapid increase of the current on the surface of the insulator with increasing relative humidity as well as contamination levels. Electrodes used for the measurement of leakage current show the same trend of the curve, but heavy pollution affected the value of the leakage current mainly at 60 % and 70 % relative humidity levels. Therefore, I recommend using one type of electrode for all measurements.

The measurement continues with other quantities that can be used as a tool for the identification of contamination, namely capacity, electric charge, and dielectric loss factor. All these quantities show promising results. They could be used to determine the level of pollution [16]. All quantities were measured on glass, porcelain, and composite insulators. After the measurement of multiple quantities, the research moves to the harmonic analysis of leakage current. I will examine how the layer of pollution affects harmonic compounds. Measurement will be performed at a different relative humidity.

The same electrodes, contamination levels, and insulating materials will be used for the measurement of impulse breakdown voltage. Standard impulse voltage tests are performed only in dry and wet conditions. I will examine the behavior of materials in more detail at different humidity to find out at which humidity the most significant decrease in the electrical strength of the material will occur.

REFERENCES

[1] M. Farzaneh a W. A. Chisholm, Insulators for icing and polluted environments. Piscataway, NJ : Hoboken, NJ: IEEE Press ; J. Wiley, 2009.

[2] E. Kuffel, W. S. Zaengl, a J. Kuffel, High voltage engineering: fundamentals, 2nd ed. Oxford ; Boston: Butterworth-Heinemann, 2000.

[3] C. Mayoux, "Degradation of insulating materials under electrical stress", IEEE Trans. Dielect. Electr. Insul., y. 7, no. 5, p. 590–601, oct. 2000, doi: 10.1109/TDEI.2000.879355.

[4] Al-Gheilani, W. Rowe, Y. Li, a K. L. Wong, "Stress Control Methods on a High Voltage Insulator: A Review", Energy Procedia, y. 110, p. 95–100, Mar. 2017, doi: 10.1016/j.egypro.2017.03.112.

[5] G. Carlo Montanari, P. Morshuis, P. Seri, a R. Ghosh, "Ageing and reliability of electrical insulation: the risk of hybrid AC/DC grids", High Voltage, y. 5, no. 5, p. 620–627, Oct. 2020, doi: 10.1049/hve.2019.0371.

[6] M. Váry a E. Firický, "Electrical insulation system ageing caused by electrical, multifactorial and environmental stresses", Posterus, June. 12, 2013. http://www.posterus.sk/?p=15765 (cit Jan. 31, 2022).

[7] Zhengfa L., Qing Z., Wuyang Z., Shimian L., Gaolin W., Jianlin H. et al. Study on leakage current characteristics and influence factors of 110kV polluted composite insulators. In: 2018 12th International Conference on the Properties and Applications of Dielectric Materials (ICPADM) [Internet]. Xi'an: IEEE; 2018 [cited 2021 Feb 8]. p. 896–900. Available from: https://ieeexplore.ieee.org/document/8401173/

[8] Deb S., Ghosh R., Dutta S., Dalai S., Chatterjee B. Effect of humidity on leakage current of a contaminated 11 kV Porcelain Pin Insulator. In: 2017 6th International Conference on Computer Applications In Electrical Engineering-Recent Advances (CERA) [Internet]. Roorkee: IEEE; 2017 [cited 2021 Feb 8]. p. 215–9. Available from: https://ieeexplore.ieee.org/document/8343329/.

[9] Ahmadi-Joneidi I., Shayegani-Akmal AA., Mohseni H. Leakage current analysis of polymeric insulators under uniform and non-uniform pollution conditions. IET Generation. Transmission & Distribution. 2017 Aug;11(11):2947–57.

[10] Dadashizadeh Samakosh J., Mirzaie M. Analysis of leakage current characteristics during aging process of SiR insulator under uniform and longitudinal non-uniform pollution conditions. Measurement. 2019 Dec;147:106862. doi: 10.1016/j.measurement.2019.106862.

[11] Ramirez I., Hernandez R., Montoya G. Measurement of leakage current for monitoring the performance of outdoor insulators in polluted environments. IEEE Electr Insul Mag. 2012;28(4):29–34. doi: 10.1109/MEI.2012.6232007.

[12] Salem AA., Abd-Rahman R., Al-Gailani SA., Salam Z., Kamarudin MS., Zainuddin H. et al. Risk Assessment of Polluted Glass Insulator Using Leakage Current Index Under Different Operating Conditions. IEEE Access. 2020;8:175827–39. doi: 10.1109/ACCESS.2020.3026136.

[13] Fauziah D., Alfiadi H., Rachmawati., Suwamo. The effect of coating on leakage current characteristic of coast field aged ceramic insulator. In: 2017 4th International Conference on Electrical Engineering. Computer Science and Informatics (EECSI) [Internet]. Yogyakarta: IEEE; 2017 [cited 2021 Feb 4]. p. 1–7. Available from: http://ieeexplore.ieee.org/document/8239140/

[14] Salem AA., Abd-Rahman R., Al-Gailani SA., Kamarudin MS., Ahmad H., Salam Z. The Leakage Current Components as a Diagnostic Tool to Estimate Contamination Level on High Voltage Insulators. IEEE Access. 2020;1–1. doi: 10.1109/ACCESS.2020.2993630.

[15] Pandian RG., Subburaj P. Leakage current and flash over performance analysis of 11 kv pin insulator under Bird excretion pollution. In: 2016 International Conference on Energy Efficient Technologies for Sustainability (ICEETS) [Internet]. Nagercoil. India: IEEE; 2016 [cited 2021 Feb 4]. p. 311–4. Available from: http://ieeexplore.ieee.org/document/7583771/

[16] B. Dolník, Ľ. Šárpataky, I. Kolcunová, and P. Havran, "Sensing Method Using Multiple Quantities for Diagnostic of Insulators in Different Ambient Conditions," *Sensors*, vol. 22, no. 4, p. 1376, Feb. 2022, doi: 10.3390/s22041376.

Data exchange and the emergence of new participants in the electricity market in the Slovak Republic

[1]*Jozef HUMENÍK (2nd year)*
Supervisor: [2]Jaroslav DŽMURA

[1,2]Dept. of Electric Power Engineering, FEI TU of Košice, Slovak Republic

[1]jozef.humenik@student.tuke.sk, [2]jaroslav.dzmura@tuke.sk

***Abstract*—This article researches the current reform of the electricity market. As part of this reform, European legislation is being implemented in Slovak conditions. In practice, this means an amendment to Slovak national legislation, (Act No. 251/2012 Coll. on Energy and on Amendments to Certain Acts, as amended, and Act No. 309/2009 Coll. on the promotion of renewable energy sources and high-efficiency cogeneration and on the amendment of certain laws as amended) and the emergence of new activities/processes and new market participants, which will cause a significant increase in the volume of data flows between individual participants in the electricity market.**

***Keywords*—accumulation systems, aggregation, data exchange, energetic communities, flexibility.**

I. Introduction

The Slovak Republic is currently preparing for the implementation of EU legislation into our national legislation, which is known as "Clean Energy for All Europeans" or the "Winter Energy Package". The essence of the acceptance of such legislation in a comprehensive change in the electricity market. New processes/activities will be introduced on the market, e.g., providing flexibility, aggregation, or accumulation of electricity. Consumers/customers will be given the opportunity to act on the market as an active consumer/customer, and to associate with other such customers in the so-called “civil energy communities”. With the development of the electricity market in the energy sphere, the volume of data flows - data transmission and processing - will increase. This will lead to changes and the setting up of new processes to enable new activities and activities to be created and developed, as this is the main goal of the "Winter Energy Package".

II. Project “Energy Data Centre (EDC)”

OKTE, a.s. (Organizer of the short-term electricity market), which is a subsidiary of the transmission system operator (Slovak electricity transmission system, Plc.), was entrusted by the Ministry of Economy of the Slovak Republic with the preparation of the "Energy Data Centre" project. The aim of the project is to inform market participants about market changes, adjust data flows between existing and new market participants and set up processes. The project implementers are OKTE, a.s., SEPS, a.s., ÚRSO (the Regulatory Office for Network Industries), MH SR (the Ministry of Economy of the Slovak Republic) and distribution system operators (DSO). One of the goals of the project is to analyse and prepare technical proposals for solutions for processes and activities in the electricity market. It is divided into two categories, namely new processes (flexibility aggregation, electrical energy storage, energy communities and electricity sharing) and selected existing processes (data management, smart meter data, reports, and statistics to reduce the administrative burden on the market, sharing data on the application of guarantees of origin, sharing of data on electricity production, including data on production from "RES – Renewable Energy Sources"). The aim of the project is also to prepare the possibilities of technical solutions for the solution of the actual regulation of implementing legislation and decrees regulating the details of the functioning of the electricity market. The list of system/technical functionalities:

1. Interfaces (IFs) – for data distribution
2. Processes
3. Users’ management
4. Data – database (DB)
5. Reports
6. Configuration of functionalities

For data exchange between EDC and participants will be used these channels:

1. Web services
2. Import/export
3. E-mail
4. User portal [1][2][3][4]

A. New activities/processes on the market [1][2][3][4]

Aggregation of demand side flexibility in a smart grid

The opening of the electricity market in Slovakia will mean that the main market participants will not be conventional (monopoly) power plants and large companies in terms of consumption. A few "small" producers and customers will be added, which will cause flexibility aggregation. Flexibility is key and can be found wherever it is possible to accumulate cold (e.g., freezers) and heat. It will bring the development of services in the field of consumption management, flexibility, as well as the offer of new products for the provision of support services. The potential and usability of household flexibility will be significantly increased due to the construction of AMI (Advanced Metering Infrastructure) and

systems designed to manage electricity consumption, increase the share of electromobility and develop the concept of prosumers.

Accumulation of electrical energy

The energy data centre should ensure:

1) **Central accumulation record:** record of objects where the accumulation will be assigned to the end consumer.
2) **Receipt of technical, business and identification data on accumulation from distribution system operator:** necessary data for accumulation - automatic process of the distribution system operator to the central data register (at OKTE via EDC). These are technical data on accumulations (e.g., installed capacity) and commercial data (Connection Agreement).
3) **Linking technical data on accumulation with information on the operator of the facility:** in the central register there would be linking technical data on accumulation with information on who operates the facility.
4) **Updating the self-operator**: Via EDC, it would be possible for the storage operator to update the self-operator.
5) **Assignment of a secondary accumulation meter to the consumer point and registration of a request for accumulation assessment:** the accumulation operator will have to install a secondary accumulation meter if he wishes to exercise his right to assess accumulation
6) **Consideration of accumulation of electrical energy in sent data for invoicing with the use of fictitious accounts:** performed by OKTE within the relevant evaluation processes according to the set rules.

Energy communities and electricity sharing

Within the EDC project, several solutions or main functions needed for the management of energy communities (community of active customers) were prepared on this topic. Records of individual members of the Community and their share in the production or accumulation of electricity, including records of changes in their share, information of a technical character (electrical equipment with supply to the electricity distribution network) - small power plant or storage point.

B. Existing processes on the market [1][2][3][4]

Electric vehicle charging infrastructure

In the area of Electromobility, the need arose to solve the connection with EDC in terms of processes and data structures. Because at present, information about the load and operation of the charging station is important for the DSO from the point of view of data exchange. In the future, such data will probably not be sufficient, so DSOs will need data on the current load of charging stations to ensure a proper supply of electricity. The difference and essence consist in the speed of data availability, ideally in real time. However, providing real-time data requires enormous technical and financial resources, e.g., fee for data services or operation and maintenance of servers.

Master Data of consumer points and consumers

Currently, master data can be divided into business, technical and identification. Within this section, it would be interesting or even necessary to focus on the precise analysis / research of determining the owner of such data, which will oversee collection, validation, and change (updating). Subsequently, by creating a concept of sharing this data to eligible market participants and ensuring a unique identifier of collection points (currently the EIC code is used), e.g., by modifying the structure. For comparison, as required by the situation in the field of IT with the change of IP addresses (from IPv4 to IPv6).

Meter Data of consumer points

Measured data from smart meters are also currently sent to OKTE, a.s. In this regard, only one new communication channel should be added, which will be used to send data to a central data repository - EDC. In the EDC, the data will be unified and available to all market participants according to the assigned rights.

Production Data of consumer points

Electricity producers are currently obliged to send data on production and consumption to 5 entities (OKTE, a.s., ÚRSO, MH SR, Statistical Office of the Slovak Republic, and DSO), which causes excessive administrative burden and problems with ensuring the correctness of data. The aim of the project in this direction is to simplify / reduce the number of entities to one (OKTE database -> EDC).

Reports and statistics

Reports and statistics help for correct and efficient managerial process management. Today, they are often manually created and distributed. In this regard, the future is in automating the creation of the required outputs, which will contribute to economy and process acceleration.

III. CONCLUSION

Data exchange and the emergence of new entrants to the electricity market in Slovakia represent a long process, which is currently almost in it begin. In my opinion, it is still premature to evaluate the EDC project, but the idea and its aim suggest that the project reflects market requirements and will address them broadly. However, Slovakia and the whole area of smart grids themselves are on the threshold of changes that will positively and negatively affect the entire electricity market. In this case, it is the right step forward in building an efficient, reliable, and secure model of the electricity market.

REFERENCES

[1] OKTE, a.s. 2022. *Projekt: Energetické dátové centrum (EDC)*. Bratislava, Slovakia.

[2] ENERGOKLUB (sféra, a.s.). 28.01.2022. *Prvá etapa energetického dátového centra sa má ukončiť do konca roka* [online]. Dostupné:< https://energoklub.sk/sk/clanky/prva-etapa-energetickeho-datoveho-centra-sa-ma-ukoncit-do-konca-roka/>

[3] POTOČÁR, R. 28.01.2022. *Energetické dátové centrum ovplyvní celý trh s elektrinou. Má umožniť aj úplne nové činnosti* [online]. Dostupné: <https://www.energie-portal.sk/Dokument/energeticke-datove-centrum-ovplyvni-cely-trh-s-elektrinou-ma-umoznit-aj-uplne-nove-cinnosti-107779.aspx>

[4] HUMENÍK, J. 2022. *Workshop for the professional public to the EDC project*, Author's own notes.

3D Scanning of the Indoor Environment

[1]*Stanislav ALEXOVIČ (2nd year)*
Supervisor: [2]Milan LACKO

[1,2]Department of Electrical Engineering and Mechatronics, FEI TU of Košice, Slovak Republic

[1]stanislav.alexovic@tuke.sk, [2]milan.lacko@tuke.sk

***Abstract* — The paper provides a brief description of the design, implementation, and testing of a depth sensing device able to capture a scanned indoor environment and produce a 3D point cloud. A proof of concept for building an autonomous UAV capable of exploring unknown indoor environments is outlined. The paper also focuses on the sensors essential for successful operation of an autonomous UAV.**

***Keywords*—3D scanning, visual SLAM, UAV**

I. Introduction

3D scanning refers to the technology of acquiring visual and dimensional properties from an environment or a real-world object and transferring it to a digital form. This form can be then later processed or analysed using various algorithms [1]. This technology is a part of computer vision, a thriving field which has been receiving more and more attention from researchers and developers with various technological backgrounds.

There are several methods of acquiring a reconstruction of a 3D scene, differing mainly in the input sensors used. The first one is built on stereo vision technology making use of two cameras [2], much like the principle of depth sensing used by humans. This scanning method tends to be aided by adding an active component to the scanning process. The job of such a projector is to cast random [3] or structured patterns [4] on the target surface and improve the quality of the received data.

The second method used to acquire depth information from an environment is called Time of Flight (ToF) [5]. ToF calculates depth by measuring the time which transmitted light needs to hit the scanned object and reflect to a receiver. The receiver in such scanning devices usually consists of a CMOS chip which can measure the depth of all pixels simultaneously. This makes the information about the scene depth available all at once.

The main focus of the paper is to provide a description of a standalone device capturing 3D environment with the use of cameras based on active stereo vision (see Fig. 1). The device is built from components for standard unmanned aerial vehicles (UAV) so as to test the selected sensors along with other parts. In the future, the setup will be used for building an autonomous UAV for exploration of unknown GPS-denied environments [6][7].

Fig. 1. Photos of 3D scanning device.

II. Hardware

The components which will be used in the next phases of the project need to satisfy the following hardware requirements:

- Low power consumption – all electrical parts will be powered by a local battery source; power efficiency might increase the operation time,
- Large battery capacity – the battery source needs to have enough capacity to provide a reasonable operation time,
- Wireless communication – communication with the user's computer needs to be accomplished wirelessly,
- Computing power – Local computer which will host all internal logic must be capable of processing all algorithms required to reconstruct a 3D scene.

A. Depth camera

The Intel RealSense D435i depth camera was chosen as the primary camera. It is small, lightweight, power efficient, and may be used both in indoor and outdoor environments. Depth sensing technology in D435i is based on active stereo vision which uses Infra-Red (IR) projector and two IR cameras.

B. Tracking camera

Intel RealSense T265 tracking camera was used as a complement to the primary camera. It comes with Visual SLAM on its own and its output is a precise localization of itself in the scanned environment. Its two fisheye lenses have a FOV of 163° together and its low power consumption visual processing unit can run V-SLAM algorithms. Also, the IMU is integrated which is required by V-SLAM.

C. Battery

3D scanner must be able to operate on its own. This implies that it must have its own source of energy. For this purpose, 5000 mAh Li-Polymer 3S battery was used which is commonly present in RC models or small UAVs.

D. Mission computer

Mission computer is the name for the main computer stored inside the 3D scanner. Raspberry Pi 4 model B was used in this project. This single-board computer provides sufficient computing power to run programs required for processing the input recorded by the depth camera.

III. Software

When it comes to software, all logic in the 3D scanner is concentrated in the mission computer running on Linux operation system (OS), more specifically the Ubuntu Mate 20.04 which includes the Robot Operation System (ROS) software framework.

The key software component is the RTAB-Map, a Graph-Based SLAM approach based on an incremental appearance-based loop closure detector [8]. It combines RGB-D, Stereo, and Lidar to create a 3D point cloud of the scanned environment. RTAB-Map operates in two modes, one of which is mapping, and the other is localization.

IV. Testing

Tests were conducted in a house with a stable Wi-Fi coverage. Once the scanner was turned on, RViz on the user's computer was launched, with environment variables defining the IP address of the 3D scanner with the ROS core. One of the features of the RViz is a visual representation of the stream from the depth camera. This proved useful for debugging as well as for checking whether the camera was switched on and working. The scanning output took a form of a point cloud – a set of points located in 3D space (see Fig. 2). The point cloud gradually enlarged as the scanner obtained data from new areas. Several problems arose during the testing phase.

First, the scanner had to be moved slowly and smoothly to provide the mission computer with enough time to process input depth and record streaming from the cameras.

Second, the speed of visualization depended on the Wi-Fi quality. However, the signal quality varied based on the distance from the Wi-Fi access point and the walls interfering with the signal.

Framerate of the depth camera had to be lowered to 6 FPS due to the inability of the mission computer to process input data on higher rates.

After the scanner was moved back to its initial position, a loop closure detector of RTAB-Map corrected any accumulated scanning errors from the previous scanning. This resulted in a slight shift of the affected points in the point cloud. The data was stored in the mission computer database, which made it available in the post-processing phase.

V. Conclusion

The following paper presented a 3D scanner built from components used mainly for UAVs. The primary aim of the outlined experiment was to create a real-time 3D environment scan in a standalone device. This was achieved by combining depth and tracking cameras as main input sensors, a mission computer which held all logic and processing power, both supported by a local battery and power management circuits. The software, based on ROS and RTAB-Map with RViz as the visualisation tool, was hosted on a separate computer and interacted with the scanner over the Wi-Fi.

Fig. 2. Point cloud created with 3D scanner in indoor environment.

Testing proved that the 3D scanner managed to produce a real-time 3D scene of the captured environment which might be used for various purposes later.

Findings of this paper will be used in future research which will be aimed at autonomous exploration of an unknown environment using a UAV in a GPS-denied environment.

Acknowledgments

This research paper was supported by Slovak Research and Development Agency under the APVV-18-0436 project.

References

[1] 1. S. M. Ayaz, D. Khan and M. Y. Kim, "3D handheld scanning based on multiview 3D registration using Kinect Sensing device," 2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), 2017, pp. 330-335, doi: 10.1109/MFI.2017.8170450.

[2] 2. C. Lee, H. Song, B. Choi and Y. -S. Ho, "3D scene capturing using stereoscopic cameras and a time-of-flight camera," in IEEE Transactions on Consumer Electronics, vol. 57, no. 3, pp. 1370-1376, August 2011, doi: 10.1109/TCE.2011.6018896.

[3] 3. M. Shibata and T. Honma, "3D object tracking on active stereo vision robot," 7th International Workshop on Advanced Motion Control. Proceedings (Cat. No.02TH8623), 2002, pp. 567-572, doi: 10.1109/AMC.2002.1026983.

[4] 4. M. Atif and S. Lee, "Adaptive Pattern Resolution for Structured Light 3D Camera System," 2018 IEEE SENSORS, 2018, pp. 1-4, doi: 10.1109/ICSENS.2018.8589640.

[5] 5. O. Choi and S. Lee, "Wide range stereo time-of-flight camera," 2012 19th IEEE International Conference on Image Processing, 2012, pp. 557-560, doi: 10.1109/ICIP.2012.6466920.

[6] 7. B. Yamauchi, "A frontier-based approach for autonomous exploration," in Proceedings of the IEEE International Symposium on Computational Intelligence in Robotics and Automation, 1997, pp. 146–151.

[7] 8. C. Zhu, R. Ding, M. Lin and Y. Wu, "A 3D Frontier-Based Exploration Tool for MAVs," 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI), 2015, pp. 348-352, doi: 10.1109/ICTAI.2015.60.

[8] 9. M. Labbé and F. Michaud, "RTAB-Map as an Open-Source Lidar and Visual SLAM Library for Large-Scale and Long-Term Online Operation," in Journal of Field Robotics, vol. 36, no. 2, pp. 416–446, 2019. (Wiley)

An Overview of Internet of Vehicles: Architectures and Applications

[1]*Dušan HERICH (1st year),*
Supervisor: [2]*Ján VAŠČÁK*

[1,2]Dept. of Cybernetics and Artificial Intelligence, FEI TU of Košice, Slovak Republic

[1]dusan.herich@tuke.sk, [2]jan.vascak@tuke.sk

***Abstract*—The technology of Internet of Things shows its potential also in the area of traffic systems including future use of autonomous vehicles, too. However, some modifications of this technology are necessary, which result in a new area of the so-called Internet of Vehicles. Therefore, the aim of this paper is to provide an overview of this new technology as a distributed network for connecting and managing vehicles. Here, current architectures and potential applications are described, which represent the base for next research in this area.**

***Keywords*—architecture, internet of vehicles, navigation, swarm**

I. Introduction

The ever increasing proportion of autonomously controlled vehicles renders the need for their precise and expeditious navigation. Notwithstanding the achievement of methods for sub-tasks of navigation, particularly localization, mapping, simultaneous localization and mapping (SLAM), and path planning for singular, autonomous vehicles, existing methods are frequently deemed insufficient for the purpose of navigation of multiple autonomous vehicles, including swarm formations. Due to those constraints, the current research focuses on the utilization of technologies such as the Internet of Things (IoT), particularly its application to the autonomous vehicles known as the Internet of Vehicles (IoV) in concoction with concepts of cloud, edge and fog computing to provide means for practical and safe means for the navigation.

The concept of IoV has evolved from its predecessor, vehicular Ad Hoc Networks (VANET), and contrary to its predecessor, it enables data collection, data processing and secure data sharing among objects in the network [1]. As a result of those characteristics, the IoV networks can manage and supervise vehicles effectively [2]. Moreover, the integration of IoT technologies enables intelligent management of road traffic as well as an intelligent information distribution or remote vehicle control [2]. The concepts employed in the IoV technologies have been implemented in all three subtasks of vehicle navigation, that is, in localization [3] with a focus on real-time position estimation of singular vehicles within the network, mapping [4], where edge and fog computing is utilized to aid in the map creation by offloading computation from the robot's units and path planning [5] used for traffic scheduling for the needs of emergency vehicles.

The first section offers presentation of components necessary for the operation of the network, that is, onboard, roadside and application units. Subsequently, communication models exploitation in the relation to the components is examined. Consequently, conceptual models are evaluated and generalized. The following section provides an insight on application possibilities of the IoV. The last section describes future research directions.

II. IoV architectures

The presence of IoV networks enables vehicles to communicate with pre-established infrastructure and to independently communicate with each other. Concepts of such vehicular networks evolve gradually and lack proper standardization. Hence, this section intends to describe infrastructural components employed in communication and conceptual models of IoV architectures.

A. Components of IoV

The deployment of IoV networks is dependent on supporting infrastructure to facilitate communication of vehicles with the surrounding environment, such as other vehicles, road infrastructure or people. Vehicular networks are highly dynamic and are characterized by intermittent connections caused by wireless connections to moving vehicles and heterogeneity present in those networks.

To enable IoV operation, vehicular networks as described are usually composed of several units providing sensing, computing, and communication capabilities allowing the fulfilment of tasks arising from application requirements of the IoV. Those units are commonly separated into three categories, namely roadside unit, onboard unit and application unit. The unit categorization is based on the components utilized in the VANET as a predeceasing step to the IoV. The roadside unit signifies computing devices installed on a designated location in close proximity to the expected path of vehicles [6]. The primary functionality of roadside units (RSU) is extending the range of communication of vehicles. Hence, the RSU communicates with onboard units via dedicated short-range communication (DSRC) technology, additionally providing onboard unit with internet access. Furthermore, RSUs in current IoV deployments collect, process and forward information in multiple directions, that is, from and to vehicles and cloud datacenters [7]. Therefore, RSUs usually dispose of more powerful computational resources than onboard units.

The onboard unit represents a computing device located onboard a vehicle. Consisting of a set of sensors for monitoring of speed, acceleration, fuel consumption, geographic

location, and a dedicated communication interface connecting to onboard units (OBUs) present in other vehicles, its primary functions are collecting and processing sensory information, establishing a wireless connection with nearby OBUs and RSUs, message transmission, geographic routing and information security [8].

The application unit (AU) is located inside a vehicle capable of running specific applications for the purposes of IoV. The AU is capable of communication solely via the OBU. Currently, two types of AUs are used. The first one is a dedicated device for secure applications while the second one may be a personal device such as a smartphone [9].

B. Communication models

Due to the employment of a variety of infrastructural components within the IoV, there are numerous models of communication present. Those encompass information interchange among vehicles' OBUs, RSUs or transport infrastructure, hence three categories may be derived on the basis of the three primary components within the IoV as displayed in the Fig. 1, namely the onboard communication, vehicle-to-vehicle communication (V2V) and vehicle-to-infrastructure communication (V2I), commonly referred to as the vehicle-to-everything (V2X) communication model.

The OBU and the AU are responsible for onboard communication, with the OBU providing a communication link for the AU and assuring the functionality of numerous onboard applications.

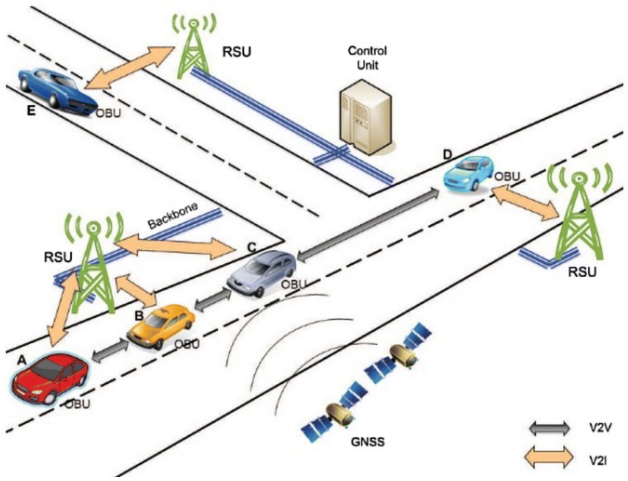


Fig. 1. Communication models within IoV network [10].

The inter-vehicle communication is deemed to be crucial for applications operating based on IoV as the employment of the V2V communication facilitates broadcasting of operational data in regards to traffic status, in particular information concerning detection of collisions or emergencies [11]. In addition, V2V communication facilitates network operations such as data forwarding to the infrastructure in case the vehicle is not located in direct coverage of the IoV infrastructure. Therefore, vehicles may operate as relay nodes in V2X communication expanding the communication range. Stemming from the ability of data sharing among vehicles, the V2V communication enables the formation of ad-hoc vehicular clouds for the purpose of computational resources pooling [11]. The formation of such vehicular clouds advances the effectiveness of the IoV system by omitting the requirement to offload the solution of a complex computational task to the remote cloud by diminishing solution delay of complex tasks such as environment map formation or path planning.

The last discussed model - V2I, defines communication among vehicles and RSUs, providing means for the information exchange among vehicles and other networks [12]. The connection establishment from a vehicle to an RSU portrays further expansion of communication range and provides comprehensive information of the transportation environment by allowing processing and integration of data from individual components.

The adoption of named communication models has rendered significant advantages in comparison to earlier vehicular networks by providing communication for vehicles out of the communication range, expanding computing and storage capabilities by resource pooling and facilitating the development of intelligent applications by processing data collected from the transport system.

C. Conceptual models

A significant portion of the work has been focused on the proposal of conceptual models for IoV networks. The Table I summarizes four selected designs. Due to lack of standardization, proposed models differ from each other significantly. A vast majority of models segregate functionality among several layers. Despite the differences in designs, it is possible to observe patterns in model' layers and their functionality, and three primary layers can be identified. It is possible to refer to those layers as sensory, communication and application.

TABLE I
OVERVIEW OF CURRENT IOV ARCHITECTURES

Reference	Layers	Applications
[13]	Data, Virtualization, Control, Application	See through, Collision Warning
[14]	Sensing, Communication, Cognition, Control, Application	Safety, Transportation Management
[15]	Perception, Communication, Application	Vehicle collaboration
[16]	User, Data Acquisition, Filtering, Communication, Control, Processing	Traffic efficiency, Safety

The lowest, sensory layer, usually encompasses all sensors located in the vehicle, collecting and preprocessing environmental data and detecting specific events of interest such as driving patterns and vehicle situations and environmental conditions.

In some architectures, such as the one exhibited in the Fig. 2, this layer may not be outlined as equipped with sensors [13], however in architectures specifying sensory apparatus [14], it serves for the collection of heterogeneous data from multiple sources such as physical space, network traffic and resource distribution in the network. Hence, the data collection is done from internal sensors in the vehicle, navigation systems, inter-vehicle communication, traffic lights or other devices located in the environment [16].

The second level, can be generalized as a communication layer with support for previously described communication models, specifically the V2X. The primary purpose of the layer is the establishment and maintenance of wireless connections to already established networks such as WiFI, Bluetooth or 5G.

The model in [2] integrates this functionality in the data layer, and wireless communication nodes are additionally equipped with resources for storage and computation while partially abstracting them as fog nodes enabling distributed provisioning of service. In [14] the layer adopts a hybrid of cloud/edge architecture for communication with usage of related wireless technologies. The objective for the adoption of cloud/edge architecture is the requirement for real-time data interchange and processing among intelligent devices. With the employment of the edge computing paradigm, the models built on this architecture are able to process time-sensitive data gathered from vehicles in motion while gradually offloading other data to the cloud environment for computing and their analysis. In [15], the layer uses traditional routing protocols dependent on the geographical location of nodes in the network, topology and vehicle clustering but also utilized graph and path planning algorithms for the shortest routing path discovery.

The third layer, frequently denoted as an application layer, bears the responsibility for storage, analysis and processing of the data gathered by the network operation [17]. For this purpose, high capacity storage and processing infrastructure are included along with tools for the analysis of data gathered from the network and its components. The primordial purpose of the layer is to provision nodes with the processing of big data from several sources for evaluation of a variety of risks and circumstances emerging in the traffic [18]. Those may include unsafe conditions for vehicle movement, congestion of traffic or emergency events .

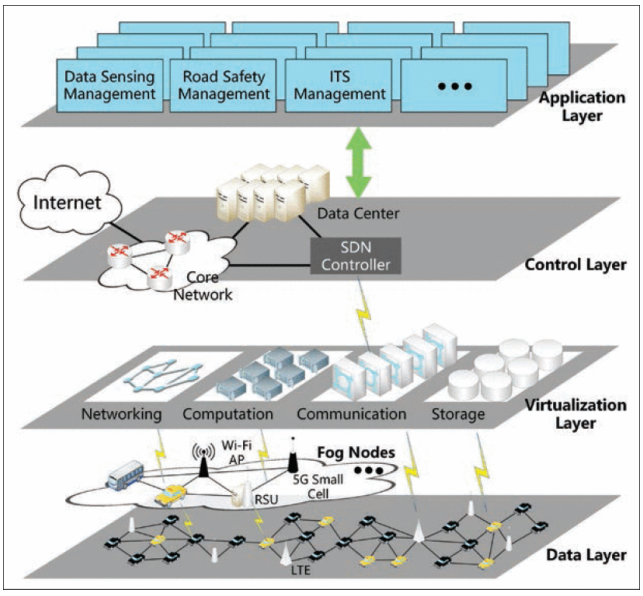


Fig. 2. Architecture of IoV model merging sensing and communication layer [13].

The model in [13] implements two services in this layer, namely the "see-through" and "collision warning" services. Operation of those is based on cloud computing technologies with adaptive scheduling of resources. In [14] services for coordination and cooperation of multiple vehicles, and automatic driving are implemented. The model enables the deployment of other services and separates them into two categories, particularly customized application services aimed at reducing the safety risks in traffic and the intelligent transportation application, including intelligent driving and intelligent transportation management.

The model in [16] uses protocols such as HTTP REST, advanced message queuing protocol or extensible messaging and presents protocol utilized in the IoT network as a demonstration of their usability in the IoV.

III. Applications

Compared to the previous VANET technology, which did not find a comprehensive utilization in the intelligent transportation system, the IoV field of application is extensive. Notable usage of IoV can be found in self-driving vehicles. The cognitive IoV [19], with architecture displayed in the Fig. 3, uses methods of artificial intelligence to enable collaborative decision making for autonomous vehicles on the road. In comparison with current methods for autonomous driving, operating mainly with singular vehicles focusing primarily on speed, steering control and obstacle detection, this application focuses on collaborative decision-making achieved by architecture designed for low-latency communication and usage of fog and cloud computing for its services.

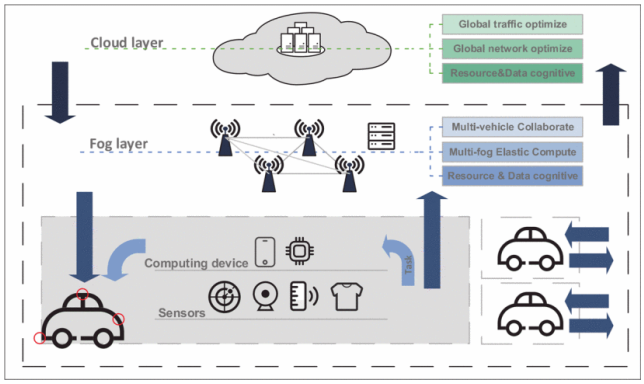


Fig. 3. Architecture of cognitive IoV [14].

Further applications of IoV [1] can be found in warning and avoidance in the presence of emergency vehicles by the guidance of non-emergency vehicles for prevention of congestion and obstructions on the emergency vehicle's route [20]. This service operates by employing the V2V communication broadcasting information about speed and direction to other vehicles in traffic.

A road warning is a service similar to the previous one. The AU onboard a vehicle may issue an alert to a driver [21]. By V2I communication, the infrastructure may broadcast messages informing about slow-speed zones, zones with high pedestrian density or traffic organization changes to a vehicle passing by. Furthermore, the V2C communication in this service allows for transmission of warning messages among cars.

The usage of IoV in intelligent traffic may be found at intelligent intersections where the need for various external signalling devices is eliminated [22]. That means the intersection may be equipped by an RSU managing an optimal traffic flow and hence increasing the road's permeability. However, in spots where an installation of an RSU is challenging to accomplish, vehicles employ a V2V communication to negotiate order of passage.

In concoction with the previous applications, intelligent traffic management aids in the improvement of the traffic flow by integrating network-wide traffic data and accordingly adjusting control signals at intersections, hence diverting traffic to less congested paths [23].

IV. Future Work

Stemming from the presented applications and emergence of cognitive IoV applying methods of artificial intelligence in the application layer, the further utilization of artificial intelligence should be examined not only for the purposes of the uppermost layer but also in all lower layers.

Due to the nature of IoV networks, including a multitude of mobile vehicles, we deem algorithms of swarm intelligence as potentially beneficial techniques aiding the improvement of services without the requirement of continuous horizontal and vertical scaling of computation, storage and communication resources. Other than swarm intelligence algorithms, methods of machine learning may be used in lower layers for the same purpose of effective enhancement without the requirement for additional resources. An example [24] may be found in path planning where machine learning is used to determine the fastest computing environment to create a path plan by estimating the algorithm's speed on edge versus cloud.

On the application layer, potential usage of swarm intelligence algorithms can be implemented in traffic management and coordination. In accordance with an application example of intersection management, some swarm-based methods may be used to effectively schedule the order of passage. Similarly, those methods might be, after some alteration to meet implementation demands, used for traffic routing throughout the available paths to aid congestion problems. Implementation of those algorithms may be based on the concepts of emerging cognitive IoV aiming to enhance cooperation amongst individual vehicles.

However, the primary focus of future work will be the solution to traffic problems via artificial intelligence without requiring more resources, novel devices serving as infrastructure may be designed to facilitate novel approaches and allow better integration of fog and edge computing in the IoV.

Acknowledgment

This publication was supported by the APVV grant ENISaC—edge-enabled intelligent sensing and computing (APVV-20-0247).

References

[1] B. Ji, X. Zhang, S. Mumtaz, C. Han, C. Li, H. Wen, and D. Wang, "Survey on the internet of vehicles: Network architectures and applications," *IEEE Communications Standards Magazine*, vol. 4, no. 1, pp. 34–41, 2020.

[2] R. Gasmi and M. Aliouat, "Vehicular ad hoc networks versus internet of vehicles-a comparative view," in *2019 International Conference on Networking and Advanced Systems (ICNAS)*. IEEE, 2019, pp. 1–6.

[3] M.-S. Gu, F. Miao, C.-B. Gao, Z.-S. He, W.-J. Fan, and L. Li, "Research of localization algorithm of internet of vehicles based on intelligent transportation," in *2018 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*. IEEE, 2018, pp. 13–16.

[4] V. K. Sarker, J. P. Queralta, T. N. Gia, H. Tenhunen, and T. Westerlund, "Offloading slam for indoor mobile robots with edge-fog-cloud computing," in *2019 1st international conference on advances in science, engineering and robotics technology (ICASERT)*. IEEE, 2019, pp. 1–6.

[5] V.-L. Nguyen, R.-H. Hwang, and P.-C. Lin, "Controllable path planning and traffic scheduling for emergency services in the internet of vehicles," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[6] S. Anbalagan, A. K. Bashir, G. Raja, P. Dhanasekaran, G. Vijayaraghavan, U. Tariq, and M. Guizani, "Machine-learning-based efficient and secure rsu placement mechanism for software-defined-iov," *IEEE Internet of Things Journal*, vol. 8, no. 18, pp. 13 950–13 957, 2021.

[7] R. Kumar, P. Kumar, R. Tripathi, G. P. Gupta, and N. Kumar, "P2sf-iov: A privacy-preservation-based secured framework for internet of vehicles," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[8] Q. Mei, H. Xiong, Y. Zhao, and K.-H. Yeh, "Toward blockchain-enabled iov with edge computing: efficient and privacy-preserving vehicular communication and dynamic updating," in *2021 IEEE Conference on Dependable and Secure Computing (DSC)*. IEEE, 2021, pp. 1–8.

[9] M. Irfan, K. Kishore, and V. A. Chhabra, "Smart vehicle management system using internet of vehicles (iov)," in *Proceedings of International Conference on Advanced Computing Applications*. Springer, 2022, pp. 73–83.

[10] J. A. Sanguesa, J. Barrachina, M. Fogue, P. Garrido, F. J. Martinez, J.-C. Cano, C. T. Calafate, and P. Manzoni, "Sensing traffic density combining v2v and v2i wireless communications," *Sensors*, vol. 15, no. 12, pp. 31 794–31 810, 2015.

[11] M. N. Peter and M. P. Rani, "V2v communication and authentication: The internet of things vehicles (iotv)," *Wireless Personal Communications*, vol. 120, no. 1, pp. 231–247, 2021.

[12] Y. Wang, X. Hu, L. Guo, and Z. Yao, "Research on v2i/v2v hybrid multi-hop edge computing offloading algorithm in iov environment," in *2020 IEEE 5th International Conference on Intelligent Transportation Engineering (ICITE)*. IEEE, 2020, pp. 336–340.

[13] K. Liu, X. Xu, M. Chen, B. Liu, L. Wu, and V. C. Lee, "A hierarchical architecture for the future internet of vehicles," *IEEE Communications Magazine*, vol. 57, no. 7, pp. 41–47, 2019.

[14] M. Chen, Y. Tian, G. Fortino, J. Zhang, and I. Humar, "Cognitive internet of vehicles," *Computer Communications*, vol. 120, pp. 58–70, 2018.

[15] L. Alouache, N. Nguyen, M. Aliouat, and R. Chelouah, "Toward a hybrid sdn architecture for v2v communication in iov environment," in *2018 fifth international conference on software defined systems (SDS)*. IEEE, 2018, pp. 93–99.

[16] J. Contreras-Castillo, S. Zeadally, and J. A. Guerrero-Ibañez, "Internet of vehicles: architecture, protocols, and security," *IEEE internet of things Journal*, vol. 5, no. 5, pp. 3701–3709, 2017.

[17] D. P. Proos and N. Carlsson, "Performance comparison of messaging protocols and serialization formats for digital twins in iov," in *2020 IFIP networking conference (networking)*. IEEE, 2020, pp. 10–18.

[18] A. Arooj, M. S. Farooq, A. Akram, R. Iqbal, A. Sharma, and G. Dhiman, "Big data processing and analysis in internet of vehicles: architecture, taxonomy, and open research challenges," *Archives of Computational Methods in Engineering*, pp. 1–37, 2021.

[19] H. Lu, Q. Liu, D. Tian, Y. Li, H. Kim, and S. Serikawa, "The cognitive internet of vehicles for autonomous driving," *IEEE Network*, vol. 33, no. 3, pp. 65–73, 2019.

[20] H.-T. Zhao, X. Zhao, L. Jian-cheng, and L.-y. Xin, "Cellular automata model for urban road traffic flow considering internet of vehicles and emergency vehicles," *Journal of Computational Science*, vol. 47, p. 101221, 2020.

[21] Z. Mahmood, "Connected vehicles in the iov: Concepts, technologies and architectures," in *Connected vehicles in the internet of things*. Springer, 2020, pp. 3–18.

[22] I. Saeed and M. Elhadef, "Performance evaluation of an iov-based intersection traffic control approach," in *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*. IEEE, 2018, pp. 1777–1784.

[23] S. A. Elsagheer Mohamed and K. A. AlShalfan, "Intelligent traffic management system based on the internet of vehicles (iov)," *Journal of advanced transportation*, vol. 2021, 2021.

[24] D. Herich, J. Vaščák, I. Zolotová, and A. Brecko, "Automatic path planning offloading mechanism in edge-enabled environments," *Mathematics*, vol. 9, no. 23, p. 3117, 2021.

Task offloading optimization in mobile edge computing architecture

[1]*Róbert RAUCH (1st year),*
Supervisor: [2]*Juraj GAZDA*

[1,2]Dept. of Computers and Informatics, FEI TU of Košice, Slovak Republic

[1]robert.rauch@tuke.sk, [2]juraj.gazda@tuke.sk

***Abstract*—While computationally intensive applications are ever increasing (such as augmented reality, self-driving cars or cloud gaming), these applications have big impact on battery life. They computation is also delay sensitive and therefore need near instant results. To tackle these demands mobile edge computing can be deployed. In this article we look at overview of what exactly mobile edge computing is and challenges it brings. Especially when it comes to the strategy of task offloading optimization.**

***Keywords*—Edge Computing, Mobile Edge Computing, Task Offloading, Deep Learning**

I. Introduction

With ever emerging compute-intensive mobile applications such as [1], mobile devices such as smart phones require big computational power while keeping power consumption low. European Telecommunications Standards Institute provided a distributed computing paradigm called multi-access edge computing[2] to provide high computational power to the end devices with low latency. Rest of this article is structured as follows: in section II we will take a look at overview of mobile edge computing and what it is. Section III is focused on what kind of task we can consider to be offloaded to the mobile edge computing server. Next section IV aims to show problems with task offloading while providing articles that aim to minimize these issues. In section V we discuss early exiting and split computing, which can be applied to improve task execution time and lower battery consumption. In section VI we will show our prepared data for task offloading optimization. Last chapter VII is conclusion to this article and future work that is planned on our prepared data.

II. Mobile Edge Computing

Mobile Edge Computing[3] (MEC) is a concept similar to that of a Cloud Computing, where User Equipment (UE) is offloading part or entire application to the remote server to be computationally processed. Difference between MEC and Cloud Computing can be seen in fig. 1. Cloud Computing is processing applications from a large number of mobile devices that are located far from UE. MEC servers are on the other hand located close to the UE. This improves connection to the server and lowers latency.

One of the use cases for offloading applications to the remote server is Cloud Gaming[5]. Basic principle of Cloud Gaming is to offload computation of rendering video games to the cloud. This means players don't need to have powerful

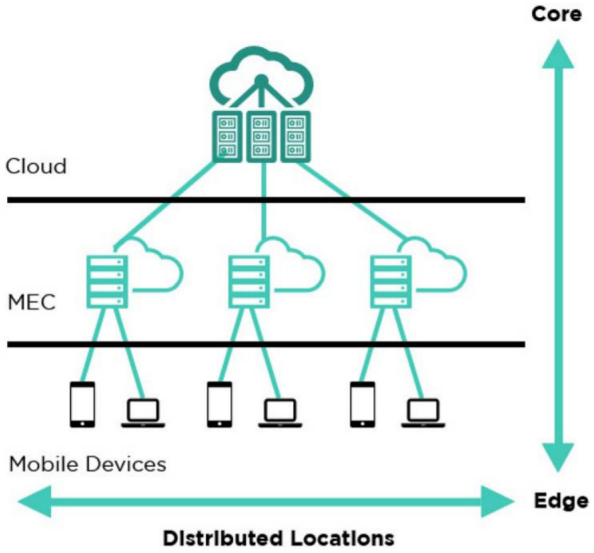


Fig. 1: Network architecture with remote cloud and MEC servers [4]

PC, console or other device for playing video games. All they need is a stable, fast internet connection and application for connecting to the remote server. With this application users can send basic inputs from their devices (i.e. keyboard and mouse on PC) to the remote server where game is processed and rendered. After rendering game audio and video feed is streamed back to the user as a video. Biggest problem with Cloud Gaming is latency problem, which means delay between server receiving users input, processing and rendering game and then streaming video of rendered game back to the user. This latency could be lowered if computational power needed for rendering game would be placed at the MEC server closer to the UE.

III. Task Definition

When UE creates connection to the MEC server new Virtual Machine (VM) or container is created. New instance of user application starts on this newly created VM or container and when UE decides to offload part of the application it is send to this VM or container to be processed. This relatively small part of the application is called task. For our needs we will think of tasks as a computer vision problem. In next subsections we will explore some possibilities for these tasks.

A. Image Classification

Image classification is a computer vision task in which high level semantic information is extracted from image features[6]. This extraction is usually done with Convolutional Neural Networks (CNN). Extracted semantic information is then run through Feed Forward Network (FFN), which gives image label which determines the category of the image.

High attention to image classification can be also seen if we look at a dataset such as CIFAR-10[1] which include 60000 small 32x32 images that are categorized into 10 categories. As seen in fig. 2 popularity of using this dataset is increasing each year.

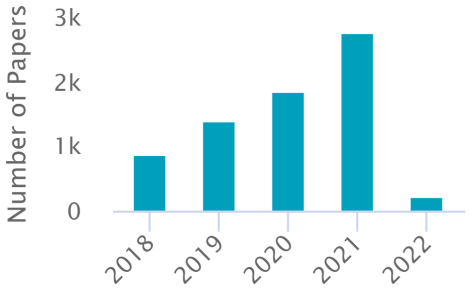


Fig. 2: Popularity of CIFAR-10 by PapersWithCode[2]

B. Image Segmentation

Image segmentation is a computer vision problem in which we are trying to identify and locate objects within an image. This is done by partitioning image into smaller segments which are determined to be similar based on their texture, color intensity etc. [7] There are several types of image segmentation. In fig. 3 we can see semantic segmentation. This segmentation doesn't recognize if two objects are different instances only that they are part of the same category.

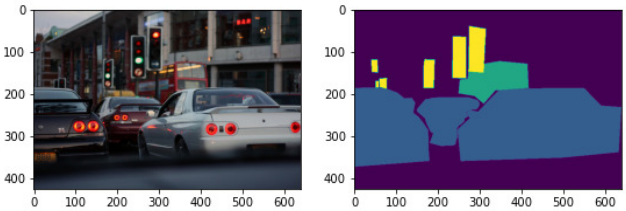


Fig. 3: Semantic segmentation on COCO[3]dataset

Instance segmentation[8] is a computer vision task which locates and identifies objects within picture just as semantic segmentation, however besides just identifying category of the image it also identifies different instances of the same category. In our example we would therefore identify all 3 cars as separate instances.

Applications for image segmentation are vast, for example crowd counting, terrain detection (identification of roads, hills etc.), face recognition and more.

[1]https://www.cs.toronto.edu/ kriz/cifar.html
[2]https://paperswithcode.com/dataset/cifar-10
[3]https://cocodataset.org/

C. Object Detection

Object detection[9] is another computer vision problem where instead of marking individual pixels of a object we create boundary box around it. Same as image segmentation with this we can identify and locate objects within a image. Fig. 4 shows object detection on single image with one subject. After identifying category and boundary box location we can draw this boundary box with label.

Process of object detection is as follows:

1) input image is ran through CNN from which feature maps are extracted
2) this feature maps are then ran through Region Proposal Network which determine which part of a image could be objects
3) last step is then to determine if proposed part of the image is object. This is done using simple image classification method which determine label for that part of the image (or denies input meaning that part of the image doesn't contain any object classifier can recognize)

Object detection has broad applications such as crowd counting or healthcare similar to those in image segmentation. Another use case are self driving cars as they need to properly recognize and identify objects around them to move through the world safely.

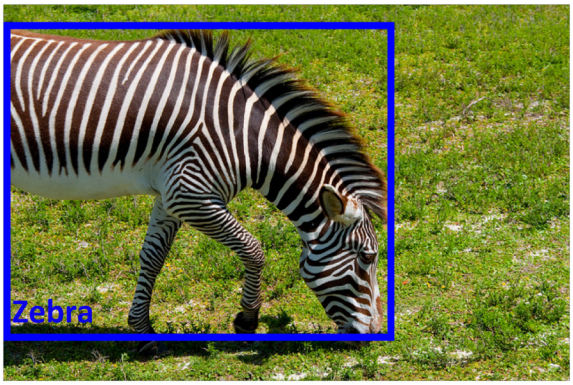


Fig. 4: Object detection on COCO dataset

IV. Task Offloading

Each time application has resource-heavy task it needs to process, application has to make near instant decision on where it wants to process this task. It can process this task in UE which will drain more battery from UE (if it's battery powered) and could take significantly longer to process compared to processing same task on the server. Another option is to therefore offload this task to a server. First it can send it to the decentralized MEC server that is near UE or send it through backhaul to the cloud. We can assume that MEC server has less latency with a UE but less computational power while remote cloud has higher computational power but also higher latency. In next subsections we will discuss couple of methods for offloading strategies.

A. Heuristic Methods

These methods implement more traditional optimization approaches to determine if and where to offload tasks. While these methods are efficient, they are high in complexity. For example, Chen et. al. [10] proves that it is NP-hard to compute

a centralized optimal solution. Therefore they adopt a game theoretic approach to efficiently offload tasks in distributed manner. Efficiency is proven because game theory they adopt always achieves Nash equilibrium.

B. Deep Learning Methods

These methods implement traditional deep learning algorithms to train model to make optimal offloading decisions. Article [11] considers autonomous driving cars as a use case for MEC. Where resource constrained vehicles can leverage from MEC architecture and send task intensive tasks to MEC servers. They implement deep Q-learning algorithm to determine where (edge server or remote cloud) and which exact MEC server to offload to. This means they also consider mobility of self-driving cars. As proven by their results deep Q-learning approach outperformed that of a game theoretic approach considered in heuristic methods. These results can be seen in fig. 5.

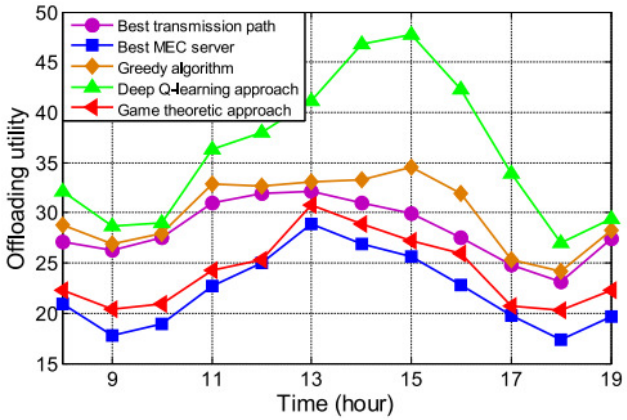


Fig. 5: Offloading utility for taxis between 8 in the morning and 19 evening [11]

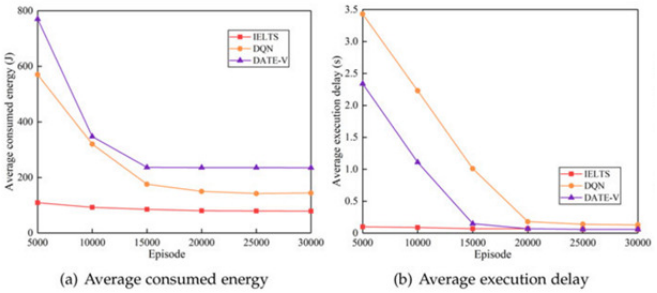
The biggest problem with deep learning methods as stated in [12] is that they have low sample efficiency and need full retraining to learn update policies for new environments. This article solves this problem by implementing meta reinforcement learning method which can adopt to new environment with less gradient steps than other reinforcement learning methods. They achieve this by implementing algorithm they call Meta Reinforcement Learning based Computation Offloading (MRLCO). This algorithm uses Proximal Policy Optimization (PPO) objective function.

Besides deep Q-learning and PPO there are a lot of state-of-the-art deep learning algorithms that were applied in mobile edge computing task offloading problem. These algorithms include SARSA [13], DDPG [14], MADDPG and SAC [15] etc.

C. Imitation Learning Methods

Imitation learning is area of deep learning where we collect experts demonstration on how we want our agent to behave and consider these demonstrations as optimal solution. After collection of demonstration, we can train deep neural network to map states s to a distribution over actions a, to mimic an expert policy[16].

We can therefore assume that learning from expert demonstration yields faster convergence than deep learning where agent is trained by trial and error. As proven by [17], using imitation learning for training of neural network to make optimal task offloading decision speeds up training process. This is particularly shown in fig. 6, where we can see that compared to deep Q-learning (DQN), Imitation Learning enabled Task Scheduling (IELTS) achieved near optimal task offloading decision right at the beginning of the training.



(a) Average consumed energy (b) Average execution delay

Fig. 6: Energy consumption and execution delay during training of IELTS, DQN and DATE-V. [17]

V. Split Computing and Early Exiting

If we think of task as deep learning problem such as image classification, image segmentation or object detection we can implement two techniques to further lower battery consumption and improve task execution time. These techniques are early exiting and split computing[18]. Early exiting is a technique where instead of going through the entire deep neural network we have an option to exit early. This exiting in earlier layers has impact on accuracy of the outcome but is executed faster. Lastly, we can fine tune our decision making to create trade-offs depending on current network status, computing power, current battery and other factors. Split computing is a technique in which deep neural network is split into head and tail networks. Head network is then processed in UE, while tail network is executed on the edge server. The goal of split computing is to distribute computing load between UE and edge server. Second goal is a task-oriented compression to reduce data transfer delays.

VI. Dataset Creation

For our experiments we will be using image classification as tasks to be considered for offloading to MEC or remote cloud servers. We will use VGG16[19] architecture, which was developed in 2014 to be used for ImageNet dataset. It achieved 92.7% accuracy which was top 5 accuracy in the world. VGG16 is CNN architecture, where number 16 implies it consists of 16 layers. Another version of this architecture is VGG19 which uses 19 layers, however we will be using this architecture on CIFAR-10 dataset where 16 layers is sufficient. CIFAR-10 consists of 60000 32x32 images which are subset of Tiny Images dataset. Each image is labeled with one of the 10 categories.

While we are using VGG16 there are 2 modifications we did to the architecture. First is usage of 32x32 images. VGG16 is expecting image size of 224x224 and since CIFAR-10 is using images of size 32x32 we had to either change resolution of the images by upscaling them to 224x224, however this would only increase training time and wouldn't add any new information to the image. Therefore, we changed input images

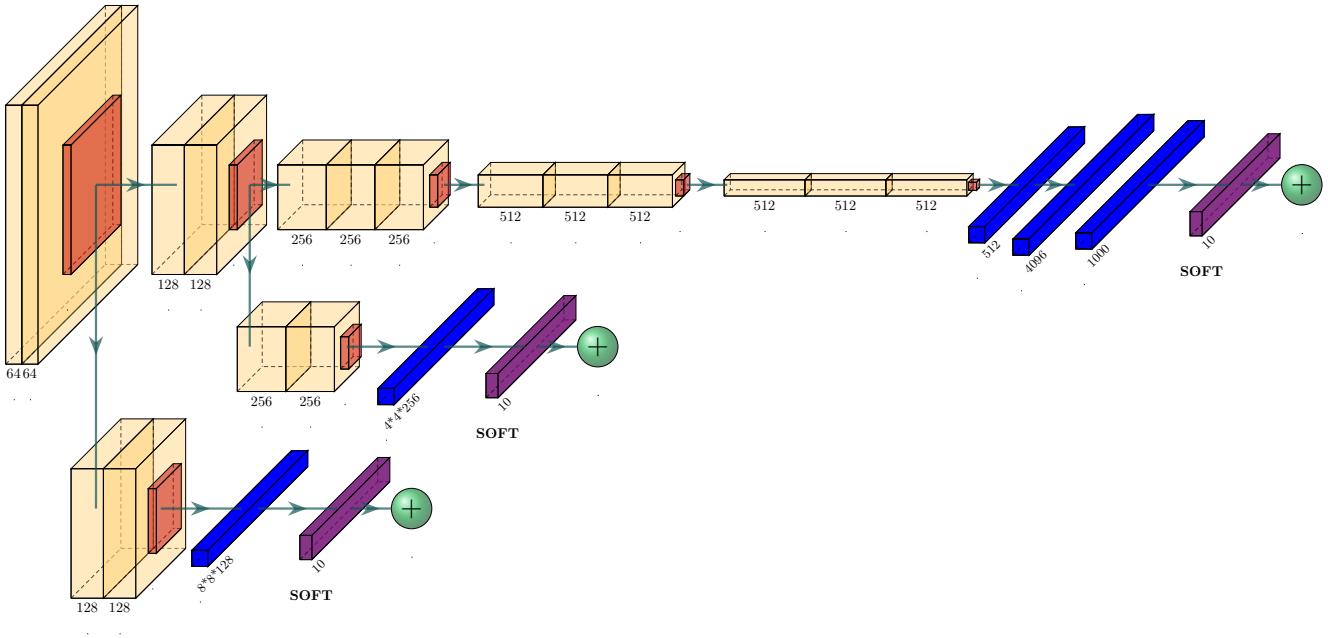


Fig. 7: VGG16 architecture with early exits and 32x32 input images

to 32x32 which is not a problem since VGG16 is using "same size" convolutions which means they don't change resolution of the images only number of output channels. This was achieved by using convolution with kernel size 3, stride 1 and padding 1. Second modification is adding early exits to the architecture. We added 2 early exits. First early exit is after second convolutional layer and first max pool. Second early exit is after fourth convolutional layer and second max pool. Both early exits than have 2 extra convolutional layers, one max pool, one fully connected layer and softmax function. We can see entire version of this VGG16 architecture in fig. 7. Each yellow square shows convolutional layer while number at the bottom indicates output channels of convolutional layer. For example, first convolutional layer takes input image which are resolution 32x32x3, where 3 indicates that it is RGB image and therefore one for each channel. Output of first convolution is then 32x32x64. Each orange square shows max pool function with kernel size 2, which halves resolution of the image. Blue squares show fully connected layers, while number under fully connected layer show how many neurons this fully connected layer has. Purple square represents softmax function. Circle with plus sign represents output of softmax function, which is number between 0 and 9.

As expected, we achieved highest accuracy if image goes through entire VGG16 architecture (no early exit is used). Each early exit then lowers accuracy of the model. If image goes through main network (entire VGG16 architecture without early exits) accuracy is 84.4%. First early exit (that is first from input layer) achieved accuracy of 79.6%. Second early exit achieved accuracy of 82.8%.

To make optimal offloading decisions we also need to know how computationally heavy task will be on the UE/MEC server. To measure how much resources, we need for computing task we will use instruction count. To our knowledge there is no way of measuring instruction count within programming language Python and therefore we extracted layout of our model to .pt file and used C++ PyTorch library to execute our model within C++. This allowed us to use framework valgrind[20] and tool within valgrind called callgrind. With callgrind we can execute specific application and measure instruction count this application needed to complete. We can alternatively also stop instruction count at the beginning and within application tell callgrind where it should start/stop counting instructions. For main network we measured that forward function (function that takes input image and returns its category) within PyTorch library needed approximately 270 million instructions to finish, while first early exit needed 204 million instructions and second early exit needed 214 million instructions. We also calculated instruction count if we would split our network on head (first part which would execute on UE) and tail (second part which would execute on MEC server). We added 2 splits. First split is right before first early exit and after first max pool and second split is right before second early exit and second max pool. Number of instructions for each split can be seen in table I.

TABLE I: Instruction count for network with split computing

network	Split 1		Split 2	
	Head	Tail	Head	Tail
Main	187	246	200	234
Early Exit 1	187	187	x	x
Early Exit 2	187	200	200	187

Each number is average of 10 runs.
Instructions are in millions.

Lastly, we might consider image compression since we will be sending images over network, and we want to minimize time that is spent communicating with MEC server. This is of course not considering split computing since in case of split computing we would be sending data generated by network. This data collection was done on CIFAR-10 images. In table II we can see each quality level (lower quality level means higher compression), accuracy for each network (main, first early exit and second early exit) and average byte size of images.

TABLE II: Accuracy with image compression and byte size of images after compression

Quality Level	Main	Early Exit 1	Early Exit 2	Byte size
100	84.4	79.6	82.8	1382.42
95	83.8	79	82.5	961.3
90	83.3	77.8	81.2	786.26
85	82.6	77	80.7	697.69
80	81.8	76	79.7	643.70
75	81	74.9	78.8	603.86
70	80.4	74.3	77.5	577.33

VII. Conclusion & Future Work

In this article we explained what mobile edge computing concept is, what kind of tasks we can consider to be offloaded to the MEC server, possible methods of optimizing task offloading process, early exits and split computing concepts. Lastly, we looked at process of our experiments and data creation which will be used in our future work to optimize task offloading process. For this data we considered both early exit and split computing concepts, while also considering image compression for faster communication with MEC server.

Now that we acquired all necessary data that we want to consider while optimizing task offloading, we will use either deep reinforcement learning or imitation learning method in our future work to optimize task offloading. We will let our training model decide whether to offload, which layers it will go through, if it should split computation load and if it should compress image. We will also be considering mobility of the users and predicting their movement to know which server is best to offload task to.

Acknowledgement

This work was supported by The Slovak Research and Development Agency project no. APVV-18-0214.

References

[1] T. Olsson and M. Salo, "Online user survey on current mobile augmented reality applications," in *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, 2011, pp. 75–84.

[2] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing—a key technology towards 5g," *ETSI white paper*, vol. 11, no. 11, pp. 1–16, 2015.

[3] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1628–1656, 2017.

[4] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, "Mobile edge computing: A survey," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 450–465, 2017.

[5] R. Shea, J. Liu, E. C.-H. Ngai, and Y. Cui, "Cloud gaming: architecture and performance," *IEEE network*, vol. 27, no. 4, pp. 16–21, 2013.

[6] J. Wu, V. S. Sheng, J. Zhang, H. Li, T. Dadakova, C. L. Swisher, Z. Cui, and P. Zhao, "Multi-label active learning algorithms for image classification: Overview and future promise," *ACM Computing Surveys (CSUR)*, vol. 53, no. 2, pp. 1–35, 2020.

[7] R. Dass and S. Devi, "Image segmentation techniques 1," 2012.

[8] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.

[9] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," *Advances in neural information processing systems*, vol. 26, 2013.

[10] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Transactions on Networking*, vol. 24, no. 5, pp. 2795–2808, 2016.

[11] K. Zhang, Y. Zhu, S. Leng, Y. He, S. Maharjan, and Y. Zhang, "Deep learning empowered task offloading for mobile edge computing in urban informatics," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 7635–7647, 2019.

[12] J. Wang, J. Hu, G. Min, A. Y. Zomaya, and N. Georgalas, "Fast adaptive task offloading in edge computing based on meta reinforcement learning," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 1, pp. 242–253, 2021.

[13] T. Alfakih, M. M. Hassan, A. Gumaei, C. Savaglio, and G. Fortino, "Task offloading and resource allocation for mobile edge computing by deep reinforcement learning based on sarsa," *IEEE Access*, vol. 8, pp. 54 074–54 084, 2020.

[14] H. Zhang, Y. Yang, X. Huang, C. Fang, and P. Zhang, "Ultra-low latency multi-task offloading in mobile edge computing," *IEEE Access*, vol. 9, pp. 32 569–32 581, 2021.

[15] H. Lu, C. Gu, F. Luo, W. Ding, S. Zheng, and Y. Shen, "Optimization of task offloading strategy for mobile edge computing based on multi-agent deep reinforcement learning," *IEEE Access*, vol. 8, pp. 202 573–202 584, 2020.

[16] K. Arulkumaran and D. O. Lillrank, "A pragmatic look at deep imitation learning," *arXiv preprint arXiv:2108.01867*, 2021.

[17] X. Wang, Z. Ning, S. Guo, and L. Wang, "Imitation learning enabled task scheduling for online vehicular edge computing," *IEEE Transactions on Mobile Computing*, vol. 21, no. 2, pp. 598–611, 2022.

[18] Y. Matsubara, M. Levorato, and F. Restuccia, "Split computing and early exiting for deep learning applications: Survey and research challenges," *arXiv preprint arXiv:2103.04505*, 2021.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[20] N. Nethercote and J. Seward, "Valgrind: A program supervision framework," *Electronic notes in theoretical computer science*, vol. 89, no. 2, pp. 44–66, 2003.

Innovative approach design of interaction with virtual reality systems proposal

[1]*Miriama MATTOVÁ (1st year)*
Supervisor: [2]Branislav SOBOTA

[1,2]Dept. of Computer and Informatics, FEI TU of Košice, Slovak Republic

[1]miriama.mattova@tuke.sk, [2]branislav.sobota@tuke.sk

***Abstract*—The work proposes a system to a innovative approach design of using and interacting with virtual reality. Proposed system is for training persona model via biometric and electroencephalographic data with a help of virtual reality technologies. Purposed system will test a capability to stimulate emotions due to the trained model and lastly the real-time data collected from biometric and electroencephalographic measurement will interact with virtual environment on a principle of trained model.**

***Keywords*— Biometric data, Electroencephalography, Interaction, Stimulation, Virtual reality**

I. INTRODUCTION

It begins to be common knowledge about how technology is advancing fast. It is possible to observe almost on daily basis new improved version of technology then it was day before. Virtual reality (VR) still belongs into technology that is new. As we are living in pandemic era, VR and related technologies are penetrating almost every aspect of our lives [1]. VR was proven to have high level of interactivity and is one of the most developing technologies [2]. It can help disabled people [3] as well as educate [4] or help in medicine field [5].

This paper is focused on how to improve interaction with VR technologies and how these technologies affect a person. More specifically, how biometric and electroencephalography (EEG) data affects VR and vice versa how the VR can affect these data.

It is possible to consider biometric data such as pressure, heartbeat, sugar level, muscle impulses, brain impulses, where a person is looking, skin conductance and so on. These data can be measured for example via pressure gauge, heart rate monitor devices, wearable glucometer, Myo armband, EEG devices, eye trackers and devices for measuring electrodermal activity. Paper proposes a system to train a model of a persona for biometric data, how to stimulate these data and lastly how to consciously control VR environment via these data. EEG data will be used as an example, where such system will be explained.

II. RELATED WORK

This section is dedicated to related works towards the proposed system in this article. Here is a list of the works which partially deal with some sections of proposed system.

- *A low-cost, open-source, BCI-VR prototype for real-time signal processing of EEG to manipulate 3D VR objects as a form of neurofeedback* [6] – This work presents low-cost and open-source brain-computer interface virtual reality prototype for real-time signal processing of EEG event related desynchronization and synchronization changes within the Precentral Gyrus which allows the user to manipulate a 3D object within a VR environment.
- *Using EEG to Decode Subjective Levels of Emotional Arousal During an Immersive VR Roller Coaster Ride [7]* -This work tested the association between continuously changing states of emotional arousal and oscillatory power in the brain during a VR roller coaster experience. It used novel spatial filtering approaches to predict self-reported emotional arousal from the electroencephalogram (EEG) signal of 38 participants. Work demonstrates a new approach to decode states of subjective emotional arousal from continuous EEG data in an immersive VR experience.
- *Emotion Recognition Using Frontal EEG in VR Affective Scenes [8]*- This paper provides a new framework for emotion recognition using frontal EEG and VR affective scenes. The mean accuracy of framework achieved about 81.30%, which exhibited better performance compared with relevant studies. The framework proposed in this work can be well applied to wearable device for EEG emotion recognition in VR scenes.
- *FACETEQ interface demo for emotion expression in VR [9]* – Faceteq prototype v.05 is a wearable technology for measuring facial expressions and biometric responses for experimental studies in Virtual Reality. Emteq founded the Faceteq project with the aim to provide a human-centered additional tool for emotion expression, affective human-computer interaction and social virtual environments.

III. DESIRED DATA

With today's technologies it is possible to collect biometric data such as face silhouette, fingerprints or voice pattern. All these data can be considered static as they will not change over time or by conscious or nonconscious trigger. And therefore, for this case it is required to focus on data such as pulse, pressure and skin conductivity. There is also EEG – Electroencephalography method to collect activity on the scalp that represent the macroscopic activity of the surface layer of the brain underneath. It can also record the blink of an

eye or where the person is looking, sledge clamp or facial muscles in general, but it all depends on how advanced the device is, used for data collection. In order to create the most accurate system, it is necessary to collect as much data as possible. But this data should be changeable and not static, such as fingerprint.

On the other hand, it is needed to think about limitations with such data collection devices. For example, if it is desired to collect data via EEG method, virtual reality headsets should not be used as an output device, where simulation scene would be displayed. It is because EEG system must be placed on the head of the subject same as virtual reality headset. EEG device has very sensitive link with a scalp and VR headset's parts, that keeps it on a person's head can break the link. This will lead to data loss and inaccurate measurement after all and in worst scenario, it will not measure data at all. Therefore, designed scenes, which will trigger some state of a person, needs to be presented on a such output device, that will not be attached on a person in any or minimal way.

IV. Display devices

Primary display device will be LIRKIS-CAVE shown in a figure 1. System consists of 20 LCD displays, 7 cameras on top, that helps determine the user position, and control unit. This system can provide full VR experience without a need to attach anything on a head of a person except the 3D glasses that are very similar to normal eyeglasses. Another advantage over the virtual helmet is, that this system offers full peripheral vision. VR headsets have peripheral vision limited which force users to move whole head if they want to look slightly around.

Figure 1. LIRKIS-CAVE system at the Technical University of Košice

As it was outlined in introduction, VR systems do not only have high level of interactivity, but such devices are also affecting users more then for example a 2D displays such as monitors on a desktop or mobile device. As it is desired to trigger in a person the most intense data output from biometric and EEG measurement, this VR system can be considered as promising.

V. Data collection – persona model training

To collect and train data, we need to redistribute them. To do so, it is possible to divide them into emotions. But it is not necessary to categorize them as such. It could be categorized also into indicators such as stress, adrenalin, excitement and so on. And therefore, this part should be consulted with experts in a field to give these data a meaning. But let's say, emotion categorization will do just fine.

Training scene will consist of several scenes that should trigger certain emotion in a person. That way it is possible to determine, that those specific data belong to certain emotion. Each scenario needs to be consulted with an expert in a field of human emotions. Here is a list of assumed scenarios designed with experts:

Scenarios A – Several scenarios that usually trigger neutral emotion (e.g., breathing exercise, meditating)

Scenarios B - Several scenarios that usually trigger surprise emotion (e.g., interactive game with unexpected behaviour output)

Scenarios C - Several scenarios that usually trigger joy emotion (e.g., funny videos, puppies, good deeds videos).

Scenarios D - Several scenarios that usually trigger sadness emotion (e.g., sad videos)

Scenarios E - Several scenarios that usually trigger fear emotion (e.g., displaying most common phobias, interactive horror game)

Scenarios F - Several scenarios that usually trigger anger emotion (e.g., videos representing violence, interactive game where it is impossible/almost impossible to finish the task, bad traffic game)

Scenarios G - Several scenarios that usually trigger disgust emotion (e.g., videos representing not so pleasant situations).

Table I. shows assumed emotions to be measured together with scenario type, which represent if it should be just something to watch on, or if it should be interactive like games for example.

TABLE I.
TRAINING SCENARIO DESIGN

State / Emotion	Scenario type	Scenario
Neutral/Calm	Static	A
Surprise	Interactive	B
Joy	Static	C
Sadness	Static	D
Fear	Static/Interactive	E
Anger	Static/Interactive	F
Disgust	Static	G

According to table I it is assumed that every emotion will have its own type and scenario. These scenarios should be so called colourful, as people react on situations differently. There should be an output of data for every scenario. These data will be averaged and considered as ideal data state of a

person for specific scenario. Output of a completed training session of a person should be a model of person's biometric and EEG data for certain situations. One person should have a possibility to train several models that belongs to his/her usual data output from measuring devices. Explanation for this purpose is described more in the chapter VII.

VI. Data Interaction – Stimulator

This part is focusing on how to interact with a real time biometric and EEG data of a person. Let's say, there is a person with a certain state and giving a certain type of data. Due to the model from training session, it is possible to approximately determine what state is person in. This person is sitting in a LIRKIS-CAVE. First scene that LIRKIS-CAVE will display will be due to his current state. For example, if data are showing that person is in a state of neutral/calm state, system will show the scenario to trigger some other state. If the data changes, system will capture these changes and compare it to a known state. If state/emotion is recognised, scene will be transformed to stimulate current state into other state due to the model of the person. Following diagram in Figure 2. presents system flow of described situation.

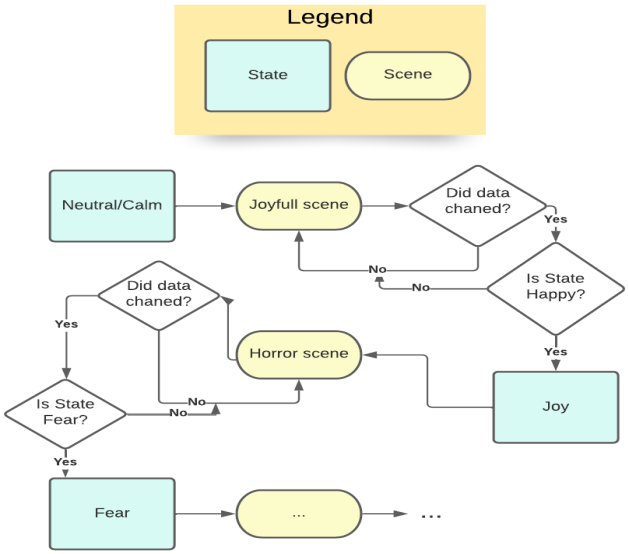


Figure 2. System flow diagram for emotion stimulation.

It is possible, to leave desired emotion stimulus up to a person using the system. For example, if user wants to stimulate his/her emotion into sadness, he/she can choose it via interface more described in chapter VIII.

VII. Data Interaction – Conscious Control

As the process of training persona model was proposed in the chapter V., it is possible to say that model can be trained by user's likings. Let's say, that user wants to call a state A - *focusing* and he/she wants to attach this state to an action - *go forward.* By this defined state and action, he/she should have a possibility to create custom model, so the conscious control interaction will go in a smoothest way possible for specific user, as many people send various data output. Following table II. shows purposed state-action assignment.

TABLE II.
State Triggering Action

State	Action
focusing	Go forward
…	…

It is possible to observe from a table II., that it is not necessary to use emotions to control the environment. Such control system does not need to be effective after all. Of course, emotion assignment can be set to a predefine action by a default, but when a user will decide to choose another approach to controlling environment, he/she can do so. Simply said, if user have trained the state A called e.g., *focusing* to an action e.g., *go forward*, avatar in virtual environment should go forward every time user give output data from measuring devices that define his/her being in a state of focusing. Following figure 3. presents a diagram of system flow for triggering action by a state.

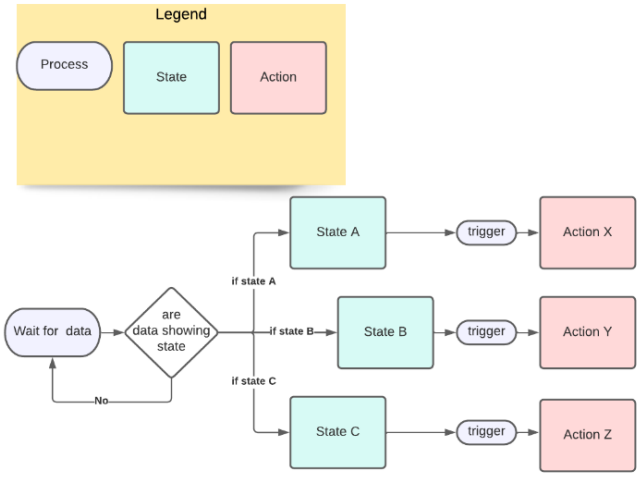


Figure 3. System flow diagram for triggering action by a state.

As it was mentioned above, system will wait for on output data from measuring devices (biometric and EEG), then it will check if data corresponded for any trained state and if state is recognized, it will trigger an action attached to a state.

VIII. System Architecture Design

System will be placed in LIRKIS-CAVE mentioned in chapter IV. Displaying as well as computing will be processed there. Once the proposed system will be launched, there will be an application executable on mobile devices. This application will connect to a system and user will log via his/her credentials. After successful authentication and authorization, if user have trained models already stored in mobile device, these models will be loaded in the system. Before he/she start any action, Analytics interface will pop up to give a message about the measuring devices. This interface should give a user an information, if measuring devices are working properly. If measuring devices are placed on user correctly, user may continue to use the system. Interface to control system will have an option to choose *Modelator*, *Stimulator* or *Interactor*.

Modelator – is a subsystem that will manage the model creation on principle mention in a chapter V. User should have a possibility to create emotion-based persona model via user interface on mobile device, where every emotion has predefined scenes to trigger such emotion. But also, he/she can create custom model of his/her states and assign these states to the predefined actions. In this option user should have possibility to display and analyze existing models created by him/her.

Stimulator – is a subsystem that will contain static scenes to stimulate states of a user. This subsystem is proposed in the chapter VI. User can choose if he/she wants to stimulate emotions due to the emotion-based persona model or practice his/her trained states on an action. The possibility to choose between these two scenes should be via user interface on a mobile device.

Interactor – is a subsystem proposed in chapter VII. This subsystem will contain interactive scenes where user can control them via measured data due to the trained model. One set of scenes will be designed to control specifically by emotion trigger and other set of scenes will be designed to control by custom state trigger.

Following figure 4. represents such architecture.

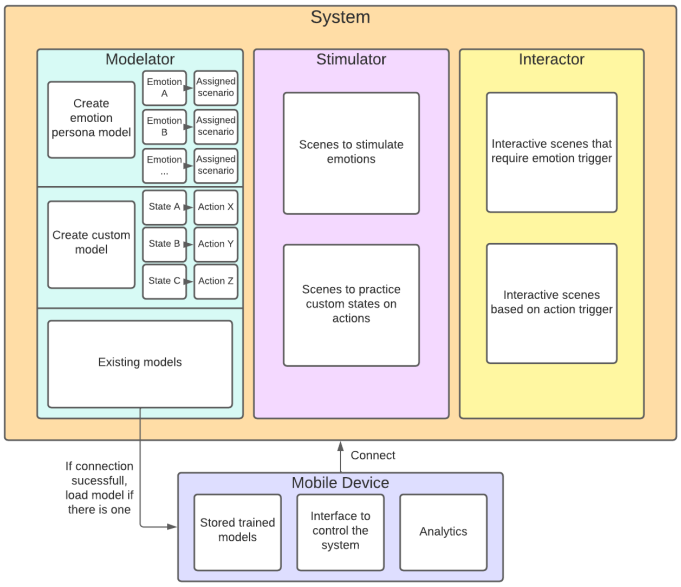


Figure 4. System flow diagram for triggering action by a state.

As it is possible to observe from figure 4., mobile device will behave as controller for a system. Mobile devices have a great ability to store digital data unlike joysticks or other controllers. Via mobile device it is possible to design user interface into any likings. Cable independence is also needed, when user have all measuring devices connected, he/she will have limited moving options as was mentioned in a chapter III.

IX. Conclusion

This work proposed a different approach of how to interact with virtual reality environment. As related works proves, it is possible to create interaction via EEG data and implement emotion recognition via EEG within VR. This work will focus on joining those works and creating a system capable of training a persona model of a specific person, possibility of virtual environment to interact with real-time EEG and biometric data and vice versa to interact via biometric and EEG data with virtual environment.

Acknowledgment

This work has been supported by the KEGA grant no. 048TUKE-4/2022:"Collaborative virtual reality technologies in the educational process"

References

[1] Sobota, B.; Korečko, Š.; Hrozek, F.: On building an object-oriented parallel virtual reality system; In: Central European Journal of Computer Science. Vol. 2, no. 3 (2012), p. 261-271. - ISSN 2081-9935

[2] Hudák, M.;Korečko, Š.; Sobota, B. : Enhancing Team Interaction and Cross-platform Access in Web-based Collaborative Virtual Environments, In: Proceedings of 2019 IEEE 15th International Scientific Conference on Informatics, IEEE, 2019, pp. 160-16.

[3] J. M. L. de Ipina et al, "Virtual reality: A Tool for the Disabled People Labour Integration," in Proc. of Challenges for Assistive Technology, IOS Press, 2007, pp. 141-145.

[4] BOGUŠČIAK, Jakub. Wheelchair simulator in web virtual reality: diploma thesis. Košice: TU, 2021 62 s. ISBN

[5] H. İ. ÖZDEMİR, B. ÇAKMAK, Ş. YOL, H. E. ÖZDEMİR and N. C. ÖZDEMİR, "Virtual Reality-Based Medical Device Education," 2020 Medical Technologies Congress (TIPTEKNO), 2020, pp. 1-4, doi: 10.1109/TIPTEKNO50054.2020.9299234.

[6] M. McMahon and M. Schukat, "A low-cost, open-source, BCI-VR prototype for real-time signal processing of EEG to manipulate 3D VR objects as a form of neurofeedback," 2018 29th Irish Signals and Systems Conference (ISSC), 2018, pp. 1-6, doi: 10.1109/ISSC.2018.8585373

[7] F. Klotzsche, A. Mariola, S. Hofmann, V. V. Nikulin, A. Villringer and M. Gaebler, "Using EEG to Decode Subjective Levels of Emotional Arousal During an Immersive VR Roller Coaster Ride," 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), 2018, pp. 605-606, doi: 10.1109/VR.2018.8446275.

[8] T. Xu, R. Yin, L. Shu and X. Xu, "Emotion Recognition Using Frontal EEG in VR Affective Scenes," 2019 IEEE MTT-S International Microwave Biomedical Conference (IMBioC), 2019, pp. 1-4, doi: 10.1109/IMBIOC.2019.8777843

[9] Mavridou et al., "FACETEQ interface demo for emotion expression in VR," 2017 IEEE Virtual Reality (VR), 2017, pp. 441-442, doi: 10.1109/VR.2017.7892369.

Liver tumor segmentation using 3D convolutional heuristic u-net

[1]*Matej GAZDA (3rd year),*
Supervisor: [2]*Ján PLAVKA*

[1,2]Dept. of Mathematics and Theoretical Informatics, FEI TU of Košice, Slovak Republic

[1]matej.gazda@tuke.sk, [2]jan.plavka@tuke.sk

***Abstract*—Liver cancer is statistically one of the most dangerous type of cancer. Automatic segmentation of a liver tumor is often an effective tool during radiotherapy treatment. We evaluate a model for liver and liver lesions segmentation in computed tomography volumes. The model consists of an encoder and a decoder trained in end-to-end approach. The network is evaluated on a dataset used in Liver Tumor Segmentation Challenge hosted by MICCAI conference in 2017. Our method achieved good performance for both segmentation of the liver and the liver tumor. In the end, we outline future work on this topic and promising research directions.**

***Keywords*—encoder-decoder networks, medical imaging, semantic segmentation, tumor segmentation**

I. Introduction

The liver, with its detoxification function, is an essential organ in the human body. From a medical point of view, liver cancer is responsible for the second most deaths of all the cancers and is the 6th most common one [1]. Radiologists use Computer Tomography (CT) and Magnetic Resonance (MRI) to examine the shape and texture of liver anomalies.

With the advances in the artificial intelligence (AI) domain, segmentations methods based on artificial intelligence are commonly used to find the contours of the lesions. Contours of the lesions are a key attribute or even prerequisite to treatments such as percutaneous ethanol injection, radiotherapy surgical resection, or arterial embolization [2].

Methods processing 3D volumes instead of 2D slices are nowadays considered state of the art, but on the other hand, they require vast computing resources.

Generative Adversarial Networks (GANs) have been explored to analyze medical images. Some of them improve segmentation with an additional adversarial objective [3], while others perform data augmentation within the training set [4]. Some works propose a weakly supervised learning relying on image-to-image translation between diseased and healthy cases [5].

Other works use u-net like architecture to perform biomedical image segmentation [6]. Original U-Net consisted of two convolutional neural networks operating across multiple resolutions of the image. U-Net gained its name based on its U shape, consisting of an encoder and a decoder with mirrored layers. The encoder first encodes the image into latent space, and then the decoder reconstructs the image based on the latent representation. Networks based on the U-Net architecture are winning at the moment segmentation-based challenges such as Kidney Tumor Segmentation Challenge [7] (KITS), Liver Tumor Segmentation Challenge [2] (LITS) and others.

Isensee et al. [8] created a framework based on the U-Net network with additional data augmentation specifically crafted for various modalities and heuristics for the parameters of the network.

II. Data

LITS, composed by Bilic et al. [2], is a dataset containing portal venous phase contrast-enhanced abdominal computer tomography (CT) scans. Liver and liver tumor boundaries are provided with pixel-level annotation. Dataset is composed out of 200 samples, 130 used for training and 70 as a hold-out set for model evaluation. Each CT belongs to a different patient.

III. Method

We utilize a U-Net like architecture based on nn-UNet to find pixel-level boundaries of tumors located in the liver.

The encoder operates on five resolutions of the image. On each resolution, two convolutional layers are used. A normalization and ReLU activation function follow each convolutional layer. Since the decoder mirrors the architecture of the encoder, similarly, it operates on five resolutions of the image. Upscaling is done through transpose convolution.

The features extracted as an output of every encoder's layer are concatenated with its respective decoder level. Note that these skip connections are called long since they connect the layers of the opposite sides of the U-shaped network. Advantages of the so called short skip connections are described in [9], however we leave that as an improvement in future work.

To reduce the potential overfit, training set was artificially inflated by data augmentation techniques. Image was randomly augmented by one or more of following transformations: (a) Rotation of any axis up to $(-\pi, \pi)$ with probability $p_r = 0.3$, (b) elastic deformation with probability of $p_e = 0.2$, (c) scaling from interval $(0.65, 1.6)$ with probability $p_s = 0.3$ and (d) zooming with zoom range $(0.5, 1)$.

Model was trained for 200 epochs with SGD with Nesterov momentum ($\gamma = 0.99$).

As a loss function, similarly to nn-UNet, we choose a combination of the crossentropy loss with dice score.

Dice score, given binary masks A and B, can be evaluated as:

$$D(A, B) = \frac{2|A \cap B|}{|A| + |B|}$$

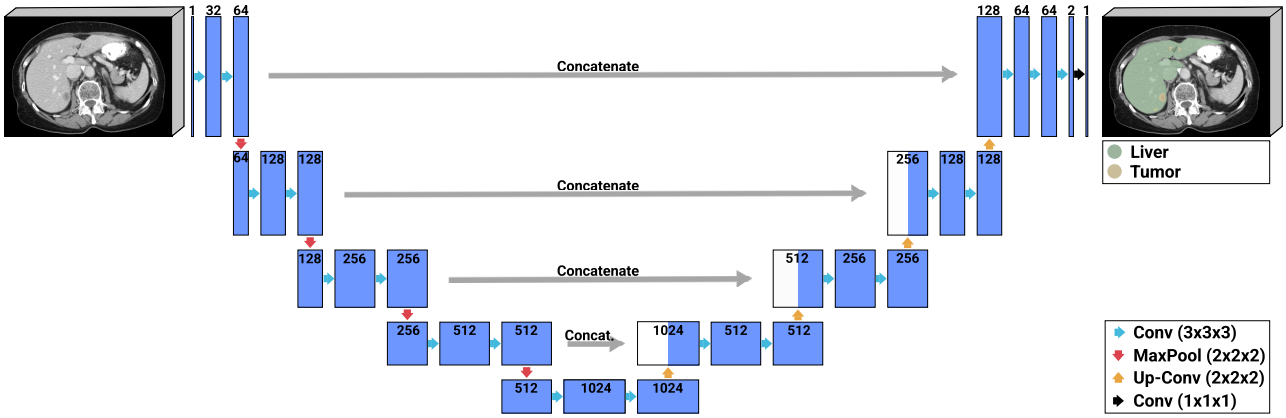


Fig. 1. U-Net architecture for liver and liver tumor segmentation

in the interval $[0, 1]$; a perfect segmentation yields a dice score of 1.

Due to hardware constraints, model was trained only on a train/validation subset without performing a K-Fold cross validation or any ensemble techniques.

IV. Preliminary results

We choose the frozen leaderboard from LITS 18 challenge leaderboards as the baseline to compare to. The results can be seen in table I.

The table shows that the objective of liver tumor segmentation is harder than segmentation of a sole liver. The winning teams employed cascading (coarse-to-fine) approach. First, they found the volume bounding box of the liver and only then they segmented the liver and liver tumors. They used u-net like architecture with added short skip connections.

Our network underperformed slightly, however without any performance boost techniques like two-stage approach, ensemble technique or a full run of K-Fold cross validation.

TABLE I
Leaderboard of LITS challenge on the day of Medical Image Computing and Computer Assisted Intervention (MICCAI) 2017

Username	Ranking	Liver tumor dice score	Liver dice score
Tian et al.	1	0.702	0.9630
Li et al.	2	0.686	0.9610
Chlebus et al.	3	0.676	0.9600
ours	–	0.6772	0.9529

A. Conclusions and future work

We showed that u-net like architecture with the specific data augmentation technique might match the results of other advanced networks with additional complexity, such as two-stage coarse-to-fine approach or short skip connections. [1]

In the future, we plan to extend this work by training the network with additional augmentation technique called mixup augmentation proposed in [10]. Gazda et al. [11] utilized a mixup method for KITS21 challenge and increased the performance by few percents for kidney, kidney cysts and kidney tumors segmentation.

As future work, we also consider extending research directions of our previous work, such as self-supervision in CNN domain [12], AI in Parkinson diagnostics [13], or feature selection in high dimensional medical data [14].

[1]Model was trained on a system with GeForce RTX 2080 GPU and AMD Ryzen Threadripper 2970WX 24-Core Processor.

Acknowledgment

This work was supported by the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic and the Slovak Academy of Sciences under contract VEGA 1/0327/20.

References

[1] S. McGuire, "World cancer report 2014. geneva, switzerland: World health organization, international agency for research on cancer, who press, 2015," *Advances in nutrition*, vol. 7, no. 2, pp. 418–419, 2016.

[2] P. Bilic, P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C.-W. Fu, X. Han, P.-A. Heng, J. Hesser *et al.*, "The liver tumor segmentation benchmark (lits)," *arXiv preprint arXiv:1901.04056*, 2019.

[3] S. Kohl, D. Bonekamp, H.-P. Schlemmer, K. Yaqubi, M. Hohenfellner, B. Hadaschik, J.-P. Radtke, and K. Maier-Hein, "Adversarial networks for the detection of aggressive prostate cancer," *arXiv preprint arXiv:1702.08014*, 2017.

[4] T. C. Mok and A. Chung, "Learning data augmentation for brain tumor segmentation with coarse-to-fine generative adversarial networks," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 70–80.

[5] E. Vorontsov, P. Molchanov, C. Beckham, W. Byeon, S. De Mello, V. Jampani, M.-Y. Liu, S. Kadoury, and J. Kautz, "Towards semisupervised segmentation via image-to-image translation. 2019," *URL http://arxiv. org/abs*, 1904.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[7] N. Heller, F. Isensee, K. H. Maier-Hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. Han *et al.*, "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge," *Medical image analysis*, vol. 67, p. 101821, 2021.

[8] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[9] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep learning and data labeling for medical applications*. Springer, 2016, pp. 179–187.

[10] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[11] M. Gazda, P. Bugata, J. Gazda, D. Hubacek, D. J. Hresko, and P. Drotar, "Mixup augmentation for kidney and kidney tumor segmentation," 2021.

[12] M. Gazda, J. Plavka, J. Gazda, and P. Drotar, "Self-supervised deep convolutional neural network for chest x-ray classification," *IEEE Access*, vol. 9, pp. 151 972–151 982, 2021.

[13] M. Gazda, M. Hireš, and P. Drotár, "Multiple-fine-tuned convolutional neural networks for parkinson's disease diagnosis from offline handwriting," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 1, pp. 78–89, 2021.

[14] P. Drotár, M. Gazda, and L. Vokorokos, "Ensemble feature selection using election methods and ranker clustering," *Information Sciences*, vol. 480, pp. 365–380, 2019.

Explainability of deep learning models in medical video processing

[1]*Michal KOLÁRIK (2nd year),*
Supervisor: [2]Ján PARALIČ, Consultant: [3]Martin SARNOVSKÝ

[1,2,3]Dept. of Cybernetics and Artificial Intelligence, FEEI TU of Košice, Slovak Republic

[1]michal.kolarik@tuke.sk, [2]jan.paralic@tuke.sk, [3]martin.sarnovsky@tuke.sk

***Abstract*—This paper describes the actual state of work in medical lung ultrasonography (USG) video data processing using explainable deep neural network models. It also presents a brief description of currently available datasets from the field of lung USG. The work also includes a description of methods for video data processing form the explainability perspective and suggestions for future research work directions.**

***Keywords*—Explainability, lung sliding, ultrasonography, XAI**

I. Introduction

Artificial intelligence (AI) has great potential to become helpful in the field of medicine. It is important that AI decisions are safe and trusted by doctors and patients. Therefore, the explainability of AI in medicine has an important role to play. Nowadays, AI is becoming more and more popular in medicine. It enables to help solve several problems, such as tumor diagnosis, fracture detection, and detection of diseases like COVID-19. Much of the data comes from imaging devices such as X-ray, computed tomography, magnetic resonance imaging, or ultrasonography. Deep neural networks outperform other methods when processing these data types (image or video data). However, the problem with deep neural networks is their complexity and the difficulty of understanding their behavior. But in medicine, it is important to explain the reasons behind decisions, especially when they directly impact human health. Moreover, there are also regulatory requirements in various parts of the world requiring transparent and trustworthy AI systems. For example, in the EU, the AI Expert Group has produced the Ethics Guidelines for Trustworthy AI [1], in China, there is China Academy of Information and Communications Technology (CAICT) that has issued a white paper on trustworthy AI [2]. Therefore, explainable methods that help explain the behavior of deep learning models and increase the transparency or trust of AI models are essential in the medical field.

II. The State of The Art

Many different techniques and methods have been developed to explain AI models recently. The different methods vary from how the results are represented through different model types or data types. The need for XAI methods is growing, especially with the increasing use of more complex machine learning models such as Ensembles, SVMs, or neural networks. These models are often referred to as black-box models [3] because they hide the logic of internal processes.

There are different taxonomies of XAI methods [4][5][6], but most of them agree on classifying the methods into four categories of problems. In the first category, methods are divided based on model design into interpretable models or ante-hoc (the explanation model is created in the AI training phase) and post-hoc (the explanation model is created only on trained models). Post-hoc methods are often used for black-box models, e.g., neural networks.

The second type of categorisation is into local and global methods. Local methods explain the prediction or decision of the model on a specific sample (outcome explanation), and global methods explain the model's behavior on the whole data set. Another categorisation divides methods based on their model specificity into model-agnostic (universal methods that are model-independent) and model-specific (methods designed for a selected black-box model architecture). Both model-agnostic and model-specific methods can be divided into local and global methods.

Another taxonomy of XAI methods is based on the data type [7], such as tabular data, image data, and text data.

An essential part of the deployment of XAI methods is their evaluation. Methods should be evaluated concerning terms such as goodness, usefulness, and satisfaction of explanations. It is also possible to compare explainable methods from the point of view of several levels. Authors in [8] propose three main levels for the evaluation of interpretability:

- Application level evaluation (real task): Implementation of models for explainability in a specific application and its testing on a real task. For example, software that will detect fracture sites based on X-ray records. The doctor could evaluate the quality of the explanations that the software offers to explain its intentions.
- Human-level evaluation (simple task): This level of explainability is also within applications, but the evaluation quality is not performed by experts, but by ordinary people - testers who are cheaper and also choose explanations according to how they help them understand at their level.
- Function level evaluation (proxy task): This level does not require people. It is suitable if the class of methods that will be used and with which the target class can work, e.g., decision tree. This model can be bounded to better explainability, e.g., using the decision tree pruning method.

III. DATA SETS

Video processing in medicine is currently used very rarely. There are several reasons for this:

- The insufficient number of datasets.
- Lack of devices that produced video recordings in medical imaging.
- Lack of physicians' time to create these video recordings or lack of time to annotate the collected videos.
- Until recently, the problem was also the lack of powerful hardware (HW) that could train a deep neural network for videos in a reasonable amount of time.

Current HW devices already offer good performance to learn such models. Gradually, datasets are also becoming available from the medical field that can be used to research this area. One of the types of video datasets available are videos in the field of lung ultrasonography. These datasets can be used to train the segmentation of areas, detect diseases such as COVID-19 and pneumonia, or detect problems after non-cardiac thoracic surgery.

A. *covid19_ultrasound*

Covid19_ultrasound [9] is a dataset from lung ultrasonography that combines data from collaborating hospitals and publicly available resources from the web (e.g., publications and educational websites). The dataset contains videos from two types of devices, 182 from convex probe and 65 from linear probe. Videos are classified as COVID-19, bacterial pneumothorax, viral pneumothorax, and healthy.

B. *COVIDx-US*

COVIDx-US [10] is an open-access benchmark dataset of COVID-19 related ultrasound imaging data. The COVIDx-US dataset was curated from multiple sources. Its current version consists of 242 lung ultrasound videos of patients with COVID-19 infection, non-COVID-19 infection, normal cases, and patients with other lung diseases. The dataset was systematically processed and validated specifically to build and evaluate artificial intelligence algorithms and models.

C. *LUS lung sliding*

These data [11] are obtained from lung ultrasonography for monitoring post-surgery complications such as pneumothorax and pleural effusion after thoracic surgery. A common approach is the use of X-ray. The currently available dataset consists of 48 videos, where 28 videos belong to the Lung Sliding Presence class and 20 videos belong to the Lung Sliding Absence class. New videos are currently being collected and annotated, which should contribute to the expansion of this dataset with new data.

IV. OBSERVED METHODS

Our previous research [12] has identified aspects that are important when applying explainability methods to AI models. We found that the most common requirements for AI systems in the field of medicine are transparency, trust, and accuracy. However, accuracy is not the most appropriate measure for medical applications. Metrics such as sensitivity, specificity, or F1-measure are more critical. Although these metrics are directly related to model performance, achieving transparency and trust in the model used requires understanding how the model works. It is dependent on the type of model used or the method used to explain it.

Our current research focuses on the processing of currently available videos of lung ultrasound for the detection of lung sliding. In this type of data, it is important to track the behavior of different regions over time, so we need to work with methods that take into account the evolution over time (3D CNN), not just the view of a single image (2D CNN). From previous research [11] that used a 2D convolutional neural network, we know that the most important region for lung sliding detection is the pleura and lung region. Therefore, we focus on the pleura and lung region in our work as well, see figure 1. Focusing on a specific region allows us to reduce the input data size and thus reduce the 3D convolutional neural network requirements.

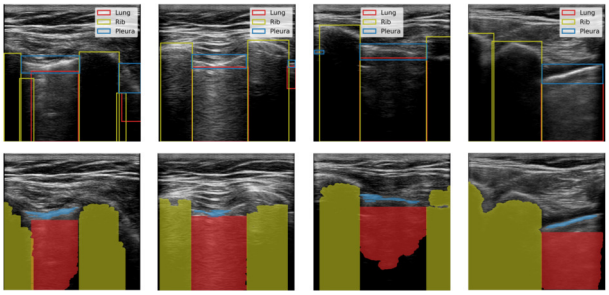


Fig. 1. Labbeled regions for the pleura, lung, and rib. Top: Bounding boxes labelled by volunteers. Bottom: Semantic masks generated from the bounding boxes. Adapted from Ref. [11]

Another area we focused on in our research is the selection of appropriate methods for the explainability of our model. Although there are many different methods and approaches for different neural network architectures, available methods are limited for 3D convolutional neural networks. The most commonly used methods are from the saliency maps category, modified from 2D convolutional neural networks to 3D convolutional neural networks. One such approach is the use of Selective Relevance [13] which allows for better motion capture on the image in the video. Existing 3D Grad CAM [13], and other approaches have their application at least in processing 3D images, for example, from MRIs of the brain [14], where the motion is not tracked but the region that most influences the model's decision is tracked.

V. EXPERIMENTS

We performed a set of experiments comparing several 3D convolutional neural network architectures as part of our work. The best results have been achieved on the Resnet3D-18 model - precision of 95%, sensitivity of 87%, and f1 of 91% while maintaining a patient-wise split. We applied the 3D GradCAM method to this model to visualize the regions on the source video that had the most significant impact on the final model decision. We also test alternative XAI methods to explain our model.

The actual results obtained are currently being consulted with physicians to verify the appropriateness of the procedure as well as the suitability of the methods used and their outputs. For this purpose, we proposed a questionnaire for physicians that has been prepared, showing them a series of original short USG videos together with two alternative explanations

and asking them for their expert feedback about provided explanations.

VI. Conclusions

In this work, we focused on reviewing currently available datasets in the field of video USG of the lung. We proposed a novel solution for processing the USG videos using a 3D convolutional neural network. We have performed several experiments while creating a 3D convolutional neural network model. We applied the explainability method to that model. However, the proposed procedures require consultation with physicians to obtain information about the suitability of the proposed approach.

In future research, we would like to focus on improving the model, implementing explainability methods to our model, and improving the model results. We want to look into the possibility of implementing methods such as 3D Grad CAM, Selective Relevance Grad CAM [13], as well as modifying the TCAV [15] method for a 3D convolutional neural network, as this method allows us to assess the quality of the explanation quantitatively. A possible approach could also be using xDNN network [16], which tries to make explanations based on similar examples already when training the network. However, this solution also requires consultation with physicians.

Acknowledgment

This work was partially supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic and Academy of Science of the Slovak Republic under grant no. 1/0685/21 and partially by The Slovak Research and Development Agency under grants no. APVV-20-0232 and APVV-17-0550.

References

[1] High-Level Independent Group on Artificial Intelligence (AI HLEG), "Ethics Guidelines for Trustworthy AI," *European Commission*, pp. 1–39, 2019.

[2] C. A. of Information and C. T. J. E. Academy, "White Paper on Trustworthy Artificial Intelligence," *China Academy of Information and Communications Technology JD Explore Academy*, no. 202106, 2021.

[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, aug 2018.

[4] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018, pp. 80–89.

[5] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[6] C. Molnar, "A guide for making black box models explainable," *URL: https://christophm. github. io/interpretable-ml-book*, 2018.

[7] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo, "Benchmarking and survey of explanation methods for black box models. arxiv e-prints," *arXiv preprint arXiv:2102.13076*, 2021.

[8] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv: Machine Learning*, 2017.

[9] J. Born, N. Wiedemann, M. Cossio, C. Buhre, G. Brändle, K. Leidermann, and A. Aujayeb, "L2 accelerating covid-19 differential diagnosis with explainable ultrasound image analysis: an ai tool," *Thorax*, vol. 76, no. Suppl 1, pp. A230–A231, 2021. [Online]. Available: https://thorax.bmj.com/content/76/Suppl_1/A230.2

[10] A. Ebadi, P. Xi, A. MacLean, S. Tremblay, S. Kohli, and A. Wong, "Covidx-us - an open-access benchmark dataset of ultrasound imaging data for ai-driven covid-19 analytics," *arXiv:2103.10003*, 2021.

[11] M. Jaščur, M. Bundzel, M. Malík, A. Dzian, N. Ferenčík, and F. Babič, "Detecting the Absence of Lung Sliding in Lung Ultrasounds Using Deep Learning," *Applied Sciences*, vol. 11, no. 15, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/15/6976

[12] M. Kolárik, J. Paralič, and M. Sarnovský, "Explainability of artificial intelligence in medical domain," *SCYR 2021: 21st Scientific Conference of Young Researchers*, pp. 97–101, 2021.

[13] L. Hiley, A. Preece, Y. Hicks, and R. Tomsett, "Explaining Motion Relevance for Activity Recognition in Video Deep Learning Models," pp. 1–9.

[14] M. Kan, R. Aliev, A. Rudenko, N. Drobyshev, N. Petrashen, E. Kondrateva, M. Sharaev, A. Bernstein, and E. Burnaev, "Interpretation of 3d cnns for brain mri data classification," in *International Conference on Analysis of Images, Social Networks and Texts*. Springer, 2020, pp. 229–241.

[15] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International conference on machine learning*. PMLR, 2018, pp. 2668–2677.

[16] P. Angelov and E. Soares, "Towards explainable deep neural networks (xdnn)," *Neural Networks*, vol. 130, pp. 185–194, 2020.

Contribution to Modeling of Distributed Control Systems for Large Physical Experiments

[1]*Milan TKÁČIK (3rd year),*
Supervisor: [2]*Ján JADLOVSKÝ*

[1,2]Dept. of Cybernetics and Artificial Intelligence, FEI TU of Košice, Slovak Republic

[1]milan.tkacik@tuke.sk, [2]jan.jadlovsky@tuke.sk

***Abstract*—The goal of this paper is to present the detector control system of the ALICE experiment at CERN, as well as to describe tools used for its modeling. The proposed model type is tested on the part of the detector control system and validated using the measurements taken on the actual detector.**

***Keywords*—Detector Control System, Distributed Control System, Finite-State Machines, Petri nets**

I. Introduction

The ALICE experiment (A Large Ion Collider Experiment) at CERN is the most complex experiment at the LHC (Large Hadron Collider), focused on a study of the quark-gluon plasma generated during Pb-Pb collisions at ultrarelativistic speeds [1]. Its modernization is currently underway, including the upgrade of software and hardware resources, with the planned commissioning in the summer of 2022 [2].

As part of the modernization of the software resources of individual detectors, the DCS systems (Detector Control Systems) are also being upgraded. The main motivation for this upgrade is to accommodate the highly increased demand on the amount and frequency of processed data.

Increased data processing requirements demand more computers and processing nodes allocated for the DCS system [3]. However, there is no exact methodology for determining the optimal distribution of software and hardware resources within the DCS system, as this distribution currently tends to be determined experimentally. This fact emphasizes the need for formal or mathematical modeling of the DCS system to determine the appropriate system distribution using optimization methods, even before its deployment.

II. Previous analysis and achieved results in research field

The DCS system is a highly distributed control system ensuring the stable and safe operation of individual detectors and the entire experiment - providing the configuration, control, monitoring, and safety of the detector.

Fig. 1 shows the structure of the new DCS system of the ITS detector (Inner Tracking System) - silicon pixel detector for particle tracking. The ITS detector consists of 192 Staves containing silicon ALPIDE chips arranged in seven barrels surrounding the particle collision point. Each Stave is operated in parallel by the DCS through the corresponding GBT link (Giga-Bit Transceiver) [4].

From the software point of view, the detector electronics are operated using sequences of commands sent via GBT links.

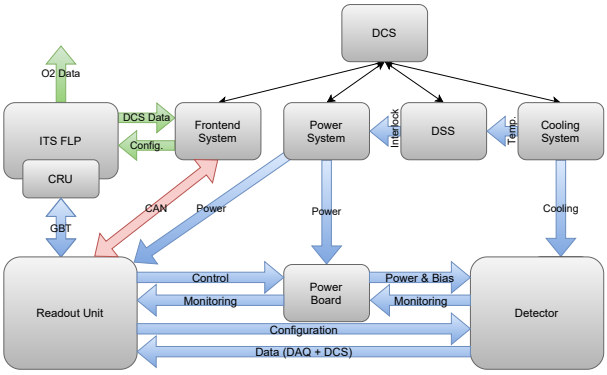


Fig. 1. DCS architecture of the ITS detector

Software distribution consists of three basic levels, all of which are interconnected using the DIM communication channels. The ALF server module (at the FLP level) provides the interface between GBT links and the DIM. The FRED server (Frontend System level) enables electronics configuration, control, and monitoring. The WinCC OA systems (SCADA/HMI at the central DCS level) ensure supervisory control of the detector and provide an HMI interface via operator panels [5].

The DCS system can be expressed as a Hybrid System, as it shows both continuous and discrete dynamics [6]. The evolution of currents, voltages, and temperatures of the detection electronics can be considered a continuous dynamics of the DCS system. The hardware and software resources of the distributed DCS system exhibit the discrete dynamics, which can be formally described using Petri nets (PNs) and Finite-State Machines (FSM).

III. Solved tasks and results

A combination of Finite-State Machines and colored timed Petri nets was chosen to describe the discrete dynamics of the DCS system. PNs are suitable for modeling the communication within distributed systems, as they provide a mechanism for modeling discrete-event concurrent systems [7]. On the other hand, FSMs can represent the states of individual subsystems and precisely define the transitions between them [8].

The timed extension of PNs allows to include the duration of individual commands in the model, which is crucial for the estimation of the overall performance of the system. The colored extension ensures the possibility of distinguishing the data being transferred within the system, e.g. sequence types or number of commands within a sequence [7]. FSMs add the

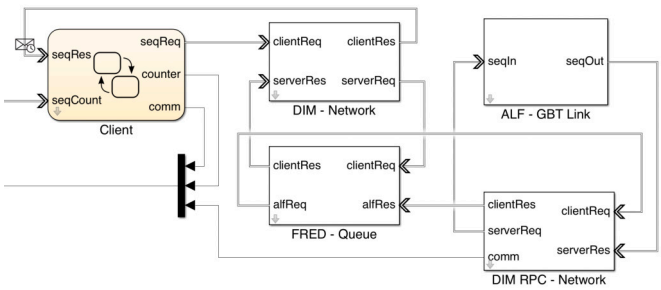


Fig. 2. Simulink schema of single ALF-FRED chain link

possibility of generating and processing sequences based on the data transferred by PNs and actual states of the system.

The DCS system model has been implemented and validated in the MATLAB/Simulink environment using the Stateflow charts. Several configurable subsystems have been created, allowing to build up the model of the entire DCS system or only a selected part of it. Parameters for the model were identified experimentally using the real detector electronics at CERN.

For the purpose of this paper, a simple test model of a single ALF-FRED chain link is used, see Fig. 2. The model includes a single *Client* responsible for request generation based on the specified parameters. The *FRED - Queue* prepares the sequences and transfers them to the *ALF - GBT Link* module when the communication channel is available. The *ALF - GBT Link* emulates the communication with detector electronics based on the received sequence of commands. The purpose of this test is to measure the time needed for the execution of a single sequence using the ALF-FRED chain based on the number of commands included in the sequence. The results are compared with the measurements acquired using the real detector electronics at CERN [9].

Fig. 3 and 4 show the test results with respect to different levels of the DCS. The results in Fig. 3 include the sequence execution emulation and the DIM RPC communication overhead, while the results in Fig. 4 include also the request/response processing, queuing, and the overhead of the DIM services communication with the client. The model of an ALF-FRED chain link exhibits the mean percentage error $MPE_{FRED} = 0.87\%$, which proves that the model is suitable for further usage.

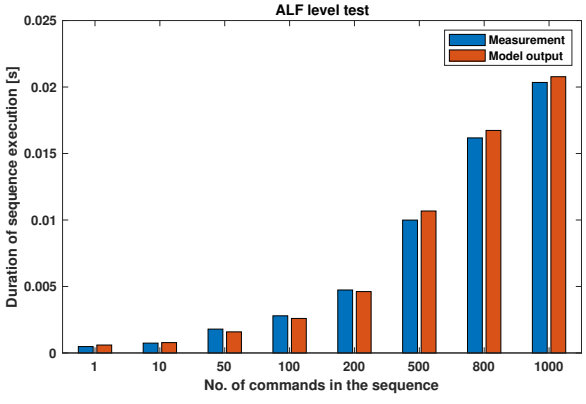


Fig. 3. Sequence duration model results from the ALF's point of view

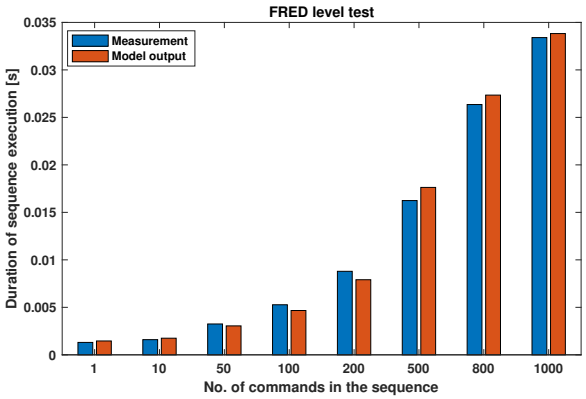


Fig. 4. Sequence duration model results from the FRED's point of view

IV. Future Research Steps

The future research steps include validating the entire DCS system model with the ITS detector at CERN. After the model validation, an optimization algorithm will be used to determine the optimal distribution of software and hardware resources within the DCS system in the ALICE experiment. It is planned to use the genetic algorithms for this purpose, due to the feasibility of describing the distributed system performance and resource usage by a fitness function.

V. Conclusion

This paper presented an overview of the newly developed DCS system and the need for its modeling. The proposed methodology for DCS modeling has been introduced, including the ALF-FRED chain link model validation with the measurements obtained using real detector electronics.

Acknowledgment

This work has been supported by the project ALICE experiment at the CERN LHC: The study of strongly interacting matter under extreme conditions (ALICE KE FEI TU 0195 / 2021).

References

[1] K. Aamodt, A. A. Quintana, R. Achenbach, S. Acounis, D. Adamová, C. Adler, M. Aggarwal, F. Agnese, G. A. Rinella, Z. Ahammed *et al.*, "The ALICE experiment at the CERN LHC," *Journal of Instrumentation*, vol. 3, no. 08, pp. S08 002–S08 002, aug 2008.

[2] M. Bernardini and K. Foraz, "Long shutdown 2@ lhc," *CERN Yellow Reports*, vol. 2, no. 00, p. 290, 2016.

[3] P. Chochula, A. Augustinus, P. Bond, A. Kurepin, M. Lechman, J. LÃ, O. Pinazza *et al.*, "Challenges of the alice detector control system for the lhc run3," *ICALEPCS Barcelona, Spain*, 2017.

[4] S. Kushpil, A. collaboration *et al.*, "Upgrade of the alice inner tracking system," in *Journal of Physics: Conference Series*, vol. 675, no. 1. IOP Publishing, 2016, p. 012038.

[5] M. Tkáčik, J. Jadlovský, S. Jadlovská, L. Koska, A. Jadlovská, and M. Donadoni, "FRED Flexible Framework for Frontend Electronics Control in ALICE Experiment at CERN," *Processes*, vol. 8, no. 5, p. 565, 2020.

[6] J. Lygeros, C. Tomlin, and S. Sastry, "Hybrid systems: modeling, analysis and control," *Electronic Research Laboratory, University of California, Berkeley, CA, Tech. Rep. UCB/ERL M*, vol. 99, 2008.

[7] O. Baldellon and J.-C. F. M. ROY, "Modeling distributed real-time systems using adaptive petri nets," *Saint-Malo, France*, pp. 7–8, 2011.

[8] F. Wagner, R. Schmuki, T. Wagner, and P. Wolstenholme, *Modeling software with finite state machines: a practical approach*. CRC Press, 2006.

[9] J. Jadlovský, A. Jadlovská, S. Jadlovská, M. Oravec, D. Vošček, M. Kopčík, J. Čabala, M. Tkáčik, P. Chochula, and O. Pinazza, "Communication architecture of the Detector Control System for the Inner Tracking System," p. THPHA208, 2018.

Deep neural networks for speech-to-text systems

[1]*Slavomír GEREG (3rd year)*

Supervisor: [2]Jozef JUHÁR

[1,2]Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

[1]slavomir.gereg@tuke.sk, [2]jozef.juhar@tuke.sk

***Abstract*—The article deals with automatic speech recognition, based on the use of time delay neural network and the experimental toolkit Kaldi. Automatic speech recognition systems usage in the modern user applications is shortly described. Deep neural networks and automatic speech recognition toolkit Kaldi are briefly introduced. Gigaspeech, the reference database of spoken English is briefly compared with the reference database of spoken Slovak. The methodology and first experimental results from comparing both databases are shortly described. Finally, the ongoing direction of research is outlined.**

***Keywords*—automatic speech recognition, deep neural networks, GigaSpeech database, Kaldi toolkit**

I. INTRODUCTION

Automatic speech recognition or speech to text transcription is the complex process of converting of audio speech signal into the it's corresponding text form. Systems and algorithms for automatic speech recognition have been important part of research for more than 6 decades. Currently, there are many commercial products and applications that are using automatic speech recognition systems and algorithms. One of the most common uses of these algorithms and systems for the general users are applied in voice control or automatic systems. For example we can use the extensive use of voice assistants in smart mobile phones (Google assistant, Siri, Cortana, Alexa, ...), smart home voice control applications and products, voice control in home gaming systems (Kinect), voice navigation and many other advanced applications and systems that use automatic speech recognition [1], [2], [3], [4], [5], [6], [7], [8].

Popular approach in modern speech recognition system these days is usage-based on deep neural networks [9].

II. MY PREVIOUS RESEARCH

In my previous work I studied Kaldi toolkit options and methodology used in our department for training of acoustic models and preparing data for training and recognition process. After I learned how to work with the system Kaldi, I have started training various acoustic models and evaluating their success. I have started with studying practical use of this system. I have gone through various stages of work with automatic speech recognition system Kaldi for example preparing of acoustic data to suitable form, finding best acoustic model for training, system training with real database, testing and evaluation of results provided by system Kaldi. I have also created a manual about practical system Kaldi training for my colleagues and for future research [14].

I have also continued my diploma experiment and investigated the impact of training data on automatic speech recognition system success.

Automatic speech recognition system used in our department is using Kaldi toolkit and it is based on time delay neural network model. The algorithm that we use in automatic speech system to prepare data and train time delay neural network model consist of 10 steps [14], [15]:

1. Speech data perturbation (speed perturbation with 0.9, 1.0 and 1.1 factors)
2. Creating of hi-resolution MFCC features
3. Extracting of low-resolution features
4. Alignment data with low-resolution features (using HMM model)
5. Diagonal UBM (Universal Background Model) training
6. i-vector extractor training
7. Extracting online i-vectors
8. Creating neural net configs using the *xconfig* parser (using HMM information)
9. Training the TDNN model
10. Online decoding

III. CURRENT STATE OF RESEARCH

For now, I am working on comparing my results for automatic speech recognition in Slovak language with published results in English collected by using one of the newest large databases designed specifically for automatic speech recognition. This database is called GigaSpeech.

Subset	Audiobook	Podcast	Youtube	Total
XL	2 655 h	3 499 h	3 846 h	10 000 h
L	650 h	875 h	975 h	2 500 h
M	260 h	350 h	390 h	1 000 h
S	65 h	87,5 h	97,5 h	250 h
XS	2,6 h	3,5 h	3,9 h	10 h

Table 1 Subsets of GigaSpeech database [16]

GigaSpeech is multidomain automatic speech recognition corpus made from 10 000 hours of transcribed audio. This database is made from multisource data, audiobooks, podcasts

and youtube audio. It contains variety of topics, such as art, science, sports, etc. Database is divided to subsets that are shown in Table 1 [16].

For testing the efficiency of their system, the research team made another evaluation datasets. These datasets consist of similar type of audio data (podcast, youtube audio) like the GigaSpeech training database. They developed 2 datasets DEV and TEST. The amount of data in these datasets is shown in Table 2.

Sets	Podcast	Youtube	Total
DEV	6,3 h	6,2 h	12,5 h
TEST	16,1 h	24,2 h	40,3 h

Table 2 GigaSpeech evaluation datasets [16]

In the Figure 1 we can see GigaSpeech creation pipeline block scheme where we can see the whole process from audio collection to evaluation.

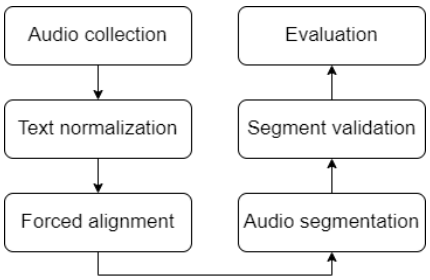


Figure 1 GigaSpeech creation pipeline

In GigaSpeech evaluation researchers published results for toolkits like Kaldi, Athena, ESPnet and Pika. For our comparation of results are relevant results for Kaldi based system. In this case researchers used Recurrent neural network language model-based algorithm.

In our work we are using Slovak language data to make similar subsets (in volume of data) like GigaSpeech subsets of data. Cause in Slovak language we don't have that big amount of data transcribed, I have managed to make subsets of data with volume 10 hours, 250 hours and 1 000 hours which are equivalent to GigaSpeech XS, S and M subsets. Our subsets are also made of multisource and multidomain audio data, such as TV shows, TV news, meeting recordings, and some other.

For HMM modeling of our acoustic models we use adaptive train process based on fMLLR (feature space maximum likelihood linear regression). For HMM training we use 12 000 HMMs and 260 000 GMM.

From language resources we used 250 000 words lexicon for training and 650 000 words lexicon for testing.

We also make similar test subset (in volume of data) for experimental evaluation and comparison of results.

First experimental results have shown that WER parameter for Slovak 1 000-hour subset is better by more than 2% against GigaSpeech 1 000-hour subset which is showed in Table 3.

Subset	Gigaspeech results (WER)	Our results (WER)
XS	-	24,63
S	22,59	21,03
M	17,96	15,08

Table 3 Comparison of Gigaspeech results with our results

IV. FURTHER RESEARCH DIRECTION

In my future work I want to continue with my previous experimental work, and I want to experimentally compare all Slovak subsets with GigaSpeech subsets. I also want compare preparing of this datasets and automatic speech recognition process of GigaSpeech authors with methods and algorithms used by me. After that I would be able compare results of our department with results of research team that designed GigaSpeech database, more complexly.

ACKNOWLEDGMENT

The research presented in this paper was supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the research projects VEGA 1/0753/20, VEGA 2/0165/21 and by Slovak Research and Development Agency project APVV SK-TW-21-0002.

REFERENCES

[1] Deng L., O'Shaughnessy D., "Speech Processing-A Dynamic and Optimization Oriented Approach," Marcel Dekker Inc., New York, 2003.

[2] Huang X., Acero A., Hon H.W., "Spoken Language Processing," Prentice-Hall, Upper Saddle River, NJ, 2001b.

[3] Baker J., Deng L., Glass J., Khudanpur S., Lee C.H, Morgan N. et al., "Research developments and directions in speech recognition and understanding," Part I. IEEE Signal Process. Mag. 26 (3), 2009a ,75-80.

[4] Davis K.H., Biddulph R, Balashek S., "Automatic recognition of spoken digits," J. Acoust. Soc. Am. 24 (6), 1952, 627-642.

[5] He X., Deng L., "Speech-centric information processing: An optimization-oriented approach," Proc. IEEE 101 (5), 2013, pp. 1116-1135.

[6] Li J., Deng L., Haeb-Umbach R., Gong Y., "Robust automatic speech recognition: A bridge to practical applications," 2015. ISBN 978-0-12-802398-3.

[7] Juhár J. et al., "Rečové technológie v telekomunikačných a informačných systémoch." Košice: Equilibria, s.r.o., 2011. ISBN 978-80-89284-75-7.

[8] Bundzel M., Sincak P., "Combining gradient and evolutionary approaches to the artificial neural networks training according to principles of Support Vector Machines." In: Proceedings of the IEEE International Joint Conference on Neural Network, Vancouver, Canada, July 16-21, 2006, DOI: 10.1109/IJCNN.2006.246976.

[9] Fohr D., Mella O., Illina I., "New Paradigm in Speech Recognition: Deep Neural Networks." IEEE International Conference on Information Systems and Economic Intelligence, Apr 2017, Marrakech, Morocco. 2017.

[10] Povey D., Ghoshal A. et al., "The Kaldi Speech Recognition Toolkit." In: IEEE 2011 Workshop on Automatic Speech Recognition anvUnderstanding, Hilton Waikoloa Village, Big Island, Hawaii, US, 2011. ISBN 978-1-4673-0366-8

[11] Smit P., Virpioja S., Kurimo M., "Improved Subword Modeling for WFST-Based Speech Recognition." In: Annual Conference of the International Speech Communication Association (INTERSPEECH), Stockholm, August 2017, pp 2551–2555.

[12] Alyousefi S. H., "Digital Automatic Speech Recognition using Kaldi." MSc. Thesis, Florida Institute of Technology, Melbourne, Florida, May 2018.

[13] Enarvi S., Smit P., Virpioja S., Kurimo M., "Automatic Speech Recognition with Very Large Conversational Finnish and Estonian Vocabularies." IEEE/ACM Transactions on Audio, Speech, and Language Processing, volume 25, issue 11, November 2017, pages 2085–2097.

[14] Povey Daniel, et al. "The Kaldi speech recognition toolkit." IEEE 2011 workshop on automatic speech recognition and understanding. No. CONF. IEEE Signal Processing Society, 2011.

[15] P. Vijayaditya, D. Povey, and S. Khudanpur. "A time delay neural network architecture for efficient modeling of long temporal contexts," Proc: Interspeech, 2015.

[16] G. Chen, S. Chai, G. Wang, J. Du, W. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Y. Wang, Z. You, and Z. Yan, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," in Interspeech, 2021.

Analysis of The Electric Drive Gear for the use in Load Torque Emulator

[1]*Jozef IVAN (3rd year)*
Supervisor: [2]František Ďurovský

[1,2] Dept. of Electrical Engineering and Mechatronics, FEI TU of Košice, Slovak Republic

[1]jozef.ivan@tuke.sk, [2]frantisek.durovsky@tuke.sk

***Abstract*— This paper presents an investigation of gearboxes behavior implemented in the system that will serve as an emulator of dynamic loads. In the first section of paper, the proposed mechanical system is described and compared with former systems used for emulators. Next, the hardware of proposed emulator system and the design of the measurement is presented. In the last section evaluation of the data and outcomes are mentioned.**

***Keywords*— Emulator, Experimental system, Analysis of gearbox**

I. Introduction

In cases when companies want to investigate the behavior of their products in the real environment, it's advantageous to imitate this environment in laboratory conditions at first. In the field of electric drives and actuators, the most effective way of testing products is to undergo multiple loading and lifespan tests before the introduction of the product to the market. There are several ways how to load electric drives and actuators in laboratory conditions, but the most effective approach is loading a tested drive with another electric drive [1]. The newest approach in testing electric machines is presented by implementing different types of emulation models into the testing system and by applying variable dynamic loads to the shaft of the tested machine such as in [2]. However, the focus on testing the geared drives is very rare.

In this paper, the investigation of the gears in the system dedicated to the testing of actuators will be presented. It is planned to use one of the emulation methods on this system and the first stage of emulator's new design includes the collection and analysis of the data and their accurate representation. Analyzed data will be used for the design of mathematical models for the new type of emulator.

II. The New Approach to Emulator System

The hardware of proposed emulator is built on the base of testing stand originally intended for testing high precision actuators with simple loading cycles. This system consists of a Load machine (LM, PMSM) which is coupled with a planetary gearbox (PG). The output shaft of the PG is connected to Drive under test (DUT) which is presented by the high precision geared actuator with cycloidal gear and permanent magnet synchronous motor. The scheme of the proposed system for testing of high precision actuators is shown in *Figure 1*.

The motor torques are denoted as T_e and T_L and torque on the common shaft is denoted as T_{c_S}. Values T_{L_out} and T_{e_out} are values of the applied torque multiplied by the ratio of gearboxes and represent desired values of output torque.

Mostly, emulators with specific emulation methods were based on a system where two drives are connected directly with an ideally rigid shaft. The spring and damping nature of the shaft is neglected which eases the calculation.

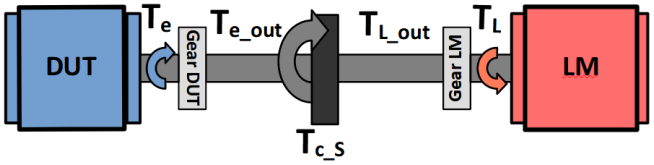


Figure 1. Structure of emulator with gears

In this system, it is requested to create/emulate the load on DUT output flange, therefore, the behavior of the gears cannot be neglected as it has a high impact on correct system estimation in the emulation algorithm. The relationship between motor torques and torque on the output shaft can be described as:

$$T_{c_S} = T_e \, i_{DUT} \, \eta_{DUT} = T_L \, i_{LM} \, \eta_{LM} \qquad (1)$$

where i_{DUT} and i_{LM} are gear ratios of the gearboxes. The variables η_{DUT} and η_{LM} represents the efficiency of the specific gearbox. To determine the value of T_{c_S}, a torque flange sensor for measuring the torque directly on the shaft is used.

The efficiencies of the gearboxes can be evaluated with the simple proportion of applied torque values and output torques on the shaft as in equation (2). The equations of the efficiencies are dependent on the operational mode of the specific drive. During the operation of the emulator, all 4 quadrant modes of the drives will be used, so in the evaluation, we have to differentiate between motor and generator modes on the drives. LM and DUT will alternate so if DUT will be in motor mode, LM will be in breaking mode and vice versa.

III. Design of the Measurement

It is important to mention that LM is working in the torque control loop and the DUT in the speed control loop. The whole emulator system is based on the Siemens hardware with the Simotion D as the control system, which is also responsible for the operation of the drives. The main task of the system in the measurement is to collect actual values of torques and speed of the drives and output shaft torque. As it is assumed that the value of the efficiency of both gearboxes will be dependent on the actual speed and torque applied to the gear, the measurement was based on the DUT operation in various speed

levels with systematically applied load torque of the LM in each level. Maximum applicable torque and speed to the system are based on the maximum allowable torques and speeds of the drives, calculated through the ratios (ω_{max_DUT} = 2000 rpm, T_{max_LM}= 5.5 Nm).

The process of the measurement is shown in Figure 2 and performed as follow:

1. DUT speeds up to the requested speed level.
2. LM raises torque from 0 to maximum torque in small steps with little steady phase between steps to record the measurement data.
3. LM lowers the torque to zero, DUT raises the speed to the next level and the torque steps are repeated.
4. The procedure is taken until the whole speed range is measured.

In every speed/torque combination, 50 values of actual values of motor speed, torque and shaft torque are measured (period 0.1 s)

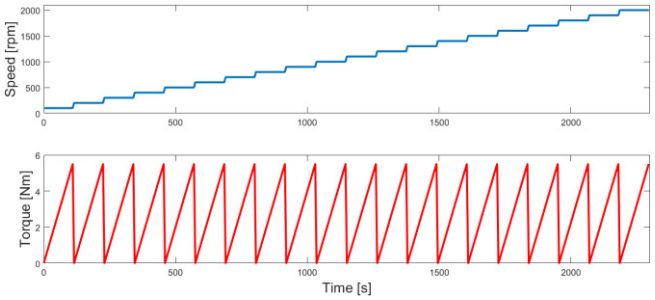


Figure 2. Description of measuring process

IV. Evaluation of the Data

Measured data were evaluated in MATLAB software according to the corresponding motor mode. In this paper, the 1st quadrant of the DUT will be shown. It has to be considered that in this motoric mode, the DUT acts as motor and LM as a break. Conversion of the torque through the DUT gearbox to the output shaft will lower the value of the torque according to gearbox efficiency. If the torque will be transferred through the LM gearbox to LM, the efficiency of the gearbox wil consume a fraction of the transferred torque. The equations for evaluation of the DUT and LM efficiencies can be denoted as:

$$\eta_{DUT} = \frac{T_{c_S}\ 100\%}{T_e\ i_{DUT}}, \eta_{LM} = \frac{T_L\ i_{LM}\ 100\%}{T_{c_S}} \qquad (2)$$

The process of the evaluation is similar for 3rd quadrant. In the 2nd and 4th quadrant, it is necessary to switch values of torques. Every combination of LM load torque and DUT speed has 50 measured values so it is necessary to average these values. Every combination of measured torque and speed is assigned to the appropriate value of efficiency as shown in Figure 3 (blue dots).

V. Results of Measurements

The main aim of measurement is to obtain an efficiency map of the gear, which describes the gearbox behavior according to applied torque and speed. Therefore, the measured points were fitted with the polynomial function.

For the evaluation of the best fit, combinations of polynomials of 2nd and 3rd degree have been tested. The best performance was acquired with the combination (torque/speed) of 3rd and 3rd degree for DUT and 3rd and 2nd degree for LM. The efficiency in the 1st quadrant for DUT changes in the range from 30 to 85%. The variation of the torque has a significantly higher influence on the efficiency of the actuator than the variation of speed.

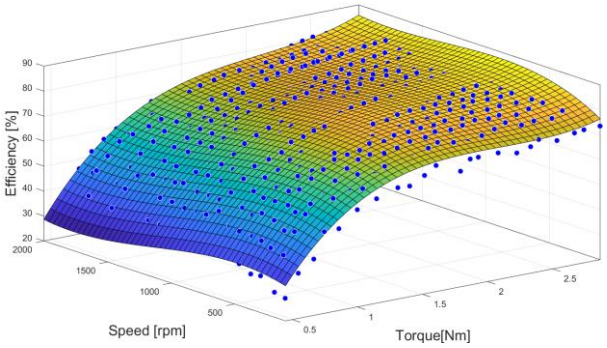


Figure 3. Comparison of measured values and polynomial

The outcome of the overall evaluation is presented in Figure 4 where the efficiency maps for all 4 quadrants are shown.

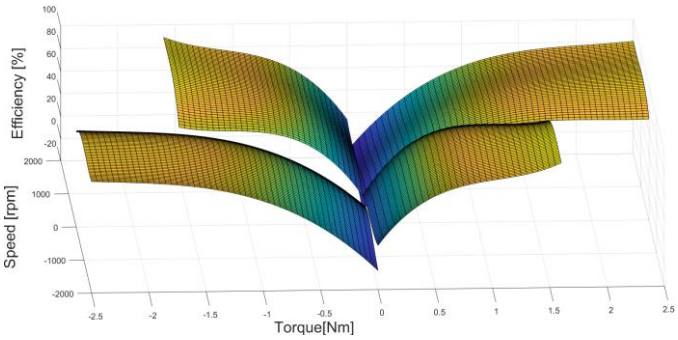


Figure 4. Polynomials of DUT efficiencies for all 4 quadrants

VI. Conclusion

This paper has described a process of investigating the behavior of gearboxes implemented in the structure of the system that will be used for the construction of an emulator of dynamic loads with a specific emulation model. Investigation of the gears plays an important role in the process of emulator design as gearboxes bring a large number of uncertainties into the system. Using these outcomes, we can identify losses in the gearboxes and take them into account in the emulator algorithm. Future work will be focused on creating the simulation model of the emulator with gearboxes, where we can simulate a real system and tune the whole emulation in the simulation environment.

Acknowledgment

This work was supported by the Slovak Research and Development Agency under contract No. APVV-15-0750, by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic under the project VEGA 1/0493/19 and by FEI TUKE under grant FEI-2021-73.

References

[1] Akpolat, Z. H., Asher, G. M., Clare, J. C.: „Emulation of high bandwidth mechanical loads using vector controlled AC dynamometer," in *Proc. 8th Int. Power Electronics and Motion Control Conf. (PEMC'98)*, 1998, vol. 5, pp. 133–138.

[2] Rodič, M., Jezernik, K., Trlep, M.: „Control Design in Mechatronic Systems Using Dynamic Emulation of Mechanical Loads." *In Proceedings of the IEEE International Symposium on Industrial Electronics*, 2005. ISIE 2005. 4 (2005): 1635-1640.

Ageing analysis of li-ion battery cells based on measured data

[1]*Juraj BIĽANSKÝ (3rd year),*
Supervisor: [2]*Milan LACKO*

[1,2]Department of Electrical Engineering and Mechatronics, FEI TU of Košice, Slovak Republic

[1]juraj.bilansky@tuke.sk, [2]milan.lacko@tuke.sk

***Abstract*—The article is focused on the analysis of measured data from long-term tests that test batteries with different discharge and charging currents in order to determine the state of health and residual capacity of the battery in different conditions.**

***Keywords*—battery, cells, testing, charging, discharging, long-term testing**

I. Introduction

In recent years, batteries have become a very important part of our lives. They provide power for the portable electronic devices that we use everyday. For example, we can mention mobile phones, laptops, watches, electric cars and much more. With increasing technological advancements in portable electronic devices, the market for affordable batteries is also growing. Many different types of batteries are available on the market today, such as nickel-cadmium (NiCd), nickelmetal hydride (NiMh) or lithium-ion (Li-ion). The latter Li-ion batteries make up the majority of the market. They are used in mobiles, laptops and many other industries. These batteries are also used in electric vehicles. Batteries used in electric vehicles are called traction batteries. Traction batteries consist of a large number of small battery cells which, by suitable series–parallel connection, form an electric vehicle battery with the required properties such as output voltage, current and capacity. Final battery quality, then, defines the basic characteristics of the vehicle such as power, range and weight. [1], [2], [3].

II. Tested Object

Fig. 1. Samsung INR18650-25R li-ion battery cell.

Item	Specification
3.1 Minimum discharge capacity	2,500mAh Charge: 1.25A, 4.2V,CCCV 125mA cut-off, Discharge: 0.2C, 2.5V discharge cut-off
3.2 Nominal voltage	3.6V
3.3 Standard charge	CCCV, 1.25A, 4.2V, 125mA cut-off
3.4 Rapid charge	CCCV, 4A, 4.2V, 100mA cut-off
3.5 Charging time	Standard charge : 180min / 125mA cut-off Rapid charge: 60min / 100mA cut-off
3.6 Max. continuous discharge (Continuous)	20A(at 25℃), 60% at 250 cycle
3.7 Discharge cut-off voltage End of discharge	2.5V
3.8 Cell weight	45.0g max

Fig. 2. Samsung INR18650-25R specifications with standard charging and discharging conditions[4].

III. Testing Device

The general purpose of our testing device is to cyclically charge and discharge the battery, using a predefined current and voltage profile for the required number of cycles. For testing, we used our developed device, which is capable of [5]:

- Charge Li-ion cells using CCCV to 10A with overcharge options
- Discharge li-ion cells up to 20A with the programable end of discharge and undercharge options.
- EIS(Electrochemical Impedance Spectroscopy)
- Device is capable to measure currents, voltage, temperature
- Logging to SD-Card or the PC using CAN-BUS communications

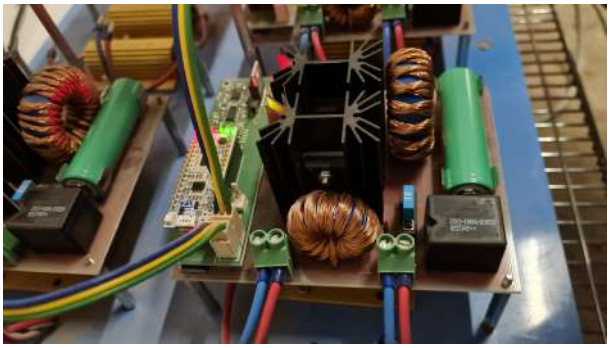


Fig. 3. Testing devices.

IV. Testing procedure

Standard charging and discharging tests have been performed on the batteries to detect changes in SOH (State of Health), which is defined as the amount of energy that the battery is able to deliver vs. the energy that the battery is able to deliver at the beginning of its life. These tests are very time consuming, as with standard charging currents, one charge and discharge cycle takes about 4 hours. Thus, a complete battery test cycle in which we perform up to 500 such cycles will take almost 84 days. After completing these tests, we examine the measured data from how the battery properties have changed. We determine the residual capacity after different cycles and from this residual capacity, we then determine the SOH batteries. In addition to capacity, we also record changes in internal resistance or battery impedances, which are directly responsible for the battery's ability to deliver/receive large currents.

1) Charging and discharging with standard currents
2) Fastcharging with 3A and discharging with 2C(5A)
3) Fastcharging with 3A and standard discharging (2,5A).
4) Standard charging and discharging with 2C(5A)

V. Test Results

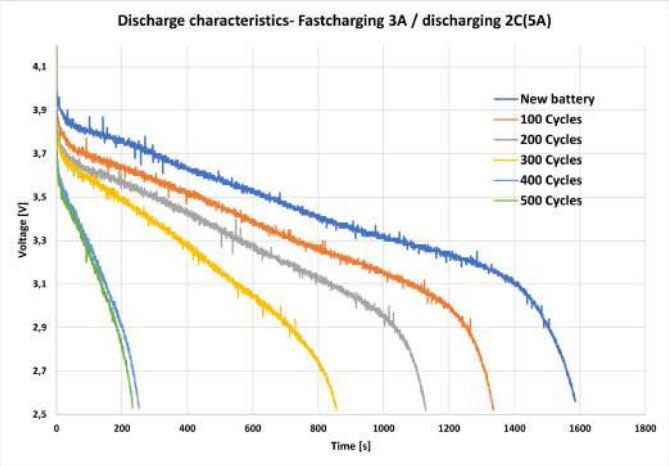


Fig. 4. Comparison of discharging characteristics using 1st testing procedure at different cycle number.

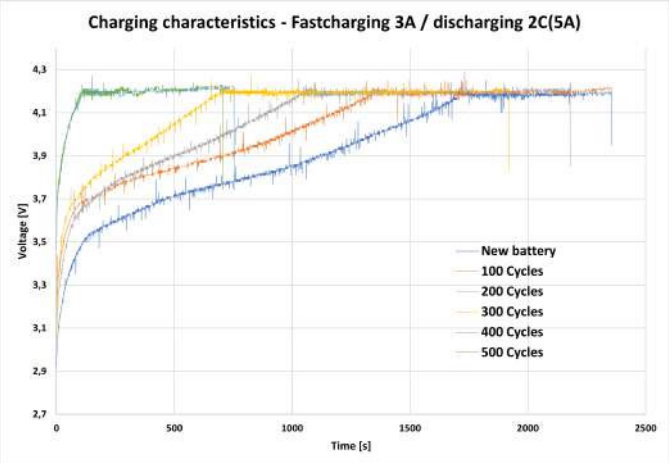


Fig. 5. Comparison in differences in charging characteristics using 1st testing procedure at different cycle number.

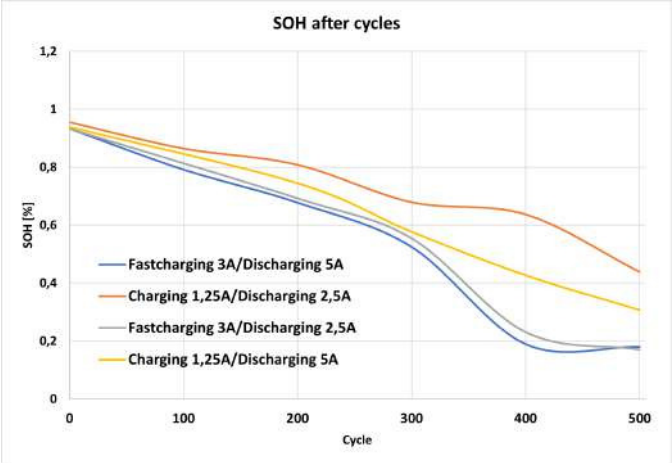


Fig. 6. Comparison of SOH changes based on number of cycles and different test conditions.

VI. Conclusion

We have subjected the Samsung INR18650-25R battery to several tests on our test devices. Selected testing methods are described in Chapter 4. The results of these tests can be seen in Figs. 4 to FIG. 6.. From FIG. 4 and FIG. 5 it can be seen those battery parameters such as OCV and maximum possible delivered capacity change significantly with an increasing number of cycles. This is due to the increasing internal resistance of the battery and the overall changes in cell impedance. This increasing internal resistance can be seen in the charging processes in FIG. 5, when the first phase, the phase of the constant current in which a larger current flows into the battery, is significantly shortened and the phase of the constant voltage, in which a smaller current flows into the battery is longer. From the waveforms in FIG. 4 and the same runs for the remaining test procedures, we have created a characteristic for determining SOH, which is shown in FIG. 6. It is clear from this characteristic that, in addition to the number of cycles, the size of the discharge and charging currents also has a significant effect on the battery parameters. At the same time, we can say that charging with higher currents than standard affects these parameters more than discharging.

Acknowledgment

This work was supported by the Slovak Research and Development Agency under Contract No. APVV-18-0436.

References

[1] J. Bilansky and M. Lacko, "Design and simulation of cyclic battery tester," *Power Electronics and Drives*, vol. 5, no. 1, pp. 229–241, 2020.
[2] J. Bil'ansky and M. Lacko, "Návrh koncepcie nabíjacej a vybíjacej časti testéra li-ion batériových článkov," *Electrical Engineering and Informatics*, vol. 11, pp. 385–389, 2020.
[3] J. Bilansky and M. Lacko, "Overview of the battery types and their testing," *SCYR 2020 - Nonconference Proceedings of Young Researchers*, pp. 143–146, 2020.
[4] *Lithium-ion rechargeable cell for power tools*, Samsung SDI Co., Ltd., 4 2014, rev. 1. [Online]. Available: https://www.tme.eu/Document/d5041798b41b6ad5e98cd9d1377d272d/INR18650-25R.pdf
[5] J. Bil'anskỳ, T. Merva, J. Ivan, A. Marcinek, and M. Lacko, "Cyclic tester of battery cells for electric vehicles," in *2021 IEEE International Workshop of Electronics, Control, Measurement, Signals and their application to Mechatronics (ECMSM)*. IEEE, 2021, pp. 1–7.

Ag_2S based thermoelectric materials for wearable electronics

Gabriela HRICKOVA (1st year)
Supervisor: Juraj DURISIN

Dept. of Technologies in Electronics, FEI TU of Kosice, Slovak Republic

gabriela.hrickova@tuke.sk, juraj.durisin@tuke.sk

***Abstract* — The paper deals with fundamentals of thermoelectric devices and with thermoelectric materials based on silver sulfide Ag_2S. It explains basic principles of thermoelectric coolers and thermoelectric generators. It discusses electric, thermoelectric and mechanical properties of Ag_2S based semiconductors, which are flexible thermoelectric materials potentially applicable in wearable electronics.**

***Keywords* — thermoelectric devices, ZT, thermoelectric materials, Ag_2S.**

I. INTRODUCTION

Thermoelectric materials have been used for many decades for various applications. They are usually employed as thermocouples for contact measurement of temperature (temperature of air, heated or cooled objects, etc.). Since thermocouples enable the conversion of the value of measured temperature directly into an electrical signal, this method of temperature measurement is suitable for electronic recording, long-term measuring or automated temperature measurement. In addition to common household or laboratory use, thermocouples are also used in industry, where they enable the control of various production processes in which it is important to know the processing temperature. They are made of different materials, due to their composition they vary in operational temperature range, accuracy and other properties [1, 2].

Another important way of using thermoelectric materials is their application as Peltier coolers. The purpose of Peltier coolers is to prevent the overheating of e.g. computer components. The thermoelectric materials used to transfer thermal energy by electric current must exhibit the highest possible value of the dimensionless figure of merit ZT; their operation shows then the highest possible efficiency. The search for new thermoelectric materials that work with the highest possible heat transfer efficiency is today's the main goal of research in this field. Semiconductor materials have shown to be the most suitable for this purpose [1, 3].

One of the today's investigated semiconductor materials is silver sulfide Ag_2S. It demonstrates (even with the content of impurities) the maximal values of ZT only around a few tenths (best thermoelectric materials exhibit ZT up to ~3), but contrary to other semiconductor thermoelectric materials it is a malleable material. It can be used in environments where increased mechanical load, shocks, etc. is expected. An example of using of Ag_2S is the conversion of waste heat from the human body into electrical energy that can be used to recharge or power a wristwatch. The properties of Ag_2S thus provide a precondition for its practical applications. The goal of my PhD study is to find new compositions and/or production methods of Ag_2S based thermoelectrics providing better properties compared to current Ag_2S based thermoelectrics [3, 2].

II. THERMOELECTRIC DEVICES

There are 2 types of thermoelectric devices: thermoelectric cooler and thermoelectric generator [4, 2, 5].

The principle of operation of a thermoelectric cooler (Fig. 1) lies in the Peltier effect. The heat transfer in the cooler is caused by the electrical current and the heat flow resulting from the temperature difference $\Delta T=T_H-T_C$ between the hot T_H and cold side T_C. The cooler is powered by a voltage source V [3, 6, 7].

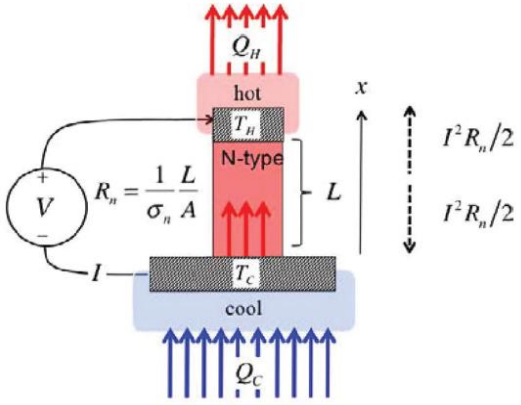


Fig.1 Simple n-type thermoelectric device with one leg operating as a Peltier cooler [3].

The overall cooling power Q_P (rate of heat flow from the cool side) of an n-type semiconductor Peltier cooler can be described as follows:

$$Q_P = \pi_n . I - \lambda_n . \frac{\Delta T . S}{\Delta x} - \frac{R_n . I^2}{2} \qquad (1)$$

where π_n is the Peltier coefficient, λ_n is the thermal conductivity (caused by electrons), S and Δx is the cross section and length of the cooler, R_n is electrical resistance of the cooler, and I is electrical current flowing through the cooler [7].

The right side of the equation represents the heat flow rate caused by the Peltier effect (1st term) minus the back diffused heat flow rate from the hot side (2nd term) and minus the Joule heating (3rd term).

The electrical current I=J.S can be obtained from the following equation for the current density J:

$$J = \frac{1}{\rho_n}.\left(E - \alpha_n.\frac{\Delta T}{\Delta x}\right) \quad (2)$$

where ρ_n is electrical intensity, α_n is the Seebeck coefficient and E is electric field (E=U/Δx). The principle of operation of a thermoelectric generator (Fig. 2) lies in the Seebeck effect.

The Seebeck effect is a conversion of the temperature difference ΔT (between the hot and cold side) into the voltage difference ΔU (Seebeck voltage) between the sides. The heat flows from the hot side to the cold side [8, 9].

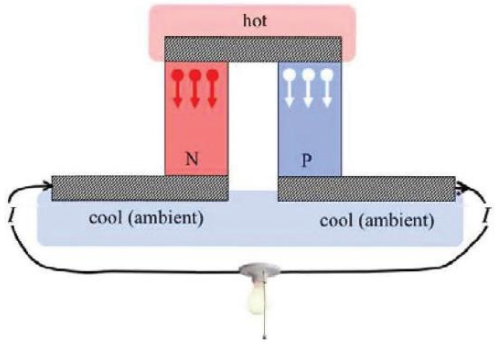


Fig.2 Schematic illustration of how a thermoelectric power generator operates. Using this device, the heat is converted into electrical power [3].

The Seebeck voltage can be described as follows:

$$\Delta U = \alpha_{pn}.\Delta T \quad (3)$$

where α_{pn} is the Seebeck coefficient of a thermoelectric power generator composed of a p-type and n-type semiconductor.

The electrical current I flowing the generator circuit can be obtained from the following equation:

$$I = \frac{\Delta U}{R_{total}} = \frac{\alpha_{pn}.\Delta T}{R_p + R_n + R_L} \quad (4)$$

where R_{total} is the total electrical resistance (of the generator electric circuit) composed of R_p (p-type semiconductor), R_n (n-type semiconductor) and R_L (load).

The power P supplied by I to the load R_L is:

$$P = R_L.I^2 \quad (5)$$

where R_L is the electrical resistance, and I is the electrical current.

III. FIGURE OF MERIT ZT

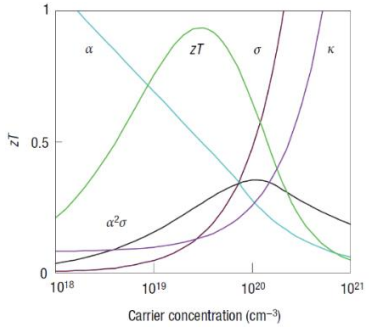


Fig.3 Seebeck coefficient α, electrical conductivity σ, thermal conductivity λ (κ), power factor PF=$\alpha^2.\sigma$ and ZT vs. carrier concentration in semiconductors [10].

For a thermoelectric cooler and thermoelectric generator, the efficiency of conversion of electrical energy into heat flow caused by electric charge carriers (and vice versa) can be described by the figure of merit ZT:

$$ZT = \frac{\alpha_{pn}^2.T}{(\lambda_p + \lambda_n + \lambda_{cl}).(\rho_p + \rho_n)} \quad (6)$$

where T is a mean temperature of a thermoelectric device, λ_p is thermal conductivity of the p-type, λ_n of n-type semiconductor and λ_{cl} is thermal conductivity of crystal lattice, ρ_p is electrical resistivity (p-type semiconductor) [11, 12, 13].

The lower electrical resistivity ρ (or higher electrical conductivity σ) reduces the losses caused by the Joule heating. The higher Seebeck coefficient α increases the amount of the transferred heat. The lower thermal conductivity λ reduces the back heat diffusion from the hot side to the cold side of a thermoelectric device (Fig. 3) [11, 12].

There are various methods of measuring ZT. Typically, one needs to measure all the parameters in the equation for ZT (α, λ, ρ). Another way is to measure ZT directly by the Harman method.

IV. THERMOELECTRIC MATERIALS

Thermoelectric materials enable a direct and reversible transformation of heat into electrical energy (and vice versa). This is the main reason why we study this type of materials. The main advantage of thermoelectric devices is the absence of moving parts, reliability, durability, quiet operation and environmental friendliness; therefore, they are used in a wide range of technical applications. Thermoelectric materials show a conversion efficiency of 5-20 %, which can be increased by doping with suitable elements, alloying, adding disperse phases to the matrix or nanostructuring [11].

Electrical conductivity of thermoelectric materials grows with the increasing number of charge carries. On the other hand, the Seebeck coefficient decreases with the growing number of these charge carries. Various materials show specific thermoelectric properties defining their application limits. Moreover, mechanical properties, the price of input raw materials, production costs and long-term stability are also essential in the application assessment of these materials. As already mentioned, the most important parameter describing thermoelectric properties of thermoelectrics is the figure of merit ZT; it represents the energy conversion efficiency of a thermoelectric and it is a function of T. According to ZT, thermoelectric materials can be classified into 2 groups [14].

The first group represents materials used for temperature measuring (thermocouples). The thermocouples are of E, J, K, M, N, T, B, R, S, C, D, G type etc . The value of ZT is of the order of 10^{-6}-10^{-3}.

The second group represents materials which are suitable for two main purposes. Firstly, they can be used as a source of electrical energy obtained by the conversion from thermal energy. Secondly, they can be applied for the transfer of heat by the electric field. These thermoelectric materials are produced by many different technologies. Their value of ZT is in the range from a few tenths to a few units. They can be inorganic or organic. Inorganic semiconductor materials are usually brittle at room temperature; this does not apply to thin inorganic thermoelectric layers applied to a substrate. On the other hand, organic semiconductor materials are naturally

flexible, but their thermoelectric properties are usually deficient.

One exception from inorganic materials is silver sulfide Ag_2S. It is a notable inorganic semiconductor material because it can be elastically deformed. It can be used there, where is expected mechanical load or vibrations.

Ag_2S alloy

The structure of Ag_2S shows polymorphism meaning its crystal structure varies with temperature. Monoclinic α-Ag_2S (acanthite, natural mineral) exists below 450 K, body centered cubic β-Ag_2S (argentite) exist in the temperature range of 451-844 K and face centered cubic γ-Ag_2S is stabile above 845 K up to the melting temperature of 1115 K. β-Ag_2S exhibits electron and ionic conductivity. The conductivity of β-Ag_2S sharply increases with the growing temperature (Fig. 4a). The sharp increase is caused by the ionic conductivity [15, 16, 17].

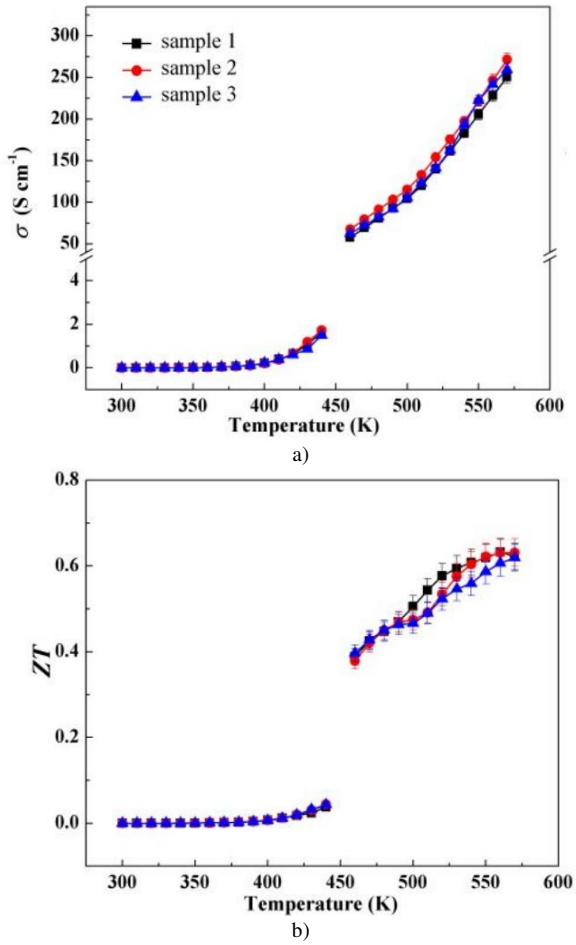


Fig.4 a) Conductivity σ of α-$Ag_{1.96}S$ and β-$Ag_{1.96}S$ vs. temperature, b) ZT for α-$Ag_{1.96}S$ and β-$Ag_{1.96}S$ vs. temperature [15] .

Monoclinic α-Ag_2S is a non-degenerated n-type semiconductor with a low electron concertation of 1.6×10^{14} cm^{-3} at 300 K and has a high negative Seebeck coefficient α at room temperature. Its electrical conductivity σ is in the range from 0.09 to 0.16 $S.m^{-1}$ and a thermal conductivity λ=0.5 $W.m^{-1}.K^{-1}$ for the polycrystalline state. The α-Ag_2S shows ZT~0 almost in the entire temperature range (300-450 K), β-Ag_2S shows ZT~0.6 at 575 K (Fig. 4b). Thus, Ag_2S without impurities is not useable as a thermoelectric material [18] .

For this reason, Se and Te (chalcogens) are added in pure Ag_2S. Se and Te exhibit similar chemical properties like sulphur, therefore they can be simply incorporated into the structure (unit cell) of Ag_2S. The value of ZT is 0.26 for $Ag_2S_{0.5}Se_{0.5;}$ 0.44 for $Ag_2S_{0.5}Se_{0.45}Te_{0.05}$ at 300 K and 0.63 for $Ag_2S_{0,8}Te_{0,2}$ at 450 K (Fig. 5d). These values are comparable or higher than those for organic thermoelectrics [15, 19].

These compositions represent a compromise between thermoelectric and mechanical properties of impurity Ag_2S. The value of the Seebeck coefficient α decreases with increasing conductivity σ, with the growing content of Se or Te and with temperature (Fig. 5b). The Seebeck coefficient α at 300 K is -123 $\mu V.K^{-1}$ for $Ag_2S_{0.5}Se_{0.5}$ and -136 $\mu V.K^{-1}$ for $Ag_2S_{0.5}Se_{0.45}Te_{0.05}$ which is approximately (for both values) one tenth of the α value for α-Ag_2S. Despite this decline of the Seebeck coefficient α compared to pure α-Ag_2S, the absolute values of α for impurity Ag_2S with the addition of Se and Te are significantly larger than the α values for organic thermoelectric materials, where the absolute values of α are at the level of several tens of $\mu V.K^{-1}$. Because the Seebeck coefficient α is negative for all the compositions, this means that electrons are the main charge carriers there [15].

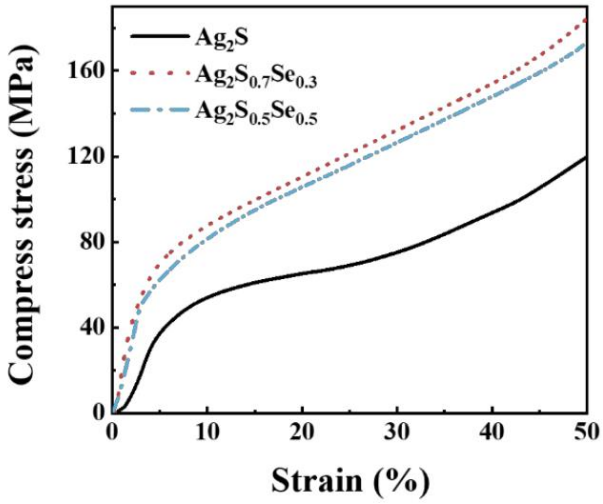


Fig.6 Compression stress vs. strain for various α-Ag_2S based samples [15].

The ratio of volumetric compressibility and shear strength is higher than 1.75 for all the compositions and the Poisson's ratio is higher than 0.26. Hence, Ag_2S based thermoelectrics are ductile materials. As shown by the following plot (Fig. 6), α-Ag_2S is more ductile compared to $Ag_2S_{0.7}Se_{0.3}$ and $Ag_2S_{0.5}Se_{0.5}$. Because of several orders of magnitude higher electrical conductivity σ of impurity Ag_2S samples compared to pure α-Ag_2S (Fig. 5a), the power factor $PF=\alpha^2.\sigma$ reaches a maximum value of 4.8 $\mu W.cm^{-1}$ for $Ag_2S_{0.5}Se_{0.5}$ and 5 $\mu W.cm^{-1}.K^{-2}$ for $Ag_2S_{0.5}Se_{0.45}Te_{0.05}$ at 300 K (Fig. 5c). This is roundly 4 orders of magnitude higher than for the pure α-Ag_2S. Compared to α-Ag_2S, the significantly higher values of ZT and power factor PF for impurity Ag_2S samples (with the addition of Se and Te) can be attributed to the growth in electron concentration. For $Ag_2S_{0.5}Se_{0.5}$ and $Ag_2S_{0.5}Se_{0.45}Te_{0.05}$, the concentration of electrons is at the level of 10^{18}-10^{19} cm^{-3} that is 4-5 orders of magnitude higher than for the pure α-Ag_2S [15].

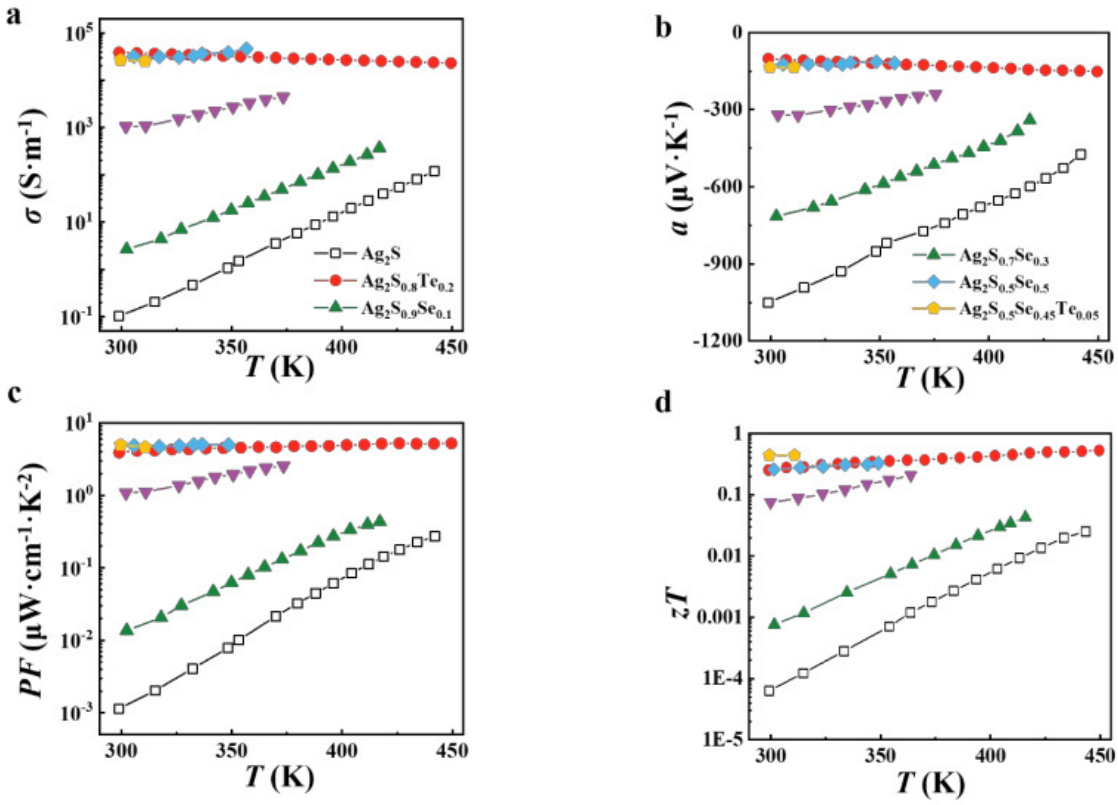


Fig.5 Electrical conductivity σ, Seebeck coefficient α, power factor PF and ZT vs. temperature of various α-Ag_2S based samples [15].

V. CONCLUSION

Silver sulfide Ag_2S is a very promising ductile thermoelectric material. It is challenge to investigate it. The purpose of my PhD. study is to improve its properties by adding impurities (Se, Te or other elements) and/or by modifying its production process (annealing of powders, rapid cooling of the melt, etc.). Addition of impurities significantly enhances thermoelectric properties (PF, ZT) of Ag_2S.

Newly developed Ag_2S based thermoelectrics can be potentially used for wearable electronics [20].

REFERENCES

[1] A. Van Herwaarden and P. Sarro, "Thermal sensors based on the seebeck effect," *Sensors and Actuators,* vol. 10, no. 3–4, pp. 321-346, 1986.

[2] H. Holdsmid, Introduction to Thermoelectricity Second Edition, Sydney: Springer Series in Materials Science, 2016.

[3] H. J. Goldsmid, Introduction to Thermoelectricity Second Edition, Berlin: Springer-Verlag Berlin Heidelberg, 2016.

[4] H. Lee, "The Thomson effect and the ideal equation on thermoelectric coolers," *Energy,* no. 56, pp. 61-69, 2013.

[5] L. HoSung, Thermoelectrics: Design and Materials, West Sussex: Wiley, 2017.

[6] H. Nettleton, "The Thomson Effect," *Proceedings of the Physical Society,* 1922.

[7] A. Ioffe, "Semiconductor thermoelement and thermoelectric cooling," *Infosearch Limited,* p. 184, 1957.

[8] H. Fujiki and Y. Amagai, "Measurement of the Thomson heat distribution in a thin-wire metal," *Precision Electromagnetic Measurements,* pp. 52-53, 2014.

[9] X. Zhang a L. Zhao, „Thermoelectric material : Energy conversion between heat and electricity," *J.Mater,* zv. 1, vyd.2, pp. 92-105, 2015.

[10] G. Snyder a E. Toberer, „Complex thermoelectric materials," *NatMater,* zv. 7, 1. vyd.2, pp. 105-114, 2008.

[11] S. Verma and J. Singh, "A Comparison of Figure of Merit for Some Common," *Global Journal of Researches in Engineering,* no. 13, 2013.

[12] M. Wolf, R. Hinterding a A. Feldhoff, „High Power Factor vs. High zT—A Review of Thermoelectric Materials for High-Temperature Application," *Entropy,* 2019.

[13] H. Jian a T. M. Tritt, „Advances in thermoelectric materials research: Looking back and moving forward," *Science,* zv. 357, vyd.6358, 2017.

[14] J. Singh and J. Verma, "A comparison of figure of merit for some common thermocouples in the high temperature range," *Global Journal of researches in engineering electrical and electronics engineering,* vol. 13, no. 11, pp. 1-7, 2013.

[15] T. Wang, , J. Liang, Q. Pengfei, Y. Shiqi, M. Chen, C. Hongyi, Q. Song, Z. Kunpeng, W. Tian-ran, R. Dudi, S. Yi-Yang, S. Xun, H. Jian a L. Chen, „Flexible thermoelectrics: from silver chalcogenides to fullinorganic devices," *Energy & Environmental Science,* 2019.

[16] A. Gusev a S. Sadovnikov, „Acanthite–argentite transformation in nanocrystalline silver sulfide and the Ag2S/Ag nanoheterostructure," *Semiconductor,* vyd.50, pp. 682-687, 2016.

[17] R. Sharma a Y. Chang, „The Ag-S (Silver-Sulfur) System," *Bulletin of Alloy Phase Diagrams,* zv. 7, . vyd.3, pp. 263-268, 1986.

[18] S. Xiaolei, Z. Jin a C. Yhi-Gang, „Advanced Thermoelectric Design: From Materials and Structures to Devices," *Chemical Reviews,* 2020.

[19] L. Lanwei , P. Chengxiao , C. Jing , M. Zheng , C. Yanqun , L. Shuyao , W. Jianli a W. Chao , „Study the effect of alloying on the phase transition behavior and thermoelectric properties of Ag2S," *Journal of Alloys and Compounds,* zv. 886, 2021.

[20] M. Malmivaara, „Wearable Electronics," *Science Direct,* 2009.

Analysis and Evaluation of Data Using Artificial Intelligence for Cybersecurity

[1]*Martin HASIN (2nd year),*
Supervisor: [2]*Martin Chovanec*

[1,2]Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Slovak Republic

[1]martin.hasin@tuke.sk, [2]martin.chovanec@tuke.sk

***Abstract*—When analyzing network traffic, you can select different methods by which network traffic can be analyzed. It is possible to implement various network analyzers that can thoroughly analyze network traffic at the layer 7 and categorize them according to predefined patterns. These applications include various Intrusion Prevention System and Intrusion detection system filters. Another method is to analyze network traffic using information about given flows. This information provides a picture of the network, where it is possible to detect who is communicating in the network and by which devices the communication takes place. Based on this information, it is possible to look for anomalies in emerging connections in this network operation. Every single device on the network corresponds to certain patterns of behavior that can be recorded and analyzed in detail using artificial intelligence. Using these models, it is possible to define possible deviations from the assumed behavior model of the respective system. Using computer learning, it is possible to create a model that describes the behavior in a given network.**

***Keywords*—Nfstream, NetFlow, Machine learning**

I. Introduction

The aim of this work is to create an automated system that will provide protection against similar incidents in the network. By analyzing network flows, be able to automatically and quickly detect potential anomalies in the network, which mark the beginning of a possible attack or an ongoing network attack. Based on this data, then offer the possibility of automatic intervention in the network to prevent the spread of the network attack [1].

When creating such a system, the data recording model that needs to be analyzed is important. The data that will enter the process comes from two main sources. The first source we can use to increase security is to record system logs from the applications that are used. The second source that enters the process is the network traffic record. The most common network traffic can be recorded using various protocols such as netflow. This data contains a thorough view of a large part of the network, as using only one of these sources we get an incomplete view of the behavior of the systems in the network. Using both system and application data, we can analyze various security breach attempts or obtain information about the spread of unwanted applications. By combining this data with network traffic data, it is possible to more thoroughly detect the sources of vulnerabilities or determine what types of attacks are carried out on a given network. By using artificial intelligence, it is possible to use fast decision-making and search for potentially unwanted connections in the network [2]. The output of this process is the implementation of a security policy that can prevent the further spread of security incidents.

II. Research environment architecture design

By introducing machine learning, it is possible to classify a given network traffic whether it is an attack or it is traffic that does not disrupt the network layer in any way. The proposed network architecture suitable for recording and analyzing network flows consists of several modules:

- Network probe module,
- Non-relational databases module,
- Database communication module,
- Data visualization and evaluation module.

These proposed modules communicate with each other and exchange the required data. The main module consists of creating a network probe that contains a machine learning model suitable for network traffic analysis. When implementing these modules, it is necessary to select a suitable location of the network probe. The connection of a network probe between the internet operator in the given network and the main firewall through which all clients connect to the given network devices seemed to be a suitable location. By placing the active probe in the environment, it is possible to ensure fast analysis without the need to reconfigure any network elements. However, such a probe increases the risk of failure and thus damage to the network. Another point is to use a monitoring port, for the creation of which it is possible to use Cisco Switched Port Analyzer (SPAN) technology, which will ensure the cloning of network traffic taking place on selected network ports of the switch or router.

III. Implementation of network probe model

The first part in creating an analysis is the implementation of a network probe. This server analyzes and decrypts the network traffic it receives on the network adapter. The network adapter connected to this server allows all network packets to be received without further processing by the operating system kernel. As a result, the packets are then automatically dropped and can only be read by the appropriate application programming (API) commands to the operating system kernel. Reading these packets from the operating system kernel uses the implemented libpcap library, which provides fast opening and analysis of received packet headers. Using the NFstream

tool, it is possible to perform network packet analysis [3]. Procedures that can be implemented in the packet analysis process can be divided into 3 categories, which are used to analyze the required data:

- nDPI packet analysis,
- Implementation of the Machine Learning model,
- Implementation of custom parameters for given models.

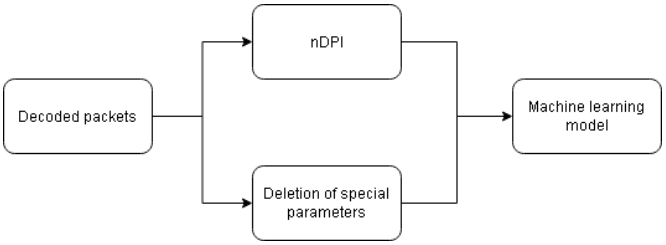


Fig. 1. Analysis procedure

In Fig. 1 is a model that analyzes packets, where the unpacked packets are compared with predefined patterns via the nDPI module, which can define the desired application. This information is then used as part of the information entering the analysis by machine learning. Packet analysis using nDPI depends to a large extent on the timeliness of the assigned pattern database, with which the required series of packets are compared. With this classification, it is possible to automatically classify the required protocols as harmful when communicating with certain Internet Protocol (IP) addresses [4].

The output data that is the result of the NFstream analysis must be stored in a fast database. The non-relational database elasticsearch was selected for such storage, which ensures document storage of data in the database. In our settings, the document means a detailed description of one network flow. The non-relational database elasticsearch enables communication using REST API HTTP commands, where it is possible to upload or otherwise edit the given records in the database [5]. Implementation into the NFstream environment can be done in two ways:

- implementation of own series of commands communicating using python environment with REST API elasticsearch database,
- implementing communication with the elasticsearch database using a python library called elasticsearch. This library allows communication with the database.

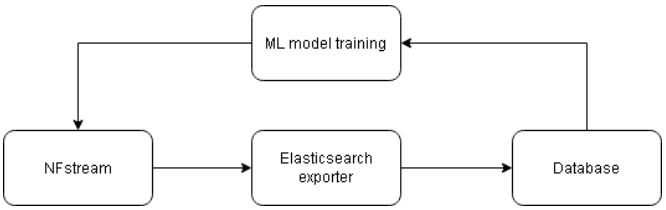


Fig. 2. Machine learning model

The nfstream application allows you to implement the created model of artificial intelligence into the analysis process. This model can classify data according to certain values and then mark them with the appropriate tag in the non-relational database. In Fig. 2 is a model of learning the ML algorithm directly from the data that were the output of the analysis of the model, where artificial intelligence is trained in its results, which enters data from external databases of known threats [6].

In Fig. 3 shows the process of dropping network packets on network adapters. Such discarding of network packets is caused by the operating system kernel, which automatically discards packets that are incorrectly shaped or corrupted.

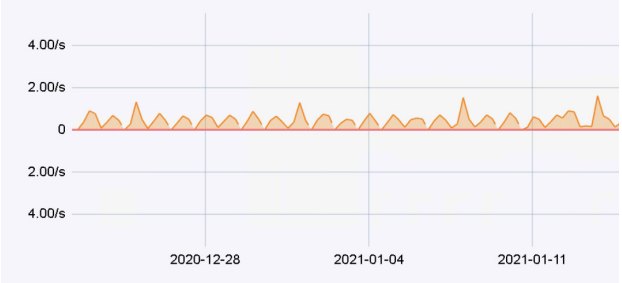


Fig. 3. Dropping network packets

Another point behind packet dropping is the impact of the topology of the libpcap library. When reading the contents of packets, this library copies the contents to the additional cache memory, where there is a higher risk of overloading or damage to the read packets. Such a large drop of network packets is also affected by the performance of the hardware with which we open the received packets. Such an analysis is based on a computation running on a single processor core. The goal is to minimize this number so that the data needed for analysis are not lost.

IV. CONCLUSION AND FUTURE WORK

Based on the analysis of methods of spreading current threats, methods were selected that can be used to implement artificial intelligence in the network analysis. These methods of artificial intelligence enable fast classification of given network flows, where the models are not dependent on the required pre-created databases with defined penetrations. By choosing such a model, there is no need to deal with lengthy and costly updating of databases of delivered applications for network analysis. External databases for detecting artificial intelligence intrusions are only supplementary data that help increase the efficiency of a given artificial intelligence algorithm. To build an experimental model of an application that performs network analysis, the nfstream element was used, which allows the implementation of machine learning in the packet group analysis process. The result of this analysis is an implemented infrastructure model based on the NetFlow python module, the output of which can then be indexed in the relevant non-relational databases. A collector was created for indexing data in the non-relational database elasticsearch, which ensures communication with nfstream and the database server using REST API commands.

The goals of the future work result from a previous analysis of current network security problems, problems with current detection and possible technologies. These objectives are defined as follows:

- Perform statistical, qualitative and quantitative data analysis on recorded network traffic.
- Select and deploy artificial intelligence models suitable for predicting and detecting network attacks resulting from the analysis of recorded data.

- Investigate the impact of configuration of flow measurement point parameters, with emphasis on early detection of anomalies in networks.

ACKNOWLEDGMENT

This publication has been published with the support of the Operational Program Integrated Infrastructure within project: Research in the SANET Network and Possibilities of Its Further Use and Development (ITMS code: 313011W988), co-financed by the ERDF.

REFERENCES

[1] Y. Fu, F. Lou, F. Meng, Z. Tian, H. Zhang, and F. Jiang, "An intelligent network attack detection method based on rnn," in *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*. IEEE, 2018, pp. 483–489.

[2] A. A. Galtsev and A. M. Sukhov, "Network attack detection at flow level," in *Smart Spaces and Next Generation Wired/Wireless Networking*. Springer, 2011, pp. 326–334.

[3] Z. Aouini and A. Pekar, "Nfstream: A flexible network data analysis framework," *Computer Networks*, p. 108719, 2022.

[4] G. B. Satrya, F. E. Nugroho, and T. Brotoharsono, "Improving network security-a comparison between ndpi and l7-filter," *International Journal on Information and Communication Technology (IJoICT)*, vol. 2, no. 2, pp. 11–11, 2016.

[5] R. Kuć and M. Rogozinski, *ElasticSearch server*. Packt Publishing Ltd, 2016.

[6] S. Qiu, Q. Liu, S. Zhou, and C. Wu, "Review of artificial intelligence adversarial attack and defense technologies," *Applied Sciences*, vol. 9, no. 5, p. 909, 2019.

Comparison of Agile Software Project Management Methods

[1]*Emira Mustafa Moamer ALZEYANI (1st year)*
Supervisor: [2]Csaba SZABÓ

[1,2]Dept. of Electrical Engineering and Informatics, FEI TU of Košice, Slovak Republic

[1]emira.mustafa.moamer.alzeyani@tuke.sk, [2]csaba.szabo@tuke.sk

***Abstract*—The Agile project management methodology has revolutionized project management. Before its emergence, many companies were facing problems in applying traditional methods of product delivery to their customers on time and without delay. This was in addition to the fact that customer requirement modifications can cause problems for the work team. The use of the agile methodology must solve these problems. But despite this remarkable development in the project management process, the data and requirements are also in continuous development, and this constitutes a new challenge in the development process and in terms of the large size of the team and their distribution over a wide geographic area. In this paper, the focus is on identifying the most important models that are dealt with in the case of the geographical distribution of teams. It will be identified in terms of the life cycle and features of the Agile methodologies and compared to the features to find the differences.**

***Keywords*—Agile methods, Scrum methodology, SAFe framework, DAD framework.**

I. INTRODUCTION

During the design or development of software applications, some steps must be passed to obtain the required product at the required speed, decorated and specified, and performs the desired requirements [7].

From this point of view, there are several methodologies or frameworks in the field of software engineering that are working under until the required application is implemented [6], and one of the most important methodologies in the field of software engineering is the agile methodology. Which is known for its flexibility in responding to changing requirements and the speed of delivery of the product to the required party [7], [12]. Also known as the style of working as a team and managing itself.

However, even the agile approach has many different methodologies or frameworks in the way of working, but they have the same goal, which is to deliver the required application in the required time [6].

II. METHOD OF THE RESEARCH

In software engineering or software development, projects have become largely dependent on the Agile methodology. Some studies show 73% of the development process is based mainly on Agile methodology [4], and the process of selecting the framework is carried out according to the project to be implemented, it is possible, after reviewing the project requirements [7], that the agile team will choose the appropriate business framework for the requirements. Some consider that Scrum is one of the most common methods of work, and in some studies, it is considered that SAFe is one of the most used methods in software development [5], especially companies or large institutions.

The most important agile approach that large organizations rely on will be addressed, and here is what is meant by large organizations that rely on a wide geographical scope in their work [13].

Scrum methodology, SAFe framework, and DAD framework are the most used in agile methods [4], [5]. All these methodologies are dealt with in the case of a large work team, and it is geographically distributed so that the team can work with the same high efficiency in the case of a small team and its geographical scope is limited [1], [15].

First, we will search of studies and understand the multiple agile methods through research on the existing studies in several scientific fields, and this will help us understand the models in the way they work and how to adapt in a particular environment. Because there are many institutions that work with agile methods and these institutions or companies all have laws and different styles of work [12]. This thing will generate a kind of knowledge of a few forms on it, and knowledge of the features in it, and through introductory books so that the life cycle of each model is understand.

1. This study focuses on three different methodologies: (Scrum methodology, Scaled Agile Framework (SAFe). Disciplined Agile Delivery (DAD)).
2. Research was carried out in several digital scientific journals by collecting the largest possible number of research papers on the topic through: (IEEE electronic library, Springer, ACM digital library, and Google Scholar).
3. It was based on certain keywords: (Agile project management, Scrum methodology, SAFe framework, DAD framework, advantage, and disadvantage of SAFe, and DAD frameworks, and the lifecycle of methods).

The questions formed by reading the articles are:

1. What are the features that distinguish all these most widely used methodologies?
2. How is it possible to collect these methodologies and design a model through the advantages obtained?

III. STUDY PLAN

In the first stage, the models will be identified in detail, as will be explained in the following sections.

The importance of this stage is due to the next stage, which is based on the application of the basic concepts of the selected models, on real data, by setting up a scenario that follows certain steps to get the desired result. The results have been analyzed and compared from several aspects that will be listed in the next stage. As shows in the Figure 1.

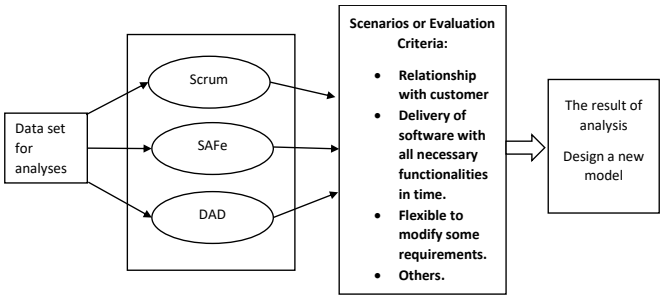


Fig. 1. The second stage of study

After the process of analysis and monitoring of results, the third stage begins, which is the stage of designing a model that depends mainly on the results obtained, focus will be on:

1. Compilation of the features found in the models on which the study was conducted.
2. Also, through the analysis process, it is expected that some defects will be revealed. which we must processed in the proposed design.

This model will be simulated on some scenarios by using on real data, until it is improved, and the desired result is reached.

IV. STUDIED METHODOLOGIES

In this section, we will describe the selected models:

A. Scrum methodology

Agile Scrum is one of the frameworks according to the agile methodology and is characterized by flexibility as it allows the customer to change requirements during implementation and start developing a **M**inimum **V**iable **P**roduct (MVP) [13].

It is based on union where team members work as one and the ability to self-motivate.

Scrum depends on dividing the work stages into small tasks and short periods called (Sprints) [10], the duration of which ranges from one to three weeks.

In Scrum, many people have important roles in implementing the needs and requirements by creating a plan to achieve the goal [13].

1. Product Owner: The person responsible for developing and defining the features of the product and its innovative ideas that are required to be available.
2. Scrum Master: He is responsible for leading the team, holding meetings, maintaining the progress of the production process, and implementing steps with accuracy, skill, and high efficiency.
3. Team Members: It consists of (Developer, Tester, Writer), and any other person who contributes to the implementation. Team members may rotate tasks, for example, the developer does the testing, and the tester completes the writing task, because the task of the team is to finish the work in the required time and in the right way.

B. Scaled Agile Framework (SAFe)

The Scaled Agile Framework (SAFe) is the most popular, most complex, and comprehensive Agile framework today [15]. The primary goal of SAFe is to facilitate the creation and growth of a Lean organization because it recognizes that many different types of companies are, in part, software companies that need to consistently deliver value in the shortest sustainable period [8].

SAFe identifies five core competencies for a Lean organization. Each competency is a set of relevant knowledge, skills, and behaviors [15]

1. Lean Leadership: Describes how leaders lead and sustain organizational change through learning, teaching, and applying the SAFe Lean-Agile mindset.
2. Team Spirit and Technology: Describes the skills, principles, and practices needed to create high-performance agile teams.
3. DevOps and Release on Demand: Describes how DevOps implementation and a continuous delivery pipeline provides organizations with the ability to issue product increments at any time necessary to meet demand.
4. Lean Business and Systems Engineering Solutions: Describes how to apply Lean-Agile principles and practices to the specification, development, deployment, and evolution of large and complex software applications.
5. Lean Portfolio Management: Balances strategy and execution through the application of lean and systems thinking approaches in strategy financing, investment finance, agile portfolio operations, and governance.

SAFe provides three organization levels as follows [15]:

1. Team Level: SAFe is based on Agile teams, each of which is responsible for defining, building, and testing stories from their backlog. Teams employ Scrum or Kanban methods, augmented by quality practices, to deliver value in a series of synchronized, fixed-length iterations.
2. Program Level: SAFe teams are organized into a virtual program structure called the “Agile Release Train” (ART). Each ART is a long-lived, self-30 organizing team of 5 to 12 Agile teams along with other stakeholders that plan, commit, execute, inspect adapt, and deliver solutions together.
3. Portfolio Level: The portfolio level organizes and funds a set of value streams. The portfolio provides solution development funding via Lean-Agile budgeting and provides necessary governance and value stream coordination.

C. Disciplined Agile Delivery (DAD)

When we are thinking about agile, it is thought of, on the basis that it is a specific model that helps us carry out specific work in a certain way and within a certain framework to help us deliver the product quickly to the desired place [1]. However, here we are talking about something that is not a

specific framework, but rather we are talking about a set of the tool kit. Here it is possible to think about how it will be an environment for work because, in the end, we need a certain structure to work under it. It gives you the basic building blocks, there are large blocks that help us create a large structure, and there are small building blocks that help us get things done on a regular basis.

In general, it is a life cycle to accomplish some tasks; it can give us multiple options for the life cycle [9], [18]. This lifecycle is like the outer layer that you can choose according to your requirements and change it according to your requirements in the same product field.

Main roles:

Principles and main components DAD have more roles than scrum and is divide into two categories of team roles. People who work with the project on an ongoing basis fill primary roles [18]. Secondary roles are usually offered on a temporary basis to assist the team with expansion or other issues. DAD has these additional roles because it handles the entire solution delivery lifecycle and because it learns about the different types of temporary and needed supporting roles that exist in the real world [9], [18].

V. COMPARISON

This section shows the comparison of the models, and we will present the features of this methodology:

A. Scrum methodology

1. The participation of all stakeholders benefiting from the product development process gives a better view of stakeholders and helps ensure that everyone's expectations are effectively manage [13], [10].
2. Transparency is available in the construction or development of the product at all stages of development [10].
3. By dividing the processes to be implemented to build a product, it is easy and quick to discover flaws in the construction process.
4. By involving stakeholders and breaking down the construction process into small parts, this helps reduce the cost of the process [3].

B. Scaled Agile Framework (SAFe)

1. It contains the most specific effective roles in the development process that help in appropriate communication and more work productivity [15].
2. Speed in the time of product access to the labor market.
3. Ease of moving the organization in the work methodology from the traditional methodology to the SAFe methodology in the case of it works in an agile way [8].
4. There is good communication between the team members [8].

C. Disciplined Agile Delivery (DAD)

1. It has flexibility and adaptability according to the size of the organization and business requirements [18].
2. The main job is to deliver the required application faster.
3. It pays attention in detail to all aspects of the institution and does not have any bias towards a particular party [8].
4. Gives good guidance on the processes that best fit the project, architecture, and development processes [8].
5. There is good communication between the team members.
6. The complexity of the work on it is considered medium, that is, the possibility of adapting the work quickly [8], [18].

VI. DATA SETS IDENTIFIED DURING THE RESEARCH

This paragraph displays the data sets, which we will use to conduct analysis on the data sets for the next stage. The goal of choosing multiple sources for a data set is to have the possibility to work in different environments of software development.

A. PROMISE Software Engineering Repository

in this library, you can get access to an open source and a variety of data sets that help in the software development process [20].

B. ISBSG

(The **I**nternational **S**oftware **B**enchmarking **S**tandards **G**roup) maintains a repository of data about completed software projects. A common use of the ISBSG dataset is to investigate models to estimate a software project's size, effort, duration, and cost. This dataset has data collected over a decade and is based on suitable metrics. About 4106 projects, with each project having a total number of 105 attributes divided into 18 sub-attributes [21].

C. TAWOS

(**T**awosi **A**gile **W**eb-based **O**pen-**S**ource) dataset. It encompasses data from 13 different repositories and 44 projects, with 508,963 issues contributed by 208,811 users. The dataset is publicly hosted on GitHub [23] as a relational database [22].

D. Kemerer

The Kemerer data set includes 15 software projects described by 6 independent attributes and one dependent attribute. The independent attributes are represented by 2 categorical and 4 numerical attributes. The effort attribute is measured by 'man-months' [24].

VII. CONCLUSION

We can say that all methods meet the same advantages in the performance of work in terms of speed of delivery of the required and communication with the customer as well, flexibility in changing requirements or developing requirements, as well as the quality of communication between team members, but there are some differences that can be answered about our previous questions:

1. Scrum is one of the simplest models that can be dealt with in terms of complexity, ease of involvement, and highly efficient at work, but at the same time it does not meet all the requirements, in terms of the large size of the team or the way to deal with it in the case of the geographical distribution of the team on a large scale.

2. SAFe framework is one of the most complex frameworks in the way of working, it has many strict rules that must be dealt with while developing a particular product, it is highly structured, which leads to increased hierarchy and decreased flexibility and agility, making it less adaptable.
3. DAD is considered one of the least complex frameworks that can be dealt with, due to the presence of more than one framework under DAD framework that can be worked on according to the system to be implemented or according to the team. And the team can work on more than one project is the same time, and each project has a special framework in line with the requirements of product.

We can say, at this stage, that the flexibility property in the DAD framework will be relied upon to design the proposed model.

Further research will be based on analysis of datasets presented in Section VI. The aim will be to detect features from these datasets to create a model with an acceptable fidelity. As the final goal is a simulation model based on the meta model in [7], the list of its parameters might be very different to the intuitive ones, namely software size, productivity, complexity, and requirement stability.

REFERENCES

[1] Edmunds, A., Olszewska, M., & Waldén, M. (2016). Using the Event-B Formal Method for Disciplined Agile Delivery of Safety-critical Systems. The Second International Conference on Advances and Trends in Software Engineering, (pp. 1-9).

[2] Carilli, J. F. (2021). THE PERCEIVED EFFECTIVENESS OF THE SCALED AGILE FRAMEWORK® IN SOFTWARE DEVELOPMENT ORGANIZATIONS. College of Technology, Indiana State University, 1-14.

[3] Larman, C., & Vodde, B. (2017). Large-Scale Scrum: more with LeSS.

[4] Remta, D., & Buchalcevova, A. (3 March 2021). Product Owner's Journey to SAFe®—Role Changes in Scaled Agile Framework. MDPI journal. doi: https://doi.org/10.3390/info12030107.

[5] Salikhov, D., Succi, G., & Tormasov, A. (2020). An Empirical Analysis of Success Factors in the Adoption of the Scaled Agile Framework – First Outcomes from an Empirical study. 46th Euromicro Conference on Software Engineering and Advanced Applications (p. 4). Portorose, Slovenia: IEEE. Doi: https://doi.org/10.48550/arXiv.2012.11144

[6] C.Goodpasture, J. (2010). Project management the agile way : making it work in the enterprise. J.Ross.

[7] Silva, I. J., Rayadurgam, S., & Heimdahl, M. P. (2015). A Reference Model for Simulating Agile Processes. ACM Digital Library, 82-91. doi: https://doi.org/10.1145/2785592.2785615.

[8] Christopher, L., & Vries, M. d. (2020). SELECTING A SCALED AGILE APPROACH FOR A FIN-TECH COMPANY. conference of the Southern African Institute for Industrial Engineering (SAIIE), (pp. 196-208). doi: https://doi.org/10.7166/31-3-2432.

[9] Kazi, L., Ivkovic, M., Radulovic, B., Bhatt, M., & Chotaliya, N. (2015). The Role of Business Process Modeling in Information System Development with Disciplined Agile Delivery Approach. ICIST 5th International Conference on Information Society and Technology, (pp. 489-522). Retrieved from http://www.eventiotic.com/eventiotic/files/Papers/URL/icist2015_90.pdf

[10] Mahalakshmi, M., & Sundararajan, D. M. (2013). Traditional SDLC Vs Scrum Methodology – A Comparative study. IJETAE, 192-196. Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.413.2992&rep=rep1&type=pdf.

[11] Mahdi, M. N., Zabil, M. H., Ahmad, A. R., Ismail, R., Yusoff, Y., & Naidu, H. H. (2021). Software Project Management Using Machine Learning Technique—A Review. MDPL, 2-39. doi: https://doi.org/10.3390/app11115183.

[12] Paasivaara, M., & Kruchten, P. (2020). Agile Processes in Software Engineering and Extreme Programming –Workshops. Springer International Publishing. Retrieved from https://library.oapen.org/viewer/web/viewer.html?file=/bitstream/handle/20.500.12657/42566/2020_Book_AgileProcessesInSoftwareEngine.pdf?sequence=1&isAllowed=y

[13] Agile Scrum Methodology. (2021, 12). Retrieved from Digite: https://www.digite.com/agile/scrum-methodology/#:~:text=Scrum%20is%20an%20agile%20development,an%20iterative%20and%20incremental%20processes.&text=The%20primary%20objective%20of%20Scrum,collective%20responsibility%20and%20continuous%20progress.

[14] Vick, K. T. (2021). An Investigation of the Scaled Agile Freamework and the Accountabitlity framework for Telehealth Digital Solution. Golden Gate University, 1-12.

[15] Yakyma, A., Knaster, R., Jemilo, D., & Oren, I. (2016). SAFe REFERENCE GUIDE. United States.

[16] Stoshikj, M., Kryvinska, N., & Strauss, C. (2013). Project Management as a Service. IIWAS '13: Proceedings of International Conference on Information Integration and Web-based Applications & Services (pp. 220-228). ACM Digital library. doi: 10.1145/2539150.2539171

[17] Kalenda, M. 2017. Scaling Agile software development in large organizations. Master's thesis in Faculty of Informatics, Masaryk University.

[18] Project Management Institute. Introduction to Disciplined Agile Delivery (DAD). Available URL: https://www.pmi.org/disciplined-agile/process/introduction-to-dad/full-delivery-lifecycles-introduction

[19] What is SAFe? (2021, 12). Retrieved from scaled agile: https://scaledagile.com/what-is-safe/

[20] Song, L., Minku, L. L., & Yao, X. (2013). The Impact of Parameter Tuning on Software Effort Estimation Using Learning Machines. Proceedings of the 9th International Conference on Predictive Models in Software Engineering, (pp. 1-10). doi: https://dl.acm.org/doi/10.1145/2499393.2499394

[21] Desai, V. S., & Mohanty, R. (2019). ANN-Cuckoo Optimization Technique to Predict Software Cost Estimation Software Cost Estimation. 2018 Conference on Information and Communication Technology (CICT'18). IEEE. doi: 10.1109/INFOCOMTECH.2018.8722380

[22] Tawosi, V., Al-Subaihin, A., Moussa, R., & Sarro, F. (2022). A Versatile Dataset of Agile Open Source Software Projects. Proceedings of the 19th International Conference on Mining Software, (pp. 1-5). Retrieved from https://arxiv.org/abs/2202.00979v1

[23] Vali Tawosi, Afnan Alsubaihin, Moussa Rebecca, and Federica Sarro. [n.d.]. The TAWOS dataset. URL: https://github.com/SOLAR-group/TAWOS.git

[24] Azzeh, M. Y. (2010). Analogy-Based Software Project Effort Estimation. PhD thesis.

Methods of data hiding in images

[1]*Samuel ANDREJČÍK (1st year)*
Supervisor: [2]Ľuboš OVSENÍK

[1,2]Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

[1]samuel.andrejcik@tuke.sk, [2]lubos.ovsenik@tuke.sk

***Abstract*— Many steganographic algorithms have been proposed until these days. They all try to hide information by relying on some of well-known techniques. In order to improve the security of steganography, new algorithms have been developed in recent years to hide information by relying on some of those techniques. However, all information hiding techniques have their advantages as well as disadvantages. In this paper, will be described image processing in steganography using the LSB method, in which the information is written to the Least Significant Bit and we compare the PSNR parameter, in which it is important to achieve the lowest possible values.**

***Keywords*—Image processing, information hiding, LSB, security, steganography**

I. INTRODUCTION

At the time we live, we use the internet daily. One of the most used functions of the Internet is communication between users. Users often share personal and copyrighted materials. The security of transmitted data is therefore a key issue today [1]. In Mahmood´s et. al. work [2] can be found that in order to declare the terms in steganography, it must be said that the original image is referred to as a cover image, while the image obtained after entering the secret data is known as a stego image. The basic condition in steganography is that a particular recipient should be able to obtain information that has been hidden. However when we look back to the 5th century BC, it can be said that steganography is one of the oldest methods of concealing secret information. According to the classical author Herodotus, it was first used by the tyrant Histiaeus, who shaved the head of a servant before tattooing a message on the slave man's scalp. At the moment, when his hair grown back, the servant was sent to deliver his message - message encouraged Aristagoras to start a revolt against the Persian king. Hidden message was revealed when the servant's head was shaved again [3]. Rabbits were also used in a similar way. The secret message was hidden on their shaved stomachs, and subsequently the message was revealed during a repeated shave [4].

Another example of steganography can be found in the story of Demeratus, This story describes alert for Sparta with message that Persian Great King Xerxes is planning invasion to them. The wax coated tablets were used to hide a message. To preserve the messages secrecy, the wax coat was removed off the tablets and the message was engraved on the underlying wood. After writing the secret message, the tablet was again covered with wax. As there were no apparent evidence for the presence of message in the tablets, without question they easily passed inspection by sentries. The embedded message can only be decoded by scraping away all of the wax [5].

Other examples include that of the ancient Chinese. The Chinese would write messages on fine silk, which was then scrunched into a tiny ball and covered in wax. The messenger would then swallow the ball of wax [6]. In the fifteenth century, the Italian scientist Giovanni Porta described how to conceal a message within a hard-boiled egg by making an ink from a mixture of one ounce of alum and a pint of vinegar. This ink was used to write on the shell of an egg. The solution penetrated the shell, and leaved a message on the surface of the hardened egg albumen, which can be read only when the shell is removed [7].

Spartan scytale was tool used for steganography purpose. It was made back in the fifth century B.C. The principle of the tool was that a message was written on a leather strip or parchment that was wrapped around the tube, as shown in Fig. 1. The message symbols were written below each other around the perimeter of the tube. Subsequently, other random symbols were added to the tape. After unwinding the strip, it was a meaningless sequence of characters. The strip could be easily transferred, and when wound on another message receiver tube with the same diameter, the message was re-exposed [8] [9].

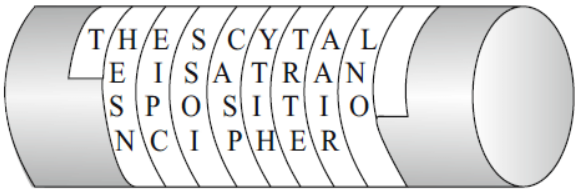


Figure 1 When strip was unwound from the sender´s scytale (tool), nobody understood sequence of characters. After rewinding strip to recipient´s scytale with correct diameter, message can be reappeared.

Even in such important incidents as terrorists attacks undoubtedly are, there can be seen signs of steganography. There is one example of steganography from recent history. In 2001 terrorists attacks were made by the militant Islamist terrorist group Al-Qaeda against the United States. In this particular case, it was a video made by terrorist Dhiren Barot, who captured places between Broadway and South Street and after that hid video into a copy of the Bruce Willis movie Die Hard: With a Vengeance [10].

Embedding messages within other text is also a way how to hide data. It called a null cipher. A null cipher is a type of hidden message where the real message is hid in an innocent sounding message. A famous example of a null cipher is one

sent by a German Spy in WWII. The message was: *"Apparently neutral's protest is thoroughly discounted and ignored. Isman hard hit. Blockade issue affects pretext for embargo on by products, ejecting suets and vegetable oils."* Innocent message by the first look, but after we will look closely and take the second letter in each word the following message emerges: *"Pershing sails from NY June 1"* [11]. A similar case in the same principle is message: *"Fishing freshwater bends and saltwater coasts rewards anyone feeling stressed. Resourceful anglers usually find masterful leapers fun and admit swordfish rank overwhelming anyday."* When we take the third letter in each word, message is: *"Send Lawyers, Guns, and Money"* [12].

Below in this work, in chapter Related Works you can find information about current situation in steganography, where will be presented important parameters and compare known methods. Chapter Future Work will discuss about our plans for near future, where we will focus on known algorithms with purpose of make them better - faster, more stable, or smoother. We will also study unsolved problems and work on them.

II. Related Works

Nowadays, the importance of image processing techniques is growing. New and faster algorithms are created mainly due to the growing demands of users who use multimedia services such as video calls, watching satellite TV or playing videos on the Internet as their standard.

One of the applications of image encryption algorithms is image preprocessing, which serves as a secret message in steganographic systems. Such systems serve to conceal the very existence of secret communication by modifying the cover data with a secret message. The block diagram of the steganographic system, where the cover data and the secret message are represented by the images, is shown in Fig. 2. Such systems are also referred to as image steganography systems.

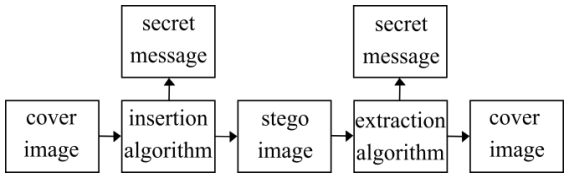


Figure 2 Block diagram of the image steganographic system.

In order to meet the security requirements of the entire system, the processing of the secret message by encrypting it was used before inserting the secret message into the cover image (see Fig. 3). Since the secret message is quite often represented by images, image encryption algorithms can be used to encrypt it. By encrypting the secret message images, in addition to the required bit distribution (described below in the F5 algorithm), the security level of the entire steganographic system is increased.

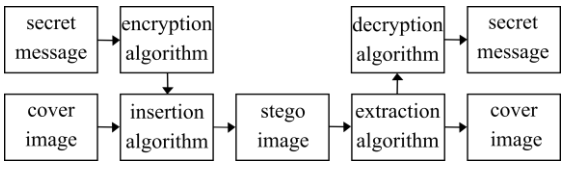


Figure 3 Block diagram of an image steganographic system with encryption.

One of the first parameters that we will examine in this work is LSB - Least Significant Bit. It is a bit whose position indicates a value which is presents by 0 or 1. The LSB is sometimes referred to as the right-most bit (see Fig. 4), due to the convention of writing less valid digits to the right. The storage of bits in the computer's RAM is defined in the reverse order of standard writing. This notation is also known as endianness. The binary notation in Matlab platform is also written in reverse order, which means that the LSB position is on the left side. For those, who made mistakes because of it Matlab provides function *flip* which turns the binary notation into the form known from practice.

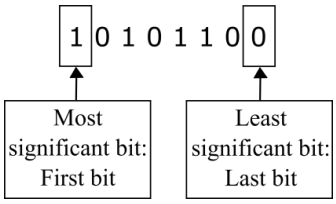


Figure 4 Binary notation of decimal number 172 with graphical highlighting of MSB and LSB.

One of the best known ways to modify cover images is to modify the LSB intensities of the pixels [13]. In the first technique, the entire pixel LSB plane is replaced with the secret message image or this plane is changed to correspond to the secret message image. This approach takes advantage of the fact that the LSB plane has little effect on the resulting pixel intensities. In addition, there were considerations of a small degree of correlation between the LSB plane and other planes, but these were refuted. Techniques working with the LSB level of pixel intensities were later analyzed and several ways to reveal their use have emerged [14].

±1 embedding technique was mentioned in Fridrich´s work [13]. It is trivial modification of LSB embedding, also known as LSB matching. This technique was discovered because of problems with LSB flipping method. It was problematic method due to its asymmetry – even values never decreased and odd values never increased during insertion. ±1 embedding method on other side is symmetrical operation. When a LSB needs to be change, this method randomly increases or decreases LSB, instead of flipping LSB. There is an exception with values 0 and 255, where those values are only increased, respectively decrease. These operations result into the LSB being modified, but other bits may be modified at the same time. There is also a case where all bits can be modified, such as when the value $127 = (01111111)_2$ is changed to $128 = (10000000)_2$. The ±1 embedding is much more difficult to attack than LSB embedding. At the time of writing Fridrich´s paper, no attacks to ± 1 embedding with comparable accuracy were known compared with LSB embedding [13].

One of the first steganographic algorithms to modify LSB of transform coefficients is the JSteg algorithm (also known as Jpeg-JSteg) by D. Upham. JSteg uses all transform coefficients except those having a value of 0 or 1 when inserting a secret message. Changing the LSB of the transform coefficients with these values would result in significant differences between the stego image and the cover image. The insertion of the secret message in the JSteg algorithm is realized by changing the LSB of transformation coefficients

so that this bit is identical with the currently read bit of the secret message. The properties of stego images created using the JSteg algorithm were later analyzed and several steganalytic techniques were developed that are able to detect the use of the JSteg algorithm [15] [16] [17].

Based on the JSteg algorithm, other steganographic algorithms were later developed, e.g. F5 embedding algorithm or OutGuess [18]. These and newer algorithms generally require an even distribution of secret message bits (approximately half the number of bits with a value of 0 and second half with value of 1).The premise of creating the F5 algorithm was to overcome the histogram attack while offering large embedding capacity. The F5 proposal contains two important components - the embedding operation and the matrix embedding. The previous operation of LSB flipping was replaced with method of decrementing the absolute value of the DCT (Discrete Cosine Transform) coefficient by one. This method can protect natural shape of DCT histogram after embedding with coefficient distribution looking as the lower quality factor was used on compressed cover image [13].

Joshi´s, Yadav´s and Allwadhi´s paper [19] describe detailed study of LSB Steganography method and analyzes the PSNR (Peak-Signal-to-Noise-Ratio) and MSE (Mean Square Error) of LSB data hiding techniques. The proposed LSB scheme works on principle that takes the first LSB bit of the gray scale image as well as first message bit of message matrix and after that those two bit embed the message into the original image. After this insertion of first message bit, pixel location of image and message is incremented by one. This process is running in algorithm until the message length is not equal to zero. Their results show that PSNR is higher for the small message and lower in case of large message size. The MSE is lower for small message size and higher in case of large data size. After increasing the size of the message the PSNR is decreased and the MSE is increased. Therefore the PSNR is inversely propositional to the MSE [19].

PSNR and MSE are the most common parameters in steganography. Those parameters are used for quality measurement of two images in steganographic system. The PSNR can provide information about similarity between two images and is reciprocal to MSE, which means that MSE is reverted value to PSNR. The higher the PSNR, the better the quality of the compressed is, or reconstructed image. The lower the value of MSE, the lower the error is. PSNR is presented in decibels [19].

Another technique should be mentioned is PVD, what is shortcut for Pixel Value Differencing. In Wu´s and Tsai´s work [20] is this steganographic method described as new and efficient method for embedding secret messages into a gray-valued cover image. A cover image is split into non-overlapping blocks of pixels which are consecutive. Two values of two pixels in each block are used for calculation of difference value. After that all difference values are classified into a number of ranges. The selection of the range intervals is based on the characteristics of human vision´s sensitivity to gray value variations from smoothness to contrast. After this process difference value is replaced by new value of hidden message. The method is overflow resistant, what means, that number of bits which we adding into pixel pairs are limited by width of the range that the difference value belongs to. This method is characterized by providing unobtrusive results compared to the LSB method. In order to demonstrate the security of this method, dual statistics attacks were made on the proposed solution. This method also offers an easy way to accomplish secrecy and can be extended to efficiently carry messages which related to a content, for example captions or annotations in audios and videos [20].

Principles described in Chang´s et. al. publication [21] talking about novel tri-way pixel-value differencing which provides larger capacity for the hidden secret information as much as an imperceptible stego-image for possibility of human vision. The original PVD method is working with one direction but solution presenting in this work is working with principle of three different directional edges for more effective tri-way PVD. In their work they presented adaptive rules and an optimal approach of selecting the reference point which leads to reducing the quality distortion of stego-image brought from setting larger embedding capacity. An advantage against original proposal is also that embedded confidential information can be extracted from stego-images without the assistance of original images. In general, we can declare that by using three different directional edges we can hide more secret data to the cover image if we compare this method with known original PVD method. Results of this method demonstrated that secret data in stego-image are invisible for human eye by comparison with covered image and after the detection test for the hidden data made by dual statistical stego-analysis proposed method survived and proved robustness to avoid the data detection [21].

III. Future Work

In the near future, we plan to explore the field of steganography in more depth; we will monitor changing parametric values, such as MSE, PSNR and others. We will be testing techniques mentioned in this paper, such as PVD, LSB flipping, LSB embedding, ±1 embedding, F5 embedding algorithm and others. Those algorithms can be used in various applications that require data representation in image format. These applications include e.g. image steganography, transmission and archiving of biometric features or provision of sensitive information in medical images. Many of the proposed solutions solve the complicated problems of hiding data in the image, there are several solutions to one problem and each offers a new perspective on the issue. Sometimes the result of the proposed solution is an algorithm that has better parameter values, other times it is an algorithm that offers a faster, more stable, smoother solution. The image encryption can be considered a promising area of research, as new approaches increase the security or speed of encryption build on new knowledge in the field of analysis of such algorithms. The further development of image encryption algorithms is also aided by the growing number of applications of image encryption algorithms, especially in steganographic systems.

In global, our task will be to come up with new proposals of image steganography algorithms and experimentally discovered characteristics and properties, as much as comparing them with already known and achieved results. We also will look up for unsolved problems of proposed methods and try to fix them for better results. In the future, we also do not rule out the possibility of combining steganography with encryption algorithms, testing them and comparing the results of the proposed solutions with existing ones.

ACKNOWLEDGMENT

This work was supported by the research project FEI-2022-84 "Data Processing Techniques in High Speed Transmission Systems".

REFERENCES

[1] A. Rehman, T. Saba, T. Mahmood, Z. Mehmood, M. Shah, A. Anjum, "Data hiding technique in steganography for information security using number theory," in. *J. of Information Science,* Saudi Arabia, 2018.

[2] T. Mahmood, Z. Mehmood, M. Shah, T. Saba, "A robust technique for copy-move forgery detection and localization in digital images via stationary wavelet and discrete cosine transform." in *J. of Vis. Commun. and Image Represent.*, Pakistan, 2018.

[3] M. R. N. Torkaman, N. S. Kazazi, A. Rouddini, "Innovative Approach to Improve Hybrid Cryptography by Using DNA Steganography," in *Int. J. on New Computer Architectures and Their Applications,* Malaysia, 2012.

[4] N. Alabdali, S. Alzahrani, "An Overview of Steganography through History," in *Int. J. of Scientific Engineering and Science,* Kingdom of Saudi Arabia, 2021, pp. 41-44.

[5] E. Zielińska, W. Mazurczyk, K. Szczypiorski, "Trends in steganography," in *Communications of the ACM,* New York, 2014, pp. 86-95.

[6] S. Arora, S. Anand, "A New Approach for Image Steganography using Edge Detection Method," in *Int. J. of Innovative Research in Computer and Communication Engineering*, India, 2013.

[7] E. Carter, G. Blank, J. Walz, "Bringing the breadth of computer science to middle schools," in *SIGCSE '12: Proceedings of the 43rd ACM technical symposium on Computer Science Education*, New York, 2012, pp. 203-208.

[8] F. Khan, A. A.-A. Gutub, "Message Concealment Techniques using Image based Steganography," in *The 4th IEEE GCC Conference,* Bahrain, 2014, pp. 11-14.

[9] S. Singh, The Code Book: How to Make It, Break It, Hack It, Crack It. New York, NY: Delacorte Press, 2002.

[10] J. Fridrich, Steganography in Digital Media. Binghamton, NY: Cambridge University Press, 2014.

[11] I. Karadoğan, R. Daş, "An Examination on Information Hiding Tools for Steganography," in *Int. J. of Information Security Science,* Turkey, 2014, pp. 200-208.

[12] M. S. Chandini, "An Overview about a Milestone in Information Security: STEGANOGRAPHY," in *Int. J. of Progressive Research in Science and Engineering,* India, 2020, pp. 33-35.

[13] J. Fridrich, "Steganography in digital media: Principles, algorithms and applications," in *Cambridge University Press*, Cambridge, 2009, pp. 466.

[14] J. Fridrich, M. Goljan, R. Du, "Reliable Detection of LSB Steganography in Color and Grayscale Images," in *Proceedings of the 2001 Workshop on Multimedia and Security MM&Sec '01,* Canada, 2001, pp. 27-30.

[15] S. K. Muttoo, S. Kumar, "Data Hiding in JPEG Images," in *BVICAM's Int. J. of Information Technology*, New Delhi, 2009, pp. 13-16.

[16] A. Westfeld, A. Pfitzmann, "Attacks on steganographic systems: Breaking the steganographic utilities EzStego, Jsteg, Steganos and S-Tools – and some lessons learned," in *Proceedings of 3rd International Workshop on Information Hiding IH'99,* Germany, 1999, pp. 61-76.

[17] T. Zhang, X. Ping, "A fast and effective steganalytic technique against JSteg-like algorithms," in *Proceedings of the 2003 ACM symposium on Applied computing SAC'03,* USA, 2003, pp. 307-311.

[18] A. Westfeld, "F5 – A steganographic algorithm: High capacity despite better steganalysis," in *Proceedings of Information Hiding,* USA, 2001, pp. 289-302.

[19] K. Joshi, R. Yadav, S. Allwadhi, "PSNR and MSE Based Investigation of LSB," in *Int. Conference on Computational Techniques in Information and Communication Technologies*, India, 2016.

[20] D.-Ch. Wu, W.-H. Tsai, "A steganographic method for images by pixel-value differencing," in *Pattern Recognition Letters,* Taiwan, 2003, pp. 1613-1626.

[21] K.-Ch. Chang, Ch.-P. Chang, P. S. Huang, T.-M. Tu, "A Novel Image Steganographic Method Using Tri-way Pixel-Value Differencing," in *J. of Multimedia,* Taiwan, 2008, pp. 37-44.

Energy consumption of UAV when hovering

[1]*Marek RUŽIČKA (3rd year),*
Supervisor: [2]*Juraj GAZDA*

[1,2]Dept. of Computers and Informatics, FEI TU of Košice, Slovak Republic

[1]marek.ruzicka@tuke.sk, [2]juraj.gazda@tuke.sk

***Abstract*—The optimization of the energy consumption of unmanned aerial vehicles (UAV) is crucial to ensure the longest possible flight time. In this paper, we analyze the scenario of hovering over a static place according to the power requirements. The knowledge concluded in this paper can later be used to optimize the energy consumption of a UAV.**

***Keywords*—UAV, energy consumption, optimization**

I. Introduction

The energy consumption optimization of battery enhanced UAV is important to improve the flight duration. In the task of UAV based coverage of end users, represented by user equipments (UE), the UAV doesn't travel much from place to place such as in other task as goods delivery or crops monitoring. The UAV often stands still - at the same place - holding only it's vertical position. This state is called hovering. UAV hovers over a static area to provide good channel gains. There was a lot of research published on the topic of UE coverage by UAVs.

We consider the UAVs to be used as relay stations such as described by Najla et al. [1]. In their paper, they provided a way to improve the channel gain between UE and static base station (SBS). In this scenario, UAV acts as a switch, that just forwards the connection between SBS and UE. It is considered that the UAV is hovering on a fixed position after the most optimal position according to the channel gain is determined. This, however, doesn't take into consideration the energy consumption of the UAV. On one hand, the energy consumption is lower if the UAV is positioned to the place where it has a good signal to both SBS and UE. On the other hand, this place can be of high wind flows, causing high energy consumption to keep on the position. Zheng et al. found out that circular flight with a constant speed can be more optimal in means of power consumption in comparison to a simple hover [2]. They have found that according to UAV speed, the required power first decreases with increasing flight speed until minimum, then it starts rising sharply. With this in mind we can see that researchers such as Xu et al. [3] or Zeng et al. [4] have found out that optimization of the trajectory itself can lead to minimization of the propulsion energy needed for the flight. However, best to our knowledge, there is so far no research on the minimizing wind effect of hovering or flying UAV in a circular trajectory.

The research aimed on the channel gain improvement or throughput increasing through more optimal UAV positioning does not seem to jointly consider the wind conditions. There is very little research done at the downside of the most optimal placement of the UAV according to the channel gain. In this paper, we analyze the energy consumption problem when hovering. In the urban environment, there is often turbulent wind flow which can affect the UAV hovering ability hard. Even if the wind velocity is low, the turbulent flows around buildings can become very fast at certain points. In case the UAV is placed at such point, the energy consumption may become too high and the battery drains very fast.

II. Energy consumption when hovering

Hovering over a given coordinates with no wind consumes relatively small amount of energy. According to [5], we need to consider different losses such as iron and copper loss connected to electricity transmission. Because these losses are not related to outer conditions, we do not need to consider them in our scenario. Authors of [6] state, that hovering is a process where work is done to overcome gravitational force $F_g = m * g$, where m is a mass of the UAV and $g = 9.8ms^{-2}$ is gravitational acceleration. According to this, to calculate the energy of hovering, only basic newton laws are needed. Then we can theoretically express the energy needed to hover the UAV when no wind is present similarly as in [6].

$$E = \frac{1}{2} m_{air} v_0^2 \tag{1}$$

We consider a UAV for FlyBS with six propellers ($c = 6$), with radius of a propeller r. The mass of air m_{air} is the air pushed through rotors of the UAV and is computed according to the air density ρ and the time t of hovering:

$$m_{air} = cAv_0 t\rho \tag{2}$$

$A = \pi * r^2$ is a disc area of a propeller and v_0 is known as mean rotor induced velocity and according to [4] is calculated as

$$v_0 = \sqrt{\frac{W}{2\rho A}} \tag{3}$$

Here W is weight of the UAV in newtons, equivalent to the F_g. The mean rotor induced velocity is the same speed as the UAV would have in case the rotor was off - the speed of free fall in given moment in opposite direction.
Combining (1) and (2), the hover energy can be calculated as

$$E = 3Av_0^3 \rho t \tag{4}$$

Finally, when the wind velocity vector $\overrightarrow{w} = (w_x, w_y, w_z)$ is known, the energy for hovering according to w_z is

$$E_h = 3A\rho t(v_0 - w_z)^3 \tag{5}$$

TABLE I
SPECIFICATIONS OF DJI UAV USED FOR CALCULATIONS WITHIN THIS PAPER

UAV Model	**DJI Spreading Wings S900**
Number of propellers	6
Weight	4.9kg
Maximum speed	57.6km/h
Power supply	1500mAh
Reference hover power consumption (6.8kg)	1kW
Propeller radius r	13.21cm
Front surface A_f*	$0.05m^2$
Bottom surface A_b*	$0.24m^2$

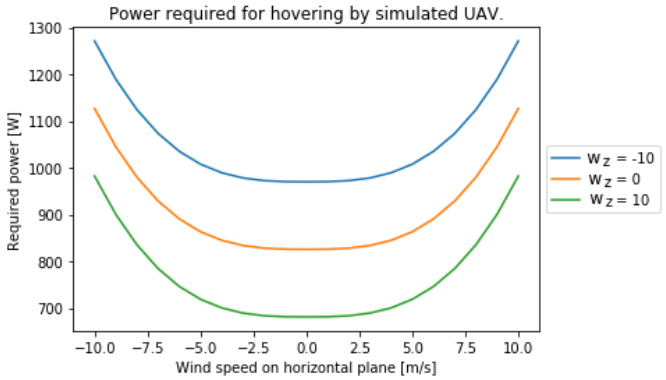


Fig. 1. Power required for hovering with different speed at horizontal plane (x and y-axis). The measurements were performed for three fixed vertical wind speeds $w_z = -10, 0, 10ms^{-1}$

In the last formula, the wind is already present. Other papers, such as [6], [7], did not take this wind velocity into the consideration. However, the force of the wind on static parts must as well be considered. There is a constant pressure by wind on the UAV, which gives the UAV kinetic energy E_k in each direction. Vectors in x and y axes can be considered together, as the kinetic energy is provided to a single surface. The shape of the UAV is symmetric, thus face surface A_f is same from any direction. The bottom and top surface A_b is as well identical. The kinetic energy transmitted from wind to the UAV during time t is:

$$E_{k_f} = \frac{1}{2} m_{air} w_{xy}^2 = \frac{1}{2} A_f \rho w_{xy}^3 t \tag{6}$$

The UAV has to spend the same energy to stay in stable state according to the first newtons rule. The velocity w_{xy} is simply calculated as $w_{xy} = \sqrt{w_x^2 + w_y^2}$. In the z axis, kinetic energy is transferred to the drone as well, however considering the surface A_b:

$$E_{k_b} = \frac{1}{2} m_{air} w_z^2 = \frac{1}{2} A_b \rho w_z^3 t \tag{7}$$

The total energy consumed over time t is the sum of these three parts:

$$E = E_h + E_{k_f} + E_{k_b} \tag{8}$$

Similarly as other researchers such as in [8], the UAV used as a reference for a FlyRS within this paper is DJI Spreading Wings S900 [9]. The specifications of the considered UAV are presented in table I. Note that values of surface marked with asterisk (*) were estimated from the sizes in user manual, as these values are not officially available.

At fig. 1 we can see the results of the simulated UAV with the values from table I. The UAV doesn't suffer much from wind speed if the speed is low. However, if the speed becomes higher than approx. $2.5ms^{-1}$, the power required starts to increase sharply. Also the wind in vertical direction on z-axis has high impact on the UAV power requirements. Low values of wind, under $|2.5|ms^{-1}$ do not impact the UAV significantly. However, if the wind velocity increases to the positive values, the power required for hovering decreases because the wind helps to keep UAV in the air. Similarly, if the wind velocity drops deep to negative values, the UAV is pushed down by air and needs more power to keep in position.

III. Conclusion and Discussion

We have provided an overview on the effect of the wind velocity on hovering UAV. The wind can significantly decrease the flight time and affect the channel gain as the UAV needs to be replaced. Because the wind flow in urban area tends to be turbulent, the possible solution to decrease its impact on power consumption of UAV is to find a better position to hover. This has to be done jointly to maximize the channel gain and minimize the power consumption. This task is left for the future research.

Acknowledgment

This work was supported by the Slovak Research and Development Agency, project number APVV-18-0214 and by the Scientific Grant Agency of the Ministry of Education, science, research and sport of the Slovak Republic under the contract: 1/0268/19.

References

[1] M. Najla, Z. Becvar, P. Mach, and D. Gesbert, "Positioning and association rules for transparent flying relay stations," *IEEE Wireless Communications Letters*, vol. 10, no. 6, pp. 1276–1280, 2021.
[2] Y. Zeng, J. Xu, and R. Zhang, "Energy minimization for wireless communication with rotary-wing uav," *IEEE Transactions on Wireless Communications*, vol. 18, no. 4, pp. 2329–2345, 2019.
[3] J. Xu, Y. Zeng, and R. Zhang, "Uav-enabled wireless power transfer: Trajectory design and energy optimization," *IEEE Transactions on Wireless Communications*, vol. 17, no. 8, pp. 5092–5106, 2018.
[4] Y. Zeng, J. Xu, and R. Zhang, "Energy minimization for wireless communication with rotary-wing uav," *IEEE Transactions on Wireless Communications*, vol. 18, no. 4, pp. 2329–2345, 2019.
[5] C. Chan and T. Kam, "A procedure for power consumption estimation of multi-rotor unmanned aerial vehicle," in *Journal of Physics: Conference Series*, vol. 1509, no. 1. IOP Publishing, 2020, p. 012015.
[6] H. V. Abeywickrama, B. A. Jayawickrama, Y. He, and E. Dutkiewicz, "Comprehensive energy consumption model for unmanned aerial vehicles, based on empirical studies of battery performance," *IEEE access*, vol. 6, pp. 58 383–58 394, 2018.
[7] A. Thibbotuwawa, G. Bocewicz, G. Radzki, P. Nielsen, and Z. Banaszak, "Uav mission planning resistant to weather uncertainty," *Sensors*, vol. 20, no. 2, p. 515, 2020.
[8] A. Fotouhi, H. Qiang, M. Ding, M. Hassan, L. G. Giordano, A. Garcia-Rodriguez, and J. Yuan, "Survey on uav cellular communications: Practical aspects, standardization advancements, regulation, and security challenges," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3417–3442, 2019.
[9] "user manual v1.2 - dl.djicdn.com." [Online]. Available: http://dl.djicdn.com/downloads/s900/en/S900_User_Manual_v1.2_en.pdf

Angle-resolved magnetoresistance of TmB_4

[1]*Július BAČKAI (2nd year)*
Supervisor: [2]Slavomír GABÁNI

[1]Department of Physics, FEI TU Košice, Slovak Republic
[1,2]Centre of Low Temperature Physics, IEP SAS Košice, Slovak Republic

[1]julius.backai@tuke.sk, [2]gabani@saske.sk

Abstract— **Angle-resolved magneto-resistance (ARMR) measurements in various magnetic fields enabled to create distributions of electrical resistivity $\Delta\rho/\rho(\varphi, H)$ in TmB_4, where φ is the angle between the sample's c axis and applied field H. These distributions reveal the charge transport anisotropy in this strongly Ising anisotropic quantum antiferromagnet with a geometrically frustrated Shastry-Sutherland lattice exhibiting fractional magnetization plateaus. While in the paramagnetic region $\Delta\rho/\rho(\varphi, H)$ reaches its maxima for $H \perp c$, below Néel temperature $T_N = 11.7$ K is the situation different. Here the main MR features appear for $H \parallel c$, i.e. along the easy axis of magnetic anisotropy, and correspond to magnetic phases and phase transitions between them. Expressive are above all the features (maxima) related with the scattering of conduction electrons on spin magnetic structure related with fractional magnetization plateaus. With increasing φ shift these MR anomalies to higher fields. Above the field of magnetic saturation, moreover, also significant MR maxima can be observed at certain angles which correspond to specific directions in the crystal lattice. These point to field directions in which the scattering of conduction electrons on the magnetic structure is the highest. Thus, ARMR appears to be a sensitive experimental tool reflecting the angular dependence of the interplay between charge carriers and magnetic structure as a function of temperature and applied magnetic field.**

***Keywords*— low temperature physics, tetraborides, frustrated systems, angle-resolved magnetoresistance**

I. Introduction

Properties of quantum spins with antiferromagnetic (AF) coupling on frustrated lattices have attracted widespread interest in recent years due to the discovery of a variety new quantum ground states as e.g. spin ice [1], quantum spin liquid-like states [2], and fractional magnetization plateaus on the Shastry-Sutherland lattice [3], among them on insulating $SrCu_2(BO_3)_2$ [4] as well as on the family of metallic and magnetic rare earth (RE) tetraborides, REB_4 [5]–[14]. In insulating $SrCu_2(BO_3)_2$ the exchange interaction is of the Heisenberg type, but on the other hand in metallic REB_4 magnets, the AF exchange interaction between their magnetic moments is of long-range Ruderman-Kittel-Kasuya-Yosida (RKKY) type mediated by conduction electrons. On the other hand, in such systems also the transport properties can be strongly influenced by the magnetic structure. This interplay between charge carriers and magnetism therefore enables to use transport experiments as an indirect probe of the magnetic structures that are present in such systems.

Probably the most investigated REB_4 is thulium tetraboride TmB_4, which orders antiferromagnetically at $T_N = 11.7$ K, has attracted attention for its rich magnetic phase diagram which is strongly biased by crystal field effects at Tm^{3+} ion sites that lift the degeneracy of the $J = 6$ multiplet and lead to a $M_J = \pm 6$ ground state doublet [5], [15]–[18]. The most distinctive features of magnetization M along the c axis are various fractional magnetization plateaus which depend on applied field H. Very recent results concerning the properties of fractional plateaus based mainly on magnetization and heat capacity measurements can be found in [19]–[21].

Nevertheless, in metallic TmB_4, as mentioned above, also conducting electrons can provide additional information about various magnetic states. Such information, based on resistivity ρ, magnetoresistance (MR) and Hall effect measurements, were recently obtained in [18], [22], [23]. They show that electronic transport as a function of temperature and magnetic field is a very sensitive probe of scattering processes on the magnetic order / disorder in these frustrated systems.

Here, we present a detailed study of angle-resolved magneto-resistance (ARMR) measurements which enable to construct complex $\Delta\rho/\rho(\varphi, H)$ distributions providing information about scattering processes of charge carriers in various magnetic phases of TmB_4 when the applied magnetic field changes its magnitude H and orientation φ (see Fig. 1), and identify field directions in which the scattering on magnetic structure is the highest.

II. Experimental

The used TmB_4 single crystalline samples were grown by inductive, crucible-free zone melting method, with residual resistivity ratio larger than 30, documenting their high quality. All used samples were cut from one large oriented single crystal. More information about sample preparation can be found e.g. in [24].

Magnetoresistance $\Delta\rho/\rho(\varphi, H) = [\rho(\varphi, H)-\rho(\varphi, H=0)]/\rho(\varphi, H=0)$ measurements were performed using a standard low-frequency ac technique in a commercial PPMS unit equipped with a sample rotation option. The current I was applied along the [110] direction of the Shastry-Sutherland (a-b) plane. The sample orientation to external magnetic field H direction was changed with a step of $\Delta\varphi = 1°$ from [001] via $[\bar{1}11]$ and $[\bar{1}10]$

to [00$\bar{1}$]. In such a case there was always $H \perp I$, but the field could change its orientation, during sample rotation from perpendicular to parallel alignment to the Shastry-Sutherland plane (for details see Fig. 1).

Since the fractional magnetization plateaus depend on field history [18], [21], the ssame measuring protocols were used for magnetoresistance as for magnetization measurements.

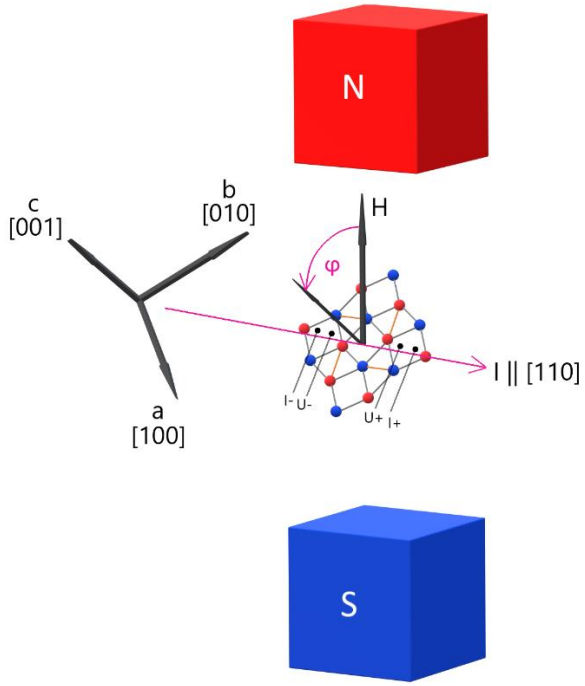


Fig. 1. Layout of angle-resolved magnetoresistance (ARMR) measurements. During measurements the electrical current I flows along crystallographic direction [110]. The orientation of applied magnetic field H, which is always perpendicular to I, changes from direction [001] via [$\bar{1}$11] and [$\bar{1}$10] to direction [00$\bar{1}$], where φ is changed from 0° to 180°. Spin configuration of the antiferromagnetic ground state of the Shastry-Sutherland lattice in the a-b plane is displayed by red (spin "up") and blue (spin "down") spheres.

III. Results and Discussion

The obtained precise angle-resolved magneto-resistance $\Delta\rho/\rho(\varphi, H)$ data set for current direction $I \parallel [110]$ can be displayed in the polar presentation $\Delta\rho/\rho = f(H, \varphi)$ which is shown in Fig. 2. One can see that at 13 K (Fig. 2a), the ARMR distribution is symmetric and reflects in higher fields only the strong Ising magnetic anisotropy in TmB_4, which is present also in the paramagnetic phase above T_N. Below T_N (see Fig. 2b) the ARMR distribution is symmetric only in low magnetic fields (H < 20 kOe), where magnetic hysteresis is not present. In higher fields significant MR anomalies can be observed at certain angles, above all at angles which correspond to [$\bar{1}$14], [$\bar{1}$11], [$\bar{1}\bar{1}$1] and [$\bar{1}$1$\bar{4}$] directions of the crystal lattice and in fields above magnetic saturation (H > 40 kOe). Thus, the obtained ARMR distribution points to field directions in which the scattering of conduction electrons on the magnetic structure is the highest. This suggests that ARMR is a rather sensitive experimental tool which provides additional information about the interplay between charge carriers and various magnetic phases as a function of temperature, applied magnetic field and crystallographic orientations.

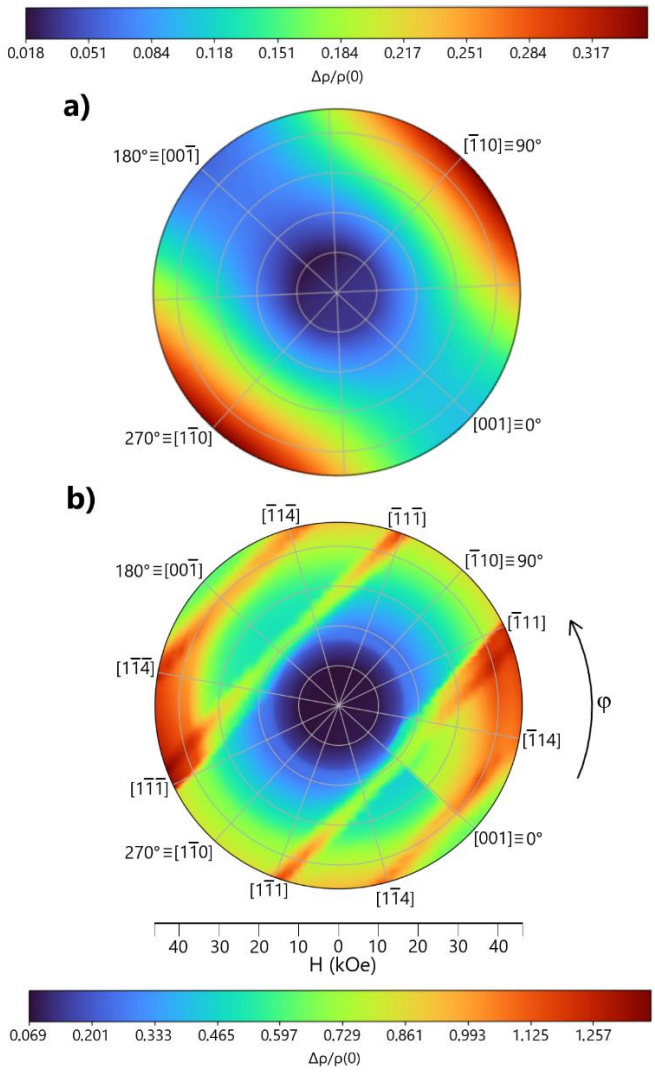


Fig. 2. Angle-resolved magnetoresistance (ARMR) of TmB_4, i.e. $\Delta\rho/\rho$ dependence of TmB_4 as a function of angle φ and applied field H in polar coordinates for current direction $I \parallel [110]$ at 13 K (a) and 2 K (b).

IV. Conclusion

Precise ARMR measurements in various magnetic fields were used to obtain complex distributions of magneto-transport in antiferromagnet with strong Ising anisotropyin TmB_4. These ARMR distributions reveal fields and crystalline directions in which anomalies (maxima) of conduction electrons scattering on the magnetic structure appear. It turned out that in the ordered state large scattering comes from fractional plateau states. In fields above magnetic saturation, significant MR maxima can be observed also at certain angles which correspond to specific directions in the crystal lattice. Thus, ARMR shows to be a sensitive tool providing additional information about the interplay between charge carriers and various magnetic phases of the system. The obtained results are in good agreement with previous results of other authors.

Acknowledgment

This work was supported by projects APVV-17-0020, VEGA 2/0032/20, DAAD-57561069, VA SR ITMS2014+ 313011W856 and by the European Microkelvin Platform. Liquid nitrogen for experiments was sponsored by U.S. Steel Košice, s.r.o.

References

[1] A. P. Ramirez *et al.*, "Zero-point entropy in 'spin ice,'" 1999, Accessed: Feb. 04, 2022. [Online]. Available: www.nature.com

[2] J. S. Helton *et al.*, "Spin Dynamics of the Spin-1/2 Kagome Lattice Antiferromagnet ZnCu_3(OH)_6Cl_2," *Physical Review Letters*, vol. 98, no. 10, Oct. 2006, doi: 10.1103/PhysRevLett.98.107204.

[3] B. S. Shastry and B. Sutherland, "EXACT GROUND STATE OF A QUANTUM MECHANICAL ANTIFERROMAGNET," 1981.

[4] H. Kageyama *et al.*, "Exact Dimer Ground State and Quantized Magnetization Plateaus in the Two-Dimensional Spin System <span class," *Physical Review Letters*, vol. 82, no. 15, p. 3168, Apr. 1999, doi: 10.1103/PhysRevLett.82.3168.

[5] K. Siemensmeyer *et al.*, "Fractional magnetization plateaus and magnetic order in the shastry-sutherland magnet TmB4," *Physical Review Letters*, vol. 101, no. 17, Oct. 2008, doi: 10.1103/PhysRevLett.101.177201.

[6] G. Will *et al.*, "Neutron diffraction studies of TbB4 and ErB4," *Journal of The Less-Common Metals*, vol. 82, no. C, pp. 349–355, 1981, doi: 10.1016/0022-5088(81)90238-1.

[7] D. Okuyama *et al.*, "Competition of Magnetic and Quadrupolar Order Parameters in HoB4," *http://dx.doi.org/10.1143/JPSJ.77.044709*, vol. 77, no. 4, Apr. 2008, doi: 10.1143/JPSJ.77.044709.

[8] J. Y. Kim *et al.*, "Anisotropic magnetic phase diagrams of HoB4 single crystal," *Journal of Applied Physics*, vol. 105, no. 7, p. 07E116, Feb. 2009, doi: 10.1063/1.3075871.

[9] R. Watanuki *et al.*, "Geometrical Quadrupolar Frustration in DyB4," *https://doi.org/10.1143/JPSJ.74.2169*, vol. 74, no. 8, pp. 2169–2172, Nov. 2013, doi: 10.1143/JPSJ.74.2169.

[10] D. Okuyama *et al.*, "Quadrupolar Frustration in Shastry–Sutherland Lattice of DyB4 Studied by Resonant X-ray Scattering," *http://dx.doi.org/10.1143/JPSJ.74.2434*, vol. 74, no. 9, pp. 2434–2437, Nov. 2013, doi: 10.1143/JPSJ.74.2434.

[11] J. Y. Kim *et al.*, "Magnetic anisotropy and magnon gap state of SmB4 single crystal," *Journal of Applied Physics*, vol. 107, no. 9, p. 09E111, May 2010, doi: 10.1063/1.3365061.

[12] P. Farkašovský *et al.*, "Numerical study of magnetization processes in rare-earth tetraborides," *Physical Review B - Condensed Matter and Materials Physics*, vol. 82, no. 5, p. 054409, Aug. 2010, doi: 10.1103/PHYSREVB.82.054409/FIGURES/6/MEDIUM.

[13] J. Strečka *et al.*, "Magnetization process, bipartite entanglement, and enhanced magnetocaloric effect of the exactly solved spin-1/2 Ising-Heisenberg tetrahedral chain," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 89, no. 2, Feb. 2014, doi: 10.1103/PHYSREVE.89.022143.

[14] S. Gabani *et al.*, "Magnetism and superconductivity of rare earth borides," *Journal of Alloys and Compounds*, vol. 821, p. 153201, Apr. 2020, doi: 10.1016/J.JALLCOM.2019.153201.

[15] F. Iga *et al.*, "Highly anisotropic magnetic phase diagram of a 2-dimensional orthogonal dimer system TmB4," *Journal of Magnetism and Magnetic Materials*, vol. 310, no. 2 SUPPL. PART 2, Mar. 2007, doi: 10.1016/j.jmmm.2006.10.476.

[16] S. Gabáni *et al.*, "Magnetic structure and phase diagram of TmB4," *Acta Physica Polonica A*, vol. 113, no. 1, pp. 227–230, 2008, doi: 10.12693/APHYSPOLA.113.227.

[17] S. Michimura *et al.*, "Complex magnetic structures of a Shastry-Sutherland lattice TmB 4 studied by powder neutron diffraction analysis," *Journal of the Physical Society of Japan*, vol. 78, no. 2, Feb. 2009, doi: 10.1143/JPSJ.78.024707.

[18] S. S. Sunku *et al.*, "Hysteretic magnetoresistance and unconventional anomalous Hall effect in the frustrated magnet TmB4," *Physical Review B*, vol. 93, no. 17, May 2016, doi: 10.1103/PhysRevB.93.174408.

[19] J. Trinh *et al.*, "Degeneracy of the 1/8 Plateau and Antiferromagnetic Phases in the Shastry-Sutherland Magnet TmB4," *Physical Review Letters*, vol. 121, no. 16, Oct. 2018, doi: 10.1103/PhysRevLett.121.167203.

[20] D. Lançon *et al.*, "Evolution of field-induced metastable phases in the Shastry-Sutherland lattice magnet TmB₄," *PhRvB*, vol. 102, no. 6, p. 060407, Aug. 2020, doi: 10.1103/PHYSREVB.102.060407.

[21] M. Orendáč *et al.*, "Ground state and stability of the fractional plateau phase in metallic Shastry–Sutherland system TmB4," *Scientific Reports*, vol. 11, no. 1, p. 6835, Dec. 2021, doi: 10.1038/S41598-021-86353-5.

[22] L. Ye *et al.*, "Electronic transport on the Shastry-Sutherland lattice in Ising-type rare-earth tetraborides," *Physical Review B*, vol. 95, no. 17, May 2017, doi: 10.1103/PhysRevB.95.174405.

[23] S. Mitra *et al.*, "Quadratic to linear magnetoresistance tuning in TmB4," *Physical Review B*, vol. 99, no. 4, Jan. 2019, doi: 10.1103/PhysRevB.99.045119.

[24] N. Shitsevalova, "Crystal Chemistry and Crystal Growth of Rare-Earth Borides," *Rare-Earth Borides*, pp. 1–243, Oct. 2021, doi: 10.1201/9781003146483-1.

Classifying heart disorders using machine learning techniques

[1]*Dávid Vaľko (1st year)*
Supervisor: [2]*Norbert Ádám*

[1,2]Dept. of Computers and informatics, FEI TU of Košice, Slovak Republic

[1]david.valko@tuke.sk, [2]norbert.adam@tuke.sk

***Abstract*— In this paper, we will focus on the use of artificial intelligence in medicine. Specifically we will discuss cardiovascular diseases. The advantage of using artificial intelligence in medicine is that it can solve specific obstacles in healthcare, such as a lack of qualified personal. Artificial intelligence itself is suitable for data processing for the purpose of not only classification but also prediction [1]. In our research, the correct prediction made by AI will protect the patient from the consequences of cardiovascular disease.**

***Keywords*— Artificial intelligence in medicine, Cardiovascular diseases, ECG, Neural Networks**

I. Introduction

Artificial neural networks serve as computer systems with the capability to learn or adapt. In our case, the correct prediction made by the AI will serve to protect the patient from the consequences of cardiovascular disease. According to the published statistics, approximately 660,000 people die of heart diseases in Europe each year. In Slovakia, more than 23,000 people die of cardiovascular disease. This number represents approximately 40% of deaths in men and 50% in women [2].

Nowadays, we need a doctor to make an accurate diagnosis or suggest a treatment for the patient. In the field of medicine, cardiologists are the specialized doctors, who deal with cardiovascular diseases. The core problem of most cardiovascular diseases is the heart. Luckily, there is a way to measure heart activity and determine, if the heart itself has any pathology. The cardiologist most often use ECG recordings to determine, wheatear the heart is working correctly or not. One of the problems that today's healthcare faces is that these specialists are hard to find. The number of these specialists are low and continuously declining.

Implementing machine learning methods in medicine could potentially solve these specific problems. In medicine, artificial intelligence will work with digitized information such as imagery, device signals, or measured values. We will also discuss forms of acquiring ECG recording from patient.

Every ECG recording has a prescribed shape and length (Fig. 1). It is singular for every health human. This means that any change in the ECG recording is somewhat easy to detect. The ECG signal consists of several segments. Segments of ECG also have a prescribed shape and duration. In each stage of human life, ECG signal could be different.

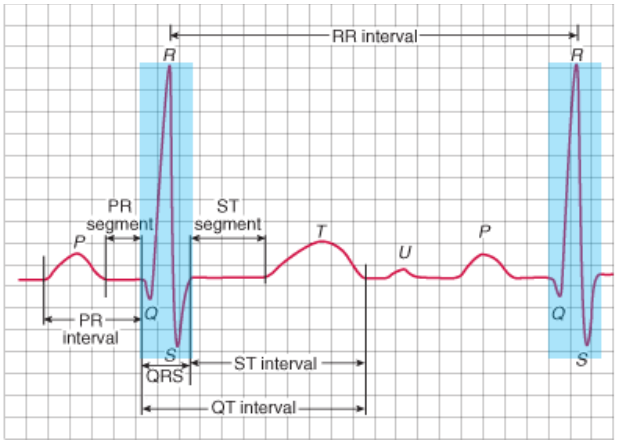


Fig. 1 ECG recording
(source: https://www.techmed.sk/qrs-komplex/)

One of the challenges in ECG preprocessing is to detect QRS complexes. There are several algorithms used for QRS detection such as Pan-Tompkins algorithm, RS slope algorithm or sixth power algorithm [3].

II. Artificial intelligence in medicine

In the field of urgent care, there is an opportunity to implement artificial intelligence in order to improve the quality of healthcare [4]. By using digitalized data in healthcare [5] we can create databases with large amount of patient's data. This way we can obtain medical data such as measured values, various laboratory tests, used medication, and many other values. Artificial intelligence can later examine these databases and find complex relationships between these data. In order to efficiently process these data by humans as well as machines, there is a need to agree on the form of these databases on the worldwide level. The usage of FAIR (findable, accessible, interoperable and reusable) principle [6] could potentially solve this problem and help us achieve our goal.

Based on patient's medical history and other measurements, artificial intelligence will look for connections between the data. After the diagnosis itself, it will choose the appropriate therapy for the patient. It is assumed that artificial intelligence can handle not only smaller therapeutic tasks, such as prescribing drugs, but also larger therapeutic procedures, such as surgeries. Artificial intelligence would still be able to train and improve itself through further data collection. This way, we are creating something which is not used often in healthcare and that is personalized patient treatment [7].

The goal of modern and fair medicine is to:

- provide professional healthcare
- provide safe and comfortable healthcare
- provide economically sustainable solutions to healthcare

III. USING NEURAL NETWORKS FOR ECG CLASSIFICATION

One of the early suggestions on how to implement artificial intelligence in medicine was a software that will communicate with the patient through text; so-called chatBot. Soon *Darwin* was created to fulfill this task. *Darwin* was using convolutional neural network created in *Python* and also uses the *Tensorflow* library. *Darwin's* goal is to understand the patient and his medical history through written text. Based on this, *Darwin* proposes specific actions to take in treating the given condition [8].

With the rapid development of neural networks, new neural network architectures are emerging. One example is *MobileNet*, which is the architecture of a convolutional neural network. This architecture has been shown to be best in classifying COVID-19. A study [9] was carried out to design a system that classifies this disease. Over 6500 X-rays were used to train CNN with the *MobileNet* architecture. Over 500 X-ray images represented patients with COVID-19 disease. The remaining 6000 X-rays represented patients who were not positive for COVID-19. From the performed experiment, they achieved an accuracy of classification around 98.6%.

By using neural networks, we can diagnose the patient's condition in real time, but we can also predict his future condition. Researchers [10] have found out that through deep learning we can predict irregularities in heart activity such as atrial fibrillation before it has any impact on patient's health. Researchers of this study used 1 million ECG recordings that did not indicate atrial fibrillation. Simultaneously they obtained another 1 million data that obtained other heart activity. By applying deep neural networks to ECG, they were given a list of patients who had increased risk of developing atrial fibrillation. Tests later showed that 1 of 3 predicted patients suffered from atrial fibrillation within one year. We have to note the fact, that these people had no signs or symptoms of upcoming heart disease.

The following sections describe different approaches used to classify ECG recordings, while these studies focused either on binary record classification, on abnormal and normal recording, or to determine the exact type of arrhythmia.

A. Recurrent neural networks

According to the research [11] using recurrent neural networks for ECG classification showed great results. Recurrent neural networks are a subset of feed-forward neural networks that use feedback loops to classify experience from previous training cycles. The accuracy for recurrent networks was 93.07%. The researchers also noted, that neural network with Radial Basis Function (RBF) provided the worst classification accuracy of 92.54%. In recurrent neural networks, they used the hyperbolic tangent (tanh) as an activation function for the network. From the given research [11] we have learned that a higher number of hidden layers could potentially give us higher accuracy percentage in classification.

B. Convolutional Neural Network

In recent time, there is a big increase in usage of convolutional neural networks in healthcare for complex tasks, such as ECG classification or segmentation of brain images. CNN is capable to extract useful predictors even when the data is noisy or unclear. CNNs also can sufficiently distinguish morphological differences in ECG signals. CNN is by far the most commonly-used algorithm for ECG arrhythmia detection nowadays [12].

Convolutional neural networks are mostly used to analyze or detect objects in an image, face recognition, or in autonomous vehicle control. This type of network contains two basic components, namely the feature identifier component and the fully connected layer component. The feature identifier component is used to automatically learn and search for features. The fully connected layer component classifies the signal based on the extracted features [13]. This technique can be very well used in the classification of ECG recordings that were transformed into a visual form; spectrogram. Convolutional neural networks are known to achieve excellent results in recognizing patterns from images.

C. Long short-term memory

Classifying ECG recordings can also be done using Long short-term memory; LSTM. This neural network is not only used to classify ECG recordings, but is also implemented in today's smart devices such as watches or mobile phones. Unlike other methods, the use of long short-term memory does not require high hardware requirements.

Compared to other neural networks, it significantly saves consumed energy. The biggest advantage of using LSTM is time complexity. This network was discovered to be the fastest compared to other neural network types [14]. In another study [15], it was reported that the success rate of classifying ECGs was only 74.15%. Moreover, they found out that their model gave poor results in identifying ST-segments. This suggests that it will not correctly classify pathologies like STEMI heart attacks, which are very common in elderly population today. Approximately 30% of the samples that contained ST-segment abnormalities were classified as physiological activity.

Despite this disadvantage, all other cardiac pathologies have been classified with high accuracy.

D. Classifying ECG spectrograms

Engineers from the Turkish University of Yaşar [16] also contributed to the issue of classifying ECGs, who proposed a new way of classifying ECG records. The idea of their design was that they would convert the ECG signal into some kind of visual form; spectrogram (Fig. 2). They will then classify the spectrogram using convolutional neural network. With this approach, they achieved an accuracy of classification of 99.67%. In their study, they worked with the MIT-BIH Arrhythmia database. The ReLU activation function was used for classification. In their work, they presented the results comparing the obtained accuracy in the classification of ECG signals and ECG spectrograms.

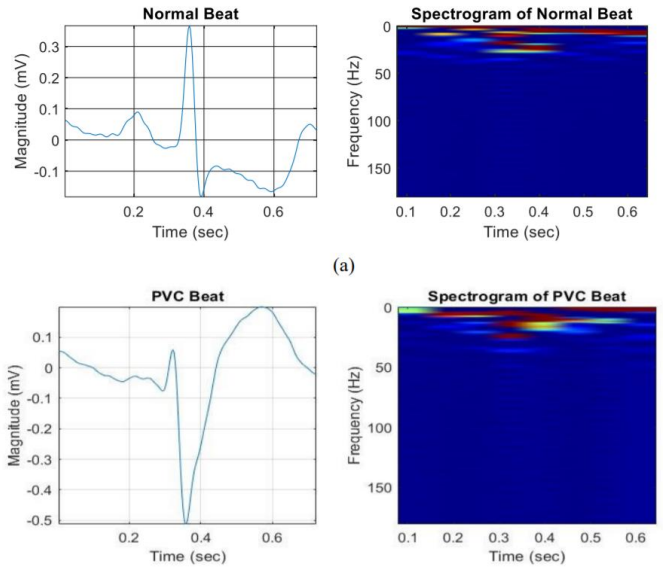


Fig. 2 Converting ECG signal into spectrogram
(source: ECG Arrhythmia Classification By Using Convolutional Neural Network And Spectrogram)

E. ECG databases

Previously mentioned scientific works had different approaches towards choosing an appropriate ECG database. Most commonly used databases were The European ST-T Database, AFPDB database as well as MIT-BIH Arrhythmia. Some of these works even created their own databases. We have concluded that the most commonly used ECG database was *MIT-BIH Arrhythmia*. This database is available to the public since 2005 and contains 48 half-hour ECG recording strips. There you can also find an annotation file for each of these 48 recordings, in which they denote each wave in the given ECG recording. The waves could be marked as normal, or abnormal. These ECG recordings were digitized at 360 samples per second per channel with 11-bit resolution over a 10mV range. Several cardiologists experts independently denoted each recording [17].

IV. Acquiring ECGs with Low-Cost Devices

In many cases, the patient requires continuous ECG monitoring. Traditional monitoring systems enable monitoring of vital functions requiring the sensors to be connected to bedside machines. With the technological advancement, it has become possible to create low-cost ECG monitoring systems in which the system records and displays the signals from the human body. Different studies have been carried out in connection with the development of remote healthcare systems, particularly heart monitoring systems [19].

Obtaining ECG from patient can be done using Arduino's AD8232 sensor (Fig. 3). This sensor has 6 connecting pins. The supply voltage of this sensor is 3.3V. Three ECG electrode pads are needed for the AD8232 to take measurements in the human body. These pads are connected to the sensor via the stereo jack.

Since the amplitude levels of biomedical signals are at μV levels, amplification is required. Therefore, another advantage of this sensor is that it contains the amplifier and filter circuit. This sensor sends the ECG signal to the ARDUINO platform as analogue form [20].

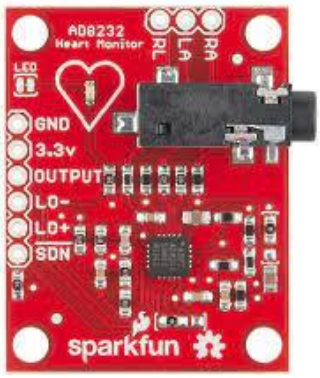


Fig. 3 AD8232 ECG sensor for Arduino
(source: https://www.antratek.com/single-lead-heart-rate-monitor-ad8232)

A. Processing ECG signal from Arduino

Once we have the ECG recorded using Arduino, we need to process it, so that it can be used further on. In this point we face two obstacles:

- noise removal a cancelation
- processing ECG signal

The first thing we need to do is noise removal. We need to record the ECG with minimal noise. There are certain techniques which ensures us, that the measured ECG of human body will not contain noise [21]. The second thing is processing the ECG signal. The ECG signal is recorded and stored usually in CSV file. Our task will be to visualize the ECG wave and delete unnecessary outliers and noise. There are many proposals on how to approach this problem. Most famous are using IIR Low-Pass Filter and db08-wavelet filter [22].

The problem of this task will be to proceed with noise removal in a way, in which we don't distort the image of the ECG, thus eliminating pathologies from it. At the same time, we need to ensure once we have the ECG, that we don't add artefacts or other pathologies to the signal while preprocessing it. Choosing a filter for ECG signal preprocessing and noise removal is very important, while different filters produce different ECG waves (Fig. 4). Some of those filters distorts the values in mV, that the wave is no longer a ECG wave [22-23].

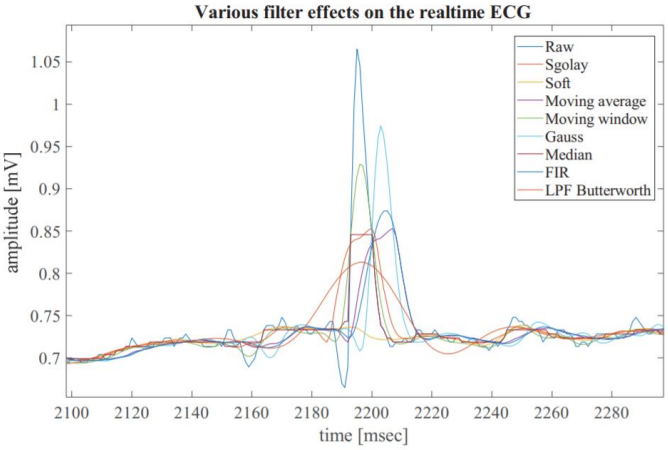


Fig. 4 Usage of different filters on ECG signal

V. Proposed experiments

We propose that we use AD8232 Arduino sensor as a low-cost device to obtain ECG from a patient. Now, the correct preprocessing of the ECG signal needs to be researched. Choosing the correct methods and filter will guarantee us smooth and readable ECG wave, which we can use later for classification. The ECG recording will be classified using neural network. Many papers suggest that convolutional neural network is the best option on how to approach this problem. The goal will be to achieve high accuracy in ECG classification.

VI. Conclusion

This paper considers problems of acquiring and evaluating ECG recordings. The correct implementation of artificial intelligence in the field of medicine is a topic that has been in discussion for a very long time. As of now, we are fully convinced that this issue will be further discussed in the upcoming years. The challenges and obstacles in the field of medicine will be discussed more intensively. We can expect that in the next several years there will be a serious growth in the use of artificial intelligence in medicine.

Acknowledgment

This research was supported by KEGA 002TUKE-4/2021 Implementation of Modern Methods and Education Forms in the Area of Cybersecurity towards Requirements of Labour Market.

References

[1] RiskCardio predicts cardiovascular death (Orig. “RiskCardio predikuje kardiovaskulárnu smrť”) [Online] Available: https://umelainteligencia.sk/riskcardio-predikuje-kardiovaskularnu-smrt/

[2] Slovakia – country with the biggest mortality rate regarding cardiovascular diseases within OECD (Orig. “Slovensko – krajina s jednou z najväčších úmrtností na kardiovaskulárne ochorenia v rámci OECD”) [Online] Available: https://zalepsiezdravotnictvo.sk/blogy/36-slovensko-krajina-s-jednou-z-najvaecsich-umrtnosti-na-srdcove-ochorenia-v-ramci-oecd

[3] E. Tiryaki, A. Sonawane and L. Tamil, "Real-Time CNN Based ST Depression Episode Detection Using Single-Lead ECG," 2021 22nd International Symposium on Quality Electronic Design (ISQED), 2021, pp. 566-570, doi: 10.1109/ISQED51717.2021.9424275.

[4] Z. Dankovičová, D. Sovák, P. Drotár, and L. Vokorokos, “Machine Learning Approach to Dysphonia Detection,” *Applied Sciences*, vol. 8, no. 10, p. 1927, Oct. 2018.

[5] B. Madoš, N. Cichovská, M. Zorkovský, and M. Fedorčák, “Computationally Intensive Medical Application Using Mobile Device and Raspberry PI,” *International Journal of Engineering and Applied Sciences (IJEAS)*, vol. 3, no. 6, pp. 97–100, 2016.

[6] Turning FAIR into reality, Final Report and Action Plan from the European Commission Expert Group on FAIR Data. [Online] Available: https://ec.europa.eu/info/sites/default/files/turning_fair_into_reality_0.pdf

[7] Trenkler Š., Artificial intelligence in medicine (Orig. Umelá inteligencia v medicíne) [Online]. Available: http://www.lf.upjs.sk/ceea/doc5/texty/19%20Trenkler%20Umela%20inteligencia%20v%20medicine%20%20CEEA%202019.pdf

[8] S. Rai, A. Raut, A. Savaliya and R. Shankarmani, "Darwin: Convolutional Neural Network based Intelligent Health Assistant," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2018, pp. 1367-1371, doi: 10.1109/ICECA.2018.8474861.

[9] B. Jabber, J. Lingampalli, C. Z. Basha and A. Krishna, "Detection of Covid-19 Patients using Chest X-ray images with Convolution Neural Network and Mobile Net," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 2020, pp. 1032-1035, doi: 10.1109/ICISS49785.2020.9316100.

[10] Artificial Intelligence examining ECGs predicts irregular heartbeat, death risk [Online] Available: https://medicalxpress.com/news/2019-11-artificial-intelligence-ecgs-irregular-heartbeat.html

[11] O. Aligholipour, M. Kuntalp and S. Sadaghiyanfam, "Silent Paroxysmal Atrial Fibrillation Detection by Neural Networks Based on ECG Records," 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), Istanbul, Turkey, 2019, pp. 1-4, doi: 10.1109/EBBT.2019.8741771.

[12] L. Wei, D. Liu, J. Lu, L. Zhu and X. Cheng, "A low-cost Hardware Architecture of Convolutional Neural Network for ECG Classification," 2021 9th International Symposium on Next Generation Electronics (ISNE), 2021, pp. 1-4, doi: 10.1109/ISNE48910.2021.9493657.

[13] D. SS and D. J. Auxillia, "Classification of ECG using convolutional neural network (CNN)," 2019 International Conference on Recent Advances in Energy-efficient Computing and Communication (ICRAECC), Nagercoil, India, 2019, pp. 1-6, doi: 10.1109/ICRAECC43874.2019.8995096.

[14] S. Saadatnejad, M. Oveisi and M. Hashemi, "LSTM-Based ECG Classification for Continuous Monitoring on Personal Wearable Devices," in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 2, pp. 515-523, Feb. 2020, doi: 10.1109/JBHI.2019.2911367.

[15] Mostayed, Ahmed & Luo, Junye & Shu, Xingliang & Wee, William. (2018). Classification of 12-Lead ECG Signals with Bi-directional LSTM Network.

[16] S. Y. ŞEN and N. ÖZKURT, "ECG Arrhythmia Classification By Using Convolutional Neural Network And Spectrogram," *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Izmir, Turkey, 2019, pp. 1-6, doi: 10.1109/ASYU48272.2019.8946417.

[17] Goldberger, A., et al. “PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals,” in Circulation [Online]. 101 (23), pp. e215–e220." (2000)

[18] P. Ghosal, D. Sarkar, S. Kundu, S. Roy, A. Sinha and S. Ganguli, “ECG beat quality assessment using Self Organizing Map,” in *2017 4th International Conference on Opto-Electronics and Applied Optics (Optronix)*, 2017, pp. 1-5, doi: 10.1109/OPTRONIX.2017.8349994.

[19] M. M. Rahman, M. A. H. Rimon, M. A. Hoque and M. R. Sammir, "Affordable Smart ECG Monitoring Using Arduino & Bluetooth Module," 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), 2019, pp. 1-4, doi: 10.1109/ICASERT.2019.8934498.

[20] M. Çelebi, "Portable ECG Monitoring Device Design Based on ARDUINO," 2020 Medical Technologies Congress (TIPTEKNO), 2020, pp. 1-4, doi: 10.1109/TIPTEKNO50054.2020.9299238.

[21] S. S. Chatterjee, R.S. Thakur, R.N. Yadav, L. Gupta and D.K. Raghuvanshi, "Review of noise removal techniques in ECG signals," In IET Signal Process., 2020, pp. 569-590. doi: https://doi.org/10.1049/iet-spr.2020.0104

[22] M. -J. Wu, S. -F. Shieh, Y. -L. Liao and Y. -C. Chen, "ECG Measurement System Based on Arduino and Android Devices," 2016 International Symposium on Computer, Consumer and Control (IS3C), 2016, pp. 690-693, doi: 10.1109/IS3C.2016.177.

[23] H. Güvenç, "Wireless ECG Device with Arduino," 2020 Medical Technologies Congress (TIPTEKNO), 2020, pp. 1-4, doi: 10.1109/TIPTEKNO50054.2020.9299248.

Tool for automated optimization and parallelization in heliospheric field

[1]*Michal SOLANIK (2nd year),*
Supervisor: [2]*Ján GENČI, Consultant:* [3]*Pavol BOBÍK*

[1,2]Dept. of Computers and Informatics, FEI TU of Košice, Slovak Republic
[3]Department of Cosmic Physics, Institute of Experimental Physics SAS Kosice, Slovak Republic

[1]michal.solanik@tuke.sk, [2]jan.genci@tuke.sk, [3]bobik@saske.sk

***Abstract*—Various simulations, like simulations of cosmic rays distribution in the heliosphere, can be demanding on computational resources. A decrease in execution time can be achieved by different platforms, like GPUs and FPGAs. However, optimizations and parallelizations are usually to the specific type of GPU or FPGA. It is possible to avoid the development of implementations for each type and automate the process of parallelization and optimization. In this paper, we will discuss available tools for automated optimization and parallelization. We will propose the design of an application aimed at automated parallelization of simulations of cosmic rays distribution in the heliosphere.**

***Keywords*—Mathematical methods, Parallelization, Acceleration, Automated parallelization**

I. Introduction

Researchers in many fields use mathematical models to simulate events in the real world. Development in computer systems in recent decades allowed a decrease in execution time for these models. Researchers are also developing more complex models, which require a higher amount of computing resources. One of the available way is to utilize computing resources more efficiently. Optimizations and parallelization of the implementations are some of the most common ways to accelerate their execution time. Application of these techniques takes some time and requires experience with such techniques. Some researchers and projects tried to overcome these disadvantages with applications that can automate optimizations and parallelization. In this paper, we will discuss such works and we will propose a generic design for such application.

II. Automated tools for optimization and acceleration

Developers of scientific applications can use various approaches for creating new accelerated and parallelized implementations. The most basic approach is to optimize and parallelize code manually, but this approach is time-consuming. Several frameworks addressed the issue of the learning curve of optimization and parallelization and time-consuming development of scientific applications. E.g. OpenMP and OpenACC use compiler directives as a marker for parallelization. We can label this approach as semi-automated. The last available approach is automated parallelization and optimization of code. Automated parallelization and optimization require advanced tools that allow the developer to generate new parallel and optimized implementation. In this section, we will discuss already existing tools focused on automated optimization or parallelization.

In our previous paper [1] one of our future research directions was the acceleration of simulations of cosmic rays modulation in heliosphere on multiple platforms. Automated parallelization and optimization is a possible way to achieve it.

OpenTuner [2] is framework for automated optimization of source code. Developers of code can define optimization settings that are used in the selection of measurements and techniques.

GeNN [3] is a domain-specific framework for generating parallel implementation for brain simulations. Developers of scientific code can use predefined interfaces to create a single-thread version of brain simulations.

Other tools for automated optimizations like Kernel Tuner [4] or CLTuner [5] focused on tuning kernels executed on GPUs. CLTuner is aimed at applications written in OpenCL while Kernel Tuner is a more generally oriented tool. Gu et al. [6] designed and implemented loop autotuner for high precision applications. Christen et al. [7] designed and implemented Patus - code generator and autotuning tool for parallel iterative stencil computations.

III. Requirements on tool for automated optimization and acceleration

Based on the presented problematic in section II, we decided to propose an tool dedicated to parallelization and optimization of single-thread implementations of models of cosmic rays modulation in the heliosphere. We analyzed and published implementation of Fokker-Plank's equation on GPU in [8] and [9].

From the analysis of existing parallel implementation we can state:

- existing implementation takes advantage of SIMD architecture,
- implementation of simulation is divided into pre-simulation, simulation, and post-simulation part,
- variables are defined in the shortest possible scope to allow the compiler to optimize the usage of registers.

Simulations of cosmic rays modulation in the heliosphere are independent and do not require operations on shared variables. Independence on each other allowed efficient implementation on GPU.

Based on the presented problematic, we can state requirements on the tool:

- tool should contain interfaces needed for the definition of the single-thread version of simulations of cosmic rays modulation in the heliosphere,
- tool should use various approaches to generate optimization recommendations,
- tool should generate code based on optimization recommendations.
- tool should select the fittest version of generated code based on analysis.

IV. Design of tool for automated optimization and acceleration

In figure 1 is the present design of the tool which requirements we stated in the previous section. The design of automated parallelization flow is based on a genetic algorithm. Genetic algorithm [10] is based on rules of natural selection.

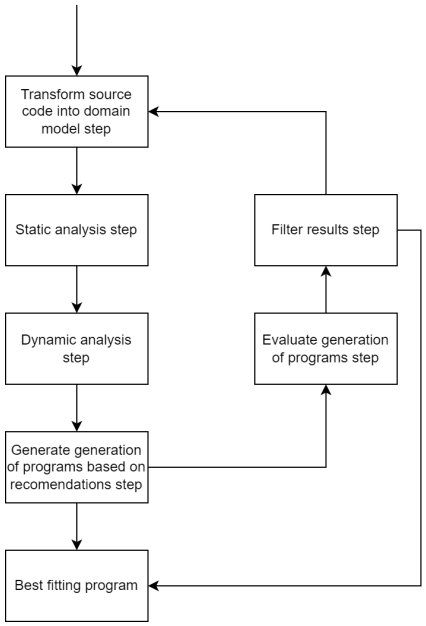


Fig. 1. Design of optimization and parallelization flow in tool for automated optimization and parallelization

Input code is transformed into a domain model. This allows easier manipulation in further steps. Before we present the analysis steps, we need to define the meaning of recommendations in our domain. A recommendation is an output object from analysis, that contains the definition of action that can improve performance or parallelize section of the source code. Action has its type and weight to determine prioritization in case of multiple recommendations applied to the same section of the source code.

Static analysis is used to generate recommendations based on the state of the input source code itself. For example, during automated parallelization on GPU, static analysis can generate a recommendation for reordering of variables for better adjustment for registers on GPU. Dynamic analysis is used to generate recommendations based on recommendations from static analysis and results from testing.

The generator used for generating generations of source codes takes recommendations and creates combinations of them. Based on the combination of recommendations, some recommendations can be conflicting. In this case, we can define incompatible combinations or filter these generated programs as unsuccessful.

Evaluation can be realized in following two forms: evaluation based on execution of generated program and evaluation based on usage of external tools. Evaluation based on execution of the generated program will use basic parameters, like execution time, memory consumption, etc. to evaluate the efficiency of generated programs. Besides this form, evaluation can take advantage of various external tools. These tools include profilers, output from compilers, benchmarking programs, etc.

The filter results step is used to filter generated programs. Input for this filter can be from the user but also based on the parameters of the target system. The filter should have access to every version of generated source code along with output from evaluation. This approach ensures that the fittest version is the most effective version.

V. Further research directions

Automated parallelization and optimization can have benefits in a specific domain, like brain simulations [3] or simulations of cosmic rays distribution in the heliosphere that we proposed in this paper.

In future research, we want to specify static and dynamic analysis phases. Both analyses, static and dynamic should include generally used techniques for optimization and should be platform dependent. We also want to identify tools and techniques that can be helpful during the evaluation and filtering steps.

One of the other research directions is optimization of proposed flow. Implementation of genetic algorithm in this case can be slow. We want to also focus on decrease of execution time during the generation of the most optimized program.

References

[1] M. Solanik, "Methods and algorithms for acceleration / parallelization of physical model calculations," in *21st Scientific Conference of Young Researchers*, 2021, pp. 127–130.

[2] J. Ansel, S. Kamil, K. Veeramachaneni, J. Ragan-Kelley, J. Bosboom, U.-M. O'Reilly, and S. Amarasinghe, "Opentuner: An extensible framework for program autotuning," in *Proceedings of the 23rd international conference on Parallel architectures and compilation*, 2014, pp. 303–316.

[3] E. Yavuz, J. Turner, and T. Nowotny, "Genn: a code generation framework for accelerated brain simulations," *Scientific reports*, vol. 6, no. 1, pp. 1–14, 2016.

[4] B. van Werkhoven, "Kernel tuner: A search-optimizing gpu code autotuner," *Future Generation Computer Systems*, vol. 90, pp. 347–358, 2019.

[5] C. Nugteren and V. Codreanu, "Cltune: A generic auto-tuner for opencl kernels," in *2015 IEEE 9th International Symposium on Embedded Multicore/Many-core Systems-on-Chip*. IEEE, 2015, pp. 195–202.

[6] R. Gu, P. Beata, and M. Becchi, "A loop-aware autotuner for high-precision floating-point applications," in *2020 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. IEEE, 2020, pp. 285–295.

[7] M. Christen, O. Schenk, and H. Burkhart, "Patus: A code generation and autotuning framework for parallel iterative stencil computations on modern microarchitectures," in *2011 IEEE International Parallel & Distributed Processing Symposium*. IEEE, 2011, pp. 676–687.

[8] M. Solanik, P. Bobík, and J. Genči, "Cosmic rays modulation in heliosphere models on GPU," in *37th International Cosmic Ray Conference*, 2021.

[9] ——, "Geliosphere – GPU implementation of cosmic rays modulation models in heliosphere," in *25th Conference of Slovak Physicists*, 2021.

[10] J. R. Koza, "Genetic programming as a means for programming computers by natural selection," *Statistics and computing*, vol. 4, no. 2, pp. 87–112, 1994.

Control of Multiport Power Converter

[1]*Adrián MARCINEK (2nd year)*
Supervisor: [2]František ĎUROVSKÝ

[1,2]Dept. of Electrical Engineering and Mechatronics, FEI TU of Košice, Slovak Republic

[1]adrian.marcinek@tuke.sk, [2]frantisek.durovsky@tuke.sk

***Abstract*— One of the ways to control multiport power converter is proposed in this paper. Multiport power converters replace a conventional topology where many DC-DC or DC-AC converters are connected to a common DC bus. Traditional topology requires multiple power conversion, which reduces the efficiency of the whole system. So that, a new topology of power converter will be presented briefly. This type of converter uses a smaller number of power conversions between ports, which increases the whole system's efficiency. Topology also allows easily add or remove next port of converter. A proposed control method is derived from a converter mathematical model. The control method will enable the transfer of energy between converter ports.**

Keywords—multiport, converter control, multiport power converter control

I. INTRODUCTION

Today, the most used method for interconnecting sources and loads uses multiple DC-DC or DC-AC converters connected to a common DC bus. So, this topology has low efficiency and increase the cost of such a system and decrease power density. Their main disadvantages are also interconnected sources and loads with different voltage levels [1] – [8].

A better solution for interconnecting several sources and loads is using multiport power converters. These converters can transform energy between ports in one conversion step. The base element of these converters is a single multi winding transformer or multi double winding transformers. Multiport converters do not have a common bus which we know from conventional topologies. Because using an energy storage system in most applications is necessary, the ports for connecting these storage systems must be bidirectional. So that, it is not needed to distinguish between input and output ports in a multiport converter, so this converter is called multiport instead of multi-input or multi-output converters. So that, all ports of the multiport converter are considered bidirectional [9] – [16]. The advantages of the multiport converter are the smaller size of the system, one conversion step for power conversion, better power density and high modularity of converter with multiple double winding transformers.

II. MATHEMATICAL MODEL OF MULTIPORT CONVERTER

A description of a multiport converter is based on electric machine theory. An equivalent circuit of a multiport converter is in Fig.1. All parameters are related to the secondary side of transformer. Parameters of transformers primary windings are recalculated to the secondary side and designated by the apostrophe. For transformer description, we use the coordinate system shown in Fig. 2. Variables in Fig. 1 are harmonic and time dependent. If we place them to the complex plane and denote the real axis as α and the imaginary axis as β, the equation for input voltage will be:

$$\mathbf{u}_{11S}(t) = u_{11\alpha}(t) + ju_{11\beta}(t) = u_{11}(t)e^{j\varepsilon}. \qquad (1)$$

The coordinate system with α and β axes we will call as a „stationary system“. Variables in this coordinate system will be labeled by index „S“. The coordinate system with axes x, y is rotating system. Variables in this system will be labeled by index „K“. The equation describes input voltage in rotating system:

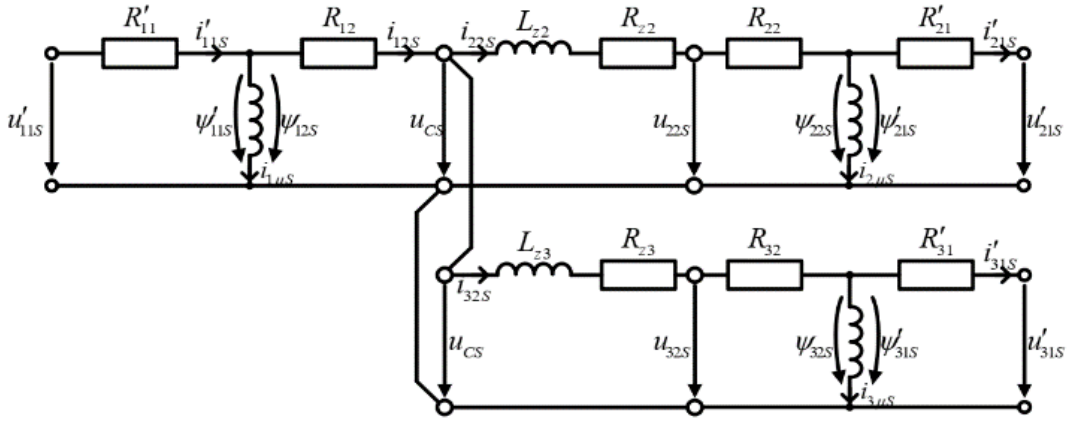


Fig. 1 Multiport converter equivalent circuit

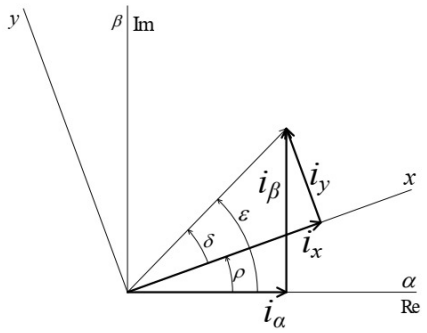


Fig. 2. Coordinate's system used for transformer description

$$\mathbf{u}_{11K}(t) = u_{11x}(t) + ju_{11y}(t) = u_{11}(t)e^{j\delta}. \quad (2)$$

For further description an indication of time-dependent variables are omitted. Therefore element *(t)* will not be use in further equations.

For mathematical model further conditions are valid:

- Parameters and variables on the primary side of transformers are recalculated to the secondary side.
- Secondary windings of transformers are connected in parallel. This connection we will call as intermediate circuit.
- System is supplied by harmonic voltages.
- Leakage inductances of transformers represent approximately 1% of transformers' overall inductance and they do not appear in scheme. They are included in inductance in the intermediate circuit.
- We will analyze the connection of three transformers.

We introduce equations for circuit on the primary side of transformers and equations for $u_{cs.}$ u_{cs} is voltage in intermediate circuit and we will describe equations for circuit between T_1 and T_2 and equations for circuit between T_1 and T_3. For circuit on the primary side of transformer T_1 the following equations are valid:

$$\mathbf{u}'_{11S} = R'_{11}\mathbf{i}'_{11S} + \frac{d\boldsymbol{\psi}'_{11S}}{dt}, \quad (3)$$

$$\mathbf{i}'_{11S} = \mathbf{i}_{12S} + \mathbf{i}_{1\mu S}, \quad (4)$$

$$\boldsymbol{\psi}'_{11S} = L'_{11h}(\mathbf{i}'_{11S} - \mathbf{i}_{12S}) = L'_{11h}\mathbf{i}_{1\mu S} = L_{12h}\mathbf{i}_{1\mu S}, \quad (5)$$

$$\frac{d\boldsymbol{\psi}'_{11S}}{dt} = \frac{d\boldsymbol{\psi}_{12S}}{dt} = \mathbf{u}'_{i11S} = \mathbf{u}_{i12S}. \quad (6)$$

Similar equations are valid for transformer T_2 and transformer T_3. For circuit between T_1 and T_2 is valid:

$$\mathbf{u}_{i12S} = R_{12}\mathbf{i}_{12S} + \mathbf{u}_{CS}. \quad (7)$$

For $\mathbf{u}_{cs}$ voltage for circuit between T_1 and T_2 the equation shown below is valid:

$$\mathbf{u}_{CS} = (R_{z2} + R_{22})\mathbf{i}_{22S} + L_{z2}\frac{d\mathbf{i}_{22S}}{dt} + \mathbf{u}_{i22S}. \quad (8)$$

If we substitute some parts of equation (7) and rewriting it, we will get

$$L_{z2}\frac{d\mathbf{i}_{22S}}{dt} = \mathbf{u}_{i12S} - \mathbf{u}_{i22S} - R_{12}\mathbf{i}_{32S} - \\ -(R_{12} + R_{22} + R_{z2})\mathbf{i}_{22S}. \quad (9)$$

Similarly, we derivate equations for a circuit between T_1 and T_3. The equation for $\mathbf{u}_{CS}$ will change:

$$\mathbf{u}_{CS} = (R_{z3} + R_{32})\mathbf{i}_{32S} + L_{z3}\frac{d\mathbf{i}_{32S}}{dt} + \mathbf{u}_{i32S}. \quad (10)$$

If we use equation (7), substitute some parts and rewriting it, we will get

$$L_{z3}\frac{d\mathbf{i}_{32S}}{dt} = \mathbf{u}_{i12S} - \mathbf{u}_{i32S} - R_{12}\mathbf{i}_{22S} - \\ -(R_{12} + R_{32} + R_{z3})\mathbf{i}_{32S}. \quad (11)$$

1. Transformation to α - β System

The final form of equations is based on variables we want to control. The principle of controlling the system is as follows:

- System enables independent control of active and reactive current components.
- The reactive components are controlled primary by port 1. Output voltage of T_1 is controlled by magnetizing current $\mathbf{i}_{1\mu}$. Since the resistance and leakage inductance of transformers are small, the difference between induced voltage and terminal voltage on the primary and secondary side are small. The induced voltage of transformer T_1 defines voltage in intermediate circuit $\mathbf{u}_C$. Current $\mathbf{i}_{1\mu}$ will be controlled by voltage $\mathbf{u}'_{11}$. Magnetizing currents for port 2 and 3 are generated by $\mathbf{i}_{22}$ and $\mathbf{i}_{32}$ controllers according to $\mathbf{u}_C$ voltage.
- The active currents of port 2 and 3 are controlled by $\mathbf{i}_{22}$ respectively $\mathbf{i}_{32}$ controllers. These currents will be controlled by voltages $\mathbf{u}'_{21}$ respectively $\mathbf{u}'_{31}$. Controllers $\mathbf{i}_{22}$ and $\mathbf{i}_{32}$ define the active current, which ports supply to or withdraw from the intermediate circuit. The difference between both active currents is supplied by port 1 as well.

Further, a transformation of T_1 primary circuit to rotating system will be described. For transformation we use equations (3), (4) and (5). We edit it for $\mathbf{i}_{1\mu}$ current:

$$\mathbf{u}'_{11S} = R'_{11}\mathbf{i}'_{11S} + \frac{d\boldsymbol{\psi}'_{11S}}{dt} = \\ = R'_{11}\mathbf{i}_{1\mu S} + R'_{11}\mathbf{i}_{12S} + L_{12h}\frac{d\mathbf{i}_{1\mu S}}{dt}, \quad (12)$$

and apply equation $\mathbf{u}_{11S} = \mathbf{u}_{11K}e^{j\rho}$ into this equation:

$$\mathbf{u}'_{11S} = \mathbf{u}'_{11K}e^{j\rho} = R'_{11}\mathbf{i}'_{11K}e^{j\rho} + \frac{d}{dt}(\boldsymbol{\psi}'_{11K}e^{j\rho}) =$$

$$= R'_{11}\mathbf{i}_{1\mu K}e^{j\rho} + R'_{11}\mathbf{i}_{12K}e^{j\rho} + L_{12h}\frac{d}{dt}\left(\mathbf{i}_{1\mu K}e^{j\rho}\right). \quad (13)$$

Transformation angle ρ is integral of rotational speed ω_K. This speed represents the frequency of transformers harmonic supply voltage. This frequency is equal for all transformers. Final equation for $\mathbf{i}_{1\mu}$ we can split into two parts. Real part

$$\frac{di_{1\mu x}}{dt} = \frac{1}{L_{12h}}u'_{11x} - \frac{R'_{11}}{L_{12h}}i_{1\mu x} - \frac{R'_{11}}{L_{12h}}i_{12x} + \omega_K i_{1\mu y}, \quad (14)$$

and imaginary part

$$\frac{di_{1\mu y}}{dt} = \frac{1}{L_{12h}}u'_{11y} - \frac{R'_{11}}{L_{12h}}i_{1\mu y} - \frac{R'_{11}}{L_{12h}}i_{12y} - \omega_K i_{1\mu x}. \quad (15)$$

Similarly, to the transformation of T_1 primary circuit, we introduce equations for the primary circuit of transformer T_2, transformer T_3, and the intermediate circuit as well. We do not describe the complete procedure, but we just mention final equations for current split into the real and imaginary part.

For transformer T_2:

$$\frac{di_{2\mu x}}{dt} = \frac{1}{L_{22h}}u'_{21x} - \frac{R'_{21}}{L_{22h}}i_{2\mu x} + \frac{R'_{21}}{L_{22h}}i_{22x} + \omega_K i_{2\mu y}, \quad (16)$$

$$\frac{di_{2\mu y}}{dt} = \frac{1}{L_{22h}}u'_{21y} - \frac{R'_{21}}{L_{22h}}i_{2\mu y} + \frac{R'_{21}}{L_{22h}}i_{22y} - \omega_K i_{2\mu x}. \quad (17)$$

For transformer T_3:

$$\frac{di_{3\mu x}}{dt} = \frac{1}{L_{32h}}u'_{31x} - \frac{R'_{31}}{L_{32h}}i_{3\mu x} + \frac{R'_{31}}{L_{32h}}i_{32x} + \omega_K i_{3\mu y}, \quad (18)$$

$$\frac{di_{3\mu y}}{dt} = \frac{1}{L_{32h}}u'_{31y} - \frac{R'_{31}}{L_{32h}}i_{3\mu y} + \frac{R'_{31}}{L_{32h}}i_{32y} - \omega_K i_{3\mu x}. \quad (19)$$

For intermediate circuit:

- for transformers T_1 and T_2:

$$L_{z2}\frac{di_{22x}}{dt} = u_{i12x} - u'_{21x} - R'_{21}i'_{21x} - R_{12}i_{32x} - R_{c2}i_{22x} + \omega_K L_{z2}i_{22y}, \quad (20)$$

$$L_{z2}\frac{di_{22y}}{dt} = u_{i12y} - u'_{21y} - R'_{21}i'_{21y} - R_{12}i_{32y} - R_{c2}i_{22y} - \omega_K L_{z2}i_{22x}. \quad (21)$$

- for transformers T_1 and T_3:

$$L_{z3}\frac{di_{32x}}{dt} = u_{i12x} - u'_{31x} - R'_{31}i'_{31x} - R_{12}i_{22x} - R_{c3}i_{32x} + \omega_K L_{z3}i_{32y}, \quad (22)$$

$$L_{z3}\frac{di_{32y}}{dt} = u_{i12y} - u'_{31y} - R'_{31}i'_{31y} - R_{12}i_{22y} - R_{c3}i_{32y} - \omega_K L_{z3}i_{32x}. \quad (23)$$

Finally, the mathematical model of the three-port multiport power converter is described by equations (14) and (15) for $i_{1\mu K}$, (16) and (17) for $i_{2\mu K}$, (18) and (19) for $i_{3\mu K}$, (20) and (21) i_{22K} and (22) and (23) for i_{32}.

III. Control of Multiport Converter

As is evident from previous chapter, the currents $\mathbf{i}_{1\mu}$, $\mathbf{i}_{22}$ and $\mathbf{i}_{32}$ will be controlled. We introduce just a simplified description of $i_{1\mu X}$ current controller design. Block diagram for $i_{1\mu x}$ control loop is shown in Fig. 3.

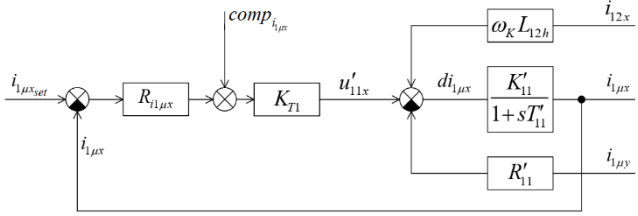


Fig. 3. Block diagram for $i_{1\mu x}$ control

K_{T1} represents the gain of converter connected to primary side of all transformers. $R_{i1\mu x}$ represents the controller of $i_{1\mu x}$ current. The current sensor transfer function is equal to 1, so it is not appeared in the block diagram. Current $i_{1\mu x}$ is got from measured currents i'_{11} and i_{12}. We can denote parameters of primary winding as gain K'_{11} and time constant as T'_{11}:

$$K'_{11} = \frac{1}{R'_{11}}, \quad (24)$$

and

$$T'_{11} = \frac{L_{12h}}{R'_{11}}. \quad (25)$$

After rewriting the equation (14) with (24) and (25) we get

$$i_{1\mu x} = \frac{K'_{11}}{1+sT'_{11}}\left(u'_{11x} - R'_{11}i_{12x} + \omega_K L_{12h}i_{1\mu y}\right). \quad (26)$$

The first element in parenthesis represents a contribution of voltage on primary winding of T_1, the second element is part of load current, and the third element expresses the *y* part of magnetizing current. We assume that all values are measurable. If we compensate the second and third parts by proper compensation signal

$$comp_{i_{1\mu x}} = \frac{1}{K_{T1}}\left(R'_{11}i_{12x} - \omega_K L_{12h}i_{1\mu y}\right), \quad (27)$$

we got a simplified transfer function. Subsequently, equation for open loop system will have a form

$$F_o = \frac{i_{1\mu x}}{i_{1\mu x_{set}}} = R_{1\mu x} K_{T1} \frac{K'_{11}}{1+sT'_{11}}. \quad (28)$$

If we use PI controller, time constant T_{R1} can compensate the system time constant ($T_{R1} = T'_{11}$). A desired dynamic of current loop can be set by controller gain. The transfer function of the system with controller will be

$$F_o = \frac{i_{1\mu x}}{i_{1\mu x_{set}}} = K_{R1} \frac{1+sT_{R1}}{sT_{R1}} K_{T1} \frac{K'_{11}}{1+sT'_{11}} =$$

$$= K_{R1} \frac{1+sT'_{11}}{sT'_{11}} K_{T1} \frac{K'_{11}}{1+sT'_{11}} =$$

$$= \frac{K_{R1}K_{T1}K'_{11}}{sT'_{11}} = \frac{1}{s\frac{T'_{11}}{K_{R1}K_{T1}K'_{11}}} = \frac{1}{sT_{n1}}. \quad (29)$$

where T_{n1} is a time constant of system with PI controller. The final transfer function of system is first order system with the time constant T_{n1}. New time constant can be changed, compared to original one, by factor x. Controller gain K_{R1} is set as follows

$$T_{n1} = \frac{T'_{11}}{K_{R1}K_{T1}K'_{11}} = \frac{T'_{11}}{x}, \quad (30)$$

and then

$$K_{R1} = \frac{x}{K_{T1}K'_{11}}. \quad (31)$$

Final transfer function of closed loop system for control $i_{1\mu x}$ current is:

$$F_W = \frac{i_{1\mu x}}{i_{1\mu x_{set}}} = \frac{1}{1+s\frac{T'_{11}}{K_{R1}K_{T1}K'_{11}}} =$$

$$= \frac{1}{1+s\frac{T'_{11}}{\frac{x}{K_{T1}K'_{11}}K_{T1}K'_{11}}} = \frac{1}{1+s\frac{T'_{11}}{x}} \quad (32)$$

Parameters for the controller of $i_{1\mu X}$ are listed in TABLE I. Calculations of the controller are made for three similar transformers. Therefore, parameters listed in TABLE I. are valid for all three transformers.

The same approach was used for design of $i_{1\mu y}$, i_{22} and i_{32} controllers too.

TABLE I.
PARAMETERS OF MODEL AND CONTROLLER

Parameter	Value
U_{1eff}	24 V
U_{2eff}	48 V
r_1	0.5
$R'_{11} = R_{12}$	0.0073 Ω
$L'_{11h} = L_{12h}$	13.5 µH
L_{z2}	5 µH
R_{z2}	0.05 Ω
R_{c2}	0.064 Ω
K_{T1}	1
T'_{11}	0.0567 s
K'_{11}	137.15
T_{n1}	0.005 s
K_{R1}	0.0827

IV. Conclusion

The paper presents a new power converter topology and its control. Proposed topology enables easy change of a number of ports. The control algorithm provides independent control of active and reactive current components and allows arbitrary power flow among the ports.

Further research will be focused on confirm control in MATLAB/Simulink simulation. The method requires measurement of primary transformers currents as well as current in intermediate circuit. The signals are used for compensation in controllers.

Acknowledgment

This work was supported by the Slovak Research and Development Agency under the contract No. APVV-15-0750.

This work was supported by the Slovak Research and Development Agency under the contract No. APVV-18-0436.

References

[1] B. Wang, M. Sechilariu and F. Locment, "Intelligent DC Microgrid With Smart Grid Communications: Control Strategy Consideration and Design," in IEEE Transactions on Smart Grid, vol. 3, no. 4, pp. 2148-2156, Dec. 2012.

[2] K. Sun, L. Zhang, Y. Xing and J. M. Guerrero, "A Distributed Control Strategy Based on DC Bus Signaling for Modular Photovoltaic Generation Systems With Battery Energy Storage," in IEEE Transactions on Power Electronics, vol. 26, no. 10, pp. 3032-3045, Oct. 2011.

[3] H. Zhou, T. Bhattacharya, D. Tran, T. S. T. Siew and A. M. Khambadkone, "Composite Energy Storage System Involving Battery and Ultracapacitor With Dynamic Energy Management in Microgrid Applications," in IEEE Transactions on Power Electronics, vol. 26, no. 3, pp. 923-930, March 2011.

[4] Hongfei Wu, Yan Xing, Yanbing Xia and Kai Sun, "A family of non-isolated three-port converters for stand-alone renewable power system," IECON 2011 - 37th Annual Conference of the IEEE Industrial Electronics Society, Melbourne, VIC, 2011, pp. 1030-1035.

[5] H. Zhu, D. Zhang, B. Zhang and Z. Zhou, "A Nonisolated Three-Port DCDC Converter and Three-Domain Control Method for PV-Battery Power Systems," in IEEE Transactions on Industrial Electronics, vol. 62, no. 8, pp. 4937-4947, Aug. 2015.

[6] M. McDonough, "Integration of Inductively Coupled Power Transfer and Hybrid Energy Storage System: A Multiport Power Electronics Interface for Battery-Powered Electric Vehicles," in IEEE Transactions on Power Electronics, vol. 30, no. 11, pp. 6423-6433, Nov. 2015.

[7] Y. M. Chen, A. Q. Huang and X. Yu, "A High Step-Up Three-Port DCDC Converter for Stand-Alone PV/Battery Power Systems," in IEEE Transactions on Power Electronics, vol. 28, no. 11, pp. 5049-5062, Nov. 2013.
[8] L. J. Chien, C. C. Chen, J. F. Chen and Y. P. Hsieh, "Novel Three-Port Converter With High-Voltage Gain," in IEEE Transactions on Power Electronics, vol. 29, no. 9, pp. 4693-4703, Sept. 2014.
[9] Z. Qian, O. Abdel-Rahman, H. Hu and I. Batarseh, "An integrated three-port inverter for stand-alone PV applications," 2010 IEEE Energy Conversion Congress and Exposition, Atlanta, GA, 2010, pp. 1471-1478.
[10] Z. Qian, O. Abdel-Rahman, H. Al-Atrash and I. Batarseh, "Modeling and Control of Three-Port DC/DC Converter Interface for Satellite Applications," in IEEE Transactions on Power Electronics, vol. 25, no. 3, pp. 637-649, March 2010.
[11] H. Wu, J. Zhang and Y. Xing, "A Family of Multiport Buck-Boost Converters Based on DC-Link-Inductors (DLIs)," in IEEE Transactions on Power Electronics, vol. 30, no. 2, pp. 735-746, Feb. 2015.
[12] H. Zhu, D. Zhang, H. S. Athab, B. Wu and Y. Gu, "PV Isolated ThreePort Converter and Energy-Balancing Control Method for PV-Battery Power Supply Applications," in IEEE Transactions on Industrial Electronics, vol. 62, no. 6, pp. 3595-3606, June 2015.
[13] Z. Ding, C. Yang, Z. Zhang, C. Wang and S. Xie, "A Novel SoftSwitching Multiport Bidirectional DC-DC Converter for Hybrid Energy Storage System," in IEEE Transactions on Power Electronics, vol. 29, no. 4, pp. 1595-1609, April 2014.
[14] L. Wang, Z. Wang and H. Li, "Asymmetrical Duty Cycle Control and Decoupled Power Flow Design of a Three-port Bidirectional DC-DC Converter for Fuel Cell Vehicle Application," in IEEE Transactions on Power Electronics, vol. 27, no. 2, pp. 891-904, Feb. 2012.
[15] V. N. S. R. Jakka, A. Shukla and G. D. Demetriades, "DualTransformer-Based Asymmetrical Triple-Port Active Bridge (DTATAB) Isolated DC-DC Converter," in IEEE Transactions on Industrial Electronics, vol. 64, no. 6, pp. 4549-4560, June 2017.
[16] Dalton Honorio, Luiz Henrique S. C. Barreto, Andrew Foote, Burak Ozipineci, João Onofre Pereira Pinto, "Modular transformer in isolated multiport power converters" in IEEE Southern Power Electronics Conference (SPEC), Puerto Varas, 4-7 Dec. 2017.

Magnetic properties of domain wall in bi-stable amorphous ferromagnetic glass coated microwires

[1]*Simeon SAMUHEL (2nd year)*
Supervisor: [2]Jozef ONUFER

[1,2]Dept. of Physics, FEI TU of Košice, Slovak Republic

[1]simeon.samuhel@tuke.sk, [2]jozef.onufer@tuke.sk

***Abstract*— This paper is a summarization of the last year of post gradual study. Amorphous microwires are perfect materials for theoretical study and for technical applications. Some magnetic properties of bi-stable microwires are described. The study of geometry and dynamics of domain wall in bi-stable fe-baseed glass coated microwires is presented.**

***Keywords*— Glass-coated microwires, domain wall dynamics, bi-stable behavior.**

I. INTRODUCTION

Magnetic materials are in great demand in many industries, such as automotive, aerospace, medicine, microelectronics, electrical engineering, home entertainment, computer science and sensors [1] - [3]. In these applications, the mechanical properties of the magnetic materials used, their corrosion resistance, small dimensions, their sensitivity to various external conditions (pressure, temperature, electric and magnetic fields) are important, and their price is an important factor for industrial applications [4] -[6].

An important group of magnetic materials are ferromagnetic amorphous glass covered microwires. They consist of a metal core with a diameter of a few micrometers to 100 μm covered with a glass envelope with a thickness of a few micrometers up to 25 μm. They are prepared by drawing a glass capillary with molten melt through a stream of water, the so-called Taylor-Ulitovsky method [7]. The rapid cooling of the metal core melt together with the glass coating is responsible for the high internal production stress in the microwire structure. The different coefficients of thermal expansion of metal and glass create a strong tensile stress present in the metal core after cooling. The combination of tensile stress and positive magnetostriction causes a large magnetoelastic anisotropy with a light axis in the axial direction.

II. THEORY

The domain wall propagation

Based on the current state of the study of domain wall (DW) dynamics in magnetic microwires, we can assume that the DW velocity can be affected by the chemical composition of the metal core as well as by heat treatment [8].

One way to obtain information about the wall geometry is to analyze the shape of the induced signals measured by a Sixtus-Tonks experiment. These signals are strongly influenced by the parameters of the pick-up coils and the stray field around the DW [9].

In most experiments, various modifications of the Sixtus-Tonks experiment are used for these studies [10]. The DW velocity is dependent on the applied magnetic field. The DW velocity is average value for a certain part of the microwires (the area between the pick-up coils) and the applied magnetic field is considered to be constant and homogeneous throughout this area. It should also be mentioned that the DW in these microwires is not a standard 180° DW, because the magnetization component perpendicular to its surface is not equal to zero. This DW is a source of stray field, and obtaining information about DW shape / size from the analysis of the signal induced in the pick-up coil [11] is becoming a complex problem. Nevertheless, experiments of this type exhibit that the DW in a bi-stable microwire is not a solid object and thus the DW shortens with increasing damping force [12]. In addition, as shown in [12] and [13], the DW in a bi-stable microwire has a non-zero inertial mass.

Geometry of domain wall

The study of DW geometry is very important due to the use of microwires in various applications as well as due to basic research. At present, there is still a lack of a relatively detailed and at the same time simple DW model, which would allow the description of the magnetization process of microwires by the movement of a single DW. Although we know of several models that try to describe the shape of the DW, they contradict each other. For example, Panina et al. considered a conical wall shape [14], or a planar wall considered by D.-X. Chen [15]. Recent analytical models performed on the basis of the Landau-Lifshitz equation consider planar and conical DW, as well as flexular planar and vortex DW [16].

III. EXPERIMENTAL

We wanted to find out what the shape of the signal is depending on the axial field affects on the DW. By repeated measurements with a gradually increasing field, we could observe different peaks on the signal (Figure 1) induced in pick-up coil 2 cm long. These velocity changes are localized on the sample and are still in the same position. Due to the value of the applied field, the velocity also changes, but it still changes in the same place. The amplitudes of the peaks also

increase with the increasing field. As can be seen in Figure 1, the DW velocity in certain areas increases (compared to the average), although it could be expected to decrease from the effect of deformations.

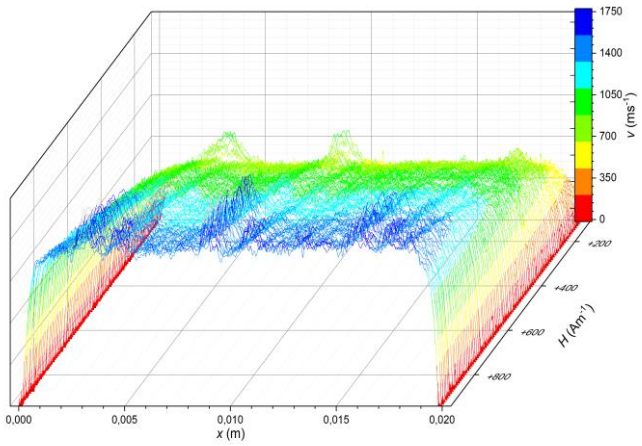


Figure 1 Dependence of the DW velocity on its position in the pick-up coil at different applied fields.

Published results [11] show that DW in a bi-stable microwire is not a solid object in the sense that it retains its shape and dimensions regardless of velocity (applied magnetic field). The conical type DW (figure 2) model with a peak at the front qualitatively explains its shortening with increasing velocity.

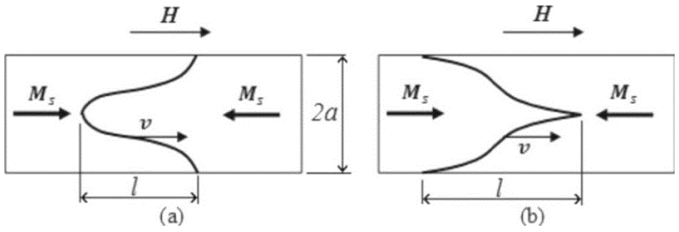


Figure 2 Two types of possible DW shape with cylindrical symmetry [8].

Assume that in the sample we have pinning centers near the surface of the axially magnetized core, which create local braking forces causing the elongation of the DW. This process does not necessarily lead to a slowing down of the DW, because as the DW lengthens, its mobility increases. At the moment of leaving the pinning center, the DW is long and its velocity begins to increase sharply, it reaches a maximum, the DW will gradually shorten and its velocity returns to its original value [17]. An interesting paradox arises when the capture center can cause a local increase in DW velocity, even if it is expected to decrease due to deformations (local disturbances and capture centers).

IV. CONCLUSION

The work summarized the primary information on the issue of amorphous ferromagnetic microwires. The main emphasis was on microwires with positive magnetostriction, which show spontaneous bi-stability. The presented results reveal interesting changes of induced signals which do not yet allow us to evaluate changes in DW geometry. The velocities of a single DW were measured using a simple Sixtus-Tonks experiment. Another interesting phenomenon observed was the occurrence of local increase in DW velocity.

ACKNOWLEDGMENT

The work is a part of the projects supported by the Slovak Research and Development Agency under contract No. APVV-16-0079, with VEGA Grant No. 1/0250/21 from the Scientific Grant Agency of the Ministry for Education of the Slovak Republic and through project No. 019/2019/1.1.3/OPVaI/DP with ITMS code 313011T557 entitled “Support of the research and development potential in the area of transportation vehicles”.

REFERENCES

[1] Zhukov, A., Ipatov, M., Corte-León, P., Gonzalez-Legarreta, L., Blanco, J. M., & Zhukova, V. (2020). Soft magnetic microwires for sensor applications. *Journal of Magnetism and Magnetic Materials*, *498*, 166180.

[2] Morón, C., Cabrera, C., Morón, A., García, A., & González, M. (2015). Magnetic sensors based on amorphous ferromagnetic materials: A review. *Sensors*, *15*(11), 28340-28366.

[3] Mohri, K., Uchiyama, T., Panina, L. V., Yamamoto, M., & Bushida, K. (2015). Recent advances of amorphous wire CMOS IC magneto-impedance sensors: Innovative high-performance micromagnetic sensor chip. *Journal of Sensors*, *2015*.

[4] Kozejova, D., Fecova, L., Klein, P., Sabol, R., Hudak, R., Sulla, I., ... & Varga, R. (2019). Biomedical applications of glass-coated microwires. *Journal of Magnetism and Magnetic Materials*, *470*, 2-5.

[5] Zhukova, V., Corte-Leon, P., Ipatov, M., Blanco, J. M., Gonzalez-Legarreta, L., & Zhukov, A. (2019). Development of magnetic microwires for magnetic sensor applications. *Sensors*, *19*(21), 4767.

[6] Zhukova, V., Corte-Leon, P., Blanco, J. M., Ipatov, M., Gonzalez, J., & Zhukov, A. (2021). Electronic Surveillance and Security Applications of Magnetic Microwires. *Chemosensors*, *9*(5), 100.

[7] Ulitovski, A. V., & Avernin, N. M. (1964). Method of fabrication of metallic microwire. USSR Patent, No. 161325. *Bulletin*, (7), 14.

[8] Onufer, J., Ziman, J., Duranka, P., & Kladivová, M. (2018). The influence of annealing on domain wall propagation in bistable amorphous microwire with unidirectional effect. *Physica B: Condensed Matter*, *540*, 58-64.

[9] Kladivová, M., Ziman, J., & Duranka, P. (2020). Axial Domain Wall Dimension in Bistable Glass-Coated Microwire. *Acta Phys. Pol. A*, *137*, 855-857.

[10] Onufer, J., Ziman, J., & Kladivová, M. (2013). Dynamics of closure domain structure in bistable ferromagnetic microwire. *Journal of magnetism and magnetic materials*, *344*, 148-151.

[11] Horniaková, J., Onufer, J., Ziman, J., Duranka, P., & Samuhel, S. (2021). Changes in geometry of propagating domain wall in magnetic glass-coated bistable microwire. *Journal of Magnetism and Magnetic Materials*, *529*, 167846.

[12] Vázquez, M., Basheed, G. A., Infante, G., & Del Real, R. P. (2012). Trapping and injecting single domain walls in magnetic wire by local fields. *Physical review letters*, *108*(3), 037201.

[13] Ziman, J., Šuhajová, V., & Kladivová, M. (2012). Magnetic domain wall dynamics in an inhomogeneous magnetic field. *Physica B: Condensed Matter*, *407*(18), 3905-3909.

[14] Panina, L. V., Ipatov, M., Zhukova, V., & Zhukov, A. (2012). Domain wall propagation in Fe-rich amorphous microwires. *Physica B: Condensed Matter*, *407*(9), 1442-1445.

[15] Chen, D. X., Dempsey, N. M., Vázquez, M., & Hernando, A. (1995). Propagating domain wall shape and dynamics in iron-rich amorphous wires. *IEEE transactions on magnetics*, *31*(1), 781-790.

[16] Janutka, A. (2020). Effect of spin transfer via eddy current on propagating domain wall in ferromagnetic microwire. *IEEE Transactions on Magnetics*, *56*(7), 1-6.

[17] Samuhel, S., Horniaková, J., Duranka, P., Onufer, J., & Ziman, J. (2021, November). Dynamics of domain wall in rapidly-changing magnetic field. In *AIP Conference Proceedings* (Vol. 2411, No. 1, p. 050013). AIP Publishing LLC.

Experimental study of the vortex lattice in strong disordered ultrathin 3 nm Mo_2N film

[1]*Marek KUZMIAK (3rd year)*
Supervisor: [2]Pavol SZABÓ

[1]Dept. of Physics, FEI TU of Košice, Slovak Republic
[1]Centre of Low Temperature Physics, Institute of Experimental Physics of Košice, Slovak Academy of Sciences, Slovak Republic

[1]marek.kuzmiak@tuke.sk, [2]pszabo@saske.sk

***Abstract*—The main deal of my PhD thesis is the study of the superconducting properties of ultrathin films. I study physical properties of highly disordered superconducting ultrathin Mo_2N films near the superconductor-insulator transition (SIT). The disorder is introduced into the samples by reducing their thickness. This paper is dealing with the experimental study of the vortex lattice in a strongly disordered ultrathin 3 nm Mo_2N films, which are near the critical disorder in the vicinity of the SIT, and still exhibit superconducting properties. For our studies, we have used low temperature STM/STS measurements at various magnetic fields above the Meissner state. We observed that the vortices in our samples are forming strongly distorted hexagonal lattice, which is typical for systems with strong disorder. The research of the vortex structure, as well as the vortices themselves, can be a source of new information about the role of disorder at the forming of the superconducting condensate.**

***Keywords*—vortex lattice, scanning tunneling microscopy (STM), scanning tunneling spectroscopy (STS), superconductivity, superconductor-insulator transition (SIT), molybdenum nitride (Mo_2N).**

I. Introduction

The applied magnetic field penetrates the type-II superconductor above the Meissner state in the form of quantized flux lines, where each flux line is carrying a magnetic flux quantum of ϕ_0. Around the flux line, the diamagnetic response involved with the supercurrents is pressing out the applied magnetic field, forming a vortex. The local value of the magnetic field decreases exponentially from the vortex core (VC), where the superconducting order parameter Δ goes to zero. In this region, the superconductor behaves like a normal metal. Moving away from the vortex core the value of Δ increases to its zero field value at the coherence length, which defines the lateral dimension of the superconducting vortex. The vortices are strongly interacting, forming a solid hexagonal lattice. When the applied field is enhanced, the number of vortices increases, and the field, at which the normal vortex cores overlap defines the value of the upper critical magnetic field H_{C2} [1]. The schematic views of the *H-T* phase diagram of a type-II superconductor, the vortex lattice and the evolution of the local values of the magnetic field, coherence length and superconducting order parameter through a vortex are shown in Figure 1.

The hexagonal lattice of vortices can be disturbed under the influence of high magnetic fields, thermal fluctuations [1], and pinning effects of intrinsic disorder [2–4]. It has been shown [2–4], that in disordered films the increase of disorder breaks the vortex lattice which transforms into a vortex glass structure below the critical disorder. The STM experiments, used for these studies have shown, that the local density of states (LDoS's) measured at the vortex cores reveal strong reductions at the zero bias, which can be even comparable to the superconducting gap structure. This effect strongly affects the determination of the VC positions in disordered systems. The origin of this reduced normal state LDoS is still an unanswered question.

The present paper is dealing with the STM study of vortices in strongly disordered Mo_2N thin films with 3 nm thickness. The obtained vortex images show, that vortices in our thin films are forming a strongly distorted hexagonal lattice, which transforms into a glass structure.

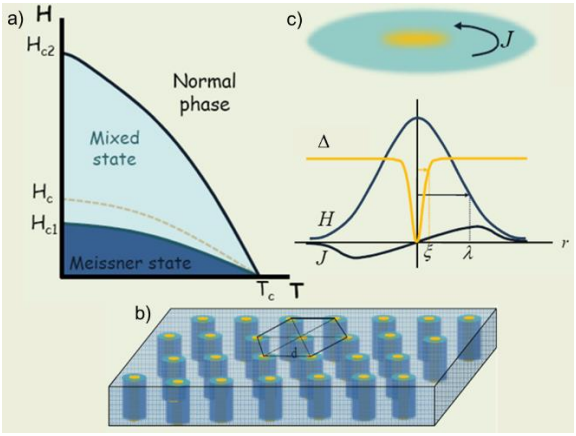


Fig. 1. **Type-II superconductor in the presence of applied magnetic field.** Schematic pictures of the a) *H-T* phase diagram and b) the triangular vortex lattice b), where *d* represents the intervortex distance. c) Evolution of the coherence length ξ, penetration depth λ and the superconducting order parameter Δ through a vortex.

II. Experiment

Our Mo_2N films of a 3 nm thickness were prepared by reactive magnetron sputtering onto a sapphire substrate in the argon-nitrogen mixture. The subsequent crystallographic structure of the Mo_2N was characterized by XRD measurements. It was detected that our films form a

stoichiometric γ-Mo_2N phase with a NaCl structure [5]. Our transport measurements, performed on Mo_2N thin films with various thicknesses showed, that with reducing sample thickness the normal state sheet resistance increases, the superconducting T_c is suppressed and transition from the superconducting to the insulating state takes place at thicknesses between 3 nm and 2 nm [6]. Thus, the 3 nm thin films are close to the critical disorder. The critical temperature of the samples T_C = 2 K was determined from transport measurements and verified with low temperature STM measurements [6].

The local superconducting properties of the samples have been studied by low-temperature STM system developed at the Center of Low Temperature Physics in Košice. This system enables STM experiments down to 300 mK and magnetic fields up to 8 T. A gold STM tip was used to form the tunneling junctions, where the local *I-V* curves have been recorded. The tunneling conductance *G(V)*, which is directly proportional to local density of states (LDoS) has been calculated by numerical differentiation of the *I−V* curves *dI(V)/dV*.

Tunneling spectra, measured deeply in the superconducting state at T = 0.5 K and small magnetic fields of H = 0.5 T, 1 T, and 2 T have been used to determine the vortex structure of our samples. The tunneling curves have been measured on flat surfaces in a grid of 128 × 128 points (CITS technique [7]). The vortex structure has been determined from the conductance maps, constructed from locally measured tunneling conductance values at zero-bias $G(0)$. When determining the positions of vortices, we assumed, that inside the vortex the system is in the normal state and the value of the zero bias conductance *G(0)* is much higher, then far from the vortex core, where the system is superconducting with the superconducting gap in the LDoS .

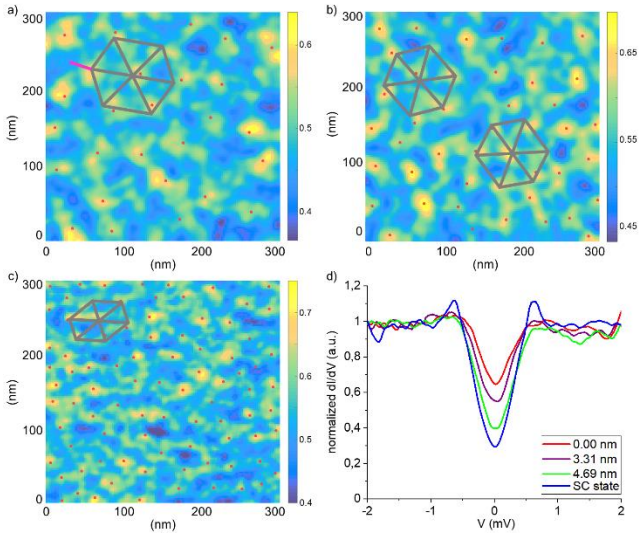


Fig. 2. **Vortex images on Mo_2N** constructed from zero bias conductance (ZBC) maps at 0.5 T (a), 1 T (b), and 2 T (c) at T = 0.5 K. The STM measurements were performed on a surface area of 300×300 nm^2, CITS technique at measuring parameters V = 3 mV, I = 0.4 nA, and T = 500 mK. The blue areas represent the superconducting state of the Mo_2N sample. The red dots show the position of the vortices obtained from the locally highest values of the ZBC's. The image d) shows evolution of the point contact spectra through a vortex at B = 0.5 T marked in (a) with pink line.

III. Results and Discussion

Figure 2 in panels (a-c) show the zero-bias conductance (ZBC) maps, constructed from tunneling conductances, measured at T = 0.5 K temperature on the same 300 × 300 nm^2 large surface at magnetic fields of 0.5 T (a), 1 T (b), and 2 T (c). Areas plotted with blue color (BA) and its tones show parts of the surface where ZBC's have the lowest values around ~ 0.4. The yellow-green areas (YA) allocate parts with the highest ZBC's, which are protruding up to the value of 0.7. Figure 2 (d) plots tunneling spectra measured through the YA marked in Fig. 2 (a) with a pink line. The blue spectrum, which has an evident superconducting gap structure with symmetrically positioned coherence peaks has been measured 30 nm from the middle of the YA. It represents a typical spectrum measured in BA. Approaching the middle of YA the superconducting gap structure is suppressing and the coherence peaks are disappearing. The red spectrum has been measured at the middle of this YA. The reduction at zero bias to the value *G(0)* ~ 0.7 and the missing coherence peaks are typical signatures of the tunneling spectra measured at the middle points of YA's. Analyzing the YA's in the ZBC maps shown in Fig. 2 (a-c) one can see, that some parts reveal "quasi" hexagonal structures, marked with gray lines. It indicates that the YA's are vortices, which are in strongly distorted lattices. The local values of the lattice parameters of the hexagonal structures sketched in Fig. 2 (a-c) are ~ 60 nm, ~ 40 nm, and ~ 30 nm for magnetic fields 0.5 T, 1 T, and 2 T, resp. Following from the Abrikosov – Gorkov (AG) theory [1] the vortex lattice parameter d (distance between two vortices) can be defined with the expression:

$$d = \left(\frac{4}{3}\right)^{1/4} \sqrt{\frac{\phi_0}{\mu_0 H}}, \qquad (1)$$

where ϕ_0 represents the quantum of the magnetic flux, μ_0 the vacuum permeability, and H the magnitude of the applied magnetic field. The calculated values of the intervortex distances are 70 nm, 49 nm, and 35 nm for magnetic fields 0.5 T (a), 1 T (b), and 2 T (c), resp. This is in good agreement with the inter-vortex distances determined from our local analysis. However, the distorted hexagonal lattices determined at 0.5T and 1T magnetic fields (Fig. 2 (a, b)) transform to the evident glass structure at 2 T field. The vortex origin of the YA's at all measured magnetic fields can be shown by comparing the number of YA's with the theoretically calculated number of vortices at given magnetic fields. The number of vortices can be calculated from the AG theory as

$$n = \frac{S_{\text{sample}}}{S_{\text{vortex}}} = \frac{S_{\text{sample}}}{\frac{\sqrt{3}}{2} d^2}, \qquad (2)$$

where the surface of the sample is S_{sample} = 90000 nm^2 and the surface to which one vortex belongs S_{vortex} is proportional to the magnitude of the applied magnetic field H. For a 300 × 300 nm^2 surface in magnetic fields of 0.5 T (a), 1 T (b), and 2 T (c), we obtained the vortex numbers 21, 43, and 86, resp. These values are comparable to the number of YA's (marked in the images with red points), which are 28, 45, and 88, for 0.5 T, 1 T, and 2 T fields. Thus, we can conclude, that the yellow islands shown in Fig. (a-c) are the vortices of our sample. Due to strong disorder the vortices in our samples are forming a strongly distorted hexagonal lattice, which at H = 2 T magnetic field transforms into the glass structure.

It is important to notice, that the tunneling conductance measured in the VC (red curve in Fig. 2 (d)) is not constant, as we would expect for a metallic normal state. The reduction of the LDoS at the zero bias can be connected with the presence

of a superconducting pseudogap [2,4] or VC fluctuations [2,3]. Our preliminary results strongly support the importance of Coulomb interaction effects at the VC's [8].

IV. Conclusion

This paper is dealing with the experimental study of the vortex structure in strongly disordered 3 nm ultrathin Mo_2N films, which are near the critical disorder. Our low temperature STM/STS data detected the presence of strongly distorted hexagonal lattice, which in increased magnetic field transforms to the glass structure. The tunneling conductances measured in the vortex cores revealed strongly reduced non metallic character. The detailed analysis of this effect will be one of the most important themes of my PhD thesis.

Acknowledgment

This work was supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic and the Slovak Academy of Sciences under contracts No. VEGA 1/0743/19, VEGA 2/0058/20, APVV-18-0358, the European Microkelvin Platform, the COST action CA16218 (Nanocohybri), and by the U.S. Steel Košice.

References

[1] M. Tinkham, *Introduction to superconductivity (2nd ed.)*. McGraw Hill Book Co., NY, USA: University Science Books, 2004.

[2] Y. Noat, V. Cherkez, C. Brun, T. Cren, C. Carbillet, F. Debontridder, K. Ilin, M. Siegel, A. Semenov, H.-W. Hübers, and D. Roditchev, "Unconventional superconductivity in ultrathin superconducting NbN films studied by scanning tunneling spectroscopy," *Phys. Rev. B*, vol. 88, p. 014503, Jul 2013. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.88.014503

[3] S. Dutta, I. Roy, J. Jesudasan, S. Sachdev, and P. Raychaudhuri, "Evidence of zero-point fluctuation of vortices in a very weakly pinned *a*-moge thin film," *Phys. Rev. B*, vol. 103, p. 214512, Jun 2021. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.103.214512

[4] R. Ganguly, I. Roy, A. Banerjee, H. Singh, A. Ghosal, and P. Raychaudhuri, "Magnetic field induced emergent inhomogeneity in a superconducting film with weak and homogeneous disorder," *Phys. Rev. B*, vol. 96, p. 054509, Aug 2017. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevB.96.054509

[5] M. Kozejova and et al., "Evaluation of hydrogen evolution reaction activity of molybdenum nitride thin films on their nitrogen content," *Electrochimica Acta*, vol. 315, pp. 9–16, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0013468619310308

[6] M. Kuzmiak and et al., "Superconducting properties of strongly disordered molybdenum nitride ultrathin films in the vicinity of superconductor-insulator transition," *The Scientific Conference of Young Researchers (SCYR)*, pp. 39–41, 2021.

[7] R. J. Hamers, R. M. Tromp, and J. E. Demuth, "Surface electronic structure of Si (111)-(7×7) resolved in real space," *Phys. Rev. Lett.*, vol. 56, pp. 1972–1975, May 1986. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.56.1972

[8] M. Kuzmiak and et al., to be published.

High frequency oscillators for Ultra-Wideband systems

[1]*Patrik Jurík (1st year),*
Supervisor: [2]*Pavol Galajda*

[1,2]Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

[1]patrik.jurik@tuke.sk, [2]pavol.galajda@tuke.sk

***Abstract*—An oscillator is an essential component as the generator of clock signal in radar technologies. In all microwave systems, oscillators represent one of the main components, and all the performance depends on quality of oscillator. This article describes the oscillators in general, and shows basic circuits implementations. Paper will discuss about designing high frequency oscillators, general classification of oscillators, their properties and also informs about the basic parameters of oscillators as resonant frequency, phase noise, amplitude, frequency regulation, resonator types and more. In conclusion, paper describes the current state of available and proposed oscillators types at the Department, which can be used for Ultra Wide-band sensor systems based on M-sequence.**

***Keywords*—Oscillators, RF, DRO, VCO, UWB**

I. Introduction

Oscillators form a separate group of electrical circuits, unlike other electrical circuits they do not process any input signal. Vice versa, they themselves are an alternating electrical signal generator for other circuits. Based on this, we refer to oscillators as autonomous circuits because they generate a signal without external excitation, only on the basis of power supply. The oscillator converts current from the power source to an alternating current with the required frequency. Usually oscillators generate harmonic signals and harmonic signal oscillators are used in many applications, for example in many consumer electronics applications such as radios, televisions and others, but also in measuring devices, wireless systems, radar systems and various other fields. Non-sinusoidal oscillators with a rectangular waveform or a sawtooth waveform have found their application in timing and control applications. Pulse oscillators are commonly used as a clock signal source in digital systems.

II. High frequency oscillators

Oscillators operating at frequencies in the hundreds of megahertz are high frequency oscillators. This frequency band is divided into several frequency sub bands, for example the band of meter waves (300MHz), decimeter waves (3GHz), centimeter waves (30GHz), millimeter waves (300GHz). In the frequency range up to about 30 GHz, the active element of the oscillators is usually a semiconductor, such as a Bipolar Junction Transistor (BJT) [1], Field-Effect Transistor (FET), Complementary Metal Oxide Semiconductor (CMOS) [2] or a more complex structures. The techniques used to design oscillators at low frequencies are different than the techniques used at higher frequencies. The observed performance in the simulations of low frequencies oscillators (in the order of megahertz) is very similar to the performance of the practical circuit implementation. However, by increasing the frequency, the practical implementation is influenced by the design and parasitic properties of the used electrical elements. Also with increasing frequency is necessary to used smaller components, which leads to the implementation oscillators on chip. In such cases, the simulation should provide a starting point for the related practical implementation. Modern development tools such as CST, COMSOL, ADS, CADENCE offer designers the possibility of simulating electronic circuits within the scheme, but also within the physical design, including embedded parasitic properties of components against the printed circuit board, the surrounding space and other variables. This greatly facilitates circuit debugging and prototyping. Recently, circuits operating at frequencies around 100 GHz have been gradually applied in various ways, especially in very fast communication technologies [3], radar technologies [4]. The performance of transistors has increased significantly in recent years, such as the cut-off frequency f_t and the maximum oscillation frequency f_{max} have also improved significantly [5]. Advances in research have therefore made it possible to produce oscillators far exceeding the 100 GHz limit. For example, an oscillator manufactured in InP HBT 0.25μm technology achieved an adjustable resonant frequency of 310-340 GHz, with a maximum output power of -6.5 dBm and a consumption of 13.5 mW [6]. In SiGe BiCMOS 0.13 μm technology, was created a voltage-controlled oscillator with a resonant frequency of 200 GHz [7] or in 65nm LP Bulk CMOS technology, was created a VCO with 290 GHz resonant frequency [8].

III. Classification of oscillators

Classification of oscillators can be useful for designers if it informs about the basic properties of an oscillator. Oscillator classification can be based on various parameters. For example, on the basic characteristics of the oscillator (frequency, tuning range or noise power), or on the functionality (single-phase output or multi-phase output [9], [10]), or classification based on a field pattern of the timing reference. Important information about the oscillator provides the principle of oscillator implementation. The principle of oscillator implementation and the main components of the oscillator determine its type and classification. Practical oscillators can operate strictly on one implementation principle, for example an oscillator

constructed with a lumped resonator. However, it is also possible to mix implementation principles, for example using an LC (coil and capacitor) resonator in a ring oscillator. All oscillators types can be divided into two classes of oscillator operation "continuous-time" and "discrete-time". An example of a discrete-time oscillator is a looped-back digital counter or relaxation oscillator. Continuous-time oscillators can be resonator based or non-resonator based. Fig.1 shows a non-

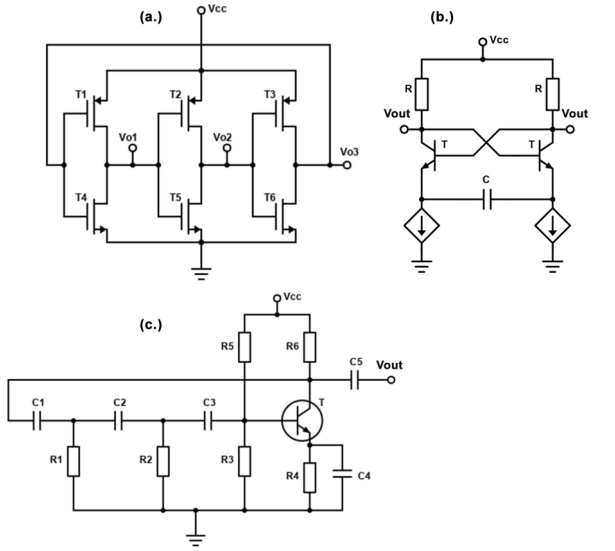


Fig. 1. (a.) Implementation of the single-ended CMOS ring oscillator, (b.) a bipolar relaxation oscillator, (c.) a phase-shift oscillator.

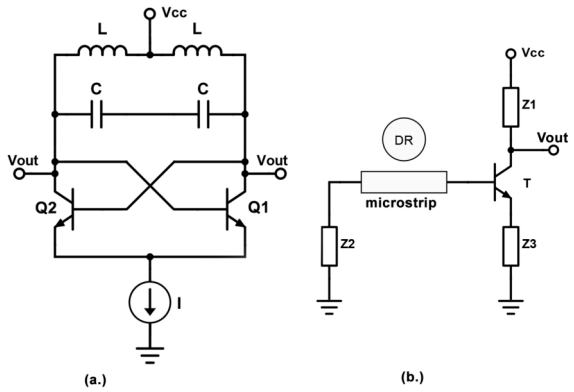


Fig. 2. Oscillators with resonators: (a.) cross coupled oscillator, (b.) oscillator with a dielectric resonator.

resonator based oscillator as a ring oscillator, relaxation oscillator and phase-shift oscillator. Feedback and resistance are required for oscillators without a resonator to function properly. Resonator based oscillators has complex poles and can be distributed or lumped. Oscillators with distributed resonators consist of various types of resonators for example dielectric resonator, crystal resonator, stripline resonator. Fig.2(b.) shows a distributed resonator based oscillator. Oscillators with crystal resonator are used to generate a clock or reference signal based on a piezoelectric crystal resonator. In a Dielectric Resonator Oscillator (DRO), a dielectric resonator (DR) is placed near microstrip lines, which implements the coupling between the resonator and the oscillator. Dielectric resonator have an excellent spectral purity and very good short-term frequency stability [11].

Classification of oscillators with lumped LC resonators is based on the number of identical stages present in an oscillator. For LC oscillators three sub-types are discussed, one stage for single-phase LC oscillators, quadrature LC oscillators with two stage, and LC oscillators with more than 3 stages. Quadrature oscillators are of high interest, since many devices use quadrature signals.

A. Oscillation conditions

Oscillator designer must predict if the oscillator design is properly designed, and oscillator will start producing a periodic signal. To fulfill the oscillation conditions at low resonant frequencies can be used method of feedback modeling [12] For representation circuit parameters at higher frequencies designers use scattering parameters. In that case negative resistance model can be regarded as a feedback model [13].

B. Parameters of oscillators

Main parameter of oscillators is an oscillation frequency. Frequency can be fixed or adjustable. Depending on the type of oscillator there are many possibilities of setting and regulating frequency. In resonator based oscillators the resonator defines the oscillation frequency. The performance of an oscillator is usually compared using a phase noise parameter at offset frequency from a carrier. Fig 3.(a.) shows the fundamental and two harmonics of a square wave with phase noise sidebands and in the time domain phase noise is referred as jitter (fig 3.(b.)). Phase noise impairs the performance of communication, sensing and radar systems. In fact, it is considered one of the most important factors that limit the quality of systems. In most applications, phase noise of oscillators or any signal-generating sources, is kept low enough while sacrificing their output power. Other important parameters for designer are output power, supply voltage, size, price, the possibility of implementation in System-in-Package (SiP) or System-on-Chip (SoC), temperature stability.

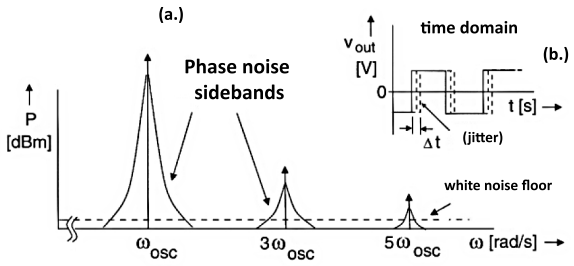


Fig. 3. Phase noise sideband around the resonant frequency.

IV. Resonators

A resonator is a frequency selective network that resonates at certain frequencies. They are basically used to store the energy of signals, because the signals contain electrical and magnetic energy. The resonator is electrically represented by a combination of inductance and capacitance for ideal resonators. For non-ideal resonators it is necessary to add a resistor. These elements represent magnetic energy, electrical energy and losses. The type of resonator is determined by its equivalent electrical model as a serial or parallel resonator. A resonator, whether concentrated, distributed, or a combination

of concentrated and distributed structures, can be classified as a series or parallel resonator depending on its equivalent electrical circuit [14]. Resonator selection is one of the most important parts of oscillator design. The properties and type of resonators strongly influence phase noise and operating frequency. For example, for the Hartley and Collpitz oscillators [15], the LC circuit with lumped elements can be used as a resonator [16]. Where resonator frequency depends on the values of capacitors and inductors to control frequency, it is necessary to regulate values of these components. This can be provided by varactor diode, or by capacitor/inductor bank (Fig. 4.). Banks can be controlled by voltage or current depending on the selected components for switching [17].

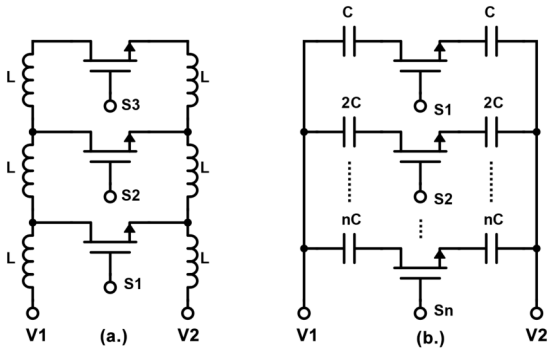


Fig. 4. Principles of regulation frequency via switching capacitance bank or inductor bank.

Preferred resonators at higher frequencies above 3 GHz are planar transmission lines, dielectric resonators, cavity resonators, tubular resonators, YIG resonators, optical resonators and ceramic resonators. For high frequency oscillators with LC resonators design is provided as an integrated circuit, where is possible to achieve very low values and sizes of components. In microwave technology, resonators are used in filters, oscillators and antennas that require frequency selective circuits. Based on many types of microwave resonators, the selection of resonator is critical before oscillator design. The requirements and application determined which type of resonator should be used. Practical resonators with concentrated or scattering parameters always have a loss that increases with increasing frequency and deteriorates the quality of the resonator. The quality (Q) of the resonators is a value characterizing the loss of the resonator and thus its usefulness as a circuit element. It is basically the most important parameter that determines how good the resonator is at storing energy (both electrical and magnetic) and actually determines how the resonator would contribute to the performance of the circuits that contain it. Oscillator noise performance and filter bandwidth depend on the quality factor.

V. Dielectric resonator

A dielectric resonator (DR) is a non-metallic object made of a dielectric material that has been developed to replace resonant cavities in microwave circuits such as filters and oscillators. Dielectric resonator is also used for tunable filters and antennas (DRA) [18]. The advantage of using a dielectric resonator in these circuits allows designers to obtain more compact devices with a higher quality factor Q and thermal stability at a low cost. A dielectric resonators are usually manufactured in the shape of a round puck or cylinder or in the shape of a block [19]. In practice, dielectric oscillators are frequency adapted for the application. Resonant frequency of manufactured resonator can be adapted to application by reducing the volume of resonator, usually with grinding top side of resonator. During grinding resonator for setting the resonant frequency, the resonance frequency of resonator is measured in the measuring cavity. There are several resonance modes of dielectric resonators such as *TE* (Transverse Electric), *TM* (Transverse Magnetic) and *TEM* (Transverse Electromagnetic). Basic mode for cylindrical resonator is $TE_{01\delta}$ [20], for rectangular resonator its *TEM* [21]. The choice of dielectric resonator depends on the requirements of the designer and the requirements of the application. The dielectric constant ε for a resonator with $TE_{01\delta}$ mode can be from 10 to 90, for *TEM* mode a resonator with ε from 90 to 120 and more is available. The $TE_{01\delta}$ mode is the lowest order resonant

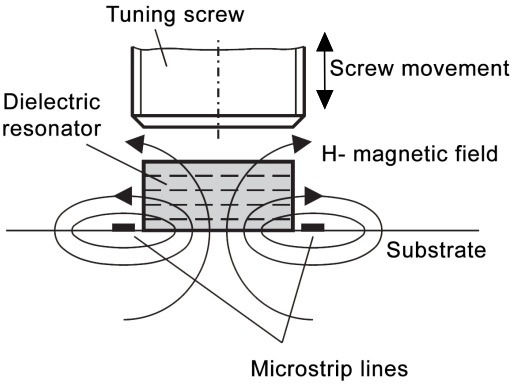


Fig. 5. Dielectric resonator placed between two microstrip lines, with screw above.

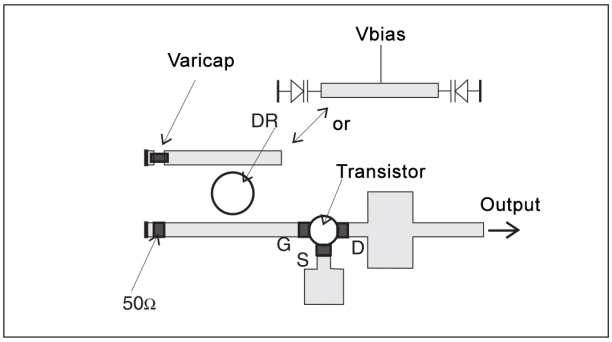


Fig. 6. Voltage controlled oscillator with a dielectric resonator.

mode. Operation in this mode will prevent the oscillator from likely operating at the higher order resonant frequency. Also for certain diameter / length ratios, this mode also achieves the lowest resonant frequency therefore i is classified as the basic mode. Thanks to these features, it is one of the most commonly used modes in many applications. The dielectric resonator captures microwave energy in an extremely small frequency band within the volume of the resonator. This energy is reflected back to the resonator due to the large permittivity gap at the resonator boundary. A small portion of this energy is still distributed in the air around the resonator. These escaping magnetic fields extend beyond the structure of the resonator, and can be used to provide coupling or frequency adjustment by distance from microstrip line [21].

Regulation of frequency can be provided in various ways, most often with a metallic or ceramic screw placed above

the dielectric resonator. Fig 5. shows placement of the tuning screw and resonator. As the screw approaches the resonator, coupling between microstrip and resonator will change [22]. Also frequency can be controlled with movement of a resonator in horizontal or vertical direction. This method is used for setting frequency not for tuning. In application where a voltage control is required, usually a varactor diode is connected to coupling microstrip lines. Varactor diodes influence electromagnetic field of a dielectric resonator with their own electromagnetic field. Circuits with varactor diodes can be placed laterally, under or on the dielectric resonator [23].

Conclusion

This article describes the general types of oscillators for high frequency applications. Paper also describe classification of oscillators, their properties, advantages and disadvantages of oscillators, frequency regulation principles and implementation technology. Based on our research for future work, oscillator with dielectric resonator was chosen, which is described in more detail in section 5. In department we developed few types of oscillators. One oscillator type were part of my diploma work [24]. Where were designed a cross-coupled voltage controlled oscillator (fig 2.(a.)) implemented as an application-specific integrated circuit (ASIC) in 0.25 µm SiGe BiCMOS technology (SG25H3 from IHP Germany). On chip were designed two oscillators with different resonant frequencies, 22 GHz and 11 GHz [25]. Fig 7. shows bonded naked die of designed oscillator in QFN16 package. In design were implemented two options for control frequency, via varactor diode and capacitor bank.

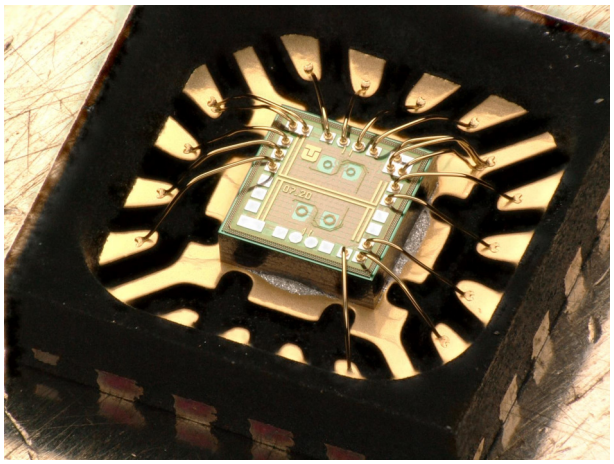


Fig. 7. Designed cross-coupled voltage controlled oscillator in 0.25 µm SiGe BiCMOS technology SG25H3.

There have been several attempts to create an oscillator with dielectric resonator. Specifically, an oscillator with a dielectric resonator in a parallel feedback configuration. Main goal was to create oscillator with oscillation frequency 21.75 GHz. For that purpose we purchased dielectric resonator manufactured for resonant frequency 21.75 GHz. Some of the layouts we designed did not work at all or only partially, but a few designs works. We found out that for better performance it is necessary to use an output stage amplifier. Finding a transistor that can reliably amplify the signal above the 20 GHz limit is difficult. In many papers researchers used naked die transistors, which are difficult to obtain. In our design we used low noise silicon bipolar transistor BFR740L3RH for the active device of the dielectric oscillator. We find out some new option for output stage amplifier like a low noise field-effect transistor CE3521M4. In the future will be tested new configuration with new transistor. Due to the knowledge gained from the creation of the oscillator on the chip and dielectric oscillator, new hybrid types of oscillator were designed and will be manufactured. Oscillator circuit will be made as system on chip. Resonator part will be manufactured on printed circuit board where dielectric resonator will be placed.

Acknowledgment

This work was supported by the Slovak Research and Development Agency under the Contract No. APVV-18-0373 and Scientific Grant Agency (VEGA) under the Contract No. 1/0584/20.

References

[1] Q. Wu, T. Quach, A. Mattamana, S. Elabd, S. R. Dooley, J. J. McCue, P. L. Orlando, G. L. Creech, and W. Khalil, "A 10mw 37.8ghz current-redistribution bicmos vco with an average fomt of −193.5dbc/hz," in *2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, 2013, pp. 150–151.

[2] S. J. Kim, Q. Khan, M. Talegaonkar, A. Elshazly, A. Rao, N. Griesert, G. Winter, W. McIntyre, and P. K. Hanumolu, "High frequency buck converter design using time-based control techniques," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 4, pp. 990–1001, 2015.

[3] Z. Niu, B. Zhang, J. Wang, K. Liu, Z. Chen, K. Yang, Z. Zhou, Y. Fan, Y. Zhang, D. Ji, Y. Feng, and Y. Liu, "The research on 220ghz multicarrier high-speed communication system," *China Communications*, vol. 17, no. 3, pp. 131–139, 2020.

[4] L. Iotti, S. Krishnamurthy, G. LaCaille, and A. M. Niknejad, "A low-power 70–100-ghz mixer-first rx leveraging frequency-translational feedback," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 8, pp. 2043–2054, 2020.

[5] IHP, "130nm sigec bicmos technology development," 2015, last accessed 14 February 2022. [Online]. Available: https://www.ihp-microelectronics.com/services/research-and-prototyping-service/mpw-prototyping-service/sigec-bicmos-technologies

[6] D. Yoon, J. Yun, and J.-S. Rieh, "A 310–340-ghz coupled-line voltage-controlled oscillator based on 0.25-um inp hbt technology," *IEEE Transactions on Terahertz Science and Technology*, vol. 5, no. 4, pp. 652–654, 2015.

[7] P.-Y. Chiang, O. Momeni, and P. Heydari, "A 200-ghz inductively tuned vco with-7-dbm output power in 130-nm sige bicmos," *IEEE Transactions on Microwave Theory and Techniques*, vol. 61, no. 10, pp. 3666–3673, 2013.

[8] Y. M. Tousi, O. Momeni, and E. Afshari, "A novel cmos high-power terahertz vco based on coupled oscillators: Theory and implementation," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 12, pp. 3032–3042, 2012.

[9] J. Gorji and M. Ghaznavi-Ghoushchi, "A 2.7 to 4.6 ghz multi-phase high resolution and wide tuning range digitally-controlled oscillator in cmos 65nm," in *2016 24th Iranian Conference on Electrical Engineering (ICEE)*. IEEE, 2016, pp. 1694–1699.

[10] R. Jiang, H. Noori, and F. F. Dai, "A multi-phase coupled oscillator using inductive resonant coupling and modified dual-tank techniques," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 9, pp. 2454–2464, 2018.

[11] R. KÖKSAL, "Phase locked dielectric resonator oscillator design and fabrication for receiver systems," Ph.D. dissertation, Middle East Technical University, 2016.

[12] J. Van Der Tang, D. Kasperkovitz, and A. H. Van Roermund, *High-frequency oscillator design for integrated transceivers*. Springer Science & Business Media, 2003, vol. 748.

[13] G. Gonzalez, *Foundations of oscillator circuit design*. Artech, 2006.

[14] C. Nguyen, *Radio-frequency integrated-circuit engineering*. John Wiley & Sons, 2015.

[15] D. Garinto, A. Syahriar, and S. Budiyanto, "A novel op-amp based lc oscillator for wireless communications," in *IECON 2020 The 46th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2020, pp. 2243–2248.

[16] M. Azadmehr, I. Paprotny, and L. Marchetti, "100 years of colpitts oscillators: Ontology review of common oscillator circuit topologies," *IEEE Circuits and Systems Magazine*, vol. 20, no. 4, pp. 8–27, 2020.

[17] F. Ullah, Y. Liu, Z. Li, X. Wang, M. M. Sarfraz, and H. Zhang, "A bandwidth-enhanced differential lc-voltage controlled oscillator (lc-vco) and superharmonic coupled quadrature vco for k-band applications," *Electronics*, vol. 7, no. 8, p. 127, 2018.
[18] J. Ashmore, "Design and analysis of a cylindrical dielectric resonator antenna array and its feed network," 2011.
[19] E. TEMEX, "Dielectric resonators," 2015, july 2015. [Online]. Available: https://www.exxelia.com/uploads/PDF/e7000-v1.pdf
[20] Q.-X. Chu, X. Ouyang, H. Wang, and F.-C. Chen, "<formula formulatype="inline"><tex notation="tex">$te_{01\delta}$</tex></formula>-mode dielectric-resonator filters with controllable transmission zeros," *IEEE Transactions on Microwave Theory and Techniques*, vol. 61, no. 3, pp. 1086–1094, 2013.
[21] S. J. Fiedziuszko and S. Holmes, "Dielectric resonators raise your high-q," *IEEE Microwave magazine*, vol. 2, no. 3, pp. 50–60, 2001.
[22] G. D. Vendelin, A. M. Pavio, U. L. Rohde, and M. Rudolph, *Microwave circuit design using linear and nonlinear techniques*. John Wiley & Sons, 2021.
[23] N. Popovic, "Review of some types of varactor tuned dros," *Applied Microwave & Wireless*, pp. 62–70, 1999.
[24] P. Jurík, "Návrh oscilátorov pre uwb senzorove systémy," 2021.
[25] P. Jurik, M. Sokol, and P. Galajda, "Design of high frequency oscillators for ultra-wideband systems," in *2021 31st International Conference Radioelektronika (RADIOELEKTRONIKA)*. IEEE, 2021, pp. 1–5.

Laboratory Model of Small Hydropower Plant for Simulation Purposes

[1]*Richard OLEXA (3rd year)*
Supervisor: [2]*Pavol FEDOR*

[1,2]Dept. of Electrical Engineering and Mechatronics, FEI Technical University of Košice, Slovak Republic

[1]richard.olexa@tuke.sk, [2]pavol.fedor@tuke.sk

Abstract – At the moment, in the world globally, there is an environmental pressure put on all of us. We do have to realize the importance of keeping our planet livable for next generation, at least at the level of today. But how to accomplish it? In the proposed article, we discuss the process of designing and creating the real laboratory model of a small hydropower plant (SHP) that will serve as an valuable tool for next development in the field of optimization of the control loops and processes used in the small hydropower plants, as well as it will provide the user the opportunity to potentially use the below-described model as an training tool for the powerplant operators.

***Keywords*—Hydraulic turbines, hydropower systems, modeling and controlling hydropower plants, small hydropower plant, asynchronous generator**

I. Introduction

Using the renewable energy sources, the list of the acknowledge renewable energy sources, keeping off the usage of the fossil fuels for energy generation, these are just few of the todays were often used phrases in our daily life. A quick look at it is enough to realize what for challenges we are going to face in the upcoming years. And this is not going to be easy at all. In combination with the gas crisis in Europe, energy prices reaching their ever maxim levels, all of us have to think about our contribution in this path. The ordinary people are separating waste, buying packaging-free items, or even thinking about not using the conventional combustion based vehicles anymore.

On the other hand, we, the researchers have some kind of the wider range of capabilities to support this matter. Therefore, we decided to keep the direction of our research in the field of simulation of the small hydropower plants (SHP), but as well to create an effective and helpful tool how to use the proposed simulation model even for those ones, who did not participate in the research, e.g. to allow it to be easily used in the SHP designing process, tuning, or even in the training process.

II. Initial status in modeling of hydropower plants

As discussed in the previous papers, there were and probably still are quite many of other researchers looking at the way how the small hydropower plants are being transferred in the simulation model. Out of these studies, there are a lot of focusing on the big scale hydropower plants mainly [1]-[2]-[3], while taking into account some of the phenomenon that are connect to their size, thus being too complex and complicated for the SHP purposes, e.g. water-hammer phenomenon.

Moreover, plenty of papers are based on the assumption of using the linearized model sub-systems, i.e. using linearized transfer functions instead of incorporating all of the non-linear features of the small hydropower plant itself. [1]-[2]-[3]-[5]-[6]-[7]

Another difference is made by the type of the generator used in the SHP. It is true, that definitely more than the half of the currently installed base is using a synchronous generator for the electricity production. On the other hand, the simple construction and the absence of the excitation part makes an asynchronous generator an attractive option to be chosen by the projecting companies. Thus, not many publications are investigating a simulation model of a SHP with the asynchronous machine as a generator part of the system. Instead, the vast majority create their models keeping the synchronous generator as a part of it. [4]-[8]-[9]-[10]-[12]

The previous article was mainly aimed to provide the summary of the work and ideas of an improvement in the field of modelling the SHP. [13] In this paper we explain the scheme we used in the process of creating the before-mention tool for the designers and users, in the for of an physical laboratory model of a SHP as well as how each of the main blocks, or in other words, subsystems [13], is being implemented in the structure of the laboratory model.

III. Proposed implementation of a small hydropower plant's simulation model into the laboratory model

As we explained in the last years paper, the proposed simulation model consist of four main subsystems; i.e. turbine-control system, or so called governor; servo drive that adjust the position of the main water inlet gate, or of the guide vanes respectively; hydraulic turbine; electric generator; plus the fifth part could be considered as an electrical grid. [13]

Detailed information about the design and purpose of each of the above-mentioned parts are summarized in the publications of [11]-[13]-[14].

In the Figure 1, there is a schematic connection between all of the subsystems depicted, as well as the names of the parts of the laboratory model are being placed next to the subsystems, to represent the element of the laboratory model that takes over its functionality.

Speaking of the elements of the proposed laboratory model of a SHP for the simulation purposes, description of theirs is needed to be provided. Therefore, the subsections *A-D* are

dedicated to give a more detailed view of how the proposed model should be operating.

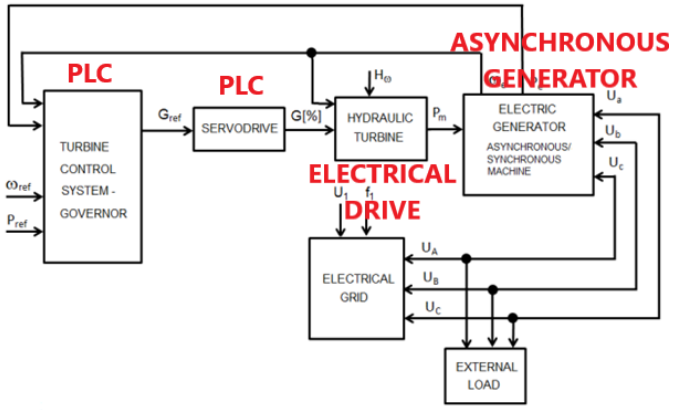


Figure 1 Block diagram of a Small Hydropower Plant with the depicted elements of the laboratory model performing their functions

A. PLC – Programmable Logic Controller – B&R Power Panel 500 series

This, all-in-one solution depicted in the *Figure 2* was chosen according to its ability to combine a controlling features with the HMI functions in one device. This programmable logic controller is used to take over the control system part, shown in the Figure 1; servo drive subsystem, as well as all the calculation connected to the operation of the simulated behavior of the hydraulic turbine.

The control mechanism is performed by using the functional block for the PID controller while using the calculated and tuned parameters obtained in MATLAB, or set up by the user. Implementation of the servo drive subsystem is done as well via programmed structure that suppose to give an input value for the hydraulic turbine part, i.e. the inlet water gate vane position G[%]. All the calculations that are performed in the MATLAB model of the hydraulic turbine subsystem is then processed by the PLC with an output of the turbine velocity [rpm] that serves as an input value for the electrical drive while using the B&R *ACOPOSinverter 8I66 series*.

Plus, a build-in TFT display serves as an HMI where the user has various options to adjust the process values and thus to affect the operation, or the simulation process respectively. Examples of these parameters are: current water head value; demanded generated power (reference value); no-load flow value, i.e. the value that determines the efficiency of the turbine; or so called turbine gain (A_t parameter). Moreover, it displays all the process data needed for the safe operation and simulation purposes.

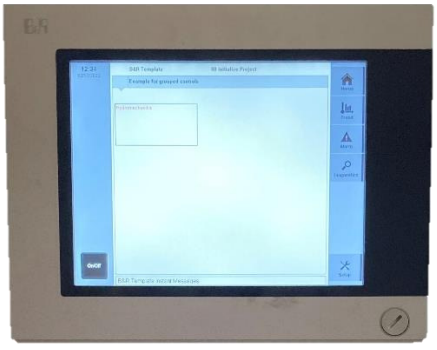


Figure 2 Laboratory Model: B&R Power Panel 500 – HMI display, placed on the front panel of the switchboard

B. Electrical drive – ABB 3GAA101520-ADE662

This three-phase squirrel cage motor replace the role of the hydraulic turbine in proposed laboratory model of a SHP, especially in terms of rotational movement of the turbine, or with the common shaft respectively. Its velocity, or the input voltage value, is calculated according to the actual demanded generated power specification, as well as according to the parameters of the exact parameters of the simulated hydraulic turbine. Since the motor shaft is fixed with the generator shaft via a fixed coupling part, the model represents and image of a real SHP operation. (see *Figure 3*)

C. Asynchronous generator – ABB 3GA101520-ADE

This similar asynchronous motor serves as a generator in the described scheme, whereas the generated power should be feed into the grid via a recuperative converter, or to the grid directly. Exact implementation of the recuperative converter into the scheme is a topic of another research that aims to upgrade the same topology and to allow new controlling techniques to be used and tested in the field of small hydropower plants. (see *Figure 3*)

Moreover the *ABB CT PRO XT 40* current transformer is used to transform primary currents to secondary currents for c.a. measurement instruments. This allows us to monitor fast changes that occur in the electrical circuit of the SHP and process them via *B&R X20AP3131* module, which measures active, reactive and apparent power individually for each phase as well as all of them collectively. In addition, this module provides the RMS values for voltage and current on the 3 phases. Definitely, for such fast changing parameters we need to secure fast processing as well. Therefore, a *B&R PLC X20 CP1686X* is included in the scheme and the communication of both PLC included in the scheme is done via OPC UA standard.

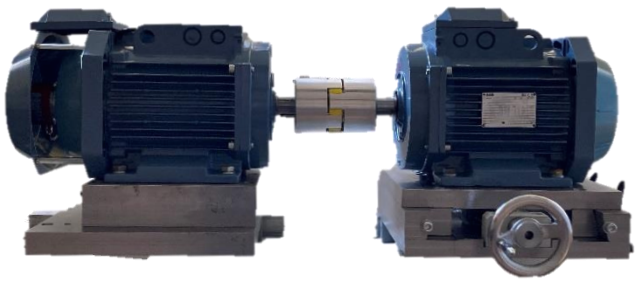


Figure 3 Laboratory Model: Controlled Electrical Drive – Hydraulic Turbine (Left) and Asynchronous generator (Right) with the fixed coupling and common shaft

IV. NEXT STEPS

Based on the above-mentioned, the phase of designing the topology, constructing and mounting of the switchboard as well as an initial programming phase is done. Currently, the next steps will definitely follow up all the previous work, mainly in terms of completing the programming phase, creating the visualization, or so called HMI, as well as the debugging and verifying process for the whole system proposed in this paper, while the documentation and description of the research will be written in a form of a dissertation thesis.

ACKNOWLEDGMENT

This work was supported by the by the APVV project APVV-19-0210.

REFERENCES

[1] IEEE Working Group., "Hydraulic turbine and turbine control models for system dynamic studies," in IEEE Trans on Power Syst., 1992, 7, pp.167–79.

[2] H.J. Wagner, J. Mathur, Introduction to Hydro Energy Systems: Basics, Technology and Operation. Heidelberg : Springer, 2011. ISBN 978-3-642-20708-2.

[3] G. Munoz-Hernandez, S. Mansoor, D. Jones, Modelling and Controlling Hydropower Plants. 1st Edition. London : Springer London Ltd, 2013. s. 299. ISBN: 978-1-4471-6221-6.

[4] D.G. Ramey, J. W. Skooglund, "Detailed hydro governor representation for system stability studies," in *IEEE Trans on Power Apparatus and Systems*, 1970, 89, pp.106–12.

[5] J. Tiwari, et al., "Modelling and Simulation of Hydro Power Plant using MATLAB & WatPro 3.0.," in *Intelligent Systems and Applications*, 2015, 08, pp. 1-8. DOI: 10.5815/ijisa.2015.08.01.

[6] G. Singh, D.S. Chauhan, D. S., "Simulation and Modeling of Hydro Power Plant to Study Time Response during Different Gate States," in *(IJAEST) International Journal of Advanced Engineering Sciences And Technologies*, 2011, Vol. No. 10, Issue No. 1, pp. 042 – 047.

[7] R.A. Naghizadeh, S. Jazebi and B. Vahidi, "Modelling Hydro Power Plants and Tuning Hydro Governors as an Educational Guideline," in *International Review on Modelling and Simulations (I.RE.MO.S)*, 2012, Vol. 5, No. 4., pp.780-1790.

[8] A. Acakpovi, E. B. Hagan and F. X. Fifatin, "Review of Hydropower Plant Models," in *International Journal of Computer Applications*, vol. 18, December 2014, pp. 33-38.

[9] A. A. Usman, A. Abubakar and R. A. Abdulkadir, "Modelling and Simulation of Micro Hydro Power Plant Using Matlab Simulink," in *Proc. 2nd International Conference on Science, Technology and Management*, New Delhi: ICSTM, 2015, pp. 1121-1133.

[10] M. Sattouf, "Simulation Model of Hydro Power Plant Using Matlab/Simulink," in *International Journal of Engineering Research and Applications*, vol. 1, Brno: IJERA, 2014, pp. 295-301. ISSN: 2248-962

[11] P. Fedor, D. Perduková, R. Olexa, "Modelling a Small Hydropower Plant". In *Proc. Energetika 2019*, Stará Lesná, pp. 352-356. ISBN: 978-80-553-3324-3

[12] C. Jaliu, I. Visa, D. "Diaconescu, Dynamic Model of a Small Hydropower Plan", In *Proc. 12th International Conference on Optimization of Electrical and Electronic Equipment*, 2010.

[13] R. Olexa, "Modelling and Controlling of Small Hydropower Plants". In *Proc. SCYR 2021: 21st Scientific Conference of Young Researchers*, Košice, pp. 183-185. ISBN: 978-80-553-3904-7

Capsule Neural Networks - Future of Deep Learning?

[1]*Dominik VRANAY(1st year),*
Supervisor: [2]*Peter SINČÁK*

[1,2]Department of Cybernetics and Artificial Intelligence, FEI TU of Košice, Slovak Republic

[1]dominik.vranay@tuke.sk, [2]peter.sincak@tuke.sk

***Abstract*—There are multiple advancements made in the field of Deep Learning and Image Classification and Capsule Neural Networks are one of them. The idea proposed by professor Hinton introduces a new structure called "capsules" into Neural Network topology. In addition to capsules, there are more differences to the traditional Convolutional Neural Networks (CNN). In this paper, I will talk about the main differences Capsule Neural Networks have and compare them to CNNs. I will go through the use cases of Capsule networks and possible research and improvements that can be made.**

***Keywords*—Capsule Networks, Classification, Deep Learning, Topology Research**

I. Introduction

In recent years, there has been a huge increase of tasks, which need to be solved by methods from Computer Vision. With this increase, we have much more data available for training, but the performance needed to accurately deal with these tasks is increasing as well. Some of these tasks require better generalisation and even training on a small number of data. Because of this a lot of research in the field of Deep Learning and Image classification is made to cope with these new and hard tasks. This increase in research and available computational power caused the introduction of new ideas on how to improve the accuracy of our Neural Networks as well as new ways to build our topologies.

One of these new ideas were Capsule Neural Networks introduced in prof. Hinton's papers in 2011 and 2017 [1] [2]. These deal with and try to fix problems of CNNs, which will be discussed later. In addition, the performance of both approaches will be shown and the improvement highlighted.

II. Main Ideas

Capsule Neural Networks are created by using multiple novel ideas[3], which will be described in depth in this section.

A. Capsule Structure

The main idea of Capsule networks is the structures called "Capsules". These are represented as vectors, where the length of the vector represents the probability that an entity represented by the capsule exists in the image. The values of the vector elements describe some information about the entity. These can be position, transformation, rotation size, and many others. The exact feature each element represents depends on the training and what our gradient descent finds as the best representation.

Procedure 1 Routing algorithm.

1: **procedure** ROUTING($\hat{\boldsymbol{u}}_{j|i}, r, l$)
2: for all capsule i in layer l and capsule j in layer $(l+1)$: $b_{ij} \leftarrow 0$.
3: **for** r iterations **do**
4: for all capsule i in layer l: $\mathbf{c}_i \leftarrow \texttt{softmax}(\mathbf{b}_i)$ ▷ softmax computes Eq. 3
5: for all capsule j in layer $(l+1)$: $\mathbf{s}_j \leftarrow \sum_i c_{ij}\hat{\mathbf{u}}_{j|i}$
6: for all capsule j in layer $(l+1)$: $\mathbf{v}_j \leftarrow \texttt{squash}(\mathbf{s}_j)$ ▷ squash computes Eq. 1
7: for all capsule i in layer l and capsule j in layer $(l+1)$: $b_{ij} \leftarrow b_{ij} + \hat{\mathbf{u}}_{j|i}.\mathbf{v}_j$
return $\mathbf{v}_j$

Fig. 1. Pseudocode of the original Hinton's dynamic routing algorithm[2]

B. Routing Algorithm

One of the most important parts of Capsule networks is the routing algorithm. This algorithm ensures that lower-level capsules are routed and vote for higher-level ones in the next layer. This is done to merge smaller entities into bigger groups of entities. One example of this could be that we have a face of a human in a picture. Then the first layer would find a circle, which would represent the iris, and another ellipse, which would be the whole eye socket. These would be merged into one entity in the second layer. As multiple merged entities would be found in the second layer like nose and mouth, then the third layer would merge them again into a face. This would mean that at the end we would find a human face in the picture and correctly classify the image as a human. To reduce the size of the input, we use this routing algorithm as well. This should ensure, that no information is lost as in pooling done in CNNs.

Part of the routing algorithm is squashing, which would, as the name suggests, squash all vectors into the range (0,1). This range is used as it is well known that networks are trained faster if the inputs are normalised[4]. In addition, this will make the length of the vector represent the probability that the entity exists. As capsule neural networks do not introduce non-linearity in any other way, squashing is used for this purpose as well.

The original Hinton's dynamic routing algorithm [2] has multiple unique features. This algorithm is shown in pseudocode in Fig. 1. The main idea of the algorithm is that the lower-level nodes are trying to predict the output of the higher levels ones in multiple iterations and from that, they will strengthen the connection to them, if the prediction is far off and loosen the connection if the prediction is spot on or close. This is done in usually 3 iterations, but this number can be changed according to the needs of the problem. The algorithm uses Coupling coefficients, which represent all the connections between the lower and higher-level nodes. For each low-level node, the sum of outgoing coupling coefficients

is 1, which is ensured by the usage of the softmax function from the probabilities. There exist multiple different routing algorithms, which can perform better on some problems and as there is not much research done on Capsule Networks, we do not know, which would be the best performing routing algorithm on average or if there is an algorithm like this. Here are some examples of other routing algorithms.

1) Variational Bayes routing, which uses priors to control capsule complexity and induces sparsity. [5]
2) Self-attention routing, which uses the different stabilising term and has only 1 iteration of routing. [6]
3) EM-Routing, which is based on GMM-EM. [7]
4) Fast dynamic routing based on weighted kernel density estimation. [8]

C. Margin Loss

As capsules are represented as vectors, the output of the network is also a vector. Because of this, capsule neural networks need a specific loss, which would work with the output of this format. Margin Loss was developed to work with capsules and fix this problem. The basics of the loss are similar to the cross-entropy loss [9] with the main difference of using λ, m^+ and m^- as can be seen from (1).

$$L_k = T_k.max(0, m^+ - ||v_k||)^2 + \\ +\lambda(1 - T_k).max(0, ||v_k|| - m^-)^2 \quad (1)$$

Another advantage of this loss is that it supports multiple true labels for one image and multi-label datasets. This means, that if we have multiple objects in an image, our network can find both of them and the result of the network would be high for both.

λ parameter is used as a helpful tool for the beginning of the training when we do not know much about other classes and taking the whole second equation would slow down the training considerably. m^+ and m^- are used to compensate for the length of the vector, as there is a low possibility that the vector's length would be 0 or 1 when it is squashed and these small numbers would help speed up the training.

D. Reconstruction

As capsule networks were created from the research of auto-encoders, reconstruction was kept as a regularisation method. This reconstruction should encourage the capsules to encode the information about input as the transformation, rotation skew, and others, which should help in generalising. As the main output of the capsule network is a number of vectors, these can be used for reconstruction loss. We would take the output for the true label and run in on the decoder part of the network. Then we would compare the reconstruction with the original input image with MSE or any other loss. This loss would be multiplied by a small number as not to dominate over our main Margin loss. In the end, it should ensure better generalisation and improvement in performance.

Using reconstruction, we can easily see what each element of the capsule represents. This was done on the MNIST dataset as can be seen in Fig. 2. The changes were created by slightly modifying one element of the capsule by small amounts to see what the element influences in the reconstruction. For small datasets such as MNIST, the thickness can be easily seen as one of the features as well as the shapes and lengths of some curves and lines for each number.

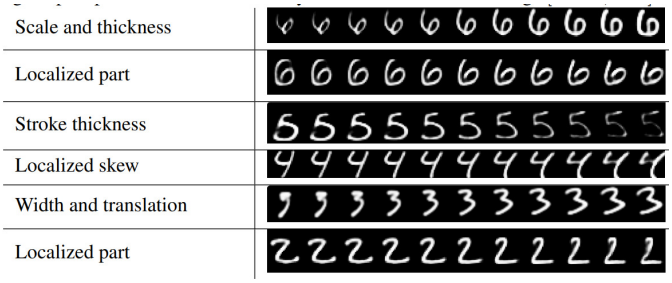


Fig. 2. Dimension perturbations of multiple digits on MNIST dataset, which showcases the meaning of some features of the capsules [2]

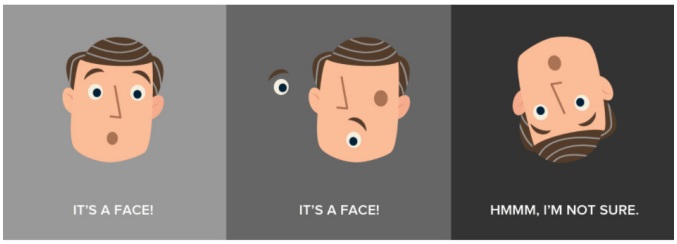


Fig. 3. Illustration of the Picasso's face problem

III. Usage cases for Capsules

To know when to use Capsule Neural Networks, we need to know first what are their advantages and disadvantages [10].

A. Advantages

- Viewpoint invariance: by using vector representation of each entity, capsules can recognise the entity regardless of the perspective of the entity in the image.
- Fewer learnable parameters: As can be seen from the experiments and benchmarks, capsule networks use fewer parameters than comparable CNNs. This is because capsules are grouped and the connections need fewer parameters.
- Better generalisation: as can be seen from the first point, Capsules are good in recognising entities regardless of perspective. This means, that it can detect entities in completely new perspectives and transformations. CNN is not able to generalise like this and needs to see the perspective during the training to correctly classify it.
- Picasso's face problem: CNNs are unable to tell, wherein the image the parts are located and they are supposed to be grouped together. Capsule networks are able to group lower-level entities better and even look at the relative position while grouping them, which can help with the problem in Fig. 3. Capsule network would correctly recognise the first and last picture as faces, but not the middle one, as one eye is not even in the face, and all parts are shuffled around. The last image would be correctly recognised as well, because of the generalisation on rotations. CNNs would make mistakes as shown in Fig. 3.
- Hard to confuse during adversarial attacks: Attacks such as Fast Gradient Sign Method can drop the accuracies of CNNs by 60%, capsule neural network is still able to perform with an only drop of 20%.

B. Disadvantages

- Slower training time per epoch: The original Hinton's dynamic routing algorithm was slow during training, as it would compute the routing iteratively. Other implementations are faster but still take a bit longer per epoch than CNNs.
- Not well explored: As capsule neural networks are a recent invention, they are not really explored. All other implementations of the routing algorithms are only scratching the surface of what the capsules are capable and if we want to use them on a different problem, we would need to experiment a lot to find the best way to use them.
- Harder to explain: CNNs can be divided into feature extraction part and classification part. It is not that easy to do with capsules as each capsule contains features and we only merge different capsules together. We cannot really say what each capsule represents except the ones in the final layer, which makes understanding the networks harder.
- Tries to account for everything: Capsules try to find every entity in the image, which means that we have a lot of information not really needed for our classification problem, and because of that, it would slow down our training by overloading capsules.
- Low performance on varied backgrounds: As was mentioned in the last point, capsules try to account for everything and if they see a different background, they will try to show that somehow in the capsules and overload them. This would cause lower accuracy as well as slower training.

IV. Comparison with CNN

The usual method for building good neural networks is to use convolutional layers and CNNs. These have a number of differences from the new capsule networks:

1) The dimensionality of the data: For image processing CNNs use 3D data: Height of the picture, the width of the picture, and the number of channels. In comparison, Capsule Networks use 4D data: height, width, number of capsules, and number of features of each capsule.
2) Size reduction of the data: CNNs use mainly MaxPooling [11] and AvgPooling to reduce the size. These use kernels of size usually 2x2 and stride 2 to half the size in each direction for the total reduction of one quarter. Capsules use the routing algorithm to reduce the size, which will reduce the number of capsules in the next layer without losing information as pooling does.
3) The class representation in the output of the network: CNNs use the traditional scalar values to represent the probability that each class exists in the image, which is calculated in the classification part of the network with fully connected layers. In comparison, each output class is represented as a capsule, which is a vector. This means that we have more information available in the output of the network that can be later used for reconstruction.
4) Better generalisation: CNNs use dropout to reduce overfitting and improve generalisation, but this would not work that easily in capsule networks so another way had to be used. This is mainly why capsules use Reconstruction loss.

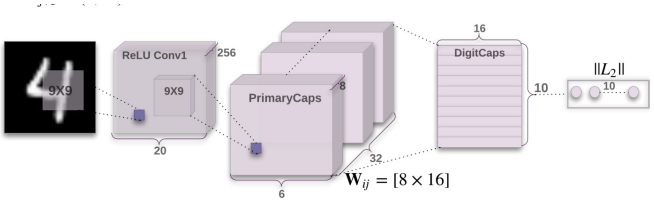


Fig. 4. A simple CapsNet with 3 layers, which is used for benchmarking on MNIST dataset [2]

TABLE I
Comparison of accuracy on MNIST testing dataset

Method	Routing	Reconstruction	MNIST (%)
Baseline	-	-	99.61
CapsNet	1	no	99.66
CapsNet	1	yes	99.71
CapsNet	3	no	99.65
CapsNet	3	yes	99.75

A. MNIST dataset

A small experiment was created, which would use baseline CNN and compare it to Capsule neural network on MNIST dataset [12]. The MNIST dataset is composed of 60000 handwritten digits of size 28x28. During training, these images were translated up to 2 pixels in each direction to help with the training and generalisation. The testing set was not using any augmentations.

The structure of the baseline CNN with total of 35.4M parameters:

- convolutional layer with 256 channels 5x5 kernel and stride 1
- convolutional layer with 256 channels 5x5 kernel and stride 1
- convolutional layer with 128 channels 5x5 kernel and stride 1
- fully connected layer with output size of 328
- fully connected layer with output size of 192
- dropout layer
- fully connected layer with output size of 10

The structure of CapsNet is computationaly similiar, but only uses 8.4M parameters. The structure is composed by following layers as well as Fig. 4:

- convolutional layer with 256 channels 9x9 kernel and stride 1
- convolutional layer with 32x8 channels 9x9 kernel and stride 2
- reshape into 32 8D capsules
- routing layer into 10 16D capsules

Reconstruction part:

- fully connected layer with output size of 512
- fully connected layer with output size of 1024
- fully connected layer with output size of 784

The testing results of these models can be seen in Table I. From the experiments we can see that Capsule network is performing better on MNIST dataset than the baseline convolutional network.

B. Other experiments

Other experiments were made, in which capsule networks outperformed baseline CNNs and sometimes even state-of-the-art models. One of the examples could be Variational Bayes

routing algorithm [5], which reached the lowest error of 1.6% on SmallNORB [13], 5.2% on Fashion-MNIST while using 170K parameters and 3.9% on SVH, 11.2% on CIFAR-10 while only using 323K parameters, which is good compared to baseline models with a small number of parameters of 2M and less.

Another Efficient-CapsNet with 150K parameters reached test errors of 0.26% on MNIST, which is a small improvement with the reduction of parameters as well as 2.54% on smallNORB, which is a bit worse than VB above, but uses a simpler routing algorithm and fewer parameters.

Other Capsule network implementations reached really good results as well compared to the CNNs with a similar number of computational parameters [14], [15], [16].

Capsule networks can be used in other fields and not just image classification and computer vision [17], [18], [19], [20].

V. Planned work

My research would be mainly into where capsules can be used. In addition to that, I would like to experiment with different topologies and algorithms to find the best for a few problems and find the main differences between them in terms of accuracy, speed of one epoch, number of epochs to finish training, and computational resources needed. The problem of creating different routing algorithm was already solved by multiple people in different ways, so the focus should mainly be on comparison and explanation of each one in a specific context. The creation of another routing algorithm should not be as important in my research as in other parts.

While I would work on this, I would like to try to find ways to mitigate some disadvantages of capsule neural networks. Mainly it would be to use segmentation maps to remove the background of the input images and if that would work, create another model, which would segment the object, and then the object would be put into the capsule neural network to maximise the accuracy.

Another part of the research would be to add some innovations used currently in state-of-the-art CNNs and transform them to work with capsule networks. Somehow adding skip connections or inception blocks at the beginning would probably have a favourable influence on the overall accuracy of capsules. This could also be part of my work, but it is highly improbable that this would work favourably.

The ideal output of my PhD. work would be to create a state-of-the-art capsule neural network on some real-world data such as some medical images or any other dataset, which would have real-world usage.

References

[1] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in *International conference on artificial neural networks*. Springer, 2011, pp. 44–51.

[2] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," *Advances in neural information processing systems*, vol. 30, 2017.

[3] M. K. Patrick, A. F. Adekoya, A. A. Mighty, and B. Y. Edward, "Capsule networks–a survey," *Journal of King Saud University-computer and information sciences*, vol. 34, no. 1, pp. 1295–1310, 2022.

[4] Z. Liao and G. Carneiro, "On the importance of normalisation layers in deep learning with piecewise linear activation units," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–8.

[5] F. D. S. Ribeiro, G. Leontidis, and S. Kollias, "Capsule routing via variational bayes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 3749–3756.

[6] V. Mazzia, F. Salvetti, and M. Chiaberge, "Efficient-capsnet: Capsule network with self-attention routing," *Scientific Reports*, vol. 11, no. 1, pp. 1–13, 2021.

[7] G. E. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with em routing," in *International conference on learning representations*, 2018.

[8] S. Zhang, Q. Zhou, and X. Wu, "Fast dynamic routing based on weighted kernel density estimation," in *International symposium on artificial intelligence and robotics*. Springer, 2018, pp. 301–309.

[9] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.

[10] C. Xiang, L. Zhang, Y. Tang, W. Zou, and C. Xu, "Ms-capsnet: A novel multi-scale capsule network," *IEEE Signal Processing Letters*, vol. 25, no. 12, pp. 1850–1854, 2018.

[11] D. Scherer, A. Müller, and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition," in *International conference on artificial neural networks*. Springer, 2010, pp. 92–101.

[12] L. Deng, "The mnist database of handwritten digit images for machine learning research [best of the web]," *IEEE signal processing magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[13] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 2146–2153.

[14] Y. Wang, L. Huang, S. Jiang, Y. Wang, J. Zou, H. Fu, and S. Yang, "Capsule networks showed excellent performance in the classification of herg blockers/nonblockers," *Frontiers in pharmacology*, vol. 10, p. 1631, 2020.

[15] E. Xi, S. Bing, and Y. Jin, "Capsule network performance on complex data," *arXiv preprint arXiv:1712.03480*, 2017.

[16] A. Jaiswal, W. AbdAlmageed, Y. Wu, and P. Natarajan, "Capsulegan: Generative adversarial capsule network," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

[17] I. J. Jacob, "Capsule network based biometric recognition system," *Journal of Artificial Intelligence*, vol. 1, no. 02, pp. 83–94, 2019.

[18] W. Wang, F. Lee, S. Yang, and Q. Chen, "An improved capsule network based on capsule filter routing," *IEEE Access*, vol. 9, pp. 109 374–109 383, 2021.

[19] Z. Jiao, H. Li, and Y. Fan, "Improving diagnosis of autism spectrum disorder and disentangling its heterogeneous functional connectivity patterns using capsule networks," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 1331–1334.

[20] D. K. Shukla, B. Keehn, A. J. Lincoln, and R.-A. Müller, "White matter compromise of callosal and subcortical fiber tracts in children with autism spectrum disorder: a diffusion tensor imaging study," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 49, no. 12, pp. 1269–1278, 2010.

New routing algorithms for cloud MANET

[1]*Natalia Kurkina (1st year),*
Supervisor: [2]*Ján Papaj*

[1,2]Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

[1]natalia.kurkina@tuke.sk, [2]jan.papaj@tuke.sk

***Abstract*—Nowadays, cloud solutions are becoming increasingly popular. Cloud MANET is a combination of mobile ad hoc networks (MANETs) and cloud technologies. One of the main problems in cloud MANETs, as in ordinary MANETs, is the issue of optimal routing. The optimal routing protocol should first of all be able to update the list of routes in use in real time. The use of machine learning methods to solve routing problems in MANETs is also becoming increasingly popular. Every year several new routing protocols are proposed for MANETs based on machine learning algorithms, but so far, they do not meet all the necessary requirements. This article contains basic information about MANETs, their characteristics, a description of Cloud MANET principles, a list of features that a routing protocol for this type of network should have, and a classification of protocols at the present time. The article also provides examples of routing protocols proposed over the past 2 years. In conclusion, the further direction of the study is described.**

***Keywords*—MANET, routing protocol, ad hoc network, machine learning**

I. Introduction

MANETs are wireless self-organizing networks without fixed infrastructure. These MANETs are organized by mobile nodes that connect to each other. A mobile node in such a kind of networks performs simultaneously the tasks of both the router, which forwards data further to the next hop node and the end station to which the data can be addressed.

Cloud MANETs are one of the subtypes of MANETs. This is a relatively new concept. It combines two areas: mobile ad hoc networks and cloud solutions [1]. In addition to mobile nodes in the network, the cloud participates in making a decision and forwarding traffic. All mobile nodes in the network are grouped into clusters. Each node can belong to only one cluster at the same time. In each cluster, one node is selected as the cluster head. It is responsible for connecting to the cloud. To transfer information from one cluster to another the information is first sent to the cloud, then the cloud sends the data to the corresponding cluster head of the other cluster. This cluster head then redirects the already given information to the destination node. The cloud MANETs scheme is shown in Fig. 1

Due to the fact that the nodes themselves are mobile and can move around within some area, enter and leave the network, and the fact they simultaneously act as routers, the network topology can be changed quickly, and ordinary routing protocols are not applicable in that case. Mobile nodes in such networks should be able to detect the presence of other devices in order to ensure communication and exchange of information. In addition, they should also be able to identify the types of services and the corresponding attributes. As the number of wireless nodes changes on the fly, the routing information also changes to reflect changes in interconnectivity.

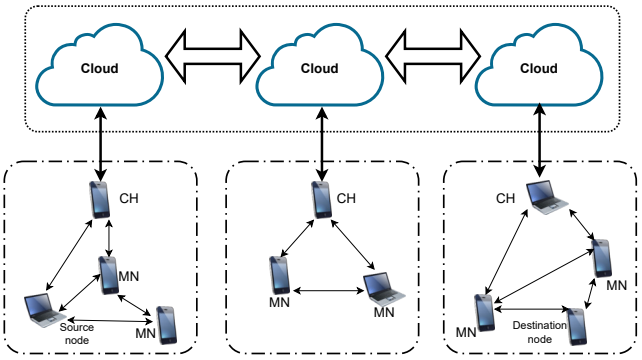


Fig. 1. Cloud MANETs

Currently, many protocols have been developed for MANETs, the most famous of them are AODV (Ad hoc On-Demand Distance Vector), OLSR (Optimized Link State Routing Protocol), DSV (Destination Sequenced Distance Vector Routing), ZPR (Zone Routing Protocol). But all of them are not applicable in cloud MANETs, due to the lack of clustering mechanisms. In the developed protocols, all nodes are considered as equal participants in the network.

As a result, there is a need for new protocols designed specifically for cloud MANETs.

II. Mobile Ad Hoc Network

The emergence of MANETs began in 1972 exclusively for military purposes. But in the mid-1990s, with the advent of commercial radio technologies, the wireless research community realized the enormous potential and advantages of mobile ad hoc networks, as evidenced by the creation of the Mobile Ad Hoc Networking Working Group within the IETF. Nowadays, research on mobile ad hoc networks is a very dynamic, actively developing, and promising area.

Deployment of mobile networks without infrastructure can be very useful in rescue operations in disaster areas where the infrastructure has been damaged and there is no way to restore it in the shortest possible time. Or when searching for injured in areas where there is no static infrastructure (for example, when searching for tourists in the mountains or the jungle). But MANETs can be used in other various situations, including in everyday life, for example, in police, firefighters and ambulances work, for highway patrol services and their traffic jam notification messages, for broadcasting information in hospitals or airports, in electronic commerce, for payment for goods and services anywhere in the world [2].

There are several types of networks such as

1) Vehicular ad hoc networks (VANETs) - mobile ad hoc network, connecting moving vehicles [3]
2) Cloud MANETs - combine two areas: mobile ad hoc networks and cloud solutions
3) SmartPhone ad hoc networks (SPANs) - connected smartphone into one network without using mobile networks, wireless networks, or traditional networks [4]
4) Internet based mobile ad hoc networks (iMANETs) - they combine mobile ad hoc networks and the Internet [5]. It can be considered this type of network as a private version of cloud MANETs, where the Internet acts as a cloud.

All MANETs are characterized by the following features [6]:

1. Dynamic network topology: nodes in the network are free to move, which means that the topology of the network can change quickly and randomly in a short amount of time.
2. Channels with limited bandwidth and variable bandwidth: wireless channels by their nature have significantly lower bandwidth compared to their wired counterparts. In addition, the bandwidth of wireless communication in a real environment is often much lower than the maximum radio transmission rate. This is due to the presence of many negative effects, such as attenuation, noise, and interference.
3. Low power consumption and resource: the network consists of mobile nodes that most likely run on batteries. Therefore, one of the main design criteria should be energy conservation.
4. Limited physical security: mobile wireless networks are more vulnerable to physical security threats than fixed cable networks. For example, there is an increased probability of sniffing, spoofing, and denial-of-service attacks, which should be carefully considered.
5. Decentralized Network Management: the decentralized nature of MANETs is one of the advantages because the failure of any one node will not lead to a stop in the operation of the whole network.

Cloud MANETs are also decentralized. Because the failure of the cluster head or any other node of the network should not lead to a network shutdown. In this case, the process of selecting a new cluster head or the process of searching for a new optimal route should be started.

Based on the above network features, the following features arise that the routing protocol should have:

1. Distribution: should not depend on a centralized control node
2. Loops free: To improve overall performance, the routing protocol must ensure that the provided routes do not have loops. This avoids bandwidth wasting and CPU consumption.
3. Work on demand: the protocol should respond only when necessary and not waste network resources by periodically sending service information.
4. Support for unidirectional communication: The radio environment can lead to the formation of unidirectional links. The use of these channels, in addition to bidirectional ones, improves the performance of the routing protocol.

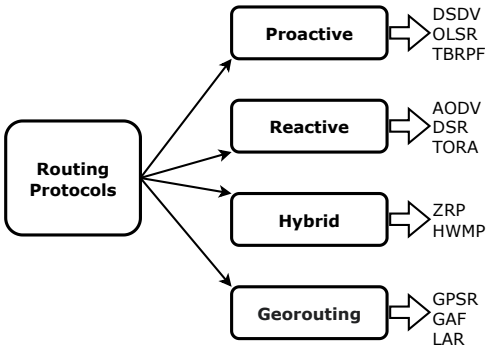


Fig. 2. MANETs Routing protocols

5. Security: in view of the increased vulnerability of the network to various attacks, the protocol must support authentication and encryption
6. Power saving: the nodes in the network can be smartphones, laptops, and other devices that have limited battery power; therefore, it is very important that the routing protocol minimizes the use of energy during operation.
7. Multiple route support: multiple routes can be used to reduce reactions to topological changes and congestion. If one route becomes invalid, it is possible that another stored route may still be valid and thus prevent the routing protocol from initiating another route discovery procedure.
8. QoS Support: protocol also should support the prioritizes traffic so that more important traffic can pass first.

Also, especially for cloud MANETs, the routing protocol must have an optimal node clustering algorithm.

III. Type of protocols for MANETs

The protocols used in MANETs are divided into several subcategories [7], Fig. 2.

1) Proactive protocols: - protocols that, like routing protocols in standard networks, store a complete routing table in memory. If there is no route in memory, then they send avalanche messages about the request/search for the destination node. Since, in fact, each node has a connectivity graph of the network, it is possible to build the shortest route, for example, using Dijkstra's algorithm. Proactive protocols include the following TBRPF (Topology Dissemination Based on Reverse-Path Forwarding), OLSR, DSDV. Due to the periodic exchange of information when using proactive protocols, some of the problems arise: loading the network with service information, a rapid decrease in the energy resources of each node.

2) Reactive protocols: - do not store the routing table in memory, and, if necessary, request/initiate the search for the required route. A node can either use an existing route to establish a connection or create a new route using information about available channels. Reactive protocols include protocols such as AODV, DSR (Dynamic Source Routing Protocol), TORA (Temporally-Ordered Routing Algorithm). The disadvantage of this group of protocols is the increase in delays in the search for the primary route with an increase in mobility and the number of nodes.

3) Hybrid protocols: - based on the name, they combine both approaches described above. As a rule, they break the network into many clusters. Within each cluster, a proactive

protocol operates, while the interaction between them is carried out by reactive methods. In large networks, this helps to reduce the size of routing tables maintained by network nodes. Hybrid protocols include protocols such as ZRP, HWMP (Hybrid Wireless Mesh Protocol). The disadvantage of hybrid protocols is the relative complexity of implementation and reduced routing efficiency associated with the need to divide the network structure into clusters.

4) Geo-routing protocols: - protocols that use data on the geographical location of nodes should be placed in a separate category. The main advantages of geo-routing protocols include the absence of a direct need to store route information at transit nodes and the ability to optimize routes based on available information about the location of nodes. As an example of protocols belonging to this class, it's possible to refer to GPSR (Greedy Perimeter Stateless Routing), GAF (Geographic adaptive fidelity), LAR (Location-Aided Routing).

At the time being, none of the existing protocols has all these properties and this is the main reason that the protocols are still under development. A large number of possible alternatives for routing protocols in MANETs has been proposed by the researching community. However, most of these solutions achieve a specific goal (for example, minimizing latency and overhead), compromising other factors (for example, scalability and reliability of the route). Thus, scientists are still searching for the optimal routing protocol that can meet most of the requirements.

IV. Routing protocols improved by machine learning in MANETs

In recent years, research in the field of MANETs has mainly focused on studying problems related to routing, clustering, power consumption, and network security [8]. Machine learning methods are used by scientists to improve any processes or tasks. In MANETs, they are often used for clustering nodes, choosing the best route, classifying nodes as safe ones and not, predicting the movement of nodes, and so on.

This article has some examples of using machine learning algorithms for routing in MANETs. In [8] authors propose a solution for detecting malicious attacks or compromised nodes in MANETs, using machine learning. They compare 2 methods, Linear regression (LR) and Support-vector machines (SVM), and come to the conclusion that LR copes better with this task than SVM.

In [9] a new routing protocol for VANETs based on machine learning is proposed (linear regression is also used). Using machine learning based on such parameters as velocity, acceleration, distance, link lifetime, and route score, researchers predict the lifetime of the link. This helps to calculate a new route, even before the existing one falls. Simulation results show that PQR is superior to existing protocols in several parameters.

Another routing algorithm for VANETs based on machine learning is proposed in [10] - PARRoT (Predictive Ad-hoc Routing). It also works based on the prediction of node mobility. But the researchers used Q-learning for that. The simulation results indicate that this protocol provides significantly higher reliability and significantly lesser end-to-end latency.

The paper [11] also proposes a routing protocol based on machine learning. In that article, Q-learning is used to determine the level of node mobility. A new metric called Q-metric is introduced for that. It combines dynamic and static routing metrics. As the result of the simulation, the authors provide evidence that this protocol allows to increase the percentage of package delivery. This protocol supports adding new metrics. And because of that, the authors plan to present a new protocol based on AQ-routing in the future, that takes into account energy costs in order to minimize them and to increase the service life of the network.

In [12] the authors propose an improved version of the AODV routing protocol, called AGEN-AODV. In their paper, the authors use a genetic algorithm to optimize AODV and a learning automat to tune GA coefficients in real time. Nodes, when exchanging messages, add three metrics to the packet: traffic rate, stability ratio, and remaining energy. This protocol allows to reduce power consumption and increase network uptime.

Many works are also devoted to the problem of network security, attack detection, isolation of malicious nodes, detection of so-called black holes in the network. For this, machine learning algorithms are also actively used [13], [14]. Machine learning is also used to classify and cluster nodes. In the papers [15], [16], [17] it can be seen that the most commonly used algorithms for this are the following: Neural Network, Q-learning, Support Vector Machine, a Decision Tree, Random Forest.

V. Comparative study of supervised machine learning technics

In addition to the description of the current routing problems in MANETs, an analysis was also made of supervised machine learning techniques that can be used to classify nodes in MANET networks. The analysis was carried out using MATLAB. Three data sets were used for the analysis:

1) Wine Quality Data Set - contains 4898 lines of data, 11 input variables, all data are classified into 6 classes.
2) Wine Data Set - contains 178 lines of data, 13 input variables, all data are classified into 2 classes.
3) Ionosphere Data Set - contains 351 lines of data, 34 input variables, all data are classified into 2 classes.

Each data set was randomly split so that 70% of the data was used for training and 30% for testing. The split of data was done with stratification using the class information in the data set. Stratification was used because the datasets are not balanced across classes [18]. For greater independence of the results from the choice of values for training and testing, 100 independent iterations were performed. The following machine learning methods were used for classification: Naive Bayes Classifier (NBC), Classification Tree (CTree), K-nearest neighbor (KNN), Multinomial logistic regression (MLG), Support vector machine (SVM), An artificial neural network (ANN), Random Forest (RForest).

For each method, the confusion matrix was calculated in each iteration. For comparison, data on the probability of error, F1 score, Precision, and Sensitivity for all classes and iterations were used. This data has been calculated from the data in the confusion matrix using the formulas given in [19].

As a result, the following experimental data are obtained, presented in Figs. 3 and Table I. As shown, the most accurate classification is given by the Random Forest algorithm. The RForest has a minimal value of the probability of error and

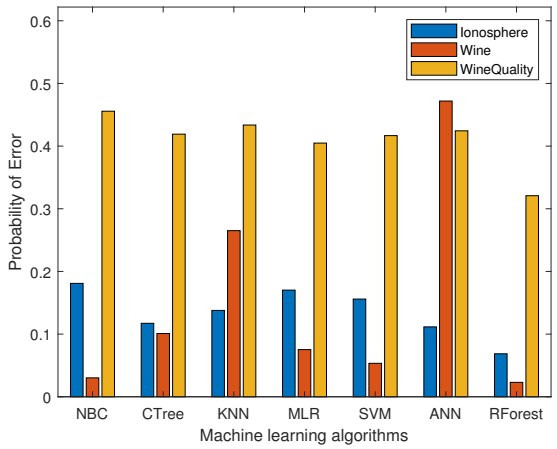


Fig. 3. Probability of Error for one hundred iterations in MANET

TABLE I
THE F1 SCORE FOR EACH MACHINE LEARNING ALGORITMS

Data Sets	Machine learning algoritms						
	NBC	*CTree*	*KNN*	*MLR*	*SVM*	*ANN*	*RForest*
F1 score							
Ionosphere	0,8113	0,8718	0,8368	0,8064	0,8175	0,8748	**0,9240**
Wine	0,9706	0,8994	0,7248	0,9255	0,9470	0,3734	**0,9774**
Wine Quality	0,3177	0,3076	0,3220	0,0111	0,2124	0,2626	**0,3557**
Precision							
Ionosphere	0,8088	0,8762	0,8874	0,8288	0,8586	0,8904	**0,9332**
Wine	0,9704	0,9056	0,7333	0,9286	0.9484	0,3304	**0,9762**
Wine Quality	0,3224	0,3108	0,3320	0,0651	0,1949	0,2788	**0,4090**
Sensitivity							
Ionosphere	0,8279	0,8708	0,8175	0,7962	0,8017	0,8663	**0,9177**
Wine	0,9728	0,9017	0,7267	0,9271	0,9491	0,4794	**0,9801**
Wine Quality	0,3258	0,3085	0,3221	0,1429	0,2347	0,2696	**0,3468**

maximal values of the F1 score, precision, and sensitivity. The probability of error rate for this technic is 0.0686, 0.0230, 0.3208, respectively for each of the data sets. The F1 score value is 0.9240, 0.9774, 0.3557, respectively. Precision is 0.9332, 0.9762, 0.4090, respectively. Sensitivity is 0.9177, 0.9801, 0.3468, respectively. In this simulation data set acts as some hypothetical parameters exchanged by mobile nodes in MANET. For research in real networks, the Random Forest algorithm looks promising and could be taken into account.

VI. Conclusion and Further Research Direction

Routing in MANETs is a key aspect in their success and propagation. Previously, MANETs were focused only on working in emergency situations when any other network was not working. Now MANETs can be used in various fields and must adapt to much larger requirements (such as QoS, high stability, bandwidth, lesser power consumption, security).

However, the dynamic nature of MANETs makes it extremely difficult to achieve the aforementioned expectations. It is necessary not only to find optimal routes from the source to the destination but also to allocate limited resources as efficiently as possible. This leads to the necessity to take into account many criteria to optimize the operation of the network.

In this connection, many scientists have turned to machine learning methods as the likely alternative to solving this situation, because machine learning offers more intuitive algorithms that allow solving problems in an optimal way with much lower computational costs compared to other optimization algorithms.

In further research, we would like to focus on finding the optimal algorithm for distributing nodes into clusters and choosing a cluster head using machine learning algorithms. Also, further research will focus on exploring the possibility of improving/developing a new more optimal routing algorithm for cloud MANETs. A comparative analysis will also be made with existing routing protocols. And as far as possible, practical tests will be carried out to verify the operation of the protocol in a real environment.

Acknowledgment

This research was funded by the Slovak Research and Development Agency, research grant no. APVV-17-0208.

References

[1] S. A. Alghamdi, "Novel trust-aware intrusion detection and prevention system for 5g manet–cloud," *International Journal of Information Security*, 2021.

[2] S. K. Singh and J. Prakash, "Energy efficiency and load balancing in manet: A survey," in *2020 6th International Conference on Advanced Computing and Communication Systems, ICACCS 2020*, 2020, pp. 832–837.

[3] S. Al-Sultan, M. M. Al-Doori, A. H. Al-Bayatti, and H. Zedan, "A comprehensive survey on vehicular ad hoc network," *Journal of Network and Computer Applications*, vol. 37, pp. 380–392, 2014.

[4] S. U. Dixit and K. M. Patil, "Smart phone wireless ad-hoc mesh network," in *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, RTEICT 2018 - Proceedings*, 2018, pp. 2346–2350.

[5] B. Kim, I. Kim, and Y. Woo, "Hop-based exclusive neighborhood caching scheme for content centric imanet," *International Journal of Engineering and Technology(UAE)*, vol. 7, no. 3, pp. 143–146, 2018.

[6] K. Ahmad, *Opportunistic Networks: Mobility Models, Protocols, Security, and Privacy*, 2018.

[7] A. Verma, P. Verma, S. Dhurandher, and I. Woungang, *Opportunistic Networks: Fundamentals, Applications and Emerging Trends*, 2021.

[8] R. Sebopelo, B. Isong, and N. Gasela, "Identification of compromised nodes in manets using machine learning technique," *International Journal of Computer Network and Information Security*, vol. 11, pp. 1–10, 01 2019.

[9] W. Xu, X. Ji, C. Zhang, B. Zhang, Y. Wang, X. Wang, Y. Wang, J. Wang, and B. Liu, "Pqr: Prediction-supported quality-aware routing for uninterrupted vehicle communication," in *2021 IEEE/ACM 29th International Symposium on Quality of Service, IWQOS 2021*, 2021.

[10] B. Sliwa, C. Schuler, M. Patchou, and C. Wietfeld, "Parrot: Predictive ad-hoc routing fueled by reinforcement learning and trajectory knowledge," in *IEEE Vehicular Technology Conference*, vol. 2021-April, 2021.

[11] A. Serhani, N. Naja, and A. Jamali, "Aq-routing: mobility-, stability-aware adaptive routing protocol for data routing in manet–iot systems," *Cluster Computing*, vol. 23, no. 1, pp. 13–27, 2020.

[12] M. Nabati, M. Maadani, and M. A. Pourmina, "Agen-aodv: an intelligent energy-aware routing protocol for heterogeneous mobile ad-hoc networks," *Mobile Networks and Applications*, 2021.

[13] J. Batra and C. Rama Krishna, "Ddos attack detection and prevention using aodv routing mechanism and ffbp neural network in a manet," *International Journal of Recent Technology and Engineering*, vol. 8, no. 2, pp. 4136–4142, 2019.

[14] P. Gulganwa and S. Jain, "Ees-wca: energy efficient and secure weighted clustering for wsn using machine learning approach," *International Journal of Information Technology (Singapore)*, 2021.

[15] T. Koshimizu, S. Gengtian, H. Wang, Z. Pan, J. Liu, and S. Shimamoto, "Multi-dimensional affinity propagation clustering applying a machine learning in 5g-cellular v2x," *IEEE Access*, vol. 8, pp. 94 560–94 574, 2020.

[16] V. Surya Narayana Reddy and J. Mungara, "Machine learning-based efficient clustering and improve quality of service in manet," *Indian Journal of Computer Science and Engineering*, vol. 12, no. 5, pp. 1392–1399, 2021.

[17] P. Tamilselvi and T. N. Ravi, "Hybridization of brownboost and random forest tree with gradient free optimization for route selection," *International Journal of Computing*, vol. 20, no. 3, pp. 400–407, 2021.

[18] A. Fernández, S. García, M. Galar, R. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*, 01 2018.

[19] S. Uddin, A. Khan, M. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making*, vol. 19, 12 2019.

A Clustering Method for One-Shot Learning

[1]*Andrinandrasana David Rasamoelina (3rd year),*
Supervisor: [2]*Peter Sinčak*

[1,2]Dept. of Cybernetics and Artificial Intelligence, FEI TU of Košice, Slovak Republic

[1]andrijdavid@tuke.sk, [2]peter.sincak@tuke.sk

***Abstract*—Learning from a limited number of instances is a fundamental unresolved topic that has been thoroughly investigated in the literature. Few-Shot Learning (FSL) tries to solve this issue. In this, work we describe approaches that perform well in a one-shot setting.**

***Keywords*—Few-Shot Learning, Computer Vision**

I. Introduction

In recent years, deep learning has become more prevalent in our everyday life. It is used in multiple applications such as voice assistants, self-driving cars [1], disease mapping [2], automated financial investing [3], and more. This became possible due to the advent of big data. As the internet and our technology grew, access to data and knowledge became cheaper [4]. However, there exists still some domains where data is scarce for example in healthcare or situation where the data is abundant but not labeled or structured. This makes the use of deep learning difficult for such problems. Few-shot learning tries to solve such problems by proposing different algorithms and approaches. Those algorithms are able to learn given very few data points. For example, One-shot learning algorithms deal with problems with only one instance per class. Generally, we use the terminology N-shot, where N is the number of instances available per class, and N varies from 0 to 25. Although there is no upper limit set in stone, 25-shot is the largest used in the literature now.

II. Related Work

There are multiple ways to solve this type of problem in deep learning. One of them is Meta-Learning, a field that took off in recent years. The idea is to mimic the way human learns. Humans learn by recognizing general concepts and patterns and exploiting their past experiences. So when we handle new unknown tasks, we rely on our knowledge even if it is a totally different domain. Our ability, to learn that task rapidly depends on how well we are able to transfer our knowledge to this new task. In simple words, we take advantage of our past knowledge to acquire a new skill.

Meta-Learning applies that principle to deep learning by training on multiple datasets of small tasks (meta-training) and also testing on multiple datasets of small tasks [5]. The learning algorithm used varies but they belong to 4 distinct groups:

- Learning to compare: the network learns to compare and predict if two instances belong to the same class (Prototypical Network [6], Matching Network [7], Siamese Neural Network [8], Relation Network [9],).
- Learning to remember: the network has an additional memory component that helps them to remember (Memory Augmented Neural Network (MANN) [10], Neural Turing Machine [11]).
- Learning to optimize: the network has a custom optimization flow that allows fast convergence on a new task (Meta Networks [12]).
- Learning to initialize: A network predicts the initialization weight of another network that is easier to optimize and train afterward (Model-Agnostic Meta-learning (MAML) [13], Reptile [14]).

III. Our Contribution

Our work solely focuses on the family of learning to compare algorithms. Those algorithms take advantage of a given encoder to produce a vector that represents the input. The outputs vectors of this encoder are compared to another known vector. If the distance is close enough we can assume that they belong to the same class. In hindsight, we can split this algorithm into 2 steps:

- Embedding generation
- High dimensional clustering.

We purpose another way for clustering and computing our distance in high dimensions. We show that our approach is effective in One-shot settings.

IV. Experiment Settings

A prototypical network has two parts: the feature extractor and a classifier. The feature extractor is 4 layers convolutional network. It produces an embedding (representation) of the input into a high-dimensional space. The classifier is a simplified version of k-nearest neighbor. It takes the embedding of a given input and computes the distance to the prototypes. The prototypes are the average of the embeddings of a given class from our training set. The given input is then assumed to belong to the same class as the closest prototype. This network is then trained end to end by minimizing the distance between the prototype and the inputs. In hindsight, we want the same class to cluster close to each other.

Formaly, Let f_θ be our model with parameters θ. f_θ can be any neural network or convolutional neural network with any arbitrary number of parameters producing an embedding vector as output. Given a task τ_i and a dataset ($\mathcal{D}^{(i)}_{\text{train}} = x_i, y_i, \mathcal{D}^{(i)}_{\text{test}} = x_i$), where i is the index of our task.

We can compute the prototype of a given class k using

$$\mathbf{p}_k = \frac{1}{|\mathcal{D}^{(i)}_{\text{train}_k}|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}^{(i)}_{\text{train}_k}} f_\theta(\mathbf{x}_i) \tag{1}$$

Methods	Accuracy
Euclidean	49.8
Our distance	53.26
Our distance + Our Clustering metric	**54.37**

TABLE I
COMPARISON BETWEEN OUR METHOD AND THE FORMER METHOD. THE METRIC IS THE AVERAGE PERFORMANCE OVER 600 TASKS.

The probability of a given item x that it belongs to a class k can be computed using:

$$prob_\theta(y = k \mid \mathbf{x}) = \frac{\exp(-d(f_\theta(\mathbf{x}), \mathbf{p}_k))}{\sum_{k'} \exp(-d(f_\theta(\mathbf{x}), \mathbf{p}_{k'}))} \quad (2)$$

where d denote any distance function and exp the exponential function.

For a normal Prototypical network, our learning is then done by minimizing the cross-entropy loss

$$J(\theta) = -\log prob_\theta(y = k \mid \mathbf{x}) \quad (3)$$

V. DISTANCE

Numerous works show the ineffectiveness of Euclidian distance in high dimensions [15]. In high dimensions, distance tends to converge and concentrate in one value. In [6], the author used the squared Euclidean distance. However, this slows down the distance concentration but does not solve the problem. We propose a way to compute the distance in another manner. We consider points in high dimensions as 2-dimensional waves. Then the distance between two waves can be expressed as the area in between the two waves.

We can argue that the area represents an individual feature and the gap between the features is what differentiates the two points. This approach satisfies all the requirements to be considered as a distance function: For any points x, y, z

- Identity of indiscernible. $d(x, x) = 0$
- Symmetry. $d(x, y) = d(y, x)$
- Triangle inequality $d(x, z) < d(x, y) + d(y, z)$

VI. CLUSTERING METRIC

We include other metrics to improve the quality of the cluster formed by our encoder. We are mainly interested in maximizing the inter-cluster variance and minimizing the intra-cluster variance. We argue that having a clearly defined cluster should improve the classification that follows. This is indeed very intuitive if we use a simple classifier like a k-nearest neighbor. We propose to use a modified version of the Davies Bouldin index [16]. Therefore, we extend the previous loss function (eq. 3) by including a metric for the quality of the individual cluster.

VII. RESULT

Our algorithm have been tested on *mini*Imagenet. It was first trained on 10 000 different one-shot task. Then we test the performance of our approach on 600 tasks. The result presented in table I are the average on 600 tasks. We see that our approach outperforms the previous method for one-shot learning.

VIII. CONCLUSION

In this paper, we briefly describe few-shot learning. We listed different methods and then we showed a definition of a Prototypical Network. We introduced two approaches to improve this well-known algorithm: a custom high dimensional distance function and an optimizable clustering score. We show that our approach outperforms the legacy method in a one-shot setting. Our future work is to generalize this approach to n-shot learning.

REFERENCES

[1] Q. Rao and J. Frtunikj, "Deep learning for self-driving cars: Chances and challenges," in *Proceedings of the 1st International Workshop on Software Engineering for AI in Autonomous Systems*, 2018, pp. 35–38.

[2] D. Ushizima, Y. Chen, M. Alegro, D. Ovando, R. Eser, W. Lee, K. Poon, A. Shankar, N. Kantamneni, S. Satrawada *et al.*, "Deep learning for alzheimer's disease: Mapping large-scale histological tau protein for neuroimaging biomarker validation," *NeuroImage*, p. 118790, 2021.

[3] W. Bao, J. Yue, and Y. Rao, "A deep learning framework for financial time series using stacked autoencoders and long-short term memory," *PloS one*, vol. 12, no. 7, p. e0180944, 2017.

[4] A. Labrinidis and H. V. Jagadish, "Challenges and opportunities with big data," *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 2032–2033, 2012.

[5] Q. Sun, Y. Liu, Z. Chen, T.-S. Chua, and B. Schiele, "Meta-transfer learning through hard tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[6] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Advances Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 4077–4087.

[7] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," in *Advances in neural information processing systems*, 2016, pp. 3630–3638.

[8] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," in *Advances in neural information processing systems*, 1994, pp. 737–744.

[9] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199–1208.

[10] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *International conference on machine learning*, 2016, pp. 1842–1850.

[11] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," *arXiv preprint arXiv:1410.5401*, 2014.

[12] T. Munkhdalai and H. Yu, "Meta networks," *Proceedings of machine learning research*, vol. 70, p. 2554, 2017.

[13] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1126–1135.

[14] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018.

[15] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *International conference on database theory*. Springer, 2001, pp. 420–434.

[16] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.

Implementation of Augmented Reality in Industry

[1]*Ivana NOVÁKOVÁ (1st year),*
Supervisor: [2]*František JAKAB*

[1,2]Dept. of Computers and Informatics, FEEI TU of Košice, Slovak Republic

[1]ivana.novakova@tuke.sk, [2]frantisek.jakab@tuke.sk

***Abstract*—Industry is a sector of continuous improvement, from small partial innovations to system solutions. The improvement of Internet of Things (IoT) and significant advances in information and communication technologies (ICT) have led to a growing demand for AR / VR / MR technologies, which has led to its choice as an adequate test environment in the dissertation. They are based on data visualization and intuitive interaction with them, which leads to progress in data acquisition, processing, and storage. Although studies show an enormous increase, the potential for their use has still not been realized due to the technological shortcomings they encounter in implementation, from creating an adequate environment in a digitized company, through technological shortcomings of AR / MR equipment to the impact on employees.**

***Keywords*—Augmented reality, HDM devices, limitations of AR, implementation of AR.**

I. Introduction

The thesis considers augmented reality as a tool that, with the right implementation, can help the industry to prosper better. In the first phase of specifying the direction of research activities, an overview of technical constraints that the selected technology is currently facing was performed. Pointing shortcomings in augmented reality display technology can be attributed to inappropriate conditions, imperfect equipment, and ever-improving technology. These factors influence the direction of development that should be performed on various HDM devices. Working with real-time data visualization in the relevant industrial environment can be considered the basic point of research.

AR combines the observed phenomena of the real world with graphically added information [1] as reported by R. Azuma. The purpose of advanced digital information is to enrich the original phenomenon with information that is useful in many aspects, including studies. Enrichment is through graphics, simulation, or a 3D model that could improve understanding [2]. AR solutions are a system with different components that can be understood as cyber-physical (production) systems (CPS/CPPS,). Cyber-physical (production) systems are considered a core element of Industry 4.0 and describe the connection of the real (physical) world with up-to-date digital (cyber) representatives [3].

With gradual digitization, increasing connectivity in modern digitized companies, and the integration of additional sensors as well as sophisticated IT systems (e.g., ERP, MES) that provide detailed information, obtaining input data for augmented reality is no longer a critical problem, as it was a few years ago.

II. Augmented Reality

The modern principle of interoperability between people and data dates to the 1950s. Interoperability between acquired and displayed data in close connection with the humanity gives digitization a better dimension for understanding deeper, more specialized data in 3D space, while allowing you to interactively control, create or share content.

AR is often and mistakenly confused with virtual reality. It is certain that the principles of augmented reality are based on the knowledge of Virtual Reality. There are several views on their relationship, as well as on the very definition of augmented reality, which leads to an incoherent definition of technology.

The first view is that augmented reality is completely different from virtual reality. As described by Milgram et al., [4] while virtual reality creates a new digital world, AR combines this world with reality. The second view of their relationship is that augmented reality is a special virtual reality. If the virtual world connects with the real world, which are two separate but complete worlds, the so-called virtual continuity, which is considered a virtual world, but changed, hybrid, and thus an augmented reality is created. Its opposite is augmented virtuality, virtual reality augmented by elements of the real world.

All these realities (real world, virtual world, augmented reality, and augmented virtuality) are part of the phenomenon, the so-called mixed reality [4] which is shown in Fig. 1. It involves projections and direct interactions with real one's objects, by extending the data through holograms in combination with smartphones or displays mounted on the head (transparent HMD) or real objects/users become part of virtual worlds through HMD or in VR room [4] [5].

In fact, it can be deduced that AR / VR / MR are interdependent technologies that are subject to continuous improvement and their preferential use is directly dependent on the implementation itself. It is helpful for industrial manufacturing companies to focus primarily on AR / MR technologies that allow interaction with the displayed content in a real environment.

The latest phenomenon can be considered XR technology, which creates an absorbing environment when it is almost impossible to distinguish simulated from real. This modern technology is an all-encompassing term for AR, VR and MR, but not the other way around. T. Andrade et al. [6] states that the most profound use is in the visualization of IoT in virtual reality.

The division of augmented reality can be understood from

Fig. 1. Milgram's reality-virtuality continuum

several perspectives, from the division according to the senses, use, space, or the chosen display technology.

A. Sensor-based AR

Mobile devices have built-in augmented reality recognition systems, where they are mediated by looking at the screen through a camera. Sensor based AR uses GPS with a built-in sensor in the device for location and orientation in space. The position sensor is the most supported GNSS receiver [7]. Several sensors, such as accelerometers, magnetometers, and gyroscopes, are available in handheld devices for orientation. The sensor AR uses a POI (points of interest) in a simple graphics interface to display information in space [7] as stated by Jens Grubert et al in a publication focused on the development of applications for Android mobile devices.

Augmented reality uses a complex of different methods [8] for the possibilities of computer vision, but the basic principle consists of two phases, namely the observation and reconstruction of the image. The current technology allows to recognize several types of planar pictorial content, such as a specifically designed marker (marker-based tracking) or marker less tracking.

B. Basic division of AR

Spatial Augmented Reality (SAR) [9] is an example of functional AR surveillance methods that use video projectors, optical elements, holograms, radio frequency markers and other surveillance technologies to be displayed in space without the need for HMD (Head Mounted Display) [7]. SAR is also called projection mapping because it extends the content through the projection of a beam projector onto a real object. The key difference in SAR is that the display is separate from the system users. Because the displays are not assigned to each user, the SAR naturally extends to user groups, allowing users to collaborate.

Simultaneous Localization and Mapping (SLAM) [10] is a simultaneous localization and mapping of 3D objects in a previously unknown environment.

Structure from Motion (SFM) [7] is a technique of finding and mapping the environment by monitoring functional points and estimating camera parameters.

C. AR interface

One of the most important aspects of augmented reality is to create proper techniques for intuitive interaction between the user and the virtual content of AR applications. There are four main ways of interaction in AR applications as reported by [8] Carmigniani et al: tangible AR interfaces, collaborative AR interfaces, hybrid AR interfaces, and the emerging multimodal interfaces.

Tangible interfaces support direct interaction with the real world using physical objects and tools. In an industrial environment, for example, there may be a choice of tool for performing assembly.

Collaborative AR interfaces include the use of multiple displays to support remote and co-located activities. Co-located sharing uses 3D interfaces to improve physical collaborative workspace. In remote sharing, AR can effortlessly integrate multiple devices with multiple locations to enhance teleconferences [8].

Hybrid interfaces combine an assortment of different, but complementary interfaces as well as the possibility to interact through a wide range of interaction devices. According to Feng Zhou at al. [11] they provide a flexible platform for unplanned, everyday interaction where itis not known in advance which type of interaction display or devices will be used.

Multimodal interfaces combine real objects input with naturally occurring forms of language and behaviours such as speech, touch, natural hand gestures, or gaze. These types of interfaces are more recently emerging [8]. MS HoloLens require a specific approach. Unlike mobile applications, HDM software solutions are more complex and require user emphasis in terms of optics.

The entry and most crucial point for the optics of the functioning of holographic glasses is the human eye, specifically the pupil of the human eye.

Interpupillary distance further IPD is the distance of the pupils measured from the center of the first to the center of the second pupil of the human eye, which is given in millimetres. The second option is to measure the IPD from the center of the pupil of one eye to the root of the nose and from the center of the other pupil to the root of the nose. In the second case, the measured value in millimetres is given in two numbers [12].

IPD resolution is an important technological data resulting from an anthropological study for the specification of the optimal range of output optics as well as the design of a wide range of binocular devices [13].

To work with augmented reality, several ways of manipulation have been defined, which are closely related to the specific device for which a particular technology was developed.

Bowman et al. [14] have designated three basic virtual object manipulation tasks for Virtual Reality (VR) and AR: selection, positioning, and rotation.

Control is possible using buttons, hand gestures as well as voice control. Today, applications allow a combination of several methods for absolute user interaction with a selected device, which can be a handheld device, smart glasses or HDM in the form of a helmet with an integrated system.

Button-based positioning uses either physical buttons or handheld touch screen buttons applicable to smartphones such as Henrysson et al. Tablets to position virtual objects. [8] used the smartphone's physical buttons for a location where different buttons are mapped to different degrees of freedom in space (DOF). Jung et al. [15] developed a system in which virtual objects can be placed in 3D by controlling one DOF at a time using a drag gesture with one or more touches. The controlled DOF is based on the position of the device compared to the base plane. The base level must be defined by the initial designation. Marzo et al. [16] used the DS3 technique [17] for 3D multi-touch gesture positioning on a smartphone. Their method displayed the shadow on the ground

plane below the virtual object as a depth indicator that adds a new dimension of reality in space to the graphical interface. Mossel et al. [18] developed a method in which positioning is performed by a sliding gesture.

III. Augmented Reality in Industry

Augmented reality has defended its strong position especially in industry, where it is preferentially used as a support and work tool that brings companies a myriad of benefits. The most common use is in the field of maintenance, assembly, and training of new employees. It is also gaining increased popularity in logistics and trade, especially in warehouses where the technology is used mainly for navigation and categorization of goods.

The dynamism of the industry offers innovative technologies such as augmented reality the space to adapt and develop, giving companies the space to respond quickly and effectively to market changes, increase the efficiency of their employees, reduce mistakes caused by new team members and create a smart digital environment with a focus on the future. Fault-free operation, long-term sustainable asset value and safety are the priorities of every business. The tools that augmented reality offers can bring companies to the desired effect with a slow, correct implementation plan.

A. Maintenance

Maintenance applications can be considered the most beneficial in industrial manufacturing companies. The AR application should conveniently expand the technician's field of vision with all the necessary information without overwhelming him. The basic input is the loading of metadata, 3D models, and media that the application supports. As mentioned above, simpler devices allow the display of plain text guides, in the form of a guide where user interaction consists of clicking out the next steps, which do not require complex calculation algorithms or expensive equipment.

Advanced UX and UI design are required to make sophisticated space available on the lenses of AR glasses or on the screen of an AR-compatible device. Multi-step user testing is needed to ensure intuitive navigation in the AR software.

Application development software is some of the most used for thoughtful purposes are Vuforia and Unity. Unity 3D is a multiplatform game engine developed in C ++ and used to create AR / VR applications in the Sharp, C ++ programming languages. Vuforia is an augmented reality software development kit for mobile devices that enables the creation of augmented reality applications. It uses computer vision technology to recognize and track planar images and 3D objects in real-time.

Essential things are important for AR maintenance application developers:

- type of resulting devices (HDM, tablet, mobile)
- number and scope of user scenarios (such as audit, preventive maintenance, corrective maintenance, regular maintenance and combinations thereof)
- 3D machine models and a range of details that will be interactive
- machine learning
- integration of machine with equipment, integration with existing ERP company.

Maintenance applications are often also used for employee training. Animation and X-ray vision features in augmented reality training allow participants to look inside, visualize hidden components, and better understand how systems work. As these trainings are not linked to actual equipment, the training may take place even when the physical equipment not available is new, decommissioned or out of service. Many multinational companies have already understood that the benefits of augmented reality are undeniable, especially in maintenance and assembly. BOSCH [19] states that their car technicians now spend 15% less time assembling components, using AR navigation.

Boeing [20] uses AR for repair and preventive maintenance of power lines in its aircraft. Engineers who previously used a 20-foot discontinuous 2D drawing as a reference now rely on the AR application, which expands the view of a real aircraft using a dynamic 3D engagement model. The innovative solution has resulted in a 30% reduction in time and up to a 90% improvement in the quality of work performed by beginning technicians.

B. Limitations that AR brings to industry

Augmented reality, however, is not an isolated technology and for its successful incorporation into the company, it is necessary to create a suitable environment that will allow the collected data to be further processed interactively.

A suitable environment soon could be the introduction of a 5G network, which companies consider to be the second most beneficial technology after cloud solutions. It is expected that manufacturing companies will prefer private 5G networks, as this innovation in the world of networks brings more efficient interconnection of frequency bands, lower latency, and the ability to connect more devices, which means that thousands of devices communicate through a single transmitter. With peak data rates of 20 Gbps, 5G is up to 20 times faster than 4G, can transmit data with an extremely low latency of one millisecond (essentially without delay) and is almost as reliable as cable data transmissions with a high reliability of up to 99.9999 percent.

Another challenge where there is room for improvement is the software solutions themselves for companies, which so far have more limitations than benefits for companies and their deployment is usually only partial, isolated for one task or focus. The complex potential of utilization requires access to sensitive company data, connectivity with analytical IT solutions as well as a 3D model from machine and equipment manufacturers. Industry requires better and more open collaboration between machinery and equipment manufacturers, businesses, software companies and designers.

When deploying an AR that contains the digital twin of a real machine, it is necessary to model it, which will scan the reference mark, which can be in the form of a simple QR code or company logo, to launch an interactive application. A possibility for simpler applications is to scan its outline, which is currently the most widely used way to run applications for assembly tasks, maintenance tasks or training. But this solution is not the most suitable eventually.

More complex operations use the positioning of holograms in space to run applications, which is possible, for example, with Hololens.

Limitations in the devices and their disadvantages are described in the chapter below. In addition to technological

shortcomings, businesses face a general problem such as the number of devices. Ideally, each employee should be assigned their own equipment, which would significantly reduce hygiene constraints, the need to adapt the equipment for the eyes in the case of smart glasses, as well as monitoring assigned tasks and evaluating employee efficiency. Hygiene measures are a fundamental change today, which has not received much attention so far. Especially shared HDM devices require regular cleaning with alcohol products that prevent the spread of viruses and bacteria. Frequent application of disinfectant sprays can cause stinging of the eyes, nausea from inhalation of alcohol fumes or dizziness, which is another undesirable side effect.MS tried to solve this problem in the Hololens 2 models by assigning a role, scanning the user's pupil, which allows multiple users to use 1 device under different roles.

C. DM devices and their limitations

The market offers a narrow range of products that could be described as the most suitable equipment for industrial use, and each of the products has its advantages and disadvantages.

The usability of augmented reality as a supportive and working tool requires consideration of external factors that negatively affect the potential that technology brings.

Compromises, given the high prices, are not as welcome by companies as producers would imagine. An example is the Hololens device from MS, which developed an HDM device that was supposed to occupy a leading position in the industry, but these glasses proved to be unbearable for a long time, as they weighed almost 1 kg, which led to a discrepancy of workers. The second row of the Hololens 2 is much more comfortable and ergonomic, it has the option of tilting the display upwards, which the first row did not allow, but the quality of the display went much lower. There are several professional critical articles that rely on Microsoft, which has promised more than it has been able to sell, and the implementation of augmented reality HDM Hololens appears to be more than disadvantageous for several factors.

Oqmented, a start-up company, develops glasses and uses a variety of 3D imaging and imaging solutions based on MEMS[1]. Their modules are characterized using the Lissajous scanning process, where the scanning direction is both horizontal and vertical. The Lissajou principle is used to fine-tune and adjust the phase relationship between a known reference signal and the signal to be tested. The process of converting a normal image to a Lissajous scan is complicated, leading to the problem that the original rectilinear image must be mapped to an ever-changing Lissajous scan[2]. Scanning moves at a variable speed based on two different sine functions, and speed changes must be compensated using drive control. All these scaling and colour resolution and fidelity compensation, leading to a significant reduction in colour gamut and a reduction in the display of only low-resolution numbers for this module, which cannot be considered the final product.

Oqmented, Displelix in partnership with OSRAM and many other companies are part of the LaSAR alliance[3] which is an effective way to eliminate the problems facing AR in the industry. By creating an ecosystem, they jointly face the challenges of improving development, their common feature being LBS (laser beam scanning). Most laser scanning screens (LBS) have so far used a raster scanning approach. Horizontal scanning is usually a much faster sine scan in kilohertz (typically 5 kHz to 54 kHz depending on the resolution) and a slower, slightly linear, driven vertical scan with a fast return. LBS is the most efficient way to develop unbearable devices, but from a physics point of view, developers still face several recurring problems, which in turn refutes this claim.

Problems associated with their use in AR glasses include:

1) Optics directing laser light to the eye and the problems it causes with vision
2) Electromechanical scanning with low resolution and low frame rate
3) Extremely small eye box / pupil (image disappears unless it is perfectly aligned with the eye)
4) Complexity of combining three (or more) lasers into one highly coaxial and tight set of colour beams
5) Cost of lasers compared to LEDs
6) Highly variable speed beam brightness control
7) Throwing shadows on the retina due to "floats" in the eye

For this reason, it is possible to judge why AR glasses with the use of LBS technology are used for simple guidance using text characters, unlike more complex graphics, which are necessary for industrial use.

At the SPIE AR / VR / MR 2022 conference in the United States earlier this year, Kevin Curtis (Vice President of Optical Engineering Magic Leap, presentation "Unveiling the Advanced AR Platform and Revolutionary Magic Leap 2 Optics" that After evaluating all the alternatives, Magic Leap decided that LCOS (Liquid Crystal on Silicon) is the best imaging technology for their application.

Liquid crystal on silicon (LCoS or LCOS) is a microdisplay technology that uses a liquid crystal layer on top of a silicon backplane. The technology, a kind of 'spatial light modulator', (SLM), provides high resolution, contrast, and black levels compared to competing technologies including liquid crystal display (LCD) and digital light processing (DLP). As there is an ever-growing demand for high-resolution displays, the LCoS segment of the market is projected to increase 32.25% CAGR through 2024 [21].

To produce full-color images, LCoS displays, and projectors utilize three LCoS chips: one each in the red, green, and blue channels (like how LCD projectors use three-color LCD panels). LCoS now dominates the market niche of *pico-projectors*, which are distinguished by their small size and low power consumption [22].

As a form of SLM, LCoS has a broad range of potential applications, including:

- Industrial projection (fringe/pattern projection – metrology, 3D-sensor, rapid prototyping, lithography)
- AR and VR applications
- Head-up displays (HUD) and head-mounted displays (HMD) in automotive, airborne, and defense industries
- Holographic projection and storage
- Industrial imaging (data displays, medical, simulation)

LCoS micro displays, such as those in AR / VR devices, consist of a liquid crystal layer between a silicon semiconductor with a reflective surface and one transparent thin film transistor (TFT). The light source shines through a polarizing filter and on the device and the liquid crystals function as

[1]Source: https://oqmented.com/technology
[2]Source: https://www.britannica.com/science/Lissajous-figure
[3]Source: https://lasaralliance.org

gates or valves that control the amount of light that reaches the reflecting surface. The more voltage the crystal of a particular pixel receives, the lighter the crystal transmits [21].

D. Implementation of augmented reality in a digitized company

Visualization has become important aspect in a comprehensive data analysis as Yang et al. [23] describes that can effectively combine machine intelligence with human intelligence. In an AR aid system, understanding the cognitive process of the assembly procedure is a prerequisite for appropriately using the situational data provided by the AR technology.

The industrial use of augmented reality in the maintenance and provision of remote support is different. In the AR assistance system, understanding the cognitive process of the assembly process is a prerequisite for the correct use of situational data provided by AR technology. Solutions for smartphones and tablets are the dominant representation, as they do not need increased equipment costs. Stoessel et al. [24] studied the cognitive process and possible cognitive barriers in the assembly task. They believed that attention should be focused on three aspects:

- selective visual attention,
- multitasking performance,
- mental strain.

The study is based on the concept that environmental stimuli are received through organs and can be processed continuously in several specified phases until one sensitive stimulus action is finally generated. information processing during the assembly task includes all cognitive functions, from perception, attention and memory to the planning and implementation of actions.

In the commissioning task, participants should show the relevant assembly part in the assembly instructions and find and hold it in a specific storage box. The interactivity of the ability to perform a particular task varies due to the size of the display devices. fields, videos to animations in the form of simulations of the production process.

The extent of interactivity and the ability to work with data is directly proportional to the equipment used to perceive augmented reality [24].

The augmented reality guidance process includes manuals, videos, and visual references to help with maintenance, which are dynamically displayed directly in the field of view of the technician performing maintenance and repairs in the AR glasses or using a handheld device. The input information is therefore already prepared text and enclosed documents, videos, and simulations, which will be displayed after starting the application by direct selection, loading the outline of the interactive centre or markers, for example in the form of QR codes [24].

In addition to preventive and corrective maintenance or installation, this principle is also applicable in the search and recruitment of new employees and start them in a real environment task that will later be assigned to them.

As shown in Fig. 2 [25] the scope of augmented reality implementation in a digitized enterprise is extensive. The model points to the diversity of tasks that can be performed with smart glasses, based on the creation of an adequate intelligent environment that allows interoperability between machines, people and services via the Internet. Assuming the introduction of a 5G network in a digitized enterprise, the transformation of digitized enterprises into the factories of the future is expected to accelerate.

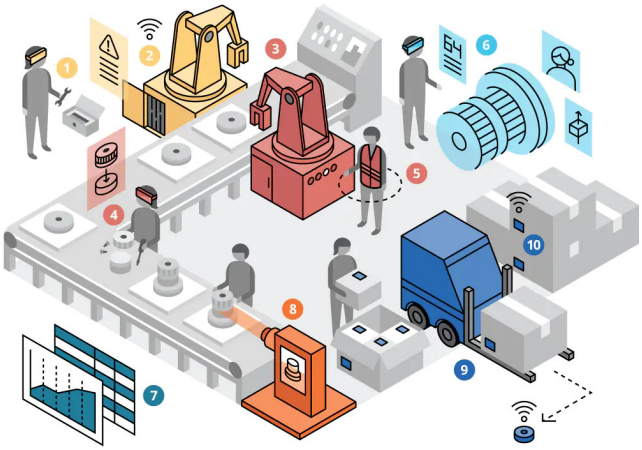


Fig. 2. Scheme of a digitized company using AR [25]

Fig. 2 [25] shows the implementation scheme of AR in the company in various positions.

1) Smart glasses provide calculations, instructions, and input information as well as instructions for remote maintenance support.
2) Sensors on the devices generate diagnostic data and machine learning helps predict potential failures and improve productivity.
3) Automated production of intelligent robots.
4) Smart glasses help new employees to manage the training process into production thanks to sensors without the need for supervision of a responsible employee.
5) The implementation of intelligent technologies in vests and helmets provides an up-to-date overview of all activities in the workplace.
6) Assembly simulation through augmented reality provides possible remote cooperation in testing and ergonomic design based on the entire movement of the body within the company.
7) Data analysis and report - all devices are connected to a process management system that accumulates and evaluate [25] data, which can predict demand and optimize production well in advance.
8) The stands use sensors and computer vision to control product quality in accordance with the standard, with 3D models making all specifications / product / product definitions available to the production operator throughout production.
9) Dispatch of the warehouse by automated controlled vehicles.
10) Product sensors provide an overall view of the entire sales chain [9] experimental methods.

IV. Conclusion and Future Plan

The initial phase of the research was a general survey of the current shortcomings faced by augmented reality as a working and support tool in industry. We have found that the original intention of real-time IOT data visualization research on the MS Hololens platform is not a suitable tool for further research. The next planned step will be the selection of a suitable working tool, where we are considering the use of Magic Leap 2, which in the survey appears to be a suitable

tool for further research. The basic goal was to point out the technological shortcomings, the elimination of which can lead to the acceleration of the progress of efficient use in the industrial environment.

The specialization of development in the near future should be the specialization of IOT data visualization research on HMD devices of graphically demanding applications.

References

[1] R. T. Azuma, "A Survey of Augmented Reality," *Presence: Teleoperators and Virtual Environments*, vol. 6, no. 4, pp. 355–385, 1997.

[2] H. Salmi, H. Thuneberg, and M.-P. Vainikainen, "Making the invisible observable by augmented reality in informal science education context," *International Journal of Science Education, Part B*, vol. 7, no. 3, pp. 253–268, 2017.

[3] H. Kagermann, J. Helbig, A. Hellinger, and W. Wahlster, "Recommendations for implementing the strategic initiative industrie 4.0. final report of the industrie 4.0 working group," 2013. [Online]. Available: https://en.acatech.de/publication/recommendations-for-implementing-the-strategic-initiative-industrie-4-0-final-report-of-the-industrie-4-0-working-group/

[4] P. Milgram and F. Kishino, "A taxonomy of mixed reality visual displays," *IEICE Transactions on Information and Systems*, vol. 77, pp. 1321–1329, 1994.

[5] M. Juraschek, L. Büth, G. Posselt, and C. Herrmann, "Mixed reality in learning factories," *Procedia Manufacturing*, vol. 23, pp. 153–158, 2018.

[6] T. Andrade and D. Bastos, "Extended reality in iot scenarios: Concepts, applications and future trends," in *5th Experiment International Conference*, 2019, pp. 107–112.

[7] J. Grubert and R. Grasset, *Augmented Reality for Android Application Development*. Packt Publishing, 2013.

[8] J. Carmigniani and B. Furht, *Augmented Reality: An Overview*. Springer New York, 2011, pp. 3–46.

[9] I. Jančíková, "Využitie rozšírenej reality v strojárenskej praxi," 2018, diploma thesis.

[10] Z. Wang, S. Huang, and G. Dissanayake, "Simultaneous localization and mapping: Exactly sparse information filters," *New Frontiers in Robotics*, vol. 3, 2011.

[11] F. Zhou, H. B.-L. Duh, and M. Billinghurst, "Trends in augmented reality tracking, interaction and display: A review of ten years of ismar," in *7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, 2008, pp. 193–202.

[12] M. K. Park, K. J. Lim, M. K. Seo, S. J. Jung, and K. H. Lee, "Spatial augmented reality for product appearance design evaluation," *Journal of Computational Design and Engineering*, vol. 2, no. 1, pp. 38–46, 2015.

[13] A. Szajna, R. Stryjski, W. Woźniak, N. Chamier-Gliszczyński, and M. Kostrzewski, "Assessment of augmented reality in manual wiring production process with use of mobile ar glasses," *Sensors*, vol. 20, no. 17, 2020.

[14] D. Bowman, E. Kruijff, J. LaViola, and I. Poupyrev, *3D User Interfaces: Theory and Practice*, 2005.

[15] J. Jung, J. Hong, S. Park, and H. S. Yang, "Smartphone as an augmented reality authoring tool via multi-touch based 3d interaction method," in *Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*. ACM, 2012, pp. 17–20.

[16] A. Marzo, B. Bossavit, and M. Hachet, "Combining multi-touch input and device movement for 3d manipulations in mobile augmented reality environments," in *Proceedings of the 2nd ACM Symposium on Spatial User Interaction*. ACM, 2014, pp. 13–16.

[17] A. Martinet, G. Casiez, and L. Grisoni, "Integrality and Separability of Multi-touch Interaction Techniques in 3D Manipulation Tasks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 3, pp. 369–380, 2012.

[18] A. Mossel, B. Venditti, and H. Kaufmann, "3dtouch and homer-s: Intuitive manipulation techniques for one-handed handheld augmented reality," in *Proceedings of the Virtual Reality International Conference: Laval Virtual*. ACM, 2013.

[19] H. Dreiner, "Augmented reality applications accelerate motor-vehicle repairs and support technical trainings," 2017. [Online]. Available: https://www.bosch-presse.de/pressportal/de/en/augmented-reality-applications-accelerate-motor-vehicle-repairs-and-support-technical-trainings-130688.html

[20] "Boeing tests augmented reality in the factory," 2018. [Online]. Available: https://www.boeing.com/features/2018/01/augmented-reality-01-18.page

[21] Mordor Intelligence LLP, "Liquid crystal on silicon (lcos) display market - growth, trends, and forecast (2019 - 2024)," 2019. [Online]. Available: https://www.reportlinker.com/p05786669/Liquid-Crystal-on-Silicon-LCoS-Display-Market-Growth-Trends-and-Forecast.html

[22] Industry Research, "Global lcos projector market insights, forecast to 2025," 2019. [Online]. Available: https://www.industryresearch.co/global-lcos-projector-market-13766848

[23] Z. Yang, J. Shi, W. Jiang, Y. Sui, Y. Wu, S. Ma, C. Kang, and H. Li, "Influences of augmented reality assistance on performance and cognitive loads in different stages of assembly task," *Frontiers in Psychology*, vol. 10, 2019.

[24] C. Stoessel, M. Wiesbeck, S. Stork, M. F. Zaeh, and A. Schuboe, "Towards optimal worker assistance: Investigating cognitive processes in manual assembly," in *Manufacturing Systems and Technologies for the New Frontier*. Springer London, 2008, pp. 245–250.

[25] T. Trends, "Mixed reality: Experiences get more intuitive, immersive, and empowering," 2017. [Online]. Available: https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/technology/deloitte-uk-tech-trends-2017-mixed-reality.pdf

Use of blockchain technology in the routing process for multi-hop networks

[1]*Maros BAUMGARTNER (2nd year)*
[2]*Jan PAPAJ*

[1,2]Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

[1]maros.baumgartner@tuke.sk, [2]jan.papaj@tuke.sk

***Abstract*—Research and development in the field of mobile ad hoc networks address various challenges. One of the most important areas of research is the robustness of data transmission. Due to the dynamic nature of multi-hop networks, nodes are more vulnerable to attacks by untrusted nodes because there is no fixed infrastructure and a central node in such networks to determine which nodes will be part of the network. This document describes a proposed model that increases the robustness of data transmission in multi-hop networks using blockchain technology. The achieved results were analyzed from the point of view of the blockchain as it forms the basis of our model. The achieved results show that the blockchain technology implemented in the MANET network ensures the elimination of untrusted nodes in the network, based on changed or discarded data packets by individual nodes. These parameters have a great influence on the robustness of data communication.**

***Keywords*—Blockchain, MANET, network, simulation**

I. Introduction

Mobile wireless technologies and wireless communication are currently the most dynamically evolving communication methods used by people around the world. Every portable device, such as a mobile phone, is equipped with Wi-Fi or Bluetooth technologies, which are designed to communicate over short distances, but with the advancement of telecommunications technologies, these devices become part of mobile wireless networks called MANET (Mobile Ad-Hoc Network) [1]. to connect to the Internet and communicate with anyone regardless of distance. In a MANET network, each subscriber, which in this terminology is called a node, can be a source, destination or some kind of transfer station through which data flows from the source node to the destination [2].

Based on current knowledge about the issue of data transmission resilience in multi-hop networks in 5G and 6G networks, a new approach was proposed in this article, which uses blockchain technology to ensure data transmission resilience [3][4]. Blockchain is a database that differs from the classic data database in several fundamental respects. It stores information about routing, trust, and forwarded nodes in blocks. Such blocks form a mutual connection, by means of which the transparency of the stored data is ensured. . Such a blockchain database is shared among all network participants and the principle applies that everyone knows everything about everyone. In the event that a untrusted node appears between the nodes, its credibility will decrease and it will be excluded from the communication process [5]. In this way, it is possible to prevent the participation of any intruders in the data transmission and thus significantly improve the very resilience of the network [6].

II. Implementation and Simulation Results

This chapter describes the implementation of the proposed model of robust data transmission, which consists of blockchain technology, which is implemented in a multi-hop network and the analysis of the achieved results. The proposed model of robust data transmission was simulated using the Python programming language based on the simulation parameters described in the TABLE I.

TABLE I Simulation parameters

Variable	Value
Simulation area [m]	2000 x 2000
Type of simulated area	Ideal
Number of nodes	200, 400, 600, 800
Radio range [m]	50
Simulation time [s]	120
Transmission technology	Wi-Fi
Mobility	Random
Nodes speed [m/s]	1 – 5
Mobility model	Random waypoint
Number of simulation run	100

A new scheme has been proposed that allows blockchain technology to be implemented in MANET networks (Fig. 1). In this scheme each node that enters the network automatically downloads or updates the blockchain because it works on the principle of a distributed database between all nodes in the network. The node participating in the information transmission sends a message with specific parameters to its neighboring node within the blockchain network. The nodes that receive the message verify it based on the time trace, hash value, and packet header parameters. At the same time, each node verifies the specific information transmitted in order to prevent attacks on the system in the form of flooding the network with messages, etc. If the message is true, such a message is written to the local blockchain network. If the message is false, it is rejected and excluded from the blockchain network to avoid re-authentication and unnecessary network congestion. The true message then reaches the authentication node or nodes, which create a block in memory and verify the

message via the PoW (Proof-of-Work) consensus. If the message has not been verified by consensus, a record of such a message is assigned to the node that transmitted the message, then the trust level is updated and the message is rejected.

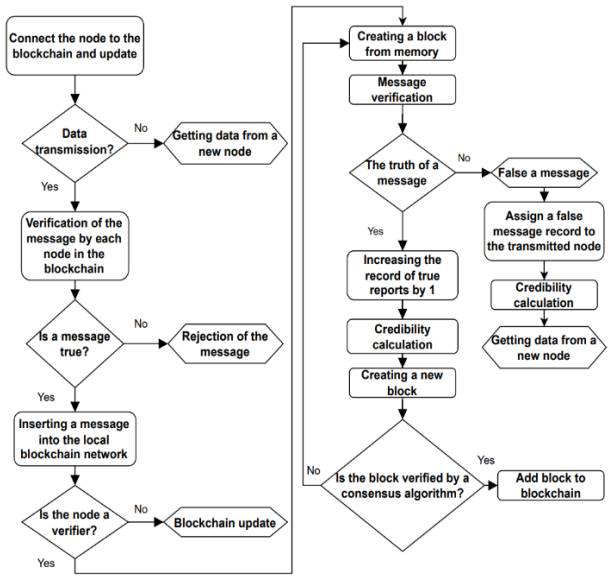


Fig. 1 Implementation model of blockchain technology for MANET

The results obtained in Fig. 2 describe the parameter of the altered data packets that were sent from the source to the destination node. At the beginning of the simulations, the rate of changed packets was the highest, with the gradual generation of the blockchain, these values continuously decreased. This phenomenon is caused by generating blocks and updating the trustworthiness of nodes in data communication. As a result, in the more advanced phases of the simulation, with a higher number of successfully or unsuccessfully transmitted packets, the network becomes more resilient. Due to the growing amount of information about network nodes stored in the blockchain, the number of changed and dropped packets decreased by an average of 40.24-50.37%.

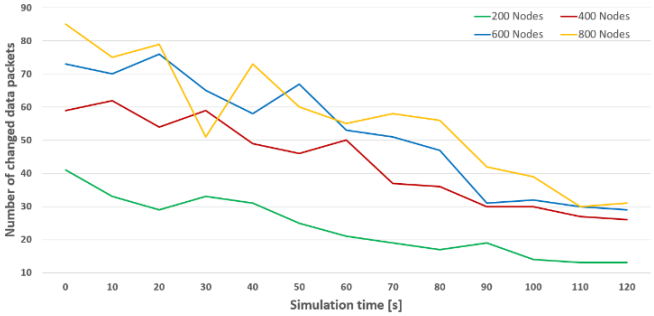


Fig. 2 Number of changed data packets for all simulated scenarios

We recorded similar results for the parameter of the number of dropped packets. The achieved results of the number of dropped packets are shown in Fig. 3.

Number of untrusted nodes in the network for all simulated scenarios are shown in Fig. 4. The higher the rate of untrusted nodes, the higher the probability of their negative intervention in communication. During the simulations, their number gradually decreased due to the gradual update of their credibility.

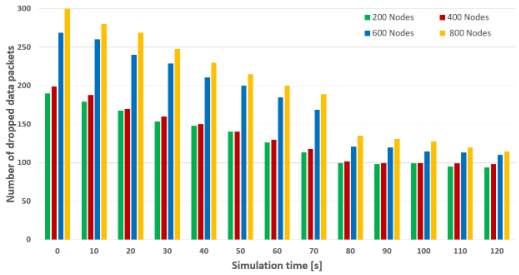


Fig. 3 Number of dropped data packets for all simulated scenarios

Nodes that appeared to be harmful after re-communication were listed in the blockchain as problematic and, based on this information, were not selected as routing nodes in subsequent communications.

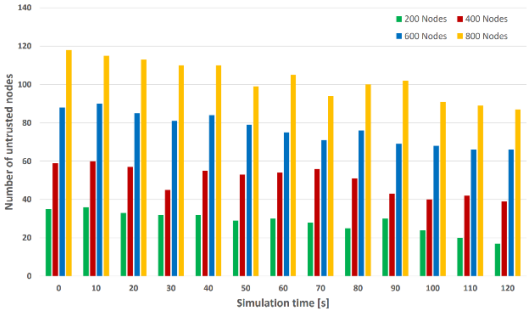


Fig. 4 Number of untrusted nodes for all simulated scenarios

Due to the growing amount of information about network nodes stored in the blockchain, the number of untrusted nodes involved in data transmission decreased by an average of 18.92 - 27.35%.

III. Conclusion

In this article we designed and implemented blockchain technology into the MANET network and their possibility of use for 5G and 6G networks. Based on the achieved results, we can state that blockchain technology significantly increases the robustness of the network, especially in the area of security. We are currently working on a qualitative expression of robustness from the point of view of the blockchain, as so far no parameter has been proposed that would be able to identify it qualitatively.

Acknowledgment

This research was funded by the Slovak Research and Development Agency, research grant no. APVV-17-0208 and APVV-21-0399.

References

[1] P. S. R. Henrique, R. Prasad, "6G The Road to the Future Wireless Technologies 2030" River Publisher, 2021, pp. 19 – 75. ISBN:9788770224390.
[2] J. HE, S. JI, Y. PAN, Y. LI, 2014. Wireless Ad Hoc and Sensor Networks Management, Performance, and Applications. CRC Press. ISBN 13: 978-1-4665-5695-9.
[3] S. BASAGNI, et al., 2013. MOBILE AD HOC NETWORKING Cutting Edge Directions. Second edition, IEEE PRESS, Mobile & Communication, Wiley. ISBN: 978-1-118-08728-2.
[4] M. LI, et al., 2020. LEChain: A blockchain-based lawful evidence management scheme for digital forensics. ELSEVIER. Future Generation Computer Systems 115 (2021) 406-420, https://doi.org/10.1016/j.future.2020.09.0.
[5] H. CHEN, et al., 2020. 6G Wireless Communications: Security Technologies and Research Challenges. IEEE. International Conference on Urban Engineering and Management Science (ICUEMS).
[6] K. ZHANG, et al., 2019. Edge intelligence and blockchain empowered 5G beyond for industrial Internet of Things. IEEE Network Magazines.

Self-organization in nanocomposites based on liquid crystals

[1]*Dmytro Miakota (1st year)*
Supervisor: [2]Natália TOMAŠOVIČOVÁ

[1,2]Institute of Experimental Physics of the Slovak Academy of Sciences, Košice, Slovak Republic

[1]dmytro.miakota@tuke.sk, [2]nhudak@saske.sk

***Abstract*— Liquid crystals are anisotropic fluids with long-range orientational order, which combine the fluidity of ordinary liquids with the directional dependence of electric and optical properties of crystalline solids. Currently, one of the hot topics of the worldwide research is to design nanomaterials that are capable to assemble into functional superstructures in multiple direction. Liquid crystals themselves are prominent example of materials in which the self-organization (self-assembly) appears spontaneously on different scales. Besides the local ordering on the molecular level, they may form micro/macroscopic superstructures via the appearance of topological defects. The research targets the exploration of physical properties of anisotropic composite systems based on liquid crystals doped with magnetic particles and understanding of phenomena related to the self-assembly.**

***Keywords*—Liquid crystals, memory effect, nanoparticles, liquid crystal-based composites, ferronematics.**

I. INTRODUCTION

Liquid crystals have a big interest in fundamental and applied research communities by using them for electronic devices [1,2]. Liquid crystals are organic materials. The most of liquid crystals molecules has rod like shape and consist of rigid and flexible part. Rigid core consists of several benzene rings that are linearly connected and thus the molecule has an elongated shape. Flexible tails are on the both sides of molecules and they are made out of alkyl chains. Typical length of such molecules is approximately 3 nm and width 0.5 nm [3]. For this research 4-cyano-4'-pentylbiphenyl (5CB) and 4-cyano-4'-hexylbiphenyl (6CB) liquid crystals been used. 5CB and 6CB are nematic liquid crystals with nematic-isotropic phase transition temperatures of approximately *22.5 ^{O}C – 25 ^{O}C* [4] and *13.5 ^{O}C – 30 ^{O}C* [5] respectfully.

The liquid crystals can be oriented when a sufficiently large field, called critical or threshold field, is applied. This effect is known as a *Fréedericksz* transition. The dielectric anisotropy is large (in order of 1) and voltages in the order of volts are sufficient to reorient the molecules. On the other hand, the diamagnetic susceptibility anisotropy is small ($\chi_a \approx 10^{-7}$). Therefore, large magnetic fields are necessary to reorient the molecules. The magnetic field in order of tesla ($B \approx 1\ T$) is needed [6].

In effort to enhance the magnetic susceptibility of liquid crystals, the idea of doping them with magnetic nanoparticles was theoretically introduced by Brochard and de Gennes [7].

For experiments will be used composites with the spherical magnetic Fe_3O_4 nanoparticles with the mean diameter 5nm, 10nm, 15nm, 20nm; Silicon dioxide (SiO_2) nanoparticles; Goethite (α-FeO(OH)) nanoparticles.

II. ANALISIS OF THE TOPIC

Liquid crystal materials in isotropic phase behave as any liquid. However, when they are cooled down, they pass through one or several mesophases before reaching crystal phase. In mesophase these materials are liquid, but their molecules have a certain degree of arrangement. Liquid crystals that pass into mesophase due to changes in temperature are called thermotropic liquid crystal. Another option for reaching mesophase is changing the concentration of amphiphilic molecules. Such liquid crystals are known as lyotropic. In mesophase, liquid crystal molecules can be ordered in various ways. The simplest case is nematic phase in which elongated molecules forming liquid crystal are pointing to one direction. To describe this direction a unit vector $\vec{n}$ (Fig. 1), so called director, is used [8]. Optical, electric and magnetic properties, such as birefringence, dielectric permittivity, electric conductivity, diamagnetic susceptibility or viscosity, are anisotropic due to molecule arrangement in mesophase [9].

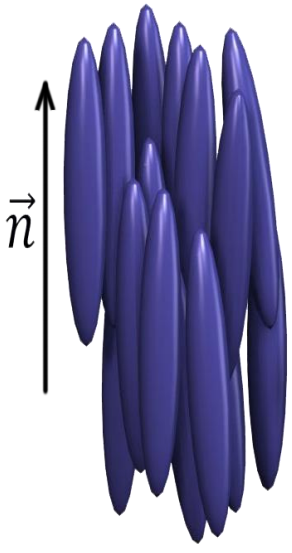


Fig 1. Alignment of the molecules in a nematic phase and director $\vec{n}$

An electric field $\vec{E}$ applied to liquid crystal produces a polarization $\vec{P}$, i.e., dipole moment per unit volume. Directions of the electric field $\vec{E}$ and the polarization $\vec{P}$ are in general different due to liquid crystal anisotropy.

Application of a magnetic field to a liquid crystal sample induces magnetization

$$\vec{M} = \chi_{m\parallel}\vec{H}$$

if the magnetic field is applied parallel to the director $\vec{n}$ and

$$\vec{M} = \chi_{m\perp}\vec{H}$$

if the magnetic field is applied perpendicular to the director $\vec{n}$. $\chi_{m\parallel}$ and $\chi_{m\perp}$ are diamagnetic susceptibilities for the magnetic field parallel and perpendicular to $\vec{n}$ and $\vec{H}$ denotes magnetic field strength.

Response to External Field

Response of external electric or magnetic field is usually observed in a cell consisting of two separated glasses with polymer layer ensuring required alignment on top of indium tin oxide (ITO) layer that serves as a transparent electrode. The scheme of the liquid crystal cell is shown on the Figure 2.

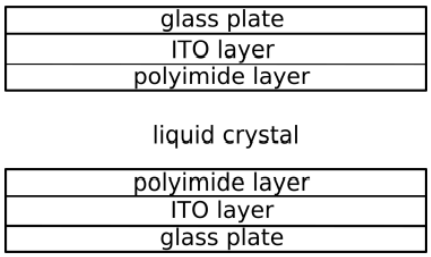


Fig. 2. Liquid crystal cell for optical and dielectric measurements

The external field applied to the liquid crystal tends to orient molecules parallel to the director when the anisotropy is positive (Fig. 3) and perpendicular to the director when the anisotropy is negative (Fig. 4). This effect is known as a *Fréedericksz* transition.

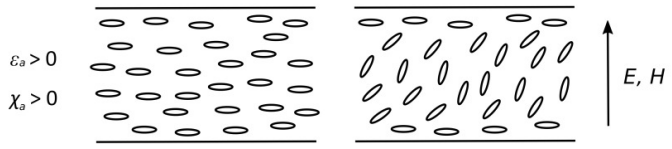


Fig. 3. Orientation of liquid crystal molecules in applied electric or magnetic field for liquid crystal with positive dielectric permittivity anisotropy ε_a and diamagnetic susceptibility anisotropy χ_a

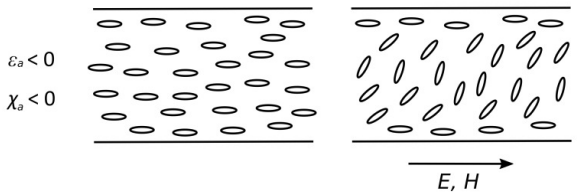


Fig. 4. Orientation of liquid crystal molecules in applied electric or magnetic field for liquid crystal with negative dielectric permittivity anisotropy ε_a and diamagnetic susceptibility anisotropy χ_a

Isotropic-Nematic Phase Transition

The phase transition from isotropic to nematic phase is a transformation from disordered phase to nearly ordered phase at critical value of temperature. Increasing temperature leads to gradual disruption of liquid crystal ordering. A liquid crystal may pass through one or several phases between isotropic and crystal phase. Cooling liquid crystal material can cause passing to nematic phase, then smectic phases without arrangement inside layers, smectic phases with the hexagonal arrangement, smectic phases with fishbone structure and finally to solid phase. Until now however, there is no known material that goes through all phases [10]. In general, to describe the change at the phase transition, the order parameter that reaches value zero at phase at higher temperature and non-zero values at phase at lower temperature is used. Orientational order parameter meets this requirement [11]. It was also experimentally showed that in mesophase with increasing temperature, order parameter monotonically decreases and suddenly drops when the material passes to the isotropic phase [10]. Therefore, it is appropriate for describing the isotropic-nematic phase transition. The order parameter changes at the transition discontinuously and thus the isotropic-nematic phase transition is a first-order phase transition. In case of a second-order transition the order parameter is changing continuously while passing through a transition. There are several theoretical papers describing the transition of thermotropic liquid crystals from isotropic to nematic phase and nematic phase near the phase transition. One of the most commonly used theory is the phenomenological theory of Landau and de Gennes [10]. It describes the phase transition from isotropic to nematic phase rather than calculate the behavior from molecular point of view. The advantage of this theory is its simplicity and simultaneous capturing of the most important transition elements. Maier-Saupe theory is on the other hand molecular theory. The theory uses the description of single molecules to find out the macroscopic phase characteristics. The number of molecules is very large and thus approximations and assumptions with calculations are used to predict the liquid crystal phase behavior. A model that assumes an average potential for all molecules and that to every molecule happens what happens to one molecule on average, is known as mean-field theory [11].

Temperature of isotropic to nematic phase transition is one of the key parameters of liquid crystals. In numerous works, it has been observed that the addition of particles to liquid crystal caused shifting of isotropic to nematic phase transition temperature. The reduction of isotropic-nematic transition temperature was shown in liquid crystal-based composites containing spherical particles [12]-[14]. Carbon nanotubes (CNT) [15] and anisotropic or chain-like particles [16] on the other hand caused increasing of the transition temperature. Furthermore, nanoparticles dispersed in liquid crystal cause the broadening of the transition temperature range [13]. The shift in isotropic nematic phase transition temperature was studied theoretically by Gorkunov and Osipov [17]. They developed the molecular-field theory for nematic liquid crystals containing nanoparticles. The theory was developed for relatively small nanoparticle concentration that allows neglecting of direct nanoparticle interactions. For spherical nanoparticles the isotropic to nematic phase transition occurs at the temperature

$$T_{FN} = T_{LC}(1 - \varphi), \tag{1}$$

where T_{LC} is isotropic to the nematic transition temperature of pure liquid crystal without dispersed nanoparticles and φ is the volume concentration of nanoparticles in the composite. The transition temperature is shifted by spherical nanoparticles to lower values by the factor $(1 - \varphi)$ that reflects the dilution of liquid crystal by dispersed nanoparticles, i.e. distance between liquid crystal molecules is larger on average.

Ferronematics

In general liquid crystals are diamagnetics with low anisotropy of diamagnetic susceptibility. Therefore, large magnetic fields are needed to reorient their molecules. To decrease the critical magnetic field for *Fréedericksz* transition liquid crystals could be doped with magnetic nanoparticles.

The initial orientation of nanoparticles in liquid crystal is either with unit vector of their magnetization $\vec{m}$ parallel or perpendicular to the liquid crystal director $\vec{n}$. Moreover, the anchoring of liquid crystal in an experimental cell can be planar or homeotropic and magnetic field can be applied perpendicular or parallel to the cell surface. In addition, liquid crystal with positive or negative diamagnetic susceptibility anisotropy can be doped by various concentrations of magnetic particles of different sizes and shapes. Depending on the experimental geometry, properties of liquid crystal and

properties of particles the critical magnetic field required for the reorientation of the liquid crystal molecules can be reduced or raised because of nanoparticles that "help" or prevent liquid crystal molecules reorientation.

Consider liquid crystal with the positive diamagnetic susceptibility and parallel initial orientation $\vec{m} \parallel \vec{n}$. Magnetic nanoparticles cause decrease of the threshold magnetic field when the magnetic field is applied perpendicular to cell surface. Magnetic nanoparticles "help" with the reorientation due to anchoring between magnetization of nanoparticles and liquid crystal director [18], [19].

On the other hand, for a liquid crystal with the negative diamagnetic anisotropy and initial orientation $\vec{m} \parallel \vec{n}$, the threshold magnetic field increases with respect to threshold field of pure liquid crystal when the magnetic field is applied parallel to the cell surface with planar alignment. In this case in magnetic field higher than threshold magnetic field the LC molecules reorient perpendicular to the direction of the applied magnetic field due to the negative diamagnetic susceptibility anisotropy while the nanoparticle direction stays the same. Coupling between magnetic particles and liquid crystal molecules prevents the reorientation of molecules.

Materials

1) Fe_3O_4 Nanoparticles

Magnetite (Fe_3O_4) is a material with cubic inverse spinel structure [20]. Oxygen ions are arranged in a cubic lattice and Fe^{3+} ferrite ions are placed at octahedral and tetrahedral sites and Fe^{2+} on octahedral sites [21]. Magnetite is ferrimagnetic below Curie temperature 850 K. Properties of Fe_3O_4 nanoparticles differs from bulk properties. Lattice parameter and volume of unit cell are larger [22], [23]. Crystal structure of nanoparticles stays the same, unit cell is face centred cubic, but the concentration of oxygen in Fe_3O_4 nanoparticles decrease with reducing the nanoparticle diameter. Although the effect of reduced concentration on structural properties is negligible, to some extent it affects the magnetic properties [20]. As the nanoparticle size is decreasing, spherical Fe_3O_4 nanoparticles under diameter 128 nm become single domain [24]. Ferrimagnetic behaviour begin to be suppressed under the diameter and superparamagnetic behaviour is preferred [20]. Nanoparticles in composites have a tendency to aggregate. Therefore, the particles are often covered by surfactant to prevent aggregation. As the surfactant oleic acid [25]-[27], sodium citrate [28] synthetic polymers [29], [30] or inorganic materials (silica, gold) [31] can be used. Oleic acid is commonly used as the surfactant for magnetite nanoparticles because of their affinity [27].

2) SiO_2 Nanoparticles

Silicon dioxide (SiO_2), referred to also as silica, is a natural compound of silicon and oxygen [32]. Molecule of SiO_2 consist of silicon atom and two oxygen atoms. Each silicon atom is surrounded by four oxygen atoms via covalent bond and create three-dimensional network with tetrahedral arranged Si-O bonds. The length of the bonds between silicon and oxygen is 0.162 nm. The distance between two oxide bonds is 0.262 nm. Silicon dioxide nanoparticles have great potential in medical uses, such as drug delivery, since they are non-toxic, biocompatible and biodegradable [33], [34]. For the experiments focused on memory effect 5CB liquid crystal was doped with SiO_2 nanoparticles with diameter 7 nm.

3) Goethite Nanoparticles

Goethite represents a ferric compound that is investigated for water purification, sensing of humidity, coatings, lithium-ion batteries and for many other applications due to its chemical stability at room temperature, low cost and nontoxicity [35]-[37]. Such particles show different magnetic properties comparing to bulk materials; holding a permanent longitudinal magnetic moment, along particle long axis [38]. It was observed that these nanostructures also have a negative magnetic susceptibility along the shortest particle dimension. This leads to formation of lyotropic nematic phase that aligns in magnetic fields [39]. Thus, goethite nanoparticles orient parallel to the field at low intensities and reorient perpendicularly in case when the magnetic field passes a threshold intensity [40], [41]. A mean length of 350 ± 100 nm, width of 25 ± 7 nm and thickness of 10 ± 5 nm is estimated for goethite nanorods. [42]

Composites Preparation

Liquid crystal composites with various volume concentrations of nanoparticles were prepared as follows:

1) Nanoparticles in chloroform were admixed to the liquid crystal and the solvent was let to evaporate.
2) After evaporation, the composite with given concentration was obtained.
3) Subsequently, the ferronematic sample was diluted by addition of liquid crystal. By the dilution, the composite with lower concentration was acquired.

III. SUMMARY

At the moment, nematic-isotropic phase transition temperatures for 5 CB liquid crystal (T_t = *33.68* ^{O}C), 5CB with SiO_2 (*1 w%*) nanoparticles (T_t = *33.66* ^{O}C), 5CB with SiO_2 (*1 w%*) nanoparticles and Goethite ($\varphi = 10^{-4}$) nanoparticles (T_t = *33.60* ^{O}C), composites were found on capacitance measurements (Fig. 5). Experiments of the memory effect for electric and magnetic fields for those samples have been completed. Dependents of bias electric and magnetic field has been researched (Fig. 6-7).

It was found that heating the sample to a certain temperature ($T \approx$ *40-45* ^{O}C) "resets" its properties.

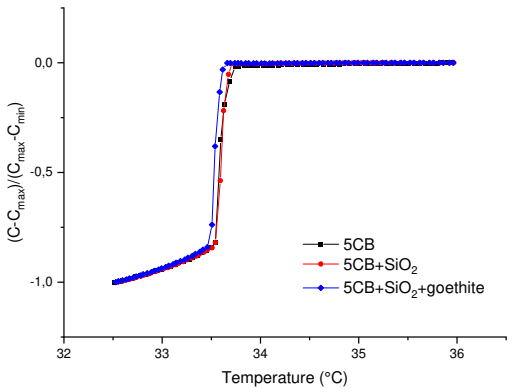


Fig. 5. The dependence of reduced capacitance and temperature for 5CB based composites

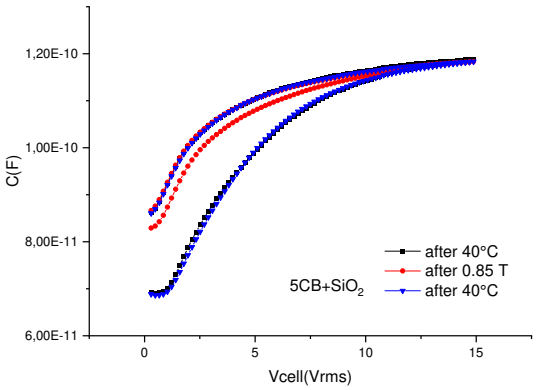


Fig 6. The dependence of capacitance on electric field for bias magnetic field ($B = 0.85\ T$) for 5CB LQ with SiO_2 nanoparticles

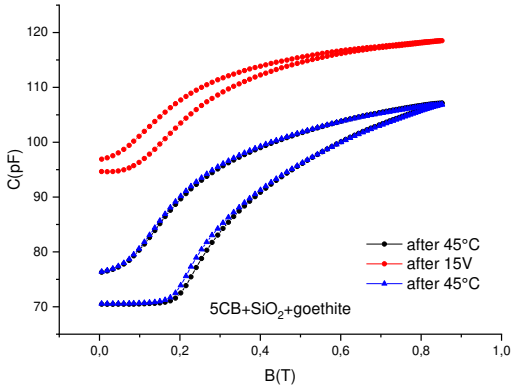


Fig 7. The dependence of capacitance on magnetic field for bias electric field ($U = 15\ V$) for 5CB LQ with SiO_2 and goethite nanoparticles

Further research will be related to liquid crystals doped with goethite nanoparticles. Aqueous suspensions of goethite nanorods, depending on volume fraction, form stable isotropic and nematic phases. The isotropic phase has very peculiar magnetic properties because goethite nanorods align parallel to a weak magnetic field but perpendicular to a strong field [43]. The main interest of these nematic suspensions rather lies in their original behavior when submitted to magnetic and electric fields. The existence of a permanent magnetic moment induces very low values of the Frederiks thresholds and the appearance of a dipolar order, which makes these suspensions somewhat in-between usual liquid crystals and ferrofluid suspensions [44].

References

[1] I. Muševic, S. Žumer, Liquid crystals: maximizing memory, Nat. Mater. 10 (2011) 266–268.

[2] H.K. Bisoyi, S. Kumar, Liquid-crystal nanoscience: an emerging avenue of soft self-assembly, Chem. Soc. Rev. 40 (2011) 306–319. H. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985, ch. 4.

[3] Čepič, M., 2014. Liquid Crystals In: Liquid Crystals Through Experiments. Morgan & Claypool Publishers. ISBN 78-1-6270-5300-6.

[4] Kuribayashi, M. and Hori, K., 1999. Crystal Structures of 4-Cyano-4-Hexylbiphenyl (6CB) and 4-Cyano-4-Heptylbiphenyl (7CB) in Relation to Odd Even Effects. In: Liquid Crystals, vol. 26, no. 6, pp. 809–815.

[5] Hadjichristov, G. B. et al., 2016. Dielectric and electrical characterization of 5CB nematic liquid crystal doped with silver nanoparticles. In: Journal of Physics: Conference Series, vol. 682, 012015.

[6] Stewart, I. W., 2004. Applications of Static Theory of Nematics In: The Static and Dynamic Continuum Theory of Liquid Crystals: A Mathematical Introduction. Taylor & Francis. ISBN 0-748-40895-9

[7] F. Brochard, P.G. de Gennes, J. Phys. (Paris) 31,691 (1970).

[8] Svoboda, J. and Glogarová, M. Kapalné krystaly. [Cit. 2019-03-14]. http://kubusz.net/Reserse/new/1/9_axial_chirality/9_9_Svoboda.pdf.

[9] Mušević, I., 2017. Nematic Liquid-Crystal Colloids. In: Materials, vol. 11, no. 24, pp. 1–27.

[10] Singh, S., 2000. Phase Transitions in Liquid Crystals. In: Physics Reports Review Section of Physics Letters, vol. 324, no. 2-4, pp. 108–269.

[11] Collings, P. J. and Hird, M., 1997. Theoretical Insights In: Handbook of Liquid Crystals, Volume 2: Low Molecular Weight Liquid Crystals I. Taylor & Francis. ISBN 0-203-21119-7.

[12] Tomašovičová, N. et al., 2017. Alternating Current Magnetic Susceptibility of a Ferronematic. In: Beilstein Journal of Nanotechnology, vol. 8, pp. 2515– 2520.

[13] Zakutanská, K. et al., 2019. Nanoparticle's Size, Surfactant and Concentration Effects on Stability and Isotropic-Nematic Transition in Ferronematic Liquid Crystal. In: Journal of Molecular Liquids, vol. 289, 111125.

[14] Jessy, P. J., Radha, S. and Patel, N., 2019. Highly Improved Dielectric Behaviour of Ferronematic Nanocomposite for Display Application. In: Liquid Crystals, vol. 46, no. 5, pp. 772–786.

[15] Duran, H. et al., 2005. Effect of Carbon Nanotubes on Phase Transitions of Nematic Liquid Crystals. In: Liquid Crystals, vol. 32, no. 7, pp. 815–821.

[16] Kopčanský, P. et al., 2011. Phase Transitions in Liquid Crystal Doped with Magnetic Particles of Different Shapes. In: International Journal of Thermophysics, vol. 32, no. 4, pp. 807–817.

[17] Gorkunov, M. V. and Osipov, M. A., 2011. Mean-Field Theory of a Nematic Liquid Crystal Doped with Anisotropic Nanoparticles. In: Soft Matter, vol. 7, no. 9, pp. 4348–4356.

[18] Kopčanský, P. et al., 2010. The Sensitivity of Ferronematics to External Magnetic Fields. In: Journal of Physics Conference Series, vol. 200, 072055.

[19] Kopčanský, P. et al., 2013. Increasing the Magnetic Sensitivity of Liquid Crystals by Rod-Like Magnetic Nanoparticles. In: Magnetohydrodynamics, vol. 49, no. 3-4, SI, pp. 586–591.

[20] Blaney, L., 2007. Magnetite (z): Properties, Synthesis, and Applications. vol. 15, pp. 33–81.

[21] Teja, A. S. and Koh, P.-Y., 2009. Synthesis, Properties, and Applications of Magnetic Iron Oxide Nanoparticles. In: Progress in Crystal Growth and Characterization of Materials, vol. 55, no. 1-2, pp. 22–45.

[22] Cornell, R. M. and Schwertmann, U., 1996. The Iron Oxides: Structure, Properties, Reactions, Occurrences and Uses. VCH Publishers, Weinheim, Germany. ISBN: 3-527-28567-8.

[23] Thapa, D. et al., 2004. Properties of Magnetite Nanoparticles Synthesized Through a Novel Chemical Route. In: Materials Letters, vol. 58, no. 21, pp. 2692–2694.

[24] Leslie-Pelecky, D. L. and Rieke, R. D., 1996. Magnetic Properties of Nanostructured Materials. In: Chemistry of Materials, vol. 8, no. 8, pp. 1770– 1783.

[25] Gass, J. et al., 2006. Superparamagnetic Polymer Nanocomposites with Uniform Fe_3O_4 Nanoparticle Dispersions. In: Advanced Functional Materials, vol. 16, no. 1, pp. 71–75.

[26] Wuang, S. C. et al., 2007. Synthesis and Functionalization of Polypyrrole Fe_3O_4 Nanoparticles for Applications in Biomedicine. In: Journal of Materials Chemistry, vol. 17, no. 31, pp. 3354–3362.

[27] Yang, K. et al., 2010. Re-examination of Characteristic FTIR Spectrum of Secondary Layer in Bilayer Oleic Acid-Coated Fe_3O_4 Nanoparticles. In: Applied Surface Science, vol. 256, no. 10, pp. 3093–3097.

[28] Wei, Y., 2012. Synthesis of Fe_3O_4 Nanoparticles and Their Magnetic Properties. In: Procedia Engineering, vol. 27, pp. 632–63

[29] Song, G. P., Bo, J. and Guo, R., 2005. Preparation of Polystyrene/Fe_3O_4 Nanoparticles in Triton X-100/Sodium Dodecyl Benzenesulfonate Mixed Surfactant System. In: Chinese Journal of Chemistry, vol. 23, no. 8, pp. 997–1000.

[30] Chastellain, M., Petri, A. and Hofmann, H., 2004. Particle Size Investigations of a Multistep Synthesis of PVA Coated Superparamagnetic Nanoparticles. In: Journal of Colloid and Interface Science, vol. 278, no. 2, pp. 353–3

[31] Laurent, S. et al., 2008. Magnetic Iron Oxide Nanoparticles: Synthesis, Stabilization, Vectorization, Physicochemical Characterizations, and Biological Applications. In: Chemical Reviews, vol. 108, no. 6, pp. 2064–2110.

[32] https://pubchem.ncbi.nlm.nih.gov/compound/Silicon-dioxide

[33] https://docbrown.info/page04/4_72bond4d.htm

[34] Plummer, J. D. and Deal, M. and Griffin, P. B., 2000. Silicon VLSI Technology - Fundamentals, Practice and Modeling. Prentice Hall Electronics and VLSI Series. Prentice-Hall Inc., Upper Saddle River, NJ. ISBN 978-0130850379.
[35] G. Zhang, S. Wang, F. Yang, J. Phys. Chem. C. 116 (2012) 3623, https://doi.org/10.1021/jp210167b.
[36] P.V. Adhyapak, U.P. Mulik, D.P. Amalnerkar, I.S. Mulla, J. Am. Ceram. Soc. 96 (2013) 731, https://doi.org/10.1111/jace.12189.
[37] J. Wang, L. Li, C.L. Wong, L. Sun, Z. Shen, S. Madhavi, RSC Adv. 3 (2013) 15316, https://doi.org/10.1039/c3ra41886c
[38] E. Van den Pol, A.A. Verhoeff, A. Lupascu, M.A. Diaconeasa, P. Davidson, I. Dozov, B.W. Kuipers, D.M. Thies-Weesie, G.J. Vroege, J. Phys.: Condens. Matter. 23 (2011) 194108, https://doi.org/10.1088/0953-8984/23/19/194108.
[39] P. Davidson, P. Batail, J.C.P. Gabriel, J. Livage, C. Sanchez, C. Bourgaux, Prog. Polym. Sci. 22 (1997) 91, https://doi.org/10.1016/S0079-6700(97)00012-9.
[40] B.J. Lemaire, P. Davidson, D. Petermann, P. Panine, I. Dozov, D. Stoenescu, J.P. Jolivet, Eur. Phys. J. E: Soft Matter Biol. Phys. 13 (2004) 309, https://doi.org/10.1140/epje/i2003-10079-5.
[41] B.J. Lemaire, P. Davidson, J. Ferr, J.P. Jamet, D. Petermann, P. Panine, I. Dozov, D. Stoenescu, J.P. Jolivet, Faraday Discuss. 128 (2005) 271, https://doi.org/10.1039/b403074e
[42] P. Kopčanský et al. The influence of goethite nanorods on structural transitions in liquid crystal 6CHBT Journal of Magnetism and Magnetic Materials (2018)
[43] Lemaire B.; Davidson P.; Ferré J.; Jamet J.; Petermann D.; Panine P.; Dozov I.; Jolivet J. Physical properties of aqueous suspensions of goethite (α-FeOOH) nanorods. Eur. Phys. J. E 2004, 13, 291–308, https://doi.org/10.1140/epje/i2003-10078-6
[44] B. J. Lemaire, P. Davidson, D. Petermann, P. Panine, I. Dozov, D. Stoenescu, J. P. Jolivet, Physical properties of aqueous suspensions of goethite (α-FeOOH) nanorods: Part II: In the nematic phase, Eur. Phys. J. E: Soft Matter Biol. Phys., 13, 309 (2004), https://doi.org/10.1140/epje/i2003-10079-5

Domain wall dynamics under the influence of low temperature in amorphous glass-coated microwire

[1]Jana HORNIAKOVÁ (3rd year)
Supervisor: [2]Jozef ONUFER

[1,2]Dept. of Physics, FEI TU of Košice, Slovak Republic

[1]jana.horniakova@tuke.sk, [2]jozef.onufer@tuke.sk

***Abstract*— An experimental method for the study of single domain wall dynamics in bistable microwires in low temperatures is presented. It enables experimenting with a single magnetic domain wall moving in a homogeneous magnetic field. Magnetoelastic anisotropy changes due to applied tensile stress recently showed additional voltage peaks that represent local increase of the domain wall velocity. Other way to modify magnetoeleastic anisotropy, is to change temperature. In presented work we lower the temperature to see if the additional voltage peaks will show up.**

***Keywords*— amorphous microwire, domain wall, temperature influence**

I. Introduction

The magnetic anisotropy in amorphous alloys is mainly determined by magnetoelastic interactions [1], which explains why the internal stresses coupled with the magnetostrictive parameter strongly affect the magnetic structure.

From the point of view of the domain structure, microwires with a positive magnetostriction constant are formed by a large inner domain with closing domains which are magnetized axially. However, in the surface layer, the direction of anisotropy is different [2] - [4]. The different direction of magnetic anisotropy in the surface layer with respect to the inner axial domain is related to stresses induced during production. In the case of glass-coated microwires also due to the presence of a glass coating having a different solidification temperature and coefficient of thermal expansion than the metal core of the microwire [3], [5]. Thus, the domain structure consists of one large axial domain in the center of the microwire, around which is a layer with a radial direction of magnetization of the domain, and closing domains have been formed to reduce the magnetostatic energy of the core. The direction of the easy axis of magnetization is therefore due to the magnetoelastic anisotropy in the direction of the axial stress.

Such a domain structure of microwires with positive magnetostriction allows the process of premagnetization by one large Barkhausen jump. As a consequence, they are said to have bistable behavior and are suitable for studying the dynamics of a single domain wall (DW) [6], [7].

The usual way to adjust the magnetoelastic anisotropy is to relax the internal stress by heat treatment. Microwire annealing leads to a reduction in internal stresses and / or the release of defects, and can increase the DW velocity at a given magnetic field as well as to increase the range of magnetic fields in which a single DW propagates [8]. Conversely, low temperature should have the opposite effect (should cause more intensity of internal stresses).

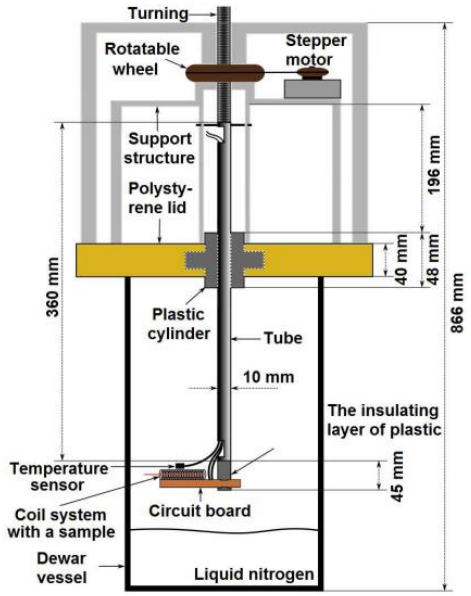


Fig. 1 Experimental setup [11].

Other way to change magnetoelastic anisotropy is a tensile stress application. In recent study of DW geometry with applied tensile stress, unexpected peaks were observed in induced voltage signals [9]. Observed peaks represent the local increase in DW velocity, which is in contradiction with predicted pinning centers function. Our motivation to perform temperature measurements was to get to know if such voltage peaks will occur too.

II. Experimental

Amorphous glass-coated $Fe_{62}Ni_{15,5}Si_{7,5}B_{15}$ microwire prepared by the Taylor–Ulitovsky method was used in the experiment. The length of the sample was 12 cm, diameter of the metallic nucleus was about 14 μm and thickness of glass coating was about 7.5 μm.

We consider experiment schematically depicted in Fig.1. The experimental setup allows to measure DW velocity versus driving field dependences at different temperatures. Measurement of DW velocity for a given value of the driving

magnetic field, was proceeded by using the Sixtus-Tonks method. The system of coils, their function and dimensions have been described in [10].

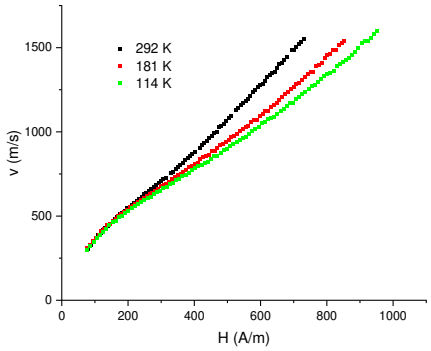


Fig. 2 Applied magnetic field dependencies of domain wall velocities as a function of temperature.

The studied sample (placed in system of coils) was inserted into a Dewar vessel which was filled with liquid nitrogen. The sample holder can be moved up or down above the liquid nitrogen surface using a computer-controlled stepper motor to stabilize or change temperature. A thermometer (based on a Pt100 resistor) is used for temperature measurement. Such experimental setup is able to measure $v(H)$ dependences at constant temperature in the interval from 100 K up to room temperature.

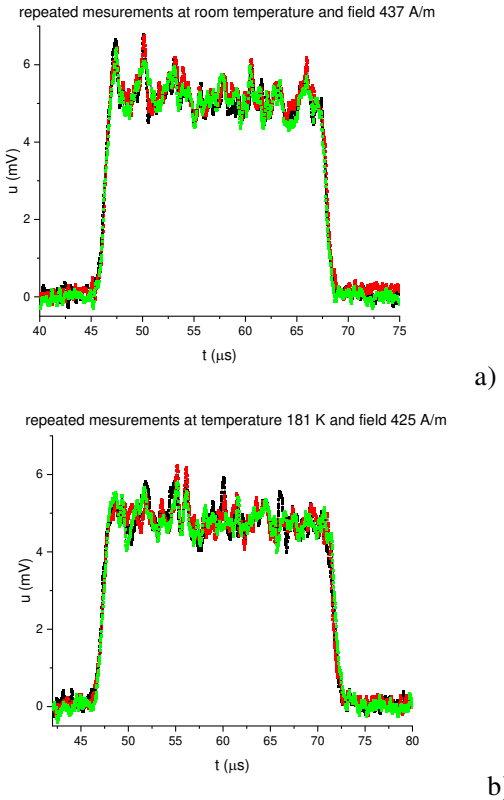


Fig. 3 Repeated time dependence of measured induced peaks from pick-up coil at room temperature a) at temperature 181 K b).

DW velocities versus driving magnetic field at three temperatures (114, 181, 292 K) are shown in Fig. 2. As the temperature decreases, the current in the excitation coil increases. Therefore, with the same measurement at a lower temperature, we can achieve larger magnetic fields. With decreasing temperature, we observe a decrease in velocity. Decrease of velocity is also observed in measurements with applied tensile stress [9]. The mechanisms of the temperature effects are related to the residual stress distribution.

Repeated time dependence of measured induced peaks from pick-up coil with applied magnetic field strength $H = 437 Am^{-1}$ at room temperature are seen in Fig. 3a. Each colour corresponds to the particular measurement. In comparison, repeated time dependence of measured induced peaks from pick-up coil with applied magnetic field strength $H = 425 Am^{-1}$ at temperature 181 K are seen in Fig. 3b. Although the peaks appear rather systematic than random at room temperature in repeated measurements, at lower temperature peaks looks rather random. In order to observe whether voltage peaks appear (an increase in peak amplitude) as the temperature decreases, we compared the signals at the same DW velocity, but different temperatures. In the studied nickel sample, these peaks were not significant enough to conclude.

III. Conclusion

The previous results show that there can be pinning centers at the surface of microwire that locally increase the domain wall velocity when applying tensile stress (additional voltage peaks). In this paper the magnetoelastic energy was changed by lowering the temperature. We cannot draw a clear conclusion by comparing the signals for the nickel sample. Further measurements on a bistable Fe-rich sample should bring answer, if we will observe similar additional voltage peaks as with tensile stress application.

References

[1] A. Zhukov, M. Ipatov, C. Garcia, J. Gonzalez, L. Panina, J. M. Blanco and V. Zhukova, Thin Soft Magnetic Amorphous Microwires for High Frequency Magnetic Sensors Applications, PIERS proceedings, Hangzhou, China, 2008. pp 650-657.

[2] K. Richter, R. Varga, A. Thiaville, Imaging the Surface Domain Structure of Amorphous Glass-Coated Microwires by Bitter Colloid. In Acta. Phys. Pol. A, Vol. 126, 2014. pp 72-73. ISSN 1898-794X

[3] M. Vázquez, A. P. Zhukov, Magnetic properties of glass-coated amorphous and nanocrystalline microwires. In J.Magn. Magn. Mater, Vol. 160, 1996. pp 223-228. ISSN 0304-8853.

[4] A. CHizhik et al., Investigation of magnetic structure in cold-drawn Fe-rich amorphous wire. In J. Magn. Magn. Mater., Vol. 279, 2004. pp 359-362. ISSN 0304-8853.

[5] H. García-Miquel et al.:Surface magnetic anisotropy in glass-coated amorphous microwires as determined from ferromagneticresonance measurements. In J. Magn. Magn. Mater., Vol. 231, 2001. pp 38-44. ISSN 0304-8853

[6] J. Onufer, Premagnetizačné procesy v amorfných feromagnetických mikrodrôtoch. Košice, 2013. pp. 81.

[7] J. Onufer, Dynamika magnetizačných procesov v bistabilných mikrodrôtoch. Košice, 2019. pp. 57.

[8] A. Zhukov, M. Ipatov, V. Zhukova, Processing magnetic microwires for magnetic bistability and magnetoimpedance. Magnetic Nano- and Microwires, Design, Synthesis, Properties and Applications, In Electronic and Optical Materials, 2015. pp 225-274

[9] S. Samuhel, J. Horniaková, J. Onufer, P. Duranka, J. Ziman, CONTRIBUTION ON STUDY OF DOMAIN WALL GEOMETRY IN BISTABLE GLASS-COATED MICROWIRES, 25th conference of Slovak physicists proceedings, 2021.

[10] J. Horniaková, J. Onufer, J. Ziman, P. Duranka, S. Samuhel, Changes in geometry of propagating domain wall in magnetic glass-coated bistable microwire. J. of Magnetism and Magnetic Materials 529, 2021.

[11] J. Onufer, J. Ziman, M. Rezničák and S. Kardoš, The Influence of Temperature on Unidirectional Effect in Domain Wall Propagation, ACTA PHYSICA POLONICA A, No. 4, Vol. 131, 2017

Electronic Structures Based on Organic Materials

[1]*Peter PROVAZEK (1st year)*
Supervisor: [2]Alena PIETRIKOVA

[1,2]Dept. of Technologies in Electronics, FEI TU of Kosice, Slovak Republic

[1]peter.provazek@tuke.sk, [2]alena.pietrikova@tuke.sk

***Abstract*—This paper presents the organic electronics, which uses organic polymer materials to create electronic devices. The first part of the paper describes the history of the organic electronics, its advantages, and disadvantages, that can be fabricated by this technology. It also includes a detailed description of technologies used to print organic structures. The second part of the paper describes the commonly used flexible substrates and their parameters. Finally, a humidity sensor fabricated by inkjet printing technology on a PET substrate is presented.**

***Keywords*—organic electronics, inkjet printing, screen printing, organic nano-inks, thick film organic paste, flexible substrate**

I. Introduction

Organic electronic devices which use organic materials such as active layers have gained interest as energy converting devices [1], light-emitting devices [2], flexible sensors [2], and many other applications. Organic materials play a key role in the managing of device functions, therefore these materials are still developing to improve the functionality [3]. New applications such as wearable electronics [4], robotics [5], and artificial intelligence required high flexibility. Therefore, traditional systems based on the rigid circuit board and hard silicon based chips cannot be used, due to new technologies and materials that have been developing [6]. The creation of inks, which are thermally compatible with polymer substrates is still a challenge [7]. Shirakawa, MacDiarmid, and Heeger discovered highly conductive polyacetylene in 1977, which led to the emergence of organic electronics and organic materials based on π-conjugated systems. The Nobel Prize in Chemistry was awarded to three scientists for this discovery in 2000 [3].

The development of nanotechnology in chemistry has significantly contributed to the development of organic electronics. These nanotechnologies make it possible to create pastes and inks based on nanoparticles and nanotubes. Without them, organic electronics would not be able to form electronic structures. The nanotubes are one of the most studied nanomaterial due to their excellent mechanical, electrical, and thermal properties [8]. With new nanotechnologies has also come the development of new printing techniques for the application of layers such as inkjet printing, screen printing, flexography printing, gravure printing, aerosol jet printing, and others [9]. For more than 20 years, printing technologies have been used in the manufacturing industry to produce mainly antennas, sensors, and membrane switches [7].

The term flexible electronics and the often-used term organic or printed electronics describe electronic devices that can be bent, twisted, compressed, folded, stretched, and deformed into different shapes. Despite mechanical stresses, it still retains high electrical performance, integration, and reliability [6].

The foundations of organic electronics consist of electrically active organic materials such as conductors, semiconductors, dielectrics, luminescent materials, and flexible substrates. The choice of material is very important to ensure the best possible performance of the devices as the properties of each material affect the performance. There is often a compromise between the best electronic features and the flexibility of the device, which is the problem of organic electronic devices. Due to its low cost, low-temperature processing, high flexibility, lightweight, and other properties, organic flexible electronics has received a good deal of attention from the academic and industrial community. The biocompatibility and biodegradability of organic materials allow the interfacing of devices with biological systems. This leads to human-friendly applications such as electronic skin [10] and smart prosthetics [3].

First electronic products based on organic electronic materials reached the market in 2005 and 2006. Organic light-emitting displays (OLEDs), organic thin-film transistors (OTFTs), organic photovoltaics (OPVs), and organic sensors are the most well-known products [11].

Despite the many advantages mentioned above, organic electronics also have its disadvantages. The future challenge in the field of organic semiconductors is to increase charge carrier mobility and reduce power consumption because these parameters have a significant impact on the device's performance [12]. Reactions of polymeric materials with water or oxygen accelerate the material aging, reduce the long-term reliability, stability, and durability of organic materials. In addition, low processing temperatures make soldering to thin flexible substrates difficult. The listed problems are challenges for the future to make organic electronics more attractive and to allow the creation of new applications [13].

This paper describes the advantages and disadvantages of printing technologies, commonly used inks and pastes based on polymer materials for printing in organic electronics. Finally, it describes a humidity sensor fabricated by inkjet printing technology on a PET substrate.

II. FABRICATION TECHNOLOGIES OF POLYMER LAYERS

Fabrication methods have a significant impact on the stability, cost, and properties of organic electronic devices. The methods of layers' deposition on the substrate are generally divided into contact and non-contact. The most commonly used technologies for the creation of layers are flexography, gravure printing, screen printing, aerosol jet, and inkjet printing [9].

Contact printing technology is based on direct contact between the mask and the substrate. These methods include screen printing, gravure printing, and flexography. Non-contact printing technology lies in the application of the ink without the direct contact between the nozzle and the substrate. Inkjet and aerosol jet printing technology belong to non-contact printing technology [9].

A. *InkJet printing*

InkJet printing technology can deposit the desired amount of material from a computer designed digital template onto a selected area of the substrate by generating droplets from a reservoir through a nozzle. Advantages of this technology are the ability to use both flexible and rigid substrates, as well as the computer-controlled process. The major disadvantage of this technology is nozzle clogging. Inks contain insoluble particles, that can agglomerate during the printing process. A factor influencing nozzle clogging is the temporal stability of the ink and the particle size. Continuous inkjet (CIJ) and Drop-on-Demand (DOD) inkjet systems are two categories of inkjet printing technology divided according to the method of droplets ejection [14].

Continuous inkjet printing system

This method is used in office printers and it is based on the continuous generation of droplets, which are controlled by high-voltage deflection electrodes. The droplet volume is defined by the piezo-element, which generates droplets at high frequency. The printing resolution is relatively low due to the large droplet diameter (approximately 40 µm) [14].

Drop-on-Demand system

Printers using the DOD system were invented by Siemens in 1977. This printing method is used in electronics to create conductive, semiconductive, and dielectric layers formed on various types of substrates using special pastes based on nanoparticles. The advantage of this system is the elimination of high voltage droplet deflection and the recirculation of unused droplets. In this method, ink droplets are ejected from the nozzles by generating pressure pulses only when are needed, as shown in Fig. 1 [15].

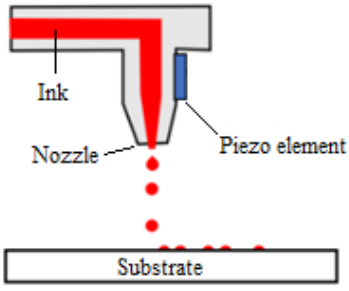


Fig. 1 Schematics of the drop-on-demand system [16]

Inkjet printers are divided into thermal, piezoelectric, electrostatic, and acoustic printers, according to the way the pressure pulse is generated. The DOD method together with nanotechnologies provides high printing accuracy which can be achieved due to the small droplets [15].

B. *Screen printing*

Screen printing technology is a simple and variable technology that is used for applying the paste on the substrate. This method is the oldest technology in printed electronics. In addition, it is the simplest and the cheapest printing technology. The screen printing process includes pushing paste through the open area of the mask using a squeegee, which is shown in Fig. 2. Thixotropic properties and viscosity of pastes are needed for this kind of printing. Paper, ceramics, or flexible foils are usually used as substrates in screen printing technology. The screen open area, thickness, as well as paste viscosity define the printing pattern. In addition, the used material, substrate roughness, screen thickness, snap off, print speed, pressure, the shape of squeegee, aspect ratio and other factors affect the printing quality. The thickness of the deposited layer is usually in the range of 10-500 µm. Polymer and cermet pastes are used in this technology. Cermet pastes and fine screens with high mesh parameter are used to achieve a resolution in the range of 50-100 µm. On the other hand, the use of polymer paste can achieve better resolution by inkjet printing technology. Cermet pastes usually include ceramic materials, therefore, the layers are fired at a temperature up to 850°C and deposited on a ceramic substrate. The polymer pastes are based on polymer materials. The formed layers are usually cured in the temperature range of 150-210°C and typically deposited on flexible foils [16], [17].

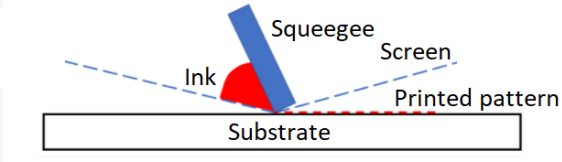


Fig. 2 Schematics of screen printing technology [17]

The screen printing technology is limited by the modest print quality. In screen printing technology two types of techniques such as flatbed screen printing and rotary screen printing are used [17].

C. *Flexography printing*

The method functions as high-speed run printing technology allowing the use of a variety of inks such as solvent-based, water-based, electron beam curing inks, UV curing inks, and two parts chemically curing inks. Low viscosity inks are more suitable for this technology. Flexographic printer usually consists of annilox cylinder that picks up the ink from the chamber. The ink is transported by annilox on a plate cylinder, which has elevated structures. The plate cylinder finally transfers the ink onto the substrates that run between the plate and the impression cylinders. The printing process is shown in Fig. 3 [18].

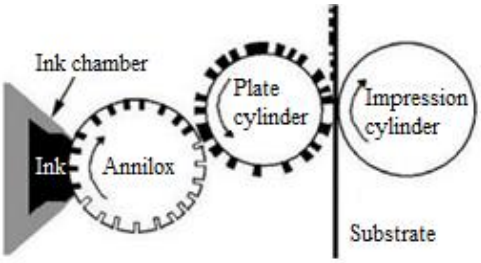


Fig. 3 Schematics of flexographic printing technology [18]

Plate cylinder can be based either on rubber or polymer materials and raised patterns are processed by photolithography. This technology allows the creation of

conductive, semiconductive, and dielectric layers for the fabrication of radio-frequency identification (RFID) antennas, displays, and sensors [18].

The advantage of this method is the possibility to create uniform layers. In addition, it offers better pattern reliability and sharper edges than gravure printing technology. The technology provides a resolution of formed layers between 50-100 µm and approximately 20 µm with the precise control of the printing processes. The disadvantage of the method is its susceptibility to errors such as open lines, overlapped lines, and edge waviness effects. This method is not suitable for the creation a narrow pattern lines on flexible substrates, due to the cylinder cells becoming clogged and the pattern not being printed. The surface irregularities, pores, non-uniform films, ragged lines, and non-availability of suitable materials are the challenges to the future in better patterning [18].

D. Gravure printing

Gravure printing is a contact printing method, in which the ink is deposited on the substrate by a gravure cylinder. Ink in gravure printing is transported from carved microcavities of the cylinder, not from its relief, as in the case of flexography. These cavities form the pattern which is to be printed. The printing process consists of rotation and partial immersion of the gravure cylinder in the ink bath, where the cavities are filled with ink, as shown in Fig. 4. The doctor blade is used to remove excess ink so that the ink remains only in the cavities. It is finally transferred from the cavity to the substrate by pressure. An electrostatic assist system (ESA) is used to improve ink transfer to the substrate. The generated electrostatic field helps to lift the ink onto the substrate. This technology is used to fabricate sensors, displays, and solar cells [16], [17].

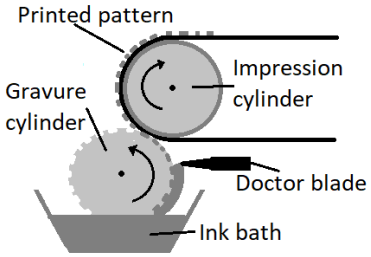


Fig. 4 Schematics of gravure printing technology

The ink viscosity, substrate speed as well as pressure applied by the impression roller are the parameters that affect the quality of the pattern. Therefore, it is necessary to choose the right processing conditions for achieving the highest possible printing quality. The low viscosity inks based on nanoparticles are suitable for the gravure process. The technology allows the creation of high-quality patterns with speed in the range of 13-16 m/s. In addition, simple processing and accurate ink application are the other benefits. Laser engraving technology significantly improved the printing quality. Gravure printing technology is less popular due to the production of a printing cylinder, which is time-consuming and expensive [16], [17].

E. Aerosol Jet printing

Aerosol jet printing is an additive digital process shown in Fig. 5 that utilizes the atomization of ink. The whole process is operated under atmospheric conditions. The ink deposition starts with the pneumatic atomization of the liquid ink into the aerosol. The small droplets have a size between 1 to 5 µm [19].

Pneumatic or ultrasonic techniques are used to atomize the ink. The droplets are delivered to the nozzle, which is attached to the printhead, with a vacuum, generated by a nitrogen gas stream and impinged as a high-velocity jet onto the substrate surface. Control of the beam is necessary for printing 3D and complex patterns, which can be executed by a shutter in front of the nozzle. In addition, the distance between the nozzle and the substrate, which should not be more than 10 mm and less than 1 mm, is crucial for the accuracy of printing. Exceeding these distances will cause overspray errors in the printing pattern [16].

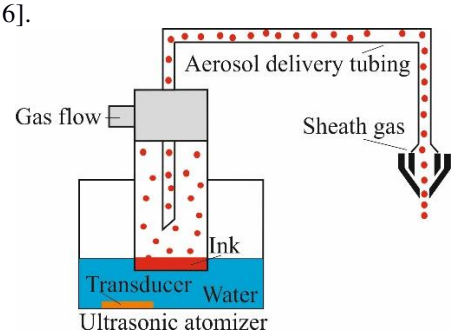


Fig. 5 Schematics of aerosol jet printing

The advantage of this printing method is a digital control of processing as in inkjet printing technology. Technology does not suffer from nozzle clogging which is its advantage over inkjet printing technology. This technology allows high resolution of the printing pattern and creation of structures' sizes as small as 10 µm [16].

III. FLEXIBLE SUBSTRATES

Various types of polymers are used such as polyimide (PI), polyethylene terephthalate (PET), polycarbonate (PC), polyethylene naphthalate (PEN), poly dimethylsiloxane (PDMS), thermoplastic polyurethane (TPU) which parameters are listed in Tab.1 [7].

TABLE 1
PARAMETERS OF FLEXIBLE POLYMERIC SUBSTRATES [7]

Substrate	**PI**	**PET**	**PC**	**PEN**	**PDMS**	**TPU**
T_g (°C)	155 - 270	70 - 110	145	120 - 155	-125	80
T_m (°C)	250 - 452	115 - 258	115 - 160	269	-40	180
Density (g/cm^3)	1,36 - 1,43	1,39	1,20 - 1,22	1,36	1,03	1,18
Volume resistivity (Ω.cm)	1,5x10^{17}	1x10^{19}	10^{12} - 10^{14}	10^5	1,2x10^{14}	3x10^{14}
Tensile strength (MPa)	231	190	60	200	3,51 – 7,65	25-50
Work temperature (°C)	up to 400	-50 +150	-40 +130	~155	-40 +200	130
CTE (ppm/°C)	8 - 20	15 - 33	75	20	310	153
Water absorption (%)	1,3 - 3,0	0,4 - 0,6	0,16 - 0,35	0,3 - 0,4	>0,1	0,2
Solvent resistance	Goog	Good	Poor	Good	Poor	Good
Dimensional stability	Fair	Good	Fair	Good	Good	Good

T_g - glass transition temperature, T_m - melting temperature, CTE - coefficient of thermal expansion.

Flexible substrates offer greater vibration resistance, weight reduction, and smaller dimensions for the equipment. Textile, paper, or polymeric films are used as substrates in printing technology. The main disadvantage of polymeric films is the low processing temperature as well as the low surface energy, which requires a surface treatment before printing [7].

Adhesion between the ink and the substrate is a factor affecting the long-time reliability of the device, so the choice of substrate and paste is a key factor. Properties such as roughness, cleanliness, flatness, and wettability are desirable for strong adhesion. The PI film has the highest quality but it is also very expensive, due to its attractive properties, therefore PEN and PET substrates are used as a cheaper alternative [7].

IV. AN EXAMPLE OF USING ORGANIC MATERIALS FOR THE CONSTRUCTION OF THE HUMIDITY SENSOR

In this chapter, a humidity sensor created by E. Starke; A. Türke; M. Schneider; W. J. Fischer using the inkjet printing technology is presented [20]. The use of PET (thickness of 36 µm) instead of PI substrate for the humidity sensor has benefit as low-cost material. The sensor works on the capacitive principle due to the lower energy consumption. The first pair of electrodes is coated with a moisture sensitive polymer (humidity electrodes-left side) and the second pair of electrodes is covered with a hydrophobic layer (reference electrodes-right side), which is shown in Fig. 6a). The comb structure of the humidity sensor is shown in Fig. 6b) [20].

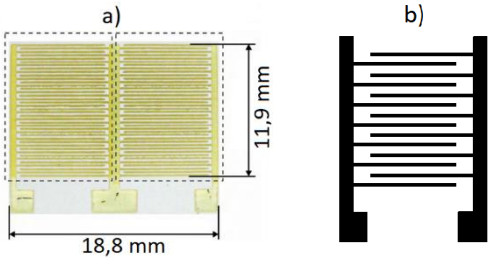


Fig. 6 a) printed humidity sensor b) comb structure [20]

Two printheads with drop volumes of 1pl and 10 pl were used. A saturated solution of silver neodecanoate in xylene was used to print the interdigital structure. Printed structures were cured in an oven at 200°C for 30 min. In the second step, sensitive polymer ink was printed [20]. The printing parameters were optimized to achieve the highest possible capacity by reducing the gaps between the electrodes, which were 240 µm and a line width of 60 µm. It has been found that the addition of sensing layers will increase the capacitance of the sensor because the surface becomes more hydrophilic [20]. The capacity of the humidity sensor increases especially at a relative humidity (RH) higher than 40%. The low-cost humidity sensor has errors below ± 7% RH in the range of 35-90% RH. Above 40 % RH, the error is even less than ± 5% RH [20].

V. CONCLUSION

This paper offers an overview of the history of organic electronics, the industries that helped the development, advantages, and disadvantages of organic electronic structures, technologies that can be used to create layers, and an example of a humidity sensor fabricated by inkjet printing technology.

My future work will be focused on the fabrication of electronic structures based on organic materials using inkjet printing and screen printing technology. The main emphasis is on using different inks, pastes, and substrates.

REFERENCES

[1] W. Tress, "Organic Solar Cells," in *Organic Solar Cells*, vol. 208, Cham: Springer International Publishing, 2014, pp. 67–214. doi: 10.1007/978-3-319-10097-5_3.

[2] T. Das, B. K. Sharma, A. K. Katiyar, and J.-H. Ahn, "Graphene-based flexible and wearable electronics," *J. Semicond.*, vol. 39, no. 1, p. 011007, Jan. 2018, doi: 10.1088/1674-4926/39/1/011007.

[3] H. Ling, S. Liu, Z. Zheng, and F. Yan, "Organic Flexible Electronics," *Small Methods*, vol. 2, no. 10, p. 1800070, Oct. 2018, doi: 10.1002/smtd.201800070.

[4] J. Kim, J. Lee, D. Son, M. K. Choi, and D.-H. Kim, "Deformable devices with integrated functional nanomaterials for wearable electronics," *Nano Convergence*, vol. 3, no. 1, p. 4, Dec. 2016, doi: 10.1186/s40580-016-0062-1.

[5] H. Ju, J. Jeong, P. Kwak, M. Kwon, and J. Lee, "Robotic Flexible Electronics with Self-Bendable Films," *Soft Robotics*, vol. 5, no. 6, pp. 710–717, Dec. 2018, doi: 10.1089/soro.2017.0141.

[6] C. W. Lee, O. Y. Kim, and J. Y. Lee, "Organic materials for organic electronic devices," *Journal of Industrial and Engineering Chemistry*, vol. 20, no. 4, pp. 1198–1208, Jul. 2014, doi: 10.1016/j.jiec.2013.09.036.

[7] S. M. F. Cruz, L. A. Rocha, and J. C. Viana, "Printing Technologies on Flexible Substrates for Printed Electronics," in *Flexible Electronics*, S. Rackauskas, Ed. InTech, 2018. doi: 10.5772/intechopen.76161.

[8] D. E. Babatunde, I. H. Denwigwe, O. M. Babatunde, S. L. Gbadamosi, I. P. Babalola, and O. Agboola, "Environmental and Societal Impact of Nanotechnology," *IEEE Access*, vol. 8, pp. 4640–4667, 2020, doi: 10.1109/ACCESS.2019.2961513.

[9] Executive Agency for Small and Medium sized Enterprises. and Technopolis Group., *Advanced technologies for industry: product watch : flexible and printed electronics.* LU: Publications Office, 2021. Accessed: Feb. 08, 2022. [Online]. Available: https://data.europa.eu/doi/10.2826/295137

[10] R. S. Ganesh, H. Yoon, and S. Kim, "Recent trends of biocompatible triboelectric nanogenerators toward self-powered e-skin," *EcoMat*, vol. 2, no. 4, Dec. 2020, doi: 10.1002/eom2.12065.

[11] G. Nisato, D. Lupo, and S. Ganz, Eds., *Organic and printed electronics: fundamentals and applications*. Singapore, 2016.

[12] C. Wang, X. Zhang, H. Dong, X. Chen, and W. Hu, "Challenges and Emerging Opportunities in High-Mobility and Low-Energy-Consumption Organic Field-Effect Transistors," *Adv. Energy Mater.*, vol. 10, no. 29, p. 2000955, Aug. 2020, doi: 10.1002/aenm.202000955.

[13] M. J. Deen, "Flexible electronics - opportunities and challenges," in *2013 IEEE International Conference of Electron Devices and Solid-state Circuits*, Hong Kong, Hong Kong, Jun. 2013, pp. 1–2. doi: 10.1109/EDSSC.2013.6628198.

[14] A. Soleimani-Gorgani, "Inkjet Printing," in *Printing on Polymers*, Elsevier, 2016, pp. 231–246.

[15] J. Li, F. Rossignol, and J. Macdonald, "Inkjet printing for biosensor fabrication: combining chemistry and technology for advanced manufacturing," *Lab Chip*, vol. 15, no. 12, pp. 2538–2558, 2015, doi: 10.1039/C5LC00235D.

[16] J. Wiklund *et al.*, "A Review on Printed Electronics: Fabrication Methods, Inks, Substrates, Applications and Environmental Impacts," *JMMP*, vol. 5, no. 3, p. 89, Aug. 2021, doi: 10.3390/jmmp5030089.

[17] B. Roth, R. R. Søndergaard, and F. C. Krebs, "Roll-to-roll printing and coating techniques for manufacturing large-area flexible organic electronics," in *Handbook of Flexible Organic Electronics*, Elsevier, 2015, pp. 171–197. doi: 10.1016/B978-1-78242-035-4.00007-5.

[18] S. Khan, L. Lorenzelli, and R. S. Dahiya, "Technologies for Printing Sensors and Electronics Over Large Flexible Substrates: A Review," *IEEE Sensors J.*, vol. 15, no. 6, pp. 3164–3185, Jun. 2015, doi: 10.1109/JSEN.2014.2375203.

[19] F. Cai, Y. Chang, K. Wang, W. T. Khan, S. Pavlidis, and J. Papapolymerou, "High resolution aerosol jet printing of D- band printed transmission lines on flexible LCP substrate," in *2014 IEEE MTT-S International Microwave Symposium (IMS2014)*, Tampa, FL, USA, Jun. 2014, pp. 1–3. doi: 10.1109/MWSYM.2014.6848597.

[20] E. Starke, A. Turke, M. Schneider, and W.-J. Fischer, "Setup and properties of a fully inkjet printed humidity sensor on PET substrate," in *2012 IEEE Sensors*, Taipei, Taiwan, Oct. 2012, pp. 1–4. doi: 10.1109/ICSENS.2012.6411259.

Biomedical and biotechnological applications of magnetic nanoparticles (including magnetoferritin)

[1]*Kristina ZOLOCHEVSKA (1st year)*
Supervisor: [2]*Peter KOPCANSKY*

[1,2]Institute of Experimental Physics of the Slovak Academy of Sciences, Košice, Slovak Republic

[1]kristina.zolochevska@tuke.sk, [2]kopcan@saske.sk

***Abstract*—Many pathological phenomena in the body are associated with structural changes in the storage protein of iron - ferritin. Magnetite formation has been observed in people with neurodegenerative, cardiovascular, and oncological diseases. The cause has not yet been fully elucidated, and detailed factors influencing this transformation have not been described. Particularly incurable neurodegenerative diseases are also associated with the formation of amyloid plaques, similar to fibrillar protein structures composed primarily of beta-pleated leaves. Magnetic iron oxide can be synthesized in the cavity of ferritin to form magnetoferritin. Magnetoferritin provides a powerful platform for tumor diagnosis and therapy, cell imaging, MRI diagnostics and targeted drug delivery, due to its peroxidase-like catalytic activity. Also, it is composed of apoferritin shell and iron-based magnetic nanoparticles, providing higher sensitivity to an applied magnetic field allow to study the hyperthermic effect in magnetoferritin aqueous colloidal solution. Currently, the use of magnetic nanoparticles, magnetic ferrofluids, magnetoferritines, combined with careful surface engineering, has led to amazing results in new therapies' approaches. However, although significant results have been observed, improvements are still needed, mainly in the employment of magnetic nanoparticles modified with biomarkers in the in vivo and clinical assay of drug delivery and magnetic hyperthermia in humans.**

***Keywords*—Biomedical applications, magnetic nanofluids, magnetic nanoparticles, magnetoferritin.**

I. INTRODUCTION

Iron is the one of essential elements for living organisms. Iron metabolic disorders are a common condition [1]. Free iron is at physiological conditions occurring in two oxidation states. The first one is a relatively soluble but highly toxic ferrous (Fe^{2+}) form, and the second one is a very insoluble but non-toxic ferric (Fe^{3+}) state [2], [3]. Both iron deficiency and iron overload carry serious risks, therefore it is stored in the cavity of a spherical protein, named ferritin that is gathering attraction in biomedicine.

Ferritin - iron storage protein that composed of an iron core and 24 subunits. It is a spherical cage that is 12 nm in diameter, composed of an apoferritin shell, which surrounds ferrihydrite (FeOOH) nanocrystals [1].

The core of native ferritin might turn from ferrihydrite to magnetite, thereby forming biogenic magnetoferritin [4]. This might be the result from iron detoxification of ferritins by oxidizing toxic Fe (II) [5]. The ferritin provides a good template for the synthesis of uniform size nanoparticles, and the protein shell also enable magnetoferritines non-interacting and well-dispersed [6].

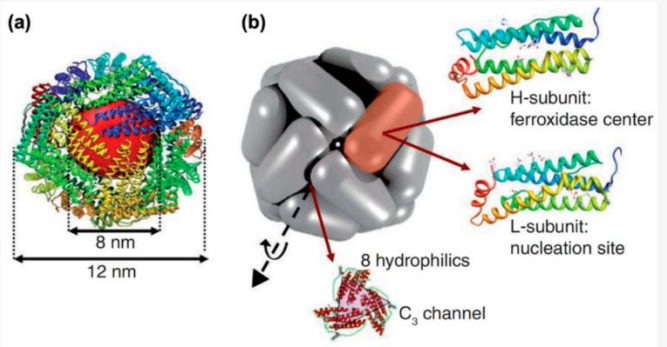


Fig. 1. The structure of native ferritin: (a) The spherical cage and iron core of ferritin; (b) ferritin is composed of H chain subunits and L chain subunits, with 8 hydrophilic channels, which formed a 3-fold axes [3].

The ferritin provides a good template for the synthesis of uniform size nanoparticles, and the protein shell also enable magnetoferritines non-interacting and well-dispersed [6]. It is composed of apoferritin shell and iron-based magnetic nanoparticles.

II. ANALISIS OF THE TOPIC

Magnetic nanoparticles

Magnetic nanoparticles are known for their low toxicity, biocompatibility, high surface area, and due to its intrinsic magnetic properties; these nanomaterials show potential for applications in many areas previously cited and, particularly, in the biomedical science [7], [8]. Unfortunately, magnetic nanoparticles are not able to recognize and correctly separate unhealthy cells from healthy cells, which can lead to a number of side effects and toxicity.

Magnetic nanoparticles surface engineering strategies have been applied to improve hyperthermia, the drug delivery system, and magnetic resonance imaging. One of the properties of superparamagnetic nanoparticles is the generation of heat through the relaxations that occur between the magnetic domain reorientation of the particles in alternating magnetic field or after its application ceases [9].

The inability of magnetic nanoparticles to differentiate unhealthy and healthy cells is one of the main challenges for

biomedical applications as drug delivery systems and magnetic hyperthermia. For this reason, many approaches have been developed aimed at the enhancement of these therapeutic strategies by using different medicines and alternating magnetic field conditions [10]. In this context, doxorubicin bound to chitosan-coated cobalt ferrite/titanium oxide nanofibers has been shown to provide rapid release in melanoma cancer cells [11].

Magnetic nanofluids

Ferrofluids (or magnetic nanofluids) are highly stable colloidal suspensions of magnetic nanoparticles dispersed into various base fluids. These stable ferrofluids possess high thermal conductivity, improved thermo-physical properties, higher colloidal stability, good magnetic properties, and biocompatibility, which are the primary driving forces behind their excellent performance, and thus enable them to be used for a wide range of practical applications. The most studied and advanced ferrofluids are based on iron oxide nanostructures especially nanoparticles, because of their easy and large-scale synthesis at low costs [12].

Ferrofluids are strongly magnetized by an external magnetic field due to good magnetic properties of the dispersed solid magnetic nanoparticles [13]. The tiny magnetic particles can be coated with surfactants such as oleic acid, sodium dodecyl sulphate, citric acid, and many other materials such as dextran, catechol, to prevent agglomeration, and can be ripped off from the homogenous colloidal mixture under the influence of the magnetic force [14], [15].

Magnetosomes

Magnetosomes are intracellular membranous structures present in magnetotactic bacteria. They contain iron-rich magnetic particles that are enclosed within a lipid bilayer membrane. Each magnetosome can often contain 15 to 20 magnetite crystals that form a chain which acts like a compass needle to orient magnetotactic bacteria in geomagnetic fields, thereby simplifying their search for their preferred microaerophilic environments [16]. The magnetosome chains consist of the mineral phase of individual ferrimagnetic crystals of magnetite (Fe_3O_4) and the organic phase, namely the biological membrane of a phospholipid bilayer. The membrane ensures chain elasticity and encapsulates the crystals, thus preventing their coagulation [17]. At the same time, it allows the binding of bioactive substances after the isolation of magnetosomes from the bacterial body [18], [19]. In general, the size of the magnetosome crystals ranges from 35 to 120 nm. It is possible to regulate their size in a targeted manner in laboratory conditions. In this size range, magnetosomes have the character of single-domain particles, which means that each crystal is a small permanent magnet [20], [21]. Due to their high abundance and diversity in marine and freshwater habitats, magnetotactic bacterias are very likely to play an important ecological role in these sediments, such as the biogeochemical cycling of iron and other elements. Magnetosome crystals remain preserved even after the death of the bacterial cell and thus can be deposited as magnetophossils, which significantly contribute to the magnetization of sediments [22].

Magnetoferritin

Magnetoferritin is a magnetic nanomaterial with good biocompatibility and flexibility for biomedical applications. Magnetoferritin with an average size of 12 nm is generally superparamagnetic due to its magnetic iron oxide core. In addition, magnetoferritin showed low toxicity and excellent biocompatibility [23], [24]. The synthesis and application of magnetoferritin was greatly expanded [25]. Magnetoferritin showed superparamagnetic behavior and magnetic anisotropy, which could be used in magnetic resonance imaging and hyperthermia [26], [27]. The shell of magnetoferritin can be functionalized by chemical and genetic engineering [25]. In addition, mineralization within the cavity can be modified to gain ideal properties. Magnetoferritin provides a powerful nanoplatform for biomedical applications [28]. This biomimetic ferritin with superparamagnetic Fe_3O_4 or γ-Fe_2O_3 core, has attracted worldwide attention [25], [29].

Artificial ferritins (magnetoferritin) belong to the bio-inspired materials group, useful for biomedical, nanotechnological or environmental engineering. Magnetoferritin can serve as an ideal biomacromolecule for pathological magnetite simulation [30]. It is prepared by an in vitro laboratory procedure using controlled thermo-oxidation conditions adapted to the formation of magnetite (Figure 3 a, b), temperature 65 °C and alkaline pH 8.6 [31]. The process of formation of magnetoferritin: penetration of Fe^{2+} ions through protein hydrophilic channels into the apoferritin shell under the action of an electrostatic gradient, oxidation of Fe^{2+} to Fe^{3+}, nucleation and growth of the inorganic core [30].

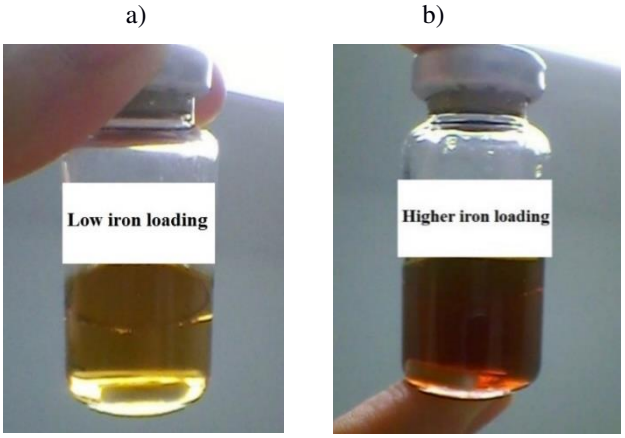


Fig. 3. The final product of the synthesis: magnetoferritin colloidal solution with protein concentration ~ 6 g/l with a) **low iron loading** ~ 168 Fe atoms per the one protein biomacromolecule; and b) **higher iron loading** ~ 532 Fe atoms per the one protein biomacromolecule [30].

Magnetoferritin can be considered as a suitable model system for in vitro studies of magnetite bio-mineralization in a constrained, uniformly sized hollow cavity. Magnetic nanoparticles prepared at 65 °C, alkaline pH, constant stirring and controlled addition of Fe/oxidant into the apoferritin nano-cage. The number of iron atoms added to the reaction affected the size of the iron core, the formation of aggregates, and the chemical composition [31].

Biomedical applications

Hyperthermia. Hyperthermia is a heat treatment technique for the organs and tissues which are infected by cancer. Therefore, ferrofluids are more susceptible to generating excess heat under an applied magnetic field and enhancing the temperature to damage the affected tissue or cells in the region of a cancer or tumor. In this technique, a ferrofluid is injected into a tissue which is infected by cancer, and then heat is generated by varying the applied magnetic field. This could

easily increase the temperature up to 42 °C – 45 °C, which is enough to destroy cancer cells [32], [33].

There is an interesting result: it can be observed that the positive effect of surface engineering on the magnetic properties of magnetic nanoparticles was synthesized biologically [34]. Magnetic nanoparticles were modified with poly-L-lysine and had its efficacy tested and compered to unmodified magnetic nanoparticles in magnetic hyperthermia for treatment of glioblastoma. In addition, poly-L-lysinemodified magnetic nanoparticles show a higher antitumoral activity with full tumor disappearances achieved in 50% of mice compared to 20% for unmodified magnetic nanoparticles [35].

In 2019 was reported that thermosensitive magnetic liposome (TML) with an anti-cancer chemotherapy drug (CPT-11) release was controlled by alternating magnetic field. Thermosensitive magnetic liposome were encapsulated with CPT-11 and acid citriccoated magnetic nanoparticles within the aqueous core and surface conjugated with antibody cetuximab (CET), for recognizing over-expressed epidermal growth factor receptors on cancer cell surface. For in vivo study, the brain tumor-bearing mice were subject to intravenous injection of test sample through the tail vein on day 11. In summary, the study groups included (I) saline (control); (II) CPT-11; (III) TML-CET; (IV) TML-CPT-11; (V) TML-CPT-11-CET; (VI) TML-CPT11-CET with magnetic guidance for 30 minutes with a magnet; (VII) TML-CPT-11-CET with magnetic guidance for 30 minutes with a magnet followed by alternating magnetic field treatment (60 A and 96 kHz in a 3.2-cm inner diameter coil) for 15 minutes. As shown in Figure 4, authors observed an increase from 1 to 7 in the study group in antitumor efficacy after treating brain tumor-bearing mice [36].

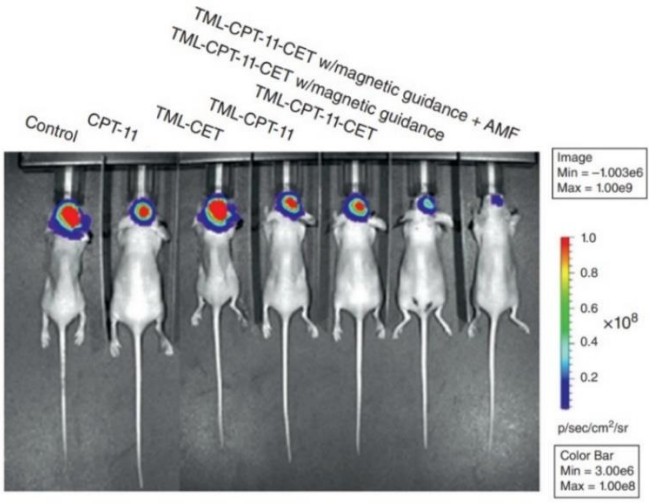


Fig. 4. Bioluminescence imaging of intracellular implanted U87 cells in the right intracranial region of each mouse obtained on day 21 [36].

Drug delivery. The research work related to the functionalization of iron oxide nanostructures, which make them a more promising candidate for drug delivery applications. Many enzymes, antibodies, antigens, genes, and drugs or drug-attracted molecules may get attached to the surface of magnetic iron oxide nanostructures, especially nanoparticles, and could be monitored through an external magnetic field. The functionalized magnet guides the dispersed nanostructure/nanoparticles into a biocompatible fluid-like serum, which enters through the intravenous system. The designated drug which is produced for the specific disease can be navigated through the strong external magnetic field and can deliver the drug to the particular portion within a given time. Although several studies have been performed using iron oxide nanoparticles, very few studies have been carried out for iron oxide nanoparticles based ferrofluids [37].

Magnetic resonance imaging. Magnetic resonance imaging is a medical imaging technique used in radiology to form pictures of the anatomy and the physiological processes of the body. For generating high-quality images, magnetic resonance imaging technology has undergone several developments, such as the application of sophisticated pulse sequences. Superparamagnetic nanoparticles enhance the proton nuclear relaxation rates, which increases the contrast of the image [38].

Magnetic separation of cells. The separation of small particles or biological cells is an important task and useful in chemical and biological applications. Various entities such as cells, bacteria, and other microorganism have a size of fewer than 5 μm [39]. Sorting and analyzing cells and molecules by size, complexity, and phenotype is critical for early detection, characterization of immune capacity, and improving infectious disease diagnosis [40]. These entities different shapes, sizes, and density can be easily separated with the help of iron-oxide based ferrofluids by applying an external magnetic field or internally induced magnetic field [41]. This technique is known as the magnetic separation or magneto-phoresis method [39], [40]. The magnetic separation performance may be affected by the size of magnetic particles, magnetic moments, the concentration of particles, the flow rate of ferrofluids [41].

III. SUMMARY

At the moment, the synthesis of magnetoferritin with different loading effects has been carried out. This experiment was done following the instructions above, but the loading effects were changed.

Fig. 5. The final product of the synthesis: magnetoferritin colloidal solution with protein concentration with **loading** ~ 550 Fe atoms per the one protein biomacromolecule; and **higher iron loading** ~ 730 Fe atoms per the one protein biomacromolecule; and **the highest iron loading** ~ 810 Fe atoms per the one protein.

As a result, a sediment was detected at a high loading effect. Sediment depends on loading factor. With higher iron loading, sedimentation is higher. About 500 and more iron atoms pre one protein led to higher sizes, and after time we can observe sediment. Sediment depends also on the electrostatic repulsion between particles and on the protein shell destroying. Protein shell (apoferritin) destroying depends on the higher iron loading, higher temperature, or extremely pH changes. Addition of NaOH can ususally partially stabilize sample and led preferable to magnetite formation. Turbidity can be in this way eliminated.

Further research and work with a solution of magnetoferritin should give more positive results. In future, these solutions may

be used as a contrast agent for magnetic resonance imaging, to obtain a brighter image.

Another goal of the work will be the study of magnetic hyperthermia and creation of samples. A number of materials is studied in the field of magnetic hyperthermia. In general, the most promising ones appear to be iron oxide particle nanosystems. On the other hand, the type of material itself provides several variations on how to tune hyperthermia indicators. Therefore, the analysis of magnetite nanoparticles in various forms is required.

REFERENCES

[1] Harrison, P.M.; Arosio, P. "The ferritins:Molecular properties, iron storage function and cellular regulation". Biochim. Biophys. Acta 1996, 1275, 161–203.

[2] Fadi Bou-Abdallah. The iron redox and hydrolysis chemistry of the ferritins. Biochim. Biophys. Acta, (2010), 1800, 719-731.

[3] L. Xue, D. Deng, J. Sun, "Magnetoferritin: process, prospects, and their biomedical applications", published online 2019 May 16. doi: 10.3390/ijms20102426.

[4] Van de Walle, A.; Plan Sangnier, A.; Abou-Hassan, A.; Curcio, A.; Hemadi, M.; Menguy, N.; Lalatonne, Y.; Luciani, N.; Wilhelm, C. Biosynthesis of magnetic nanoparticles from nano-degradation products revealed in human stem cells. Proc. Natl. Acad. Sci. USA 2019, 116, 4044–4053.

[5] Dario, F.; Paolo, A. Biology of ferritin in mammals: An update on iron storage, oxidative damage and neurodegeneration. Arch. Toxicol. 2014, 88, 1787–1802.

[6] Kostiainen, M.A.; Pierpaolo, C.; Manuela, F.; Panu, H.; Oksana, K.; Nolte, R.J.M.; Cornelissen, J.J.L.M.; Desautels, R.D.; Johan, V.L. Hierarchical self-assembly and optical disassembly for controlled switching of magnetoferritin nanoparticle magnetism. Acs Nano 2011, 5, 6394–6402.

[7] Mohammed, L., Gomaa, H.G., Ragab, D., and Zhu, J. (2016). Magnetic nanoparticles for environmental and biomedical applications: a review. Particuology 30: 1–14.

[8] Cardoso, V.F., Francesko, A., Ribeiro, C. et al. (2017). Advances in magnetic nanopartcles for biomedical applications. Advanced Healthcare Materials 7: 1–35.

[9] Deatsch, A.E. and Evans, B.A. (2014). Heating efficiency in magnetic nanoparticle hyperthermia.Journal of Magnetism and Magnetic Materials 354: 163–172.

[10] G. N. Lucena, C.C. dosSantos, G.C. Pinto, B.E. Amantéa, R.D. Piazza, M. Jafelicci Jr, R. Fernando C. Marques. (2021) Drug Delivery and Magnetic Hyperthermia Based on Surface Engineering of Magnetic Nanoparticles, ch11: 238-241.

[11] Radmansouri, M., Bahmani, E., Sarikhani, E., and Rahmani, K. (2018). Doxorubicin hydrochloride – loaded electrospun chitosan/cobalt ferrite/titanium oxide nanofibers for hyperthermic tumor cell treatment and controlled drug release. International Journal of Biological Macromolecules 116: 378–384.

[12] Mohd Imran, Adnan Mohammed Affandi, Md Mottahir Alam, Afzal Khan, Asif Irshad Khan (2021). Advanced biomedical applications of iron oxide nanostructures based ferrofluids. doi.org/10.1088/1361-6528/ac137a

[13] Shokrollahi H. (2013). Structure, synthetic methods, magnetic properties and biomedical applications of ferrofluids Mater. Sci. Eng. C 33 2476–87.

[14] Lu Y., Yin Y., Mayers B.T., Xia Y. (2002). Modifying the surface properties of superparamagnetic iron oxide nanoparticles through a sol–gel approach Nano Lett. 2:183–6.

[15] Wu W., He Q., Jiang C. (2008). Magnetic iron oxide nanoparticles: synthesis and surface functionalization strategies Nanoscale Res. Lett. 3 397.

[16] Komeili, A., Zhuo Li and D. K. Newman "Magnetosomes Are Cell Membrane Invaginations Organized by the Actin-Like Protein MamK" Science, 311, Jan. 2006, p. 242-245

[17] Pósfai, Mihály; Lefèvre, Christopher T.; Trubitsyn, Denis; Bazylinski, Dennis A.; Frankel, Richard B. (2013). "Phylogenetic significance of composition and crystal morphology of magnetosome minerals". Frontiers in Microbiology. 4. doi:10.3389/fmicb.2013.00344

[18] Ben-Shimon, Stein D., Zarivach R., "Current View of Iron Biomineralization in Magnetotactic Bacteria." Journal of Structural Biology: X, vol. 5, Elsevier, Jan. 2021, p. 100052, doi:10.1016/J.YJSBX.2021.100052.

[19] Yan, Lei, Huiyun Da, Shuang Zhang, V. M. López, Weidong Wang, "Bacterial Magnetosome and Its Potential Application." Microbiological Research, vol. 203, Urban & Fischer, Oct. 2017, pp. 19–28, doi:10.1016/J.MICRES.2017.06.005.

[20] Jacob, Jobin John, K. Suthindhiran. "Magnetotactic Bacteria and Magnetosomes – Scope and Challenges." Materials Science and Engineering: C, vol. 68, Elsevier, Nov. 2016, pp. 919–28, doi:10.1016/J.MSEC.2016.07.049.

[21] A. Hashim, M. Molcan, J. Kovac, Z. Varchulová, H. Gojzewski, M. Makowski, P. Kopcansky, Z. Tomori, M. Timko, "The Influence of Morphology on Magnetic Properties of Magnetosomes." Acta Physica Polonica A, vol. 121, no. 5–6, Państwowe Wydawnictwo Naukowe, 2012, pp. 1250–52.

[22] Lang, C.; Schüler, D. and Faivre, D. Synthesis of magnetite nanoparticles for bio- and nanotechnology: genetic engineering and biomimetics of bacterial magnetosomes. Macromol. Biosci. 2007, vol. 12, p. 144-151.

[23] Correia Carreira, S.; Armstrong, J.P.; Seddon, A.M.; Perriman, A.W.; Hartley-Davies, R.; Schwarzacher, W. Ultra-fast stem cell labelling using cationised magnetoferritin. Nanoscale 2016, 8, 7474–7483.

[24] Charlton, J.R.; Pearl, V.M.; Denotti, A.R.; Lee, J.B.; Swaminathan, S.; Scindia, Y.M.; Charlton, N.P.; Baldelomar, E.J.; Beeman, S.C.; Bennett, K.M. Biocompatibility of ferritin-based nanoparticles as targeted MRI contrast agents. Nanomedicine 2016, 12, 1735–1745.

[25] Jutz, G.; van Rijn, P.; Santos Miranda, B.; Böker, A. Ferritin: A Versatile Building Block for Bionanotechnology. Chem. Rev. 2015, 115, 1653–1701.

[26] Melníková, L.; Mitróová, Z.; Timko, M.; Kováč, J.; Avdeev, M.V.; Petrenko, V.I.; Garamus, V.M.; Almásy, L.; Kopčanský, P. Structural characterization of magnetoferritin. Mendeleev Commun. 2014, 24, 80–81.

[27] Elvira, F.; Claudia, I.; Matteo, Z.; Maria, F.; Elisabetta, F.; Miriam, C.; Valbona, S.; Lorenzo, D.C.M.; Carla, G.; Anna Maria, F. A smart platform for hyperthermia application in cancer treatment: Cobalt-doped ferrite nanoparticles mineralized in human ferritin cages. Acs Nano 2014, 8, 4705–4719.

[28] Meldrum, F.C.; Heywood, B.R.; Mann, S. Magnetoferritin: In vitro synthesis of a novel magnetic protein. Science 1992, 257, 522–523.

[29] Bulte, J.W.; Douglas, T.; Mann, S.; Frankel, R.B.; Moskowitz, B.M.; Brooks, R.A.; Baumgarner, C.D.; Vymazal, J.; Strub, M.P.; Frank, J.A. Magnetoferritin: Characterization of a novel superparamagnetic MR contrast agent. J. Magn. Reson. Imaging Jmri 2010, 4, 497–505.

[30] L. Balejčíková, M. Molčan, O. Štrbák, "Bio-inspired magnetic nanoparticles for biomedical and environmental nanotechnology development", unpublished.

[31] Wong, K.K.W.; Douglas, T.; Gider, S.; Awschalom, D.D.; Mann, S. Biomimetic synthesis and characterization of magnetic proteins (magnetoferritin). Chem. Mater. 1998, 10, 279–285.

[32] Kandasamy G., Sudame A., Bhati P., Chakrabarty A., Kale S.N., Maity D., "Systematic magnetic fluid hyperthermia studies of carboxyl functionalized hydrophilic superparamagnetic iron oxide nanoparticles based ferrofluids". J. Colloid Interface Sci. 2018. 514 534–43.

[33] Lahiri B., Ranoo S., Philip J., "Magnetic hyperthermia study in water based magnetic fluids containing TMAOH coated Fe3O4 using infrared thermography". Infrared Phys. Technol. 2017. 80 71–82.

[34] Hasany S. F., Ahmed I., Rajan J., Rehman A., "Systematic review of the preparation techniques of iron oxide magnetic nanoparticles". Nanosci. Nanotechnol. 2012. 2 148–58

[35] Wu W., Wu Z., Yu T., Jiang C., Kim W., "Recent progress on magnetic iron oxide nanoparticles: synthesis, surface functional strategies and biomedical applications". Sci. Technol. Adv. Mater. 2015. 16 43

[36] Gabbasov R., Polikarpov D., Cherepanova V., Chuev M., Mischenko I., Loginiva N., Loseva E., Nikitin M., Panchenko V., "Exogenous iron redistribution between brain and spleen after the administration of the 57Fe3O4 ferrofluid into the ventricle of the brain". J. Magn. Magn. Mater. 2017. 427 41–7.

[37] Kandasamy G., Khan S., Giri J., Bose S., Veerapu N., "One-pot synthesis of hydrophilic flower-shaped iron oxide nanoclusters (IONCs) based ferrofluids for magnetic fluid hyperthermia applications". 2019. J. Mol. Liq. 275 699–712.

[38] Bangerter N., Morrell G., Grech-Sollars M., "Bioengineering Innovative Solutions for Cancer". 2020. 2 163-194.

[39] Munaz A., Shiddiky M., Nguyen N., "Magnetophoretic separation of diamagnetic particles through parallel ferrofluid streams". Sensors Actuators B. 2018. 275 459–69

[40] Antfolk M., Laurell T., 2017. "Continuous flow microfluidic separation and processing of rare cells and bioparticles found in blood—a review". Anal. Chim. Acta 965 9–35

[41] Kuzhir P., Magnet C., Ezzaier H., Zubarev A., Bossis G. 2017. "Magnetic filtration of phase separating ferrofluids: from basic concepts to microfluidic device". J. Magn. Magn. Mater. 4

The Survey of Nonlinear Dynamical System Identification Methods

[1]*Tomáš TKÁČIK (1st year),*
Supervisor: [2]*Ján JADLOVSKÝ*

[1,2]Dept. of Cybernetics and Artificial Intelligence, FEEI TU of Košice, Slovak Republic

[1]tomas.tkacik@tuke.sk, [2]jan.jadlovsky@tuke.sk

***Abstract*—This survey provides an overview of the basic concepts, methods, and algorithms associated with the identification of nonlinear dynamical systems. It compares approaches of analytical and experimental identification in the context of white-box, grey-box, and black-box model structures. At the same time, it captures chronologically the system identification process from analytical identification to the validation of the obtained mathematical model. The main goal of my dissertation will be to create program modules for the identification of nonlinear dynamical systems.**

***Keywords*—Mathematical modeling, Nonlinear optimization, Parameter estimation, System identification**

I. Introduction

The goal of scientific research is to understand the world around us and to formally describe it using a generalized model [1]. Therefore, the purpose of the model is not necessarily to capture every detail of an object, but only the key aspects. In control engineering, the intention is to create a description of a dynamical system in the form of a mathematical model. This survey provides an overview of methods, structures, and algorithms associated with the identification of nonlinear dynamical systems. The individual sections are arranged according to the general system identification methodology. It starts with Section II, which describes the methods of analytical identification. Section III provides an overview of experimental identification methods, typical mathematical model structures, algorithms for parameter estimation, and the model validation procedure. The identification of the real (physical) system is dealt with in the final Section IV.

Presented methods, structures, and algorithms will be used in my dissertation thesis, which aims to develop and verify program modules to identify nonlinear dynamical systems. These program modules will be experimentally verified on laboratory plants with various dynamics available within the laboratories of the Center of Modern Control Techniques and Industrial Informatics (CMCT&II).

II. Mathematical modeling of Nonlinear Dynamical Systems

In our case, the subject of mathematical modeling is **dynamical systems**. A dynamical system is a real object that converts its inputs $\boldsymbol{u}$ into outputs $\boldsymbol{y}$. The output $\boldsymbol{y}$ of the system is functionally dependent on both the inputs $\boldsymbol{u}$ and the internal states $\boldsymbol{x}$ of the system. Thus, a dynamical system defines the evolution of system states $\boldsymbol{x}$ with respect to an independent variable (usually time t).

A **model** is an abstraction of a real system that describes the relationships between inputs $\boldsymbol{u}$ and outputs $\boldsymbol{y}$ of a system. The most common system representation is a **mathematical model** with input-output relationships expressed using mathematical functions. In addition, other forms of model representation exist, such as *tables* or *graphs*.

Systems can be classified into several groups based on their properties [2]. Based on the system variable types, the following classification applies:

- **Continuous:** The state values $\boldsymbol{x}$ change continuously over time $t \in \mathbb{R}$ and the system is described by differential equations.
- **Discrete:** The state values $\boldsymbol{x}$ change at discrete time intervals $k \in \mathbb{Z}$ and the system is described by a difference equation.
- **Hybrid:** Combines multiple continuous and/or discrete system dynamics that changes abruptly based on discrete events.

Systems can also be categorized according to mathematical operations that describe their input-output properties:

- **Linear:** Dependencies within the system can be described using linear functions, or their linear combination.
- **Nonlinear:** The system description consists of nonlinear functions or a nonlinear combination of functions. Real systems exhibit nonlinear behavior and thus we will solely focus on them.

Based on properties of model parameters:

- **Time-invariant:** Parameter values are fixed or their changes are insignificant.
- **Time-variant:** Parameter values evolve over time or change based on the system states.

Mathematical modeling is a set of techniques and methods with the aim to create a mathematical model of a system. The general form of the mathematical model in the state-space representation is given in (1), where $\boldsymbol{f}$ and $\boldsymbol{g}$ are arbitrary mathematical functions.

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \boldsymbol{f}\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right) \\ \boldsymbol{y}(t) &= \boldsymbol{g}\left(\boldsymbol{x}(t), \boldsymbol{u}(t)\right) \end{aligned} \tag{1}$$

The aim of mathematical modeling is to identify these two functions ($\boldsymbol{f}$ a $\boldsymbol{g}$) using physics laws. As the system can be quite complex, it is better to split it into simpler components. These can be modeled separately and subsequently connected together to obtain the overall model [3]. Standard techniques

such as impedance modeling, balance equations, or Lagrange equations can also be used. The resultant model structure is the so-called *white-box model*. The advantage of this approach is that the obtained model is parametric and the parameters have a physical interpretation. The problem arises when parameter values could not be directly measured or read from datasheets. Another disadvantage is the fact that many physical laws and equations are based on the assumptions of an *ideal environment*. These approximations may not be accurate enough in every application. Stated shortcomings of the white-box model can be mitigated by performing appropriate experiments on a real system and their subsequent analysis. Such an approach is called **experimental identification**.

III. Methods of Nonlinear System Identification

System identification is an iterative process that aims to obtain a mathematical model of the system from experimental data. More precisely from the measurable inputs $\boldsymbol{u}$ and outputs $\boldsymbol{y}$ of the observed system. The identification process can be divided into several steps [4] that are repeated when necessary:

- **Experiment design:** An initial step in which it is necessary to apply all a priori knowledge about the observed system to design an experiment, as improper design may lead to the destruction of the real system. In addition to the experiment design, this point also includes the collection of data from the system, where the set of available sensors, the method of data recording, the sampling period T_s, etc. must be taken into account. It is important to note that even though the system is continuous, the recorded data is in discrete form. This fact plays an important role in continuous model identification.
- **Structure selection:** The mathematical model can have different shapes and sizes, and the individual parameters could also have different physical interpretations in various structures. The model structure and consequently the number of parameters is closely tied to the *model complexity*. More complex models usually approximate the real system better but are associated with higher computational cost and more complex parameter estimation algorithms.
- **Parameter estimation:** This is a step in which the selected model structure adapts to the measured data by changing the parameter values. Parameter estimation is closely linked to statistical and optimization methods [5].
- **Model validation:** The last step of system identification is to verify that the identified mathematical model meets its expectations. The purposes of the mathematical models are diverse (digital twin, design of control laws, etc.) and therefore it is not possible to rely solely on qualitative or quantitative evaluation.

According to the selected model structure, the identified models can be divided into two main categories: *grey-box* and *black-box*. In the case of grey-box models, the model's structure is predetermined and only the parameters are identified. On the contrary, in the case of the black-box model both parameters and structure are identified simultaneously. In both cases, the approximation model of the system is modified according to the output prediction error $\boldsymbol{e}(k)$, (2).

$$\boldsymbol{e}(k) = \boldsymbol{y}(k) - \hat{\boldsymbol{y}}(k) \tag{2}$$

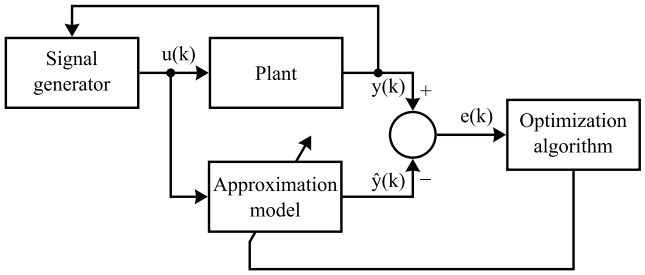


Fig. 1. General schema of system identification loop driven by output prediction error [6].

where: $\boldsymbol{e}(k)$ - output prediction error
$\boldsymbol{y}(k)$ - output of real system
$\hat{\boldsymbol{y}}(k)$ - output of approximation model

With the addition of a proper optimization algorithm, the identification structure in Fig. 1 can be used for both the online and the offline identification tasks.

A. Grey-box System Identification

Grey-box model system identification methods can be further subdivided according to whether they lean more towards white-box or black-box models. The resulting division, named by the author of [1], is called a *model's palette of grey shades*:

- **Off-white models** are derived from white-box models and the subject of identification is the parameter vector $\boldsymbol{\theta}$. This method can be challenging especially if the mathematical model (1) includes complicated functions.
- **Smoke-grey models** seek to effectively eliminate nonlinear properties of real systems and thus enabling the system to be identified in a linear structure. For this, nonlinear transformations of the measured data are used. These transformations could be based on the physical nature of the system. The described procedure is sometimes referred to as the feedback linearization method [7].
- **Steel-grey models** are based on the idea that nonlinear systems can be approximated by linear ones within the close vicinity of the operation area. By combining several linear models, it is possible to obtain one composite model that describes the dynamics of the nonlinear system. By modification of the composite model formulation, it is possible to obtain a linear time-variant model.
- **Slate-grey models** are the last stage before the black-box models. Block-oriented models are a good example of this category, where the structure of the system is composed of functional blocks. Two types of blocks are used: linear dynamical systems and static nonlinear transformations. The exact choice of suitable nonlinear transformations can also have a physical basis.

Methods based on system identification in the form of a grey-box model require designing the structure of the model. This is laborious from the user's perspective as it requires a lot of experimentation and practical experience.

B. Black-box System Identification

The use of black-box models makes it possible to identify dynamical systems without the need to define the exact structure. The structure is created automatically during the identification process using an optimization algorithm [4]. Therefore, these methods are being referred to as data-driven

methods. In (3) the general mathematical form of the model is given. The function $\boldsymbol{f}$ represents a model consisting of both the parameters and the structure of the identified system.

$$\hat{\boldsymbol{y}}(k+1) = \hat{\boldsymbol{f}}\big(\boldsymbol{y}(k), \boldsymbol{y}(k-1), ..., \boldsymbol{y}(k-n), \boldsymbol{u}(k), \boldsymbol{u}(k-1), ..., \boldsymbol{u}(k-m)\big) \tag{3}$$

This notation is used because the internal states and the parameters have no physical significance. The black-box mathematical model is usually discrete, as it was created based on experimental data obtained at discrete time intervals.

In terms of nonlinear identification in the form of a black-box model, **neural networks** are dominant category of machine learning (ML) methods. This choice is motivated by the fact that neural networks are considered a universal approximator of any mathematical function. Ignoring the random noise present in the real systems, this property is a sufficient precondition for building a system model. The most used black-box model structures are [4]:

- **Multilayer Perceptron** combined with the backpropagation optimization algorithm is one of the simplest neural network applications. In the system identification context, it provides satisfactory results, while drawbacks are associated mainly with parameter convergence, computational complexity, and the local minimum problem [8].
- **Radial Basis Function Neural Network** is a special case of a neural network with a single hidden layer that uses the Gaussian function as an activation function. The learning is split in two phases: first, the parameters of the Gaussian functions are estimated; second, the synaptic weights are estimated. The idea behind the use of the Gaussian function is to divide the nonlinear workspace into smaller areas [9] that can be described linearly.
- **Functional Link Network** uses hardcoded nonlinear input transformations to linearize the workspace.
- **Time Delay Network** has an input layer extended by historical inputs $\boldsymbol{u}$ and outputs $\boldsymbol{y}$ of the system. This makes it possible to better represent the current state of the system, which subsequently helps to improve prediction accuracy. Time Delay Network is a key concept of dynamical system identification.
- **Recurrent Neural Network** contains recurrent connections that allow the neural network to maintain the previous state of the system [10] similar to the Time Delay Network. The system state is maintained internally so there is no need to modify the input layer. An improvement to recurrent neural networks is the Long Short-Term Memory (LSTM) model, which allows longer retention of system status [11]. Recurrent networks are often associated with the stability problem.
- **Wavelet Neural Network** is an architectural design in which the activation functions of the first hidden layer are replaced by a wavelet transform. This allows for better analysis of signal properties [12] and consequently partial elimination of nonlinear properties.
- **Temporal Convolutional Network** is one of many applications of a relatively new machine learning paradigm called *Deep Learning*. Thanks to the deep and complex structures, it combines the advantages of the above approaches. The main advantage of deep learning is the automatic feature extraction from the data that makes it a very flexible structure for dynamical system identification. Authors in [13] compared this approach with *Multilayer Perceptron* and *LSTM* structures and experimentally proved its potential in system identification.

In addition to the mentioned structures of neural networks, it is possible to identify systems using other ML methods or their combinations [14]. Popular are neuro-fuzzy methods, genetic algorithms, Swarm Intelligence, etc. Although black-box models provide high flexibility in the task of system identification, their structure is often too complicated to be properly analyzed using classical methods of systems analysis.

C. Parameter estimation algorithms

Up until now, this section was focused on methods and structures of system identification. The remaining unanswered question is how to determine the values of the mathematical model's parameter vector $\boldsymbol{\theta}$. Values estimation can be formulated as an *optimization task*, that is formally written in (4).

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \boldsymbol{e}^T(i, \boldsymbol{\theta}) \boldsymbol{e}(i, \boldsymbol{\theta}) \tag{4}$$

where: $\hat{\boldsymbol{\theta}}$ - estimated parameter vector
N - number of samples

The stated task can be solved using various optimization algorithms. To make this section cleaner, we are going to list only a few of the most prevalent optimization algorithms used for dynamical system identification:

- **Gauss-Newton least squares** is a standard algorithm for the nonlinear estimation of function parameters. It is based on the assumption that the error function is a quadratic function.
- **Steepest descent** algorithm is known for estimating neural network parameters and is based on a function's gradient calculation. The values of the parameter vector $\boldsymbol{\theta}$ are literally shifted in the *downhill* direction of the gradient in each iteration.
- **Levenberg-Marquardt least squares** is an algorithm combining principles of both the Gauss-Newton and the Steepest descent algorithms. In the vicinity of optimal parameter values, it behaves similarly to Gauss-Newton and at greater distances as Steepest descent [15].
- **Evolutionary computing** is inspired by optimization observed from nature. The main advantage of this algorithm is the lower susceptibility to stuck in the local optimum, as the searched space is relatively widely covered.

The choice of the optimization algorithm is closely linked to the choice of the model structure. A combination of different algorithms or their modification can make the optimization task faster and more reliable.

Apart from the selection of optimization algorithms, the initialization of the parameter vector $\boldsymbol{\theta}$ is also important. In the case of grey-box models, where the parameters have their physical significance, it is possible to *estimate* values of unknown parameters by hand. In the case of black-box models, *random initialization* is used, as there is no connection between parameter values and the real world.

D. Model validation

The last part of the systems identification process is the validation of the identified mathematical model. The mathematical model can be validated in an open-loop or in a closed-loop. In both cases, the output of the real system is compared

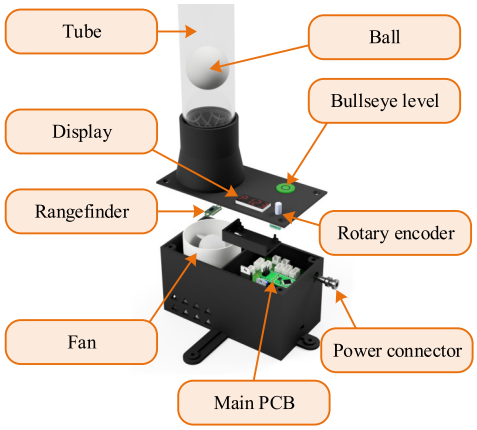


Fig. 2. Component assembly of the Aerodynamic Ball Levitation Laboratory Plant [6]. See: http://kyb.fei.tuke.sk/laboratoria/modely/al.php

with the output of the mathematical model. Historically, model validation has often been overshadowed by the identification methods [16]. Therefore, subjective evaluation methods like *by eye inspection* were mainly used to compare the outputs. From the model validity evaluation standpoint, this is still a common method [4]. The drawback of this method is sole reliance on the expert's judgment. For this reason, it is more appropriate to rely on **quantitative evaluation methods** instead. A *fit ratio* is usually used for this task. Apart from the possibility to generate negative values, it is also sensitive to the signal amplitude in the case of non-normalized values. The authors in [17] presented a modification of this method that internally uses normalization as the solution for these problems. Lastly, part of the model validation is to verify whether the model meets the author's expectations, e.g. whether it allows the design of suitable control laws.

As part of my dissertation thesis, I will use the identified mathematical model for control law design. The validation of the model will therefore be performed mainly in the closed-loop control structure.

IV. Identification of Aerodynamic Ball Levitation Laboratory Plant

Methods and techniques presented in this survey were used in our article [6] focused on the construction, mathematical modeling, experimental identification, and control of the physical dynamical system of aerodynamic levitation shown in Fig. 2. As part of analytical identification, a mathematical model was derived in the form of a system of nonlinear differential equations. Since not all parameters of the model were directly measurable, it was necessary to proceed to experimental identification. The resulting structure of the obtained model was a grey-box model. This model has been used to design control laws that were later verified in the simulation environment. Finally, control laws were applied to the real system and the outputs were compared with the simulation. The obtained results were evaluated both qualitatively and quantitatively. In conclusion, the obtained mathematical model was able to appropriately approximate key parts of the real system's dynamics. This plant will be further used in research activities of CMCT&II.

V. Conclusion

The presented survey provides an overview of key methods, structures, and algorithms used in the identification of nonlinear dynamical systems. It also compares approaches of analytical and experimental identification. I will use the presented knowledge in my dissertation thesis in the design of program modules to identify nonlinear dynamical systems. These program modules will be experimentally verified on models of physical systems at CMCT&II. In addition to our research group, the authors will explore the possibility of applying these program modules and methodology to solve modeling and identification tasks of hybrid systems in the *ALICE Experiment* at CERN that both authors are associated members of.

Acknowledgment

This work has been supported by the project ALICE experiment at the CERN LHC: The study of strongly interacting matter under extreme conditions (ALICE KE FEI TU 0195 / 2021).

References

[1] L. Ljung, "Perspectives on system identification," *Annual Reviews in Control*, vol. 34, no. 1, pp. 1–12, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1367578810000027
[2] J. Lygeros, S. Sastry, and C. Tomlin, "Hybrid systems: Foundations, advanced topics and applications," *under copyright to be published by Springer Verlag*, 2012.
[3] G. Goodwin, S. Graebe, S. GRAEBE, and M. Salgado, *Control System Design*. Prentice Hall, 2001.
[4] L. Ljung, C. Andersson, K. Tiels, and T. B. Schön, "Deep learning and system identification," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 1175–1181, 2020.
[5] J. L. Crassidis and J. L. Junkins, *Optimal Estimation of Dynamic Systems, Second Edition (Chapman & Hall/CRC Applied Mathematics & Nonlinear Science)*, 2nd ed. Chapman & Hall/CRC, 2011.
[6] T. Tkáčik, M. Tkáčik, S. Jadlovská, and A. Jadlovská, "Design of aerodynamic ball levitation laboratory plant," *Processes*, vol. 9, no. 11, p. 1950, 2021.
[7] P. Šuster and A. Jadlovská, "Modeling and control design of magnetic levitation system," in *2012 IEEE 10th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 2012, pp. 295–299.
[8] M. Mashor, "Performance comparison between hmlp, mlp and rbf networks with application to on-line system identification," in *IEEE Conference on Cybernetics and Intelligent Systems, 2004.*, vol. 1, 2004, pp. 643–648 vol.1.
[9] O. Nelles, *Nonlinear system identification: from classical approaches to neural networks, fuzzy models, and gaussian processes*. Springer Nature, 2020.
[10] W. Yu, "Nonlinear system identification using discrete-time recurrent neural networks with stable learning algorithms," *Information Sciences*, vol. 158, pp. 131–147, 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025503002032
[11] Y. Wang, "A new concept using lstm neural networks for dynamic system identification," in *2017 American Control Conference (ACC)*, 2017, pp. 5324–5329.
[12] C.-J. Lin, C.-C. Peng, C.-H. Chen, and H.-Y. Lin, "A self-organizing recurrent wavelet neural network for nonlinear dynamic system identification," *Applied Mathematics & Information Sciences, Appl. Math. Inf. Sci. 9*, no. 1L, pp. 125–132, 2015.
[13] C. Andersson, A. H. Ribeiro, K. Tiels, N. Wahlström, and T. B. Schön, "Deep convolutional networks in system identification," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 3670–3676.
[14] L. Fu and P. Li, "The research survey of system identification method," in *2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics*, vol. 2. IEEE, 2013, pp. 397–401.
[15] H. P. Gavin, "The levenberg-marquardt algorithm for nonlinear least squares curve-fitting problems," *Department of Civil and Environmental Engineering, Duke University*, vol. 19, 2019.
[16] Y. Barlas, "Formal aspects of model validity and validation in system dynamics," *System Dynamics Review: The Journal of the System Dynamics Society*, vol. 12, no. 3, pp. 183–210, 1996.
[17] H. Muroi and S. Adachi, "Model validation criteria for system identification in time domain," *IFAC-PapersOnLine*, vol. 48, no. 28, pp. 86–91, 2015.

Review of Cyber-Physical Systems and their Architectures

[1]*Zuzana Pugelová (1st year)*
Supervisor: [2]*Anna Jadlovská*

[1,2]Dept. of Cybernetics and Artificial Intelligence, FEEI TU of Košice, Slovak Republic

[1]zuzana.pugelova@tuke.sk, [2]anna.jadlovska@tuke.sk

***Abstract*— This paper deals with Cyber-Physical Systems in the context of architecture design. There is no unified concept of CPS and therefore different views on CPS are created along with several architectures aimed at different views, features, and applications. For this paper, we have chosen four architectures. Two of them are divided into unique layers that are comparable with layers of Distributed Control System. Another two are service-oriented architectures consisting of tiers, which can be associated with the DARMA system of the ALICE experiment at CERN.**

***Keywords*— Cyber-Physical System Architectures, Distributed Control System, SOA, Assembly lines**

I. Introduction

The potential of Cyber-Physical Systems (CPS) from a societal and economic point of view is greater than predicted. This has been confirmed by worldwide investments in the development of various CPS. As an engineering discipline, CPS is focused on technology. With that comes the main challenge in connecting abstractions of modeling physical processes, with computer science abstractions. The concept of CPS integrates technologies of computation, physical processes, and networking. On top of that, it incorporates the techniques of modeling, design, and analysis, which are provided by the dynamics of integrated physical, software, and networking processes [1].

CPSs are the core of my dissertational thesis *Platform for modeling and simulation of cyber-physical systems using advanced optimization methods and algorithms*. My research will be focused on the levels of CPS architecture, which contain methods, tools, and techniques of modeling, data processing, and visualization. The results of this research will be used to select appropriate methods for discrete event systems, which are incorporated into Distributed Control System (DCS). The structure of this paper is organized in this way. Section II provides a definition of CPS with a focus on various CPS architectures. Section III presents a brief description of DCS. Section IV compares different CPS architectures to DCS, which is a complex model of CPS.

II. Overview of selected Cyber-Physical System architectures

CPSs belong to the most important and significant advancements in computer science. The arise of CPSs has fascinated a significant number of researchers across the scientific community along with governments and industries [2]. The ongoing rise of CPS is pushed by numerous current trends, most notable are the availability of low-cost, low-power, high-capacity, computing devices; abundant internet bandwidth; continuous improvements in energy capacity, and others [3].

CPS flawlessly integrates computation with physical processes and therefore provides modeling, design, information, and abstraction techniques. Although CPS´s technology depends on multiple disciplines, such as computing and networking technologies, CPSs themselves range from somewhat small systems such as automobiles to large systems like a national power grid [4].

Over the duration of the past few years various definitions of CPS have been formed. A concept map of a Cyber-Physical System is shown in Fig. 1. CPS as a research area was emerging in the 1970s when the first microprocessor emerged. In the year 2006, Helen Gill named the systems that connect the physical world components with the digital as CPSs and were defined as "physical and engineered systems whose operations are coordinated, integrated and monitored by a communication and computing core" [2]. Another definition is proposed by Edward A. Lee as "the integrations of computation, networking, and physical processes. Embedded computers and networks monitor and control the physical processes, with feedback loops where physical processes affect computations and vice versa" [5]. Further definitions were proposed in [6],[7] and [8] with analogous content.

Although there is no generic or unified definition of CPS, generally accepted universal definition of CPS is **"**the fuse of the cyber world and the dynamic physical world. CPS perceive the physical world, process the data by computers, and affect and change the physical world" [9]**.**

In the following paragraphs, we will look at several CPS architectures proposed by different researchers.

A. Five-layer Cyber-Physical System Conceptual Model Architecture

Authors in [10] proposed a conceptual model of CPS consisting of five layers: *Physical*, *Network*, *Data storage*, *Processing and analytics*, and *Application*. This architecture can be applied in the manufacturing industry. Fig. 2 shows the Cyber-Physical System conceptual model.

Physical layer consists primarily of hardware such as sensors, actuators, controllers, etc., from which real-time data are processed locally or sent to the cloud. Sensors convert analog or digital events from nearby appliances into electrical signal,

which is collected via a controller. Additionally, the controller either activates devices or sends data via the *Network layer* to the *Storage layer*.

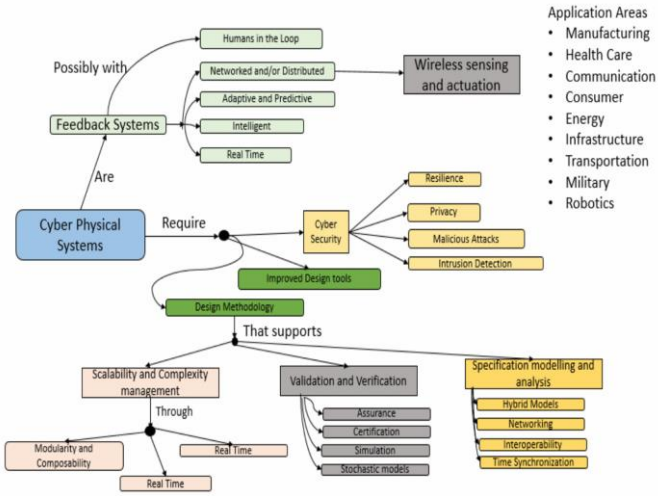


Fig. 1. Concept map of Cyber-Physical System [2].

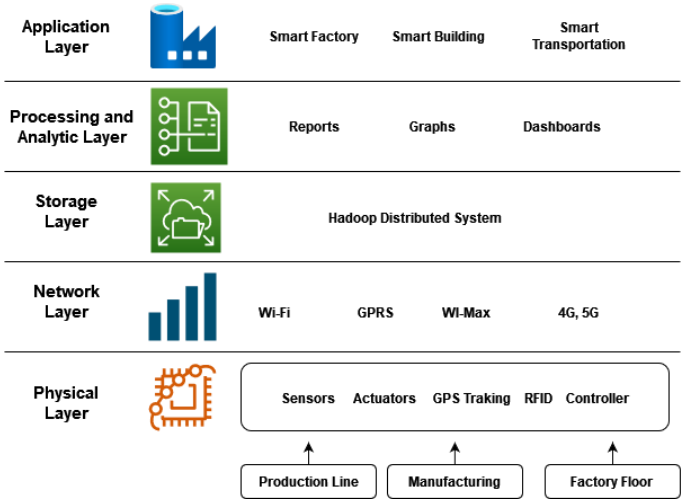


Fig. 2. Cyber-Physical System Conceptual Model [10].

Network layer uses different network protocols, which differ in transmission range, power requirement, connectivity, data rate, and other aspects. The most popular protocols are Wi-Fi, Bluetooth, UART, LTE, and others. These protocols are applied to various CPS applications, such as smart buildings, greenhouse control, and others.

Storage layer collects data mostly from the *Physical layer*. Data can be stored in the cloud (AWS, Azure, Google Cloud…) or locally on a server. To effectively store and process big datasets, systems like Hadoop are used.

Processing and analytic layer´s purpose is to process data from actuating devices, for prediction. For that, simulation models are used (e. g. map-reduce, dimensional modeling algorithm). Visualization, reports, graphs are created for evaluation purposes. Predictions from this layer can be transferred to the *Physical layer* for device maintenance.

Application layer represents interface and applications that consumers and service providers can interact with the CPS layer in a user-friendly approach based on granted privileges and user status [10].

B. CPS architecture for intelligent manufacturing

Authors in [11] proposed CPS architecture (Fig. 3) on the shop floor for intelligent manufacturing. The architecture consists of 3 layers – the *Physical connection layer*, the *Middleware layer*, and the *Computation layer*. Each layer is described as follows.

Physical connection layer uses interconnected sensors, RFIDs, and other devices for sensing the environment around them. Sensors are connected through Fieldbus protocol or industrial Ethernet.

Middleware layer acts as a link connecting the *Physical connection layer*, the *Computation layer*, and other external applications. It is required from the *Middleware layer* to support Device management (to work together with external applications flawlessly), Interface definition and, Data management (a large variety of data collected from sensors require consistent data format). The *Middleware layer* transfers data from the *Physical connection layer* to the central server for analysis. Based on the analysis, commands are sent to controllers of the *Physical connection layer*.

Computation layer is responsible for the incorporation of general knowledge with practical human experiences, thus creating a unified view of data, information, and knowledge. Thanks to this, CPS can be more intelligent when knowledge is applied to production process management. Furthermore, the *Computation layer* act as supervisory control by using batch or stream computing to process historical data in large quantities, which are then sent back to machines for maintenance purposes [11].

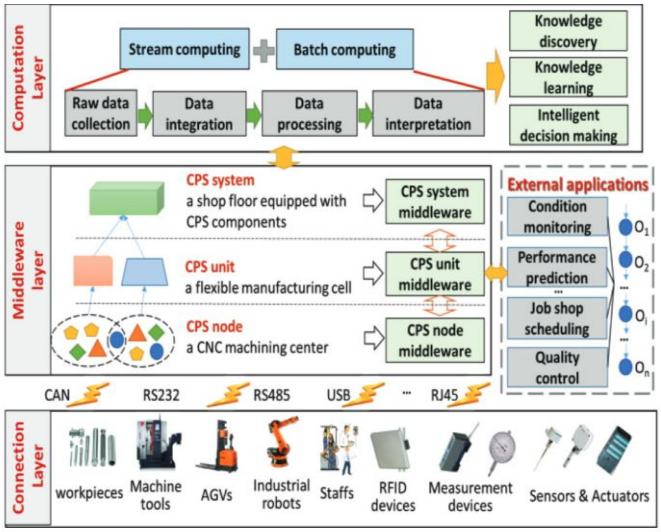


Fig. 3. Cyber-Physical System architecture for intelligent manufacturing [11].

C. Service-based Cyber-Physical System architecture

Researchers in [12] identified two reasons for research on service-based CPS architecture. The first is the accessibility of mobile networks. The second is that Service Oriented Architecture (SOA) and Cloud Computing (CC) can be solutions to the problem of limited resources on physical devices. They introduced Service based architecture CPSs based on several hypotheses and studies of what CPS is. Key benefits are presented, which give the ability to reuse services for several CPS and to separate services from physical devices. The architecture is shown in Fig. 4 and it consists of three tiers: *Environmental*, *Control*, and *Service Tier*.

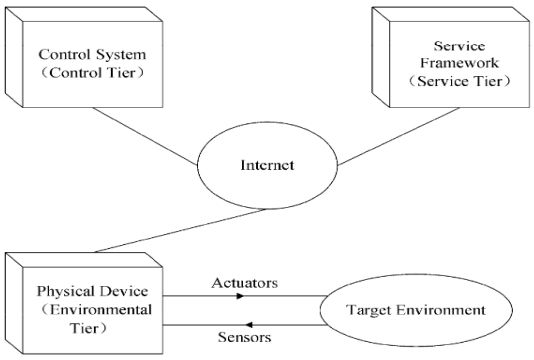


Fig. 4. Service-based Cyber-Physical System architecture [12].

Environmental Tier consists of a target environment and physical devices (typically in a form of the embedded devices). The target environment includes end-users using the devices and their associated physical environment.

Service Tier is a typical computing environment with services in SOA and CC. In that, several services are deployed in Service Repositories, and a Service Framework manages the services and interacts with service consumers.

Control Tier receives monitored data from sensors to make controlling decisions, to find the right services by consulting Service Framework, and to let the services be invoked on the Physical Device [12].

D. Cyber-Physical System architecture based on SOA

Academics in [9] proposed universal CPS architecture based on SOA. SOA is defined in [13] as "a style resulting from the use of particular policies, practices and frameworks that deliver services that conform to certain norms". The proposed architecture is shown in Fig. 5. Its main advantage is the application of services and components with great flexibility. The architecture consists of five tiers (*Perceive*, *Data*, *Service*, *Execution*, and *Security assurance tier*).

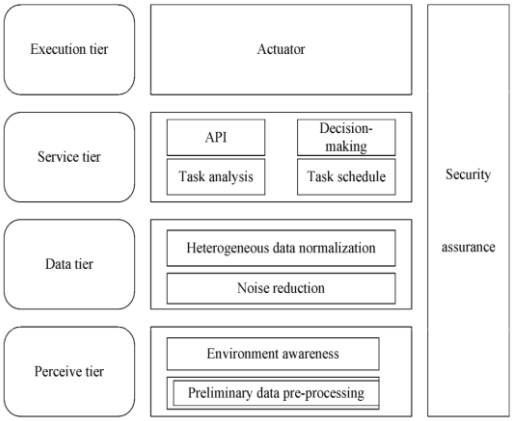


Fig. 5 SOA-based CPS architecture [9].

Perceive tier is responsible for the perception of surroundings via sensors and data preprocessing. Sensors and Wireless Sensor Networks are essential for this tier.

Data tier connects Producer with Service and provides processing (noise reduction, normalization…) of stored data. Storage devices along with computational tools are used for data storage.

Service tier oversees the deployment of services and their interaction. Additionally, it provides API, decision making, analytic software, and other tools.

Execution tier along with the *Perceive tier* interacts with the environment. Also, it executes tasks based on received commands from the system.

Security assurance tier is universal for the entire system. It is responsible for secure access to data, devices, and so on. To achieve that, a set of safety methods are used [9].

III. Distributed control system architecture

Hierarchical architecture representing the infrastructure of DCS is a complex model of CPS. One of many possible implementations of DCS is the DCS architecture at the Center of Modern Control Techniques and Industrial Informatics (CMCT&II) at the Department of Cybernetics and Artificial Intelligence (DCAI). This architecture, as an information and control system, integrates several physical plants of production lines, models of dynamical systems, and others models useable in practice [14],[15]. Fig. 6 shows DCS infrastructure.

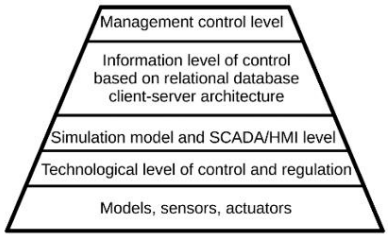


Fig. 6. Distributed Control System Infrastructure at the DCAI FEEI [14].

Level of Sensors and Actuators includes various sensors, actuators, and complex models. Hardware can be connected to a higher level by analog, digital, or various other technological interfaces.

Technological Level of Control and Regulation provides control and regulation of individual parts of the lower level and ensures communication with the second level of DCS via PLCs, process computers, and single-chip microcomputers.

Level of SCADA/HMI includes SCADA/HMI systems providing supervisory control, data acquisition, archiving, various visualizations, and simulation models. Links to upper and lower levels are provided by network interfaces using the TCP / IP protocol with various extensions.

Information Level of Control represents the level of Manufacturing Enterprise Systems used for the implementation of production support tasks. These systems are implemented using relational databases with client web applications.

Management Level of Control covers multidimensional databases with OLAP technology. This level provides the means to support strategic business planning [15], [16].

IV. Comparison of CPS architectures and DCS architecture at DCAI FEEI TUKE

CPS architectures mentioned in Section II have some things in common, e.g., the first layer and/or tier mostly consists of sensors with the aim to get data from the surrounding environment; the last layer and/or tier is in general responsible for monitoring, controlling, and function of API and whole interface. However, architectures are still different enough, that it is not possible to compared them simultaneously with a different CPS implementation. Because of that, we will compare them individually with our CPS implementation (DCS).

Architecture mentioned in Section II.A is most similar to DCS in terms of layers, application, and communication. The difference is in *Network layer*, which is not included in the DCS. Instead of a specific layer, DCS has implemented communication protocols in individual layers. Another distinction is in layers dealing with storage and analysis, which DCS implemented in *Information Level of Control*. However, II.A has distinctive layers for storage and for processing with analysis. Besides that, DCS contain individual layers for simulation and control.

Although the architecture described in II.B is also very similar to DCS in terms of application, the technology division and layer functions are different. Architecture in II.B included technologies and functions of the *Level of Sensors and Actuators*, the *Technological Level of Control and Regulation,* and the *Level of SCADA/HMI* from DCS into the *Physical connection layer*. Data and storage are presented in the *Middleware layer*. The *Computation layer* includes ERP, MES, systems like MIS, also used by DCS in the *Information Level of Control* and the *Management Level of Control*.

The architecture described in II.C was created to provide functionality through services, which is different from the DCS aim. Additionally, components are divided into three tiers, which creates different structures from the ones containing layers. The *Level of Sensors and Actuators* of DCS is incorporated into the *Environmental Tier*. The *Technological level of control and regulation* and the *Management control level* are implemented in the *Control tier*. However, other levels are not so clearly represented because the *Service tier* is an environment with SOA and CC services.

Another architecture with the aim in service, specifically in SOA, also with a structure consisting of tiers is mentioned in II.D. It has a differently composed architecture even though is also service-oriented. The *Level of Sensors and Actuators* of DCS can be seen in the *Perceive tier*. Databases from the *Information level* of DCS are implemented in the *Data tier*, along with storage and client-server architecture. Systems such as ERP, MES, and MIS from DCS are part of the *Service tier*. There is also some similarity between the *Technological level of Control and regulation* and the *Management control level* of DCS with the *Execution tier*, which interacts with sensors and executes commands. The *Security tier* is unique for DCS because DCS does not have resolved security on a specific level.

V. Conclusion and Future work

This paper presents several definitions of CPS along with an overall description, the motivation behind them, and their usability. Application of CPS is wide-range, therefore various architectures for different purposes are created. This paper presented four architectures, two of them having clear structures consisting of layers. Another two are service-oriented, consisting of tiers. Furthermore, presented CPS architectures are compared to DCS, which is a practical model of CPS.

In our future research, we will explore options for the modification of methods aimed at modeling and developing simulation tools in the context of my dissertation thesis. We will be using a model of the assembly line plant which is integrated into the infrastructure of research group CMCT&II. Architectures mentioned in Section II.A and Section II.B will be implemented on assembly lines. SOA architectures mentioned in Section II.C and Section II.D could be used in the development of system DARMA of ALICE experiment in The European Organization for Nuclear Research (CERN).

Acknowledgment

This work has been supported by the project ALICE experiment at the CERN LHC: The study of strongly interacting matter under extreme conditions (ALICE KE FEI TU 0195 / 2021).

References

[1] Cyber-Physical Systems. Retrieved from www.ptolemy.berkeley.edu: https://ptolemy.berkeley.edu/projects/cps/

[2] J. Jamwal, R. Agrawal, V. K. Manupati, M. Sharma, L. Varela, and J. Machado. “Development of cyber physical system based manufacturing system design for process optimization.” In: *IOP Conference Series: Materials Science and Engineering*, vol. 997, no. 1, p. 012048. IOP Publishing, 2020.

[3] R. Rajkumar, I. Lee, L. Sha, and J. Stankovic. "Cyber-physical systems: the next computing revolution." In *Design automation conference*, pp. 731-736. IEEE, 2010.

[4] J. WAN, et al. “Advances in cyber-physical systems research,” *KSII Transactions on Internet and Information Systems (TIIS)*, 5.11: 1891-1908, 2011.

[5] E. A. Lee. "Cyber physical systems: Design challenges." In *2008 11th IEEE international symposium on object and component-oriented real-time distributed computing (ISORC)*, pp. 363-369. IEEE, 2008.

[6] R. Baheti, H. Gill. “Cyber-physical systems.” *The impact of control technology*, 12.1: 161-166, 2011.

[7] H. Song, at al. “Cyber-physical systems: foundations, principles and applications”. Morgan Kaufmann, 2016

[8] N. Jazdi. “Cyber physical systems in the context of Industry 4.0.” In *2014 IEEE international conference on automation, quality and testing, robotics*. IEEE, p. 1-4, 2014.

[9] L. Hu, et al. “Review of cyber-physical system architecture”. In: *2012 IEEE 15th international symposium on object/component/service-oriented real-time distributed computing workshops.* IEEE, p. 25-30, 2012.

[10] A. R. Al-Ali, R. Gupta, and A. A. Nabulsi. "Cyber physical systems role in manufacturing technologies." In *AIP Conference Proceedings*, vol. 1957, no. 1, p. 050007. AIP Publishing LLC, 2018.

[11] Ch. Liu, P. Jiang, “A cyber-physical system architecture in shop floor for intelligent manufacturing.” *Procedia Cirp*, 56: 372-377, 2016.

[12] H. J. La, and S. D. Kim. "A service-based approach to designing cyber physical systems." *2010 IEEE/ACIS 9th International Conference on Computer and Information Science*. IEEE, 2010.

[13] D. Sprott, L. Wilkes. “Understanding service-oriented architecture.” *The Architecture Journal*, 1.1: 10-17, 2004.

[14] J. Jadlovský, et al. “Research activities of the center of modern control techniques and industrial informatics.” In *2016 IEEE 14th International Symposium on Applied Machine Intelligence and Informatics (SAMI).* IEEE, p. 279-285, 2016.

[15] Distributed Control System Infrastructure at the DCAI FEEI. Retrieved from www.kyb.fei.tuke.sk: http://kyb.fei.tuke.sk/laben/infdsr.php

[16] M. Tkacik. “Embedded Systems and their Implementation in Distributed Control and Monitoring Systems.” Diploma thesis, TUKE, 2019.

Recent trends in detection of hate speech and offensive language on social media

[1]*Zuzana SOKOLOVÁ (1st year)*
Supervisor: [2]Jozef JUHÁR

[1,2]Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic
[1]zuzana.sokolova@tuke.sk, [2]jozef.juhar@tuke.sk,

***Abstract*— In the article we describe a current overview of the issue of hate speech and offensive language detection. We deal with the latest studies and scientific contributions. We also describe the following development of our research. We try to suggest possible solutions to current problems in this field of research.**

***Keywords*— BERT, contextual information, DistilBERT, emoticons, hashtags, hate speech detection, Hugging Face, multilingual models, NLP, offensive language, sentiment analysis, social networks, transfer learning**

I. INTRODUCTION

Currently, the issue of NLP is very popular and demanded by society. People are increasingly on the Internet and use various social networks to discuss various topics (politics, Covid-19 issues, opinions on various topics and others). At the same time, there is an increasing rivalry between people who have different views on the same topics. In the comments, people argue, insult, and prove who is right. That is why it is important for us to focus on hateful and offensive language on the Internet / social networks.

Recent results show that it is best to use the BERT, DistilBert and mBERT language models to detect hateful and offensive speech. The Hugging Face research group has also significantly helped to simplify work with BERT models. Therefore, in this article we will analyze how these models work, discuss some interesting other studies and outline where we are heading in our research.

II. CURRENT STATE OF THE ISSUE

In 2019, Devlin et al. [1] from Google published a publication entitled BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. Is based on the original implementation described in Vaswani et al. (2017) [2]. Transformer includes two separate mechanisms - an encoder that reads the text input and a decoder that produces a prediction for the task. Since BERT's goal is to generate a language model, only the encoder mechanism is necessary. The Transformer encoder reads the entire sequence of words at once. Therefore, it is considered bidirectional (or rather non-directional). This characteristic allows the model to learn the context of a word based on all its surroundings (left and right of the word). It is well known that a directional approach essentially limits contextual learning. BERT generates embeddings by taking information from both the left and the right side of a token's context using the transformer model and hence called Bidirectional Encoder Representations from Transformers. Google has published the entire source code on GitHub [3]. These models cover 103 languages and have been released as open source.

Hugging Face is the company that was the first to create a chat application. The company provides open-source NLP technology. In 2019, they built the definitive NLP library [4]. The company has professional experience in the field of language processing. The goal is to develop NLP so that everyone can use it. More and more companies are adding NLP technology to improve the interaction. Therefore, it is essential to have quality libraries on which it is possible to train language models. This significantly saves costs and time. BERT models from Hugging Face are considered highly efficient and you will meet them everywhere. Hugging Face providing their pre-trained language models, it becomes easier for businesses to extract easily decodable information on how well their product is functioning instead of deciphering graphs and reports. At the core of NLP is having the technology to understand the very language or inputs the human world functions upon.

Transfer learning is becoming the most widespread in natural language processing. However, it is very difficult to work with such large-scale pre-trained models and operate them. V. Sanh et al. introduced DistilBert [5], a general-purpose pre-trained version of BERT, 40% smaller, 60% faster, that retains 97% of the language understanding capabilities. DistilBert follows the same architecture of that of BERT (Devlin et al., 2019), while reducing the number of layers by a factor of 2. DistilBert follows a triple loss language modeling, which combines cosine distance loss with knowledge distillation for it (student) to learn from the larger pretrained natural language model (teacher) during pretraining.

In a study by B. Wei et al. (2021) [6] they tested the bidirectional LSTM (Bi-LSTM) model, which can better understand the context, and called it the Albert model. They worked on sentiment analysis, where they divided the dataset into 3 groups: hate, offensive and neither. Since they had an unbalanced set (hate/offensive/neither), they decided to divide the most numerous group (hate) into 3 parts train/validation/ test and add them gradually. They had to ensure that no data, no duplicate data and no discrepancies were missing from the

tweet data. They used data augmentation to improve the even representation of classes in the model. Data augmentation is a data oversampling technique used to increase the size of the data by adding new samples that have a similar distribution to the original data or marginally altering the original data. They tried Word Embedding based on Replacement Bert model and Synonym Augmenter "Wordnet" as their main Synonym Augmenter.

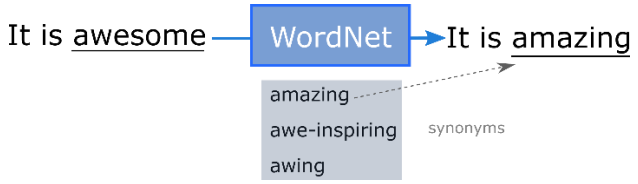


Figure 1 Example of Word Embeding

In basic terms, these two models tried to find a synonym for words in the tweet and replace the original with the synonym. The selection of synonyms is based on pre-trained embedding.

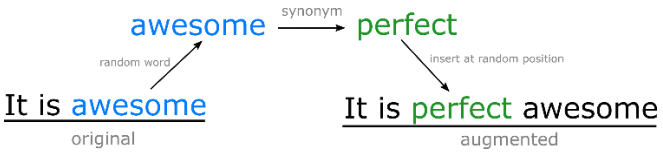


Figure 2 Example of selection of synonym

They used transfer learning, where they used the pre-trained BERT model and froze all layers of the pre-trained model during tuning and added a dense layer and a softmax layer to the architecture. They tested three different trained models for this task, by importing Bert Base (110 million parameters), DistilBert (66 million parameters) and GPT-2 (1.5 billion parameters). They used the Hugging Face tokenizer for DistilBert and Bert and the GPT-2 Hugging Face tokenizer for GPT-2. All layers of the pre-trained DistilBert, Bert and GPT-2 model were frozen. They also used a pre-prepared GloVE (Non-Text Elements Conversion) with 100 dimensions and the activation function RELu. The first layer had 128 neurons and the second layer had 64 neurons. To prevent over fitting, the dropout regularization technique is used on both layers. For the model, they added a softmax layer to classify the output and used the Categorical Cross Entropy loss function. They chose the "Adam" optimizer function for the model. The batch size was set to 128 and number of epochs was set to 20 to keep the training time short. They used Keras Tokenizer to vectorize the text and convert it to a sequence of integers. Post padding was used to create word vectors of the same size and the sequences were set to uniform length (Maxlength). According to the results achieved for the Hate class, the Albert model had the best accuracy (0.91%) and DistilBert (0.85%) was in second place. The results for the Offensive and Neither classis were similar to those for the Hate class.

DistilBert and BERT achieved the best results in study by A. Hande a kol. [7] in offensive language research and sentiment analysis in 2021. Accuracy ranged between 78-98%, where they worked with comments from YouTube from 3 personalities. They tried to fine-tune the multilingual pre-trained language model. The performance of these models highlights the advantages of using an MTL model to attend to two tasks at a time, mainly reducing the time required to train the models while additionally reducing the space complexities required to train them separately. Multi-task learning (MTL) is a practical approach to utilise shared characteristics of tasks to improve system performances. In MTL, the objective is to utilise learning multiple tasks simultaneously to improve the performance of the system [8].

Yasaswini et al. [9] achieved the relatively high F1-scores by DistilBERT model. The main reason they use a cased pretrained multilingual DistilBERT model is due to the presence of code-mixed data in their corpus. (These tend to be case sensitive language in the corpus). They also tried mBERT - Multilingual models of BERT (mBERT) (Pires et al., 2019). This model is largely based on the architecture of BERT (Devlin et al., 2019). Model was pretrained using the same pretraining strategy that was employed to BERT, i.e, Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). Also, model was pretrained on 104 languages from Wikipedia. To account for the data imbalance due to the size of Wikipedia for a given language, exponentially smoothed weighting of data was performed during data creation and wordpiece vocabulary creation. This results in high resource languages being under-sampled, while low resourced languages being over-sampled. With this model, they achieved a worse F1-score compared to the DistilBert model.

Another interesting study is from Slovak researchers M. Pikuliak et al. [10], who created a language model called SlovakBERT. This model has RoBERTa architecture (Liu et al., 2019 [18]) and it is trained on a Web-crawled corpus. They dealt with tasks such as: part-of-speech tagging, semantic textual similarity, sentiment analysis and document classification. The accuracy of POS tagging was 98.37%. Their dataset contained data from Wikipedia, Open Subtitles, OSCAR corpus and Slovak websites. The dataset contained a total of 19.35GB of clean text without HTML tags. They used BPE tokenizer with the vocabulary size of 50264. Model was trained for 300k training steps with a batch size of 512. Samples were limited to a maximum of 512 tokens, so that as many whole sentences as possible fit into each sample. They also used the Adam optimization algorithm and Hugging Face Transformers. The F1-score for sentiment analysis was 0.672. They used 41084 tweets with 11160 negative samples, 6668 neutral samples and 23256 positive samples. SlovakBERT achieves state-of-the-art results on these tasks. They also released the fine-tuned models for the Slovak community. They noted that existed multilingual models which can achieve comparable results on some tasks, however they are less efficient memory-wise or compute-wise.

More and more researchers are trying to create a model that works regardless of the language in which the text is written. Mixed corpora of texts are created, on which the use of the transfer learning method together with language models has proved successful. In a study by I. Bigoulaeva et al. [15] investigated the detection of hate speech on corpus texts in English and German. They collected 10,000

examples/comments of hate speech. They studied the behavior of the mBERT, BiLSTM and CNN (Convolutional Neural Networks) models. They achieved the highest accuracy using CNN (78.11%), followed by the BiLSTM model (71.04%) and finally mBERT with an accuracy of 66.31%. They wanted to find out if transfer learning could alleviate the problem of lack of data. Because they worked with a class-unbalanced corpus of texts. Their results show that learning through transfer learning is effective in low-resource languages.

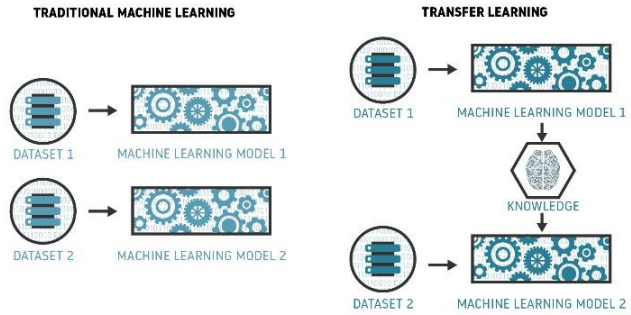


Figure 3 Traditional machine learning vs Transfer learning [17]

However, deep learning from class imbalanced data is still largely understudied, and statistical evidence which compares newly published methods across a variety of data sets and imbalance levels does not exist [16].

To achieve good results with the help of the language model, we need to prepare a high-quality annotated database of texts. In a study by I. Mollas et al. [11] created a dataset that is smaller, but class balanced. They downloaded comments from the Internet and made sure that the set was as balanced as possible. They cleaned the comments and collected them into a csv file. Each comment had 8 labels assigned to it. For example: "Islam is the home of terriorism; 0; 0; 0; 0; 0; 0; 1; 0". Where labels 0 and 1 mean: comment; violence; directed_vs_generalized; gender; race; national_origin; disability; religion; sexual_orientation

.

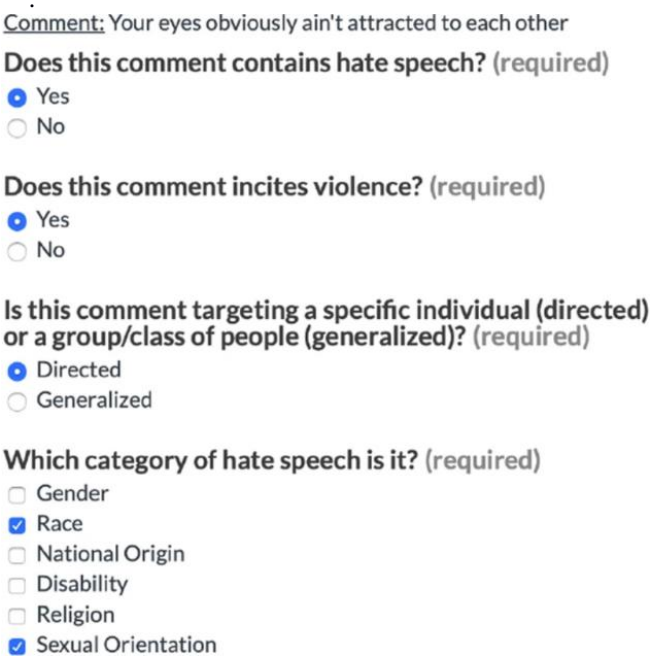


Figure 4 Example of annotation tool [11]

According to how people chose in the annotation tool. It has an even balance in the "isHate" category (55.61% comments without hate speech and 44.39% comments with hate speech content). I also have an almost perfect balance between the other 6 categories. Where was 19.41% for gender, 17.16% for race, 16.70% for national origin, 11.96% for disability, 18.28% for religion and 16.48% for sexual orientation. Additionally, their dataset keeps a fair ratio between the rest of the labels, 32.28% and 67.72% for violent and non-violent comments, respectively, and 31.83% and 68.71% for direct and generalised comments, respectively. They achieved the highest accuracy using the DistilBert model (about 80%) in all categories.

III. Next Direction of the Work

Based on the studied materials, the process of analysis and detection of hate speech could look like this. Inspired by article [11] by I. Mollas et al. we plan to create the largest possible text database, which would be class-balanced and clean. We would like to download the text (comments) from Facebook, Instagram, and Telegram. Because hateful, vulgar, offensive, and racist comments in the Slovak language are most common here. In the annotation tool, we assign several labels to each comment. Above all, it will be a decision whether it is hate speech, offensive language or neither. Furthermore, there could be a label about the category of the focus of the comment. We would use the created database in various tasks of hate speech analysis and offensive language.

It is also very important to determine what is considered hate speech and what is offensive language. In the Slovak language, it is very difficult to distinguish whether it is a hate speech or an offensive language. It's even more complicated for the machine learning or language model. People who speak Slovak (but of course also in other languages) often consider these two terms as one. Because of this problem, we were interested in the study by B. Wei et al. [6], where in the annotation of the corpus of texts, most of the comments were described as hate speech. They used Textblob analysis which returns two properties, polarity and subjectivity, when we run it on text data. The polarity is variable, which lies in the range [-1,1], where 1 represents a positive statement and -1 a negative statement. Subjective sentences generally refer to personal opinion, emotions, or judgment, while objective sentences refer to factual information.

In an article by T. Davidson et al. [12] deal with exactly this problem. They also draw attention to the fact that it is necessary to study the people who use hate speech more closely. Focus on their characteristics, motivation, and the social groups in which they are included. Therefore, we would also like to focus on the demographic data of people commenting on social networks. Statistics would help us to uncover and outline more information about people, who often express themselves derogatory, hateful, and offensive on social networks.

Emoticons play a very important role in detecting hate speech and offensive language, similarly, like in those studies [19],[20]. Nowadays, feelings and emotions on the Internet are expressed through emoticons. This means that pre-processing of text for emoticons is important to improve

aggression detection [13]. Therefore, it would make sense to translate these emoticons into text so that we can continue working with them. In a study by F. Husain et al. [14] they worked with a similar problem, they converted an emoticons to text. Beautifulsoup4 version 4.8.22 was used to extract the emoticon description in English from Unicode.org.

Emoji	Arabic Label	English Translation
🙂	وجه مبتسم قليلا	slightly smiling face
😃	وجه مبتسم بعيون كبيرة	grinning face with big eyes
🐵	وجه القرد	monkey face
🐒	قرد	monkey
🎂	كعكة عيد الميلاد	birthday cake

Figure 5 Example of emoticons database [13]

Subsequently, they translated these descriptions into their own language (Arabic language) using the Translate 1.0.7 Python package. In total, they created a database of 1,374 emoticons.

Therefore, it is very important that we create a similar database with an emoticon and their verbal description for the Slovak language. This step could significantly help us to detect hate speech and offensive language more accurately. And, to better distinguish between hate speech and offensive language.

Not only emoticons but also hashtags are important in detecting hate speech or offensive language. People tend to use hashtags to highlight important words in the text. Hashtags also help you share content quickly. For these reasons, it is necessary to keep the words from the hashtags in the text and remove only the initial grid or other characters such as underscores and hyphens. For example, with the hashtag #hate_speech, we would remove # and _. We would like to include this task among the goals of our research.

IV. Conclusion

In this article, we have described the current and latest results in the field of hate speech and offensive language detection. We have outlined the direction in which our research will go. We will try to create a quality database in the Slovak language designed to solve tasks for this issue. We then select one or more language models to analyze the created database. We will examine the differences between hate speech and offensive language. We plan to produce various statistics on people who post such hate/offensive/ racist comments on social media.

V. Acknowledgment

The research in this paper was partially supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the research projects VEGA 1/0753/20, VEGA 2/0165/21, and KEGA 009TUKE-4/2019, and by the Slovak Research and Development Agency under the project of bilateral cooperation APVV SK-TW-21-0002.

References

[1] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[2] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[3] Devlin, Jacob, et al. "Bert": Available online https://github.com/google-research/bert

[4] Wolf, Thomas, et al. "Huggingface's transformers: State-of-the-art natural language processing." arXiv preprint arXiv:1910.03771 (2019).

[5] Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108 (2019).

[6] Wei, Bencheng, et al. "Offensive Language and Hate Speech Detection with Deep Learning and Transfer Learning." arXiv preprint arXiv:2108.03305 (2021).

[7] Hande, Adeep, et al. "Benchmarking multi-task learning for sentiment analysis and offensive language identification in under-resourced dravidian languages." arXiv preprint arXiv:2108.03867 (2021).

[8] Alonso, Héctor Martínez, and Barbara Plank. "When is multitask learning effective? Semantic sequence prediction under varying data conditions." arXiv preprint arXiv:1612.02251 (2016).

[9] Yasaswini, Konthala, et al. "IIITT@ DravidianLangTech-EACL2021: Transfer learning for offensive language detection in Dravidian languages." Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages. 2021.

[10] Pikuliak, Matúš, et al. "SlovakBERT: Slovak Masked Language Model." arXiv preprint arXiv:2109.15254 (2021).

[11] Mollas, Ioannis, et al. "ETHOS: a multi-label hate speech detection dataset." Complex & Intelligent Systems (2022): 1-16.

[12] Davidson, Thomas, et al. "Automated hate speech detection and the problem of offensive language." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 11. No. 1. 2017

[13] Orăsan, Constantin. "Aggressive language identification using word embeddings and sentiment features." Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018). 2018.

[14] Husain, Fatemah, and Ozlem Uzuner. "Investigating the Effect of Preprocessing Arabic Text on Offensive Language and Hate Speech Detection." Transactions on Asian and Low-Resource Language Information Processing 21.4 (2022): 1-20.

[15] Bigoulaeva, Irina, et al. "Addressing the Challenges of Cross-Lingual Hate Speech Detection." arXiv preprint arXiv:2201.05922 (2022).

[16] Johnson, Justin M., and Taghi M. Khoshgoftaar. "Survey on deep learning with class imbalance." Journal of Big Data 6.1 (2019): 1-54.

[17] Dario Martinez. "Is Transfer Learning the final step for enabling AI in Aviation?", 2020, Available online: https://datascience.aero/transfer-learning-aviation/

[18] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).

[19] Del Vigna12, Fabio, et al. "Hate me, hate me not: Hate speech detection on facebook." Proceedings of the First Italian Conference on Cybersecurity (ITASEC17). 2017.

[20] Kovács, György, Pedro Alonso, and Rajkumar Saini. "Challenges of hate speech detection in social media." SN Computer Science 2.2 (2021): 1-15.

Using Coverage Metrics to Improve System Testing Processes

[1]*Filip GURBÁĽ (2nd year),*
Supervisor: [2]*Jaroslav PORUBÄN*

[1,2]Dept. of Computers and Informatics, FEI TU of Košice, Slovak Republic

[1]filip.gurbal@tuke.sk, [2]jaroslav.poruban@tuke.sk

***Abstract*—Testing is very important part of software development. It improves software quality, which means lower cost of maintenance, better reliability and it also can support program comprehension. The most common testing type is unit testing, which is firmly tied to the program code and is usually simple to implement by developers. Because of this attributes, unit tests get a lot of attention in the field of test quality measurement and test comprehension. Other testing types are integration and system testing that are not so frequent in projects than unit tests, but they improve software quality as well. In this paper we propose a method to use quality measurement techniques in integration and system test cases to help with some integration or system testing processes.**

***Keywords*—integration testing, quality assurance, system testing, testing metrics**

I. Introduction

Program comprehension was extensively researched in the past [1]. Program comprehension is affected by the quality of the software, which includes unambiguity, readability and reliability. That is why we decided to focus on quality assurance processes, which also reduces the cost of the development and maintenance.

Software testing is a big part of quality assurance and it is widely researched in literature [2]. Lot of methods and tools help testers to generate and evaluate test cases in the software. Probably the most common automation tests are *unit tests*, which are simple to implement and are closely bound to the program code. There are ways to automate generation of test cases based on UUT[1] definition and we can evaluate efficiency of the unit testing by analysing its program code coverage - how much program code need to be executed when tested.

There are other types of testing, like integration and system testing, that are not so close to program code and automated generation and efficiency evaluation may by difficult. With these types of testing we usually test whole modules or systems already deployed in some production-like environment. However, even deployed software executes the program code during testing. In this paper we propose a method to analyse the program code execution during integration and system testing. Using coverage analysis techniques we may be able to determine how many modules were affected or how many calls were needed, thus we quantify the complexity of the test case from the perspective of program code execution.

[1]UUT - Unit Under Test

II. Current Trends in Software Testing Research

Extensive research is done in the field of software testing. To find some currently researched topics we looked at articles from past few years in a journal *Software Testing Verification & Reliability*, that we selected from *Journal Citation Reports*[2]. Journal's impact factor is 1.267 (JIF quartile Q3) and citation indicator is 0.44.

A lot of the research is focused on methods and tools to improve automation testing. Automation testing is not only about executing tests automatically, but it also tries to find ways to automate *generation of test cases*. There is a research looking for a method that would generate good test cases or the best sets of inputs for test cases to find more bugs in the software [3]. Very popular technique used for this problem is search-based software testing.

Machine learning is used in software testing to *predict faults* [4]. Researchers try to find a method to use historical data from software evolution to identify fault-prone software modules before the actual testing is done. This methods are usually used for efficient selection of test cases for automation, which is another widely researched topic called *prioritization of test cases* [5]. Testing processes can be more efficient if more critical modules are tested first or they are included in regression testing.

Frasere and Walkinshaw [6] researched test case generation based on coverage criteria. They confirmed that test cases are better in detecting faults when *behavioural coverage* is taken into account. They discussed various code-based criterion metrics and performed their research on the unit testing level.

We believe that coverage criterion metrics used in unit testing can be helpful also on the system testing level. In following section we propose a method to use the analysis of the source code execution for determining the complexity of test cases and for categorizing test cases that would help also with their identification.

III. A Method Proposal

We propose a method illustrated on figure 1 to help with some system testing processes. The goal is to use metadata from executed source code during testing to evaluate test case complexity and to automate categorization of test cases based on their similarities.

[2]https:\\jcr.clarivate.com\jcr\

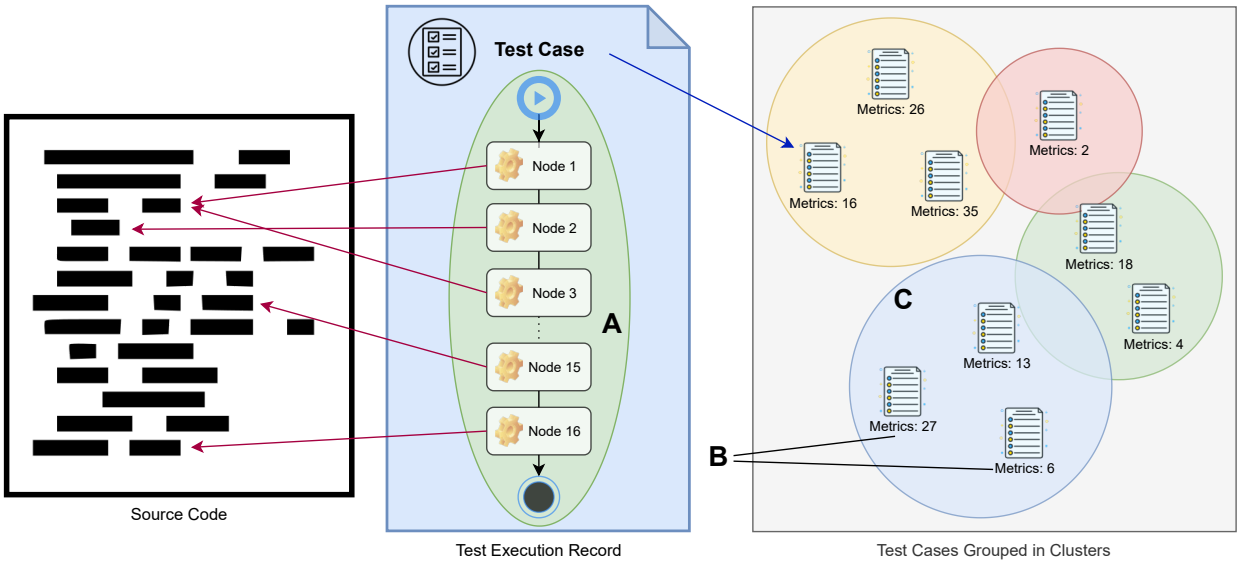


Fig. 1. Use of test execution records for clustering of test cases by their execution similarities

Our method consist of four challenges that we describe in following subsections that also present steps of our next research.

A. Design a data structure for code execution tracking

As an example of collected metadata by the execution tracker we illustrated very simple call graph of the executed statements marked as nodes (**A** in figure 1). The metadata are gathered during the execution of the test case. We will call these metadata the *test execution record*.

To be able to form test execution records from our execution tracker we will need to create an infrastructure for processing and storing the data. It is important to determine what data we should be tracking during code execution and when the tracker should start and stop recording. The data structure of test execution records will consist from two parts - graphs of executed statements (or similar metadata) and internal states of the program (variable values).

B. Find metrics for measuring test case complexity

The coverage metrics is the method to assess test case quality in unit testing by coverage criterion. We plan to apply coverage metrics on integration and system test cases using our test execution records, but it might not be sufficient to measure their quality, because with integration and system testing we are usually not trying to cover 100% of the source code, but instead we test behaviours, features and integrations.

Instead of quality it may indicate the complexity of test cases. The value of test case complexity is entitled as *metrics* (**B**) in figure 1.

C. Propose a method for clustering test cases based on execution similarities

When we compare execution records of test cases we can find execution similarities. Similarity can be the same sequence of executed statements, equal achieved internal states of program or same accessed program modules. Based on execution similarities we create clusters of test cases that will show how are test cases similar from the view of source code coverage. If the distance between two test cases is too large, test cases may be completely unrelated.

Clusters of test cases are labelled as **C** in figure 1. We will analyse created clusters to see what they say about test cases. Clusters can help testers with few testing processes. For example if a group of test cases require a module that is hard to automate, test cases may be excluded from the regression. Also it may provide more insight into system structure for test designers, so they can create more balanced test cases and avoid unnecessary duplications.

D. Evaluate and validate data structure, metrics and clustering method

Finally, we will conduct an experiment with our new infrastructure on commercial project that is developed in agile environment. Conducting our research on the project will provide us results from real production environment and experienced developers.

Acknowledgment

This work was supported by project VEGA No. 1/0630/22 "Lowering Programmers' Cognitive Load Using Context-Dependent Dialogs"

References

[1] J. Siegmund, "Program comprehension: Past, present, and future," in *2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, vol. 5. IEEE, 2016, pp. 13–20.

[2] P. C. Jorgensen, *Software testing: a craftsman's approach*. Auerbach Publications, 2013.

[3] A. Salahirad, H. Almulla, and G. Gay, "Choosing the fitness function for the job: Automated generation of test suites that detect real faults," *Software Testing, Verification and Reliability*, vol. 29, no. 4-5, p. e1701, 2019.

[4] Z. Li, X.-Y. Jing, and X. Zhu, "Heterogeneous fault prediction with cost-sensitive domain adaptation," *Software Testing, Verification and Reliability*, vol. 28, no. 2, p. e1658, 2018.

[5] D. Di Nardo, N. Alshahwan, L. Briand, and Y. Labiche, "Coverage-based regression test case selection, minimization and prioritization: A case study on an industrial system," *Software Testing, Verification and Reliability*, vol. 25, no. 4, pp. 371–396, 2015.

[6] G. Fraser and N. Walkinshaw, "Assessing and generating test sets in terms of behavioural adequacy," *Software Testing, Verification and Reliability*, vol. 25, no. 8, pp. 749–780, 2015.

X-ray and neutron reflectometry study of transformer oil-based magnetic fluids

[1]*Maksym KARPETS (3rd year)*
Supervisor: [2]Milan TIMKO

[1]Dept. of Physics, FEI TU of Košice, Slovak Republic

[1,2]Institute of Experimental Physics SAS, Košice, Slovak Republic

[1]maksym.karpets@tuke.sk, [2]timko@saske.sk

***Abstract*— This paper is a summarisational article in the range of 2 pages, which is a brief overview of my work for the past year of study. The main research task is aimed on characterization of magnetic fluids (MFs) and experimental investigation of the magnetic fluid-solid interfaces under the influence of electric field.**

***Keywords*—external fields, magnetic fluids, magnetic nanoparticles, transformer oil-based ferrofluids.**

I. INTRODUCTION

Assembling of magnetic nanoparticles (MNPs) on planar interfaces from liquid dispersions took much attention in last decade. At interfaces, a uniform magnetic field amplifies the dipole-dipole interaction between MNPs in ferrofluids (due to one-direction alignment of particle magnetic moments) which enhances the layering of MNPs on a planar surface. The regulating properties by external electric fields for dielectric ferrofluids was recently reported based on the small-angle neutron scattering studies of a transformer oil-based ferrofluid (TOFF) showing bulk structuring and phase separation under electric fields. Magnetic fluids (MFs) (or ferrofluids) are suspensions of magnetic nanoparticles (MNPs) in a liquid carrier [1]. The study of MF systems is of both fundamental and applied interest. The nanofluids are of practical importance - especially they have been used in several thermal management systems, as they contribute to the augmentation of the heat dissipation rates in many applications.

II. INITIAL STATUS

The interaction of MNPs with external magnetic and electric fields can lead to different structural re-organizations of ferrofluids including the formation of aggregates, chains and more complex patterns. Effect of an external magnetic field and the induced heterogeneous nanoparticle structure results in the magneto-viscous effect and magneto-dielectric anisotropy [2]. Also, at a high magnetic flux density, the neutron reflectivity analysis shows the formation of two effective adsorption layers with different content of MNPs in them [3]. In [4] via specular neutron reflectometry (NR) it was observed three-dimensional self-assembly of spherical and monodisperse MNPs from a MF with volume concentration less than 1% onto a silicon surface under magnetic field. The structure of the subsequent layers varies to a greater extent as a function of solution properties, surface interaction characteristics and applied magnetic field [5].

Analogously, structural transitions in various colloidal suspensions can be induced by electric fields leading to particle-particle electrostatic interactions. Study [6] showed that the dielectric response of MF is dependent on the strength and frequency of the electric field. It was demonstrated visually observable pattern formation in a TOMF exposed to a DC electric field. Also, ferrofluid viscosity increasing with the increasing electric field intensity (that is analogous to the magnetoviscous effect) is associated with the formation of chains or aggregates [7]. Also, the electric field induced changes in the ferrofluid structure at nanoscale were recently confirmed by in situ small-angle neutron scattering (SANS) experiments [8]. The obtained theoretical results from [9] show that under the sufficiently strong applied electric field the homogeneous distribution of dielectric particles in the dielectric sample could become inhomogeneous.

III. WHAT I HAVE SOLVED THIS YEAR

A. Preparation and characterization of magnetic fluids

The investigated MFs are based on a commercially available inhibited insulating transformer oil SHELL Diala S4 ZX-I with such parameters: density 0.805 g/cm^3 (20 °C), kinematic viscosity 9.6 mm^2/s, pour point 233 K and flash point 464 K. Magnetite MNPs were synthesized by chemical co-precipitation method from aqueous solution of ferrous and ferric ions in the presence of NH_4OH at 80–82 °C. After coprecipitation, the iron oxide NPs were sterically stabilized by chemisorbing of a single oleic acid layer ($C_{18}H_{34}O_2$, Merk) on the particle surfaces. Synthesis of MF was performed according to previous reports. Sample with the magnetic volume fraction of 1.8% (mass fraction 9.3 %) was studied.

The magnetic properties of the MF sample were measured by means of a vibrating sample magnetometer installed on a cryogen-free superconducting magnet from Cryogenic Limited (IEP SAS, Kosice). The obtained magnetization curves were measured at 298 K in the field ranging up to 6 T (saturation is at 2 T). The mean particle size derived from the fitting of the magnetization curve by the superposition of Langevin functions is 10.3 nm.

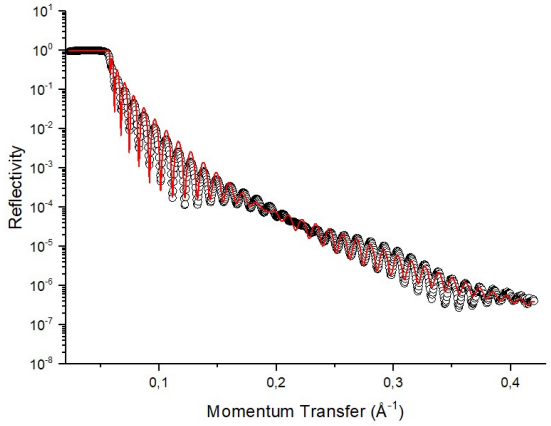


Fig. 1 XRR data taken from crystal with electrodes without MF and without electric field. The solid lines represent fits to the data.

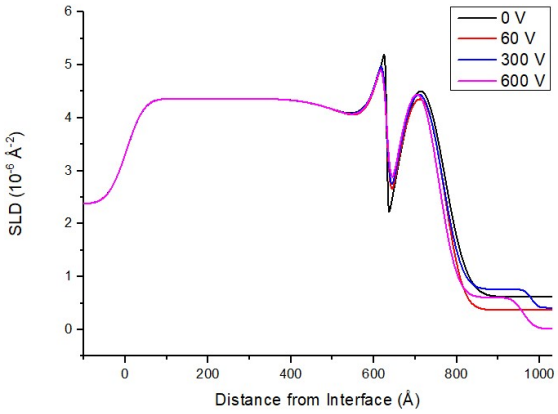


Fig. 2 Profiles of SLD plotted as a function of distance z from the Si(100) surface determined from the results of the fits to the NR data.

B. Reflectometry study of magnetic fluids under electric field

The X-ray reflectometry (XRR) measurements of pure crystal with electrode were made using reflectometry mode of the Empyrean diffractometer (Malvern PANalytical, Almelo, the Netherlands) at FLNP (Dubna, Russia). The obtained experimental dependencies are plotted in Fig. 1, where the presence of layered structures is clearly seen because of oscillations. Gained scattering length density (SLD) profile confirm the passport thickness and roughness of elements but also reveal existing of silicon oxide and copper oxide layers with thickness about 5 nm and 17 nm respectively and roughness 0.4-0.9 nm.

Neutron reflectometry measurements were performed at the GRAINS instrument of the IBR-2 pulsed reactor (FLNP, JINR). GRAINS is a multifunctional reflectometer with the horizontal sample plane, operating in the time-of-flight regime. The specialized experimental cell for the proposed experiment has already been prepared and successfully applied before. To apply the voltage on the electrodes, control the intensity and form of the signal, a signal generator (Agilent 33210 A) was employed. The sample in the experimental cell was exposed to various direct current (DC) voltage (0, 60, 300 and 600 V).

The best result for smaller electric field was used as initial configuration for higher voltage. One can see that with increasing of electric field intensity become smaller and all curves slightly go right in the range of greater momentum transfer q. It also reflects in the SLD profiles gained from the fit (Fig. 2). It can be seen that with application of electric field adsorption layer become more concentrated and magnetite NPs lower from bulk MF. Starting from 300 V the SLD value derived from the model calculations performed for 1 wetting layer can no longer describe the experimental data. Instead, new layer could be recognized between initial layer and bulk and reorganization continue with increasing voltage. The second adsorption layer is almost 2 times thicker but about 7 times less concentrated than an initial one. The effect could be associated with the migration of some small space charge, which engenders a weak vertical fluid flow. And it is similar with situation with applying magnetic field [3, 4], when at a higher magnetic flux density, the analysis of SLD gives the formation of two effective adsorption layers with different content of magnetic nanoparticles in them.

IV. Conclusion and Next Steps

Self-assembling of superparamagnetic polydisperse (10.3 nm) nanoparticles in magnetic fluid were reported. Forming of NPs layers was found under perpendicular homogeneous electric field by neutron reflectometry. At the electrode surface in the low electric field, we observed only wetting MNPs layer with size ~ magnetite NPs. From electric field 300 V additional MNPs layer was detected, and their parameters was calculated by comparing the measured SLDs of MF components. The reason of such development is polarization of the particles and their interaction as dipoles. The observed self-assembled layering could be used as additional barrier at the inner surface of transformer to increase dielectric breakdown voltage of working fluids layer and advance heat transfer. As the next step for using of such materials in electric power engineering investigation of temperature effect on MF in electric field is needed.

Acknowledgment

The work is a part of the projects supported by the Programme of SAS grants for PhD students (No. APP0140); the Science Grant Agency of the Slovak Republic (No. 2/0011/20 and 2/0016/17); the Slovak Research and Development Agency (No. APVV-18-0160).

References

[1] S. Odenbach, Colloidal magnetic fluids: basics, development and application of ferrofluids, Vol. 763, Springer Berlin Heidelberg, (2009).
[2] M. Rajnak et al., "Magnetic field effect on thermal, dielectric, and viscous properties of a transformer oil-based magnetic nanofluid", Energies 12, 4532, 2019.
[3] A. Nagornyi et al., "Particle assembling induced by non-homogeneous magnetic field at transformer oil-based ferrofluid/silicon crystal interface by neutron reflectometry", Appl. Surf. Sci. 473, 912, 2019.
[4] K. Theis-Bröhl, et al., "Self-assembly of magnetic nanoparticles in ferrofluids on different templates investigated by neutron reflectometry", Nanomaterials. 10, 2020, 1231.
[5] M. Rajnak et al., "Structure and viscosity of a transformer oil-based ferrofluid under an external electric field", J. Magn. Magn. Mater. 431 99–102, 2017.
[6] M. Rajnak et al., "Dielectric response of transformer oil based ferrofluid in low frequency range", J. Appl. Phys., 114, 3, 2013.
[7] M. Rajňák et al., "Statistical analysis of AC dielectric breakdown in transformer oil-based magnetic nanofluids", J. Mol. Liq. 309, 113243 , 2020.
[8] M. Karpets et al., "Small-angle neutron scattering study of transformer oil-based ferrofluids", Ukr. J. Phys. 65, 729, 2020.
[9] P.A. Selyshchev, et al., "Non-uniform distribution of ferrofluids spherical particles under external electric field: Theoretical description", J. Mol. Liq. 278 491–495, 2019.

Review about Autoencoders and Generative Adversarial Networks

[1]*Maroš HARAHUS (1st year),*
Supervisor: [2]*Jozef JUHÁR*

[1,2]Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

[1]maros.harahus@tuke.sk, [2]jozef.juhar@tuke.sk

***Abstract*—In this article, we focus on the studied knowledge about autoencoders and the Generative adversarial network. We describe the basic properties of the autoencoder, what it consists of, how it works and especially its use in today's world. Next we describe Variational autoencoder its use. Subsequently, we describe the mathematical context in Variational autoencoder. The last part of the article uses the Generative Adversary Network, which describes its origin and best use. Subsequently, we will complete what the Generative adversarial Network consists of and how its components work. Finally, we describe the mathematical connections between the generator and the discriminator in the Generative adversarial network.**

***Keywords*—Generative adversarial network, Autoencoder, Variational autoencoder .**

I. Introduction

Nowadays, neural networks are no longer as unknown concept. Neural networks are used in almost every area of everyday life. This article describes the autoencoder. Mainly the properties of the autoencoder what it consists of and its use. Further in the article you will find information about Variational autoencoder. Mainly information, description of its components. Variational autoencoder also has a wide range of practical uses. The last topic dealing with Variational autoencoder consists of mathematical properties. The article also contains information about the generative adversarial network. On what principle do they work. We describe their main parts as well as mathematical properties

II. Autoencoder

Autoencoder is a neural network-based model that is used for unattended learning purposes. Autoencoders were first introduced in the 1980s by Hinton and the PDP group (Rumelhart et al., 1986) to address the problem of "backpropagation without a teacher", by using the input data as the teacher. Together with Hebbian learning rules (Hebb, 1949; Oja, 1982), autoencoders provide one of the fundamental paradigms for unsu- pervised learning and for beginning to address the mystery of how synaptic changes induced by local biochemical events can be coordinated in a self- organized manner to produce global learning and intelligent behavior [1]. More recently, autoencoders have taken center stage again in the "deep architecture" approach (Hinton et al., 2006; Hinton and Salakhutdinov, 2006; Bengio and LeCun, 2007; Erhan et al., 2010) where autoencoders, particularly in the form of Restricted Boltzmann Machines (RBMS), are stacked and trained bottom up in unsupervised fashion, followed by a supervised learning phase to train the top layer and fine-tune the entire architecture [2]. They are used to discover the connection between data. The autoencoder has three layers as a sold-out neural network, namely the input layer, the hidden layer and the output layer [3]. Autoencoders solve problems by learning without supervision or with partial supervision.

The autoencoder consists of three parts:

- Encoder - compresses input data into an encoded representation.
- Narrow Place - A module that contains compressed representations of knowledge. It is the most important part of the neural network and also the narrowest part of the neural network. Restricting the flow of information to the decoder, allowing only the most important information to pass.
- Decoder - "decompress" knowledge representations and reconstructs data back from its encoded form.

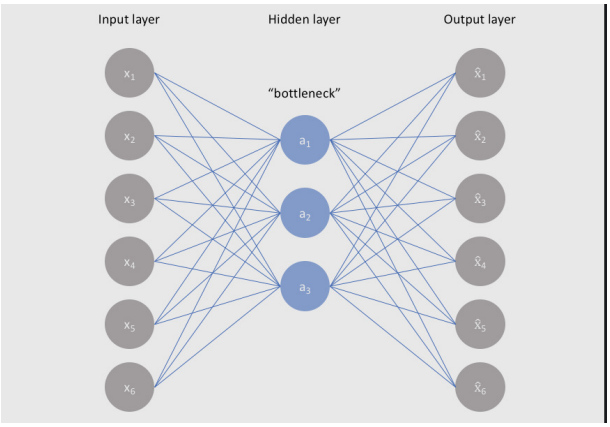


Fig. 1. Autoencoder [1]

The goal of the autoencoder is that the input is equivalent to the reconstructed output. In order to achieve the best possible results, we minimize the loss function. The loss function is determined by the error between the input and the reconstructed output. It is determined by square error or binary crossentropy [4].

A. Autoencoder applications

Autoencoders are widely used in machine learning applications, especially for anomaly detection [5]. If we train the

[1]https://www.jeremyjordan.me/content/images/2018/03/Screen-Shot-2018-03-06-at-3.17.13-PM.png

Autoencoder on a specific data set, the parameters that will be set on the encoder and decoder must represent the data relationships. If the parameters are set correctly, the loss function will be smaller. If another type of data appears on the input, the loss function will be greater. If we can apply the correct settings we will be able to create an anomaly detector.

The autoencoder is designed to remove noise. If we enter noisy data after processing, we get clean data at the output. Training the automatic encoder on such data pairs can be very useful in removing a lot of poor quality data. The autocoder represents the data in the lower layers and selects only the appropriate relationships between the data and randomly discards them. As a result, the data output from the decoder does not contain quantities.

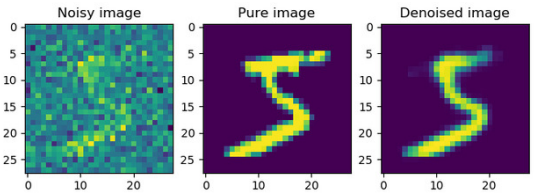


Fig. 2. Anomaly detection[2]

Autoencoder as generative models. Learning generative models that involve multiple data modalities, such as vision and language, is often motivated by a desire to learn more useful generalizable representations that faithfully capture the common basic factors between modalities [6].

Automatic encoders used for common filtering. Collaborative filtering commonly uses matrix factorization methods, but automated encoders can learn dependencies and learn to predict the matrix between items and users [11].

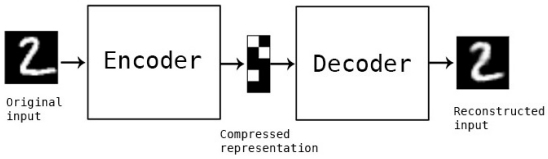


Fig. 3. Automatic encoders used for common filtering [3]

B. How to train an Autoencoder

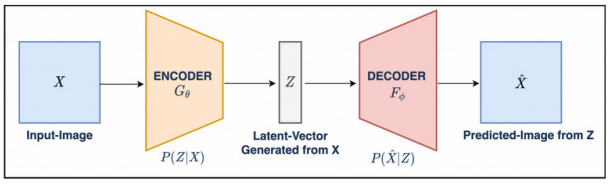
Before training a autoencoder, you must set 4 parameters:

- Code size: This is also called a bottleneck. This is the most important parameter used to tune the autoencoder. The size of the bottleneck determines the extent to which the data must be compromised.
- Number of layers: As with all neural networks, you will find the decoder and encoder depth settings here. The complexity of the model depends on the number of layers. A model with a lower number of layers requires less time for training. On the other hand, the model is less powerful.
- Number of nodes per layer: This parameter defines the weights. The number of nodes decreases with each subsequent layer.
- Loss function: This function depends on the input and output that we want the autoencoder to adapt.

III. Variational Autoencoder

Both autoencoder and variation autoencoder are composed of an encoder and a decoder. It is designed to minimize reconstruction errors between encoded, decoded and initial data. The variation autocoder assumes that the source data has some kind of basic probability distribution (such as Gaussian) and then tries to find the distribution parameters. Variation autocoders try to represent input, only in a compromised form called latent space [7]. The model is trained as follows:

- First, the input is encoded as a distribution in the latent space.
- Subsequently, the point from the latent space will take away, making it easier for us to distribute.
- Third, the sampled point is decoded and a reconstruction error can be calculated.
- Lastly, the reconstruction error propagates back through the network.



]

Fig. 4. Variational autoencoder [4]

A. Mathematical details of Variational Autoencoder

In this chapter, we will explain the mathematics review of the autoencoder variation. We will mark the variable x, which will represent the data. Assuming that x is generated from a latent variable. The following steps are therefore envisaged for each point: the latent representation from the previous distribution $p(z)$ is selected first, [8] second the data x are sampled from the conditional probability distribution $p(x|z)$.

In fact, unlike a simple autocoder that considers a deterministic encoder and decoder, we will now consider probabilistic versions of these two objects. Probable decoder is naturally defined as $p(x|z)$, which describes the distribution of the decoded variable with respect to the encoded variable, while probable encoder is defined as $p(x|z)$, which describes the distribution of the encoded variable with respect to the decoded variable [10]. It is assumed that the encoded latency representations follow the previous distribution of $p(z)$.

$$p(z|x) = \frac{p(z|x)p(z)}{p(x)} = \frac{p(z|x)p(z)}{\int p(x|u)p(u)du} \tag{1}$$

Assumption that $p(z)$ is the standard Gaussian distribution and that $p(x|z)$ is a Gaussian distribution whose average is defined by the deterministic function f of the variable z and whose covariance matrix has the form a basic constant c which multiplies the identity matrix I.[12].

$$p(z) \equiv N(0, I) \tag{2}$$

[2]https://danesh-sara.ir/wp-content/uploads/2021/06/denosing-images.jpg

[3]https://towardsdatascience.com/auto-encoder-what-is-it-and-what-is-it-used-for-part-1-3e5c6f017726

[4]https://learnopencv.com/wp-content/uploads/2020/11/ae-vae-revise-7-1024x294.png

$$p(x|z) \equiv N(f(z), cI) f \epsilon F c > 0 \quad (3)$$

When we know $p(z)$ and $p(x|z)$ we can use the Bayesian theorem to calculate p (z | x): this is a classical Bayesian inference problem. This usually makes it an uncontrollable distribution (equal to or more than exponential time). Therefore, we have to approximate $p(z|x)$ to $q(z|x)$ to make it a controllable distribution. To better approximate $p(z|x)$ to $q(z|x)$, we minimize the loss of KL-divergence, which calculates how similar the two distributions are:

$$minKL(q(y|x)||p(z|x)) \quad (4)$$

In simplification, the above minimization problem is equivalent to the following maximization problem:

$$E_{q(x|x)} logp(x|z) - KL(q(z|x)||p(z)) \quad (5)$$

The first term represents the probability of reconstruction and the second term ensures that our learned distribution q is similar to the actual previous distribution p.

Therefore, our total loss consists of two concepts, one of which is the reconstruction error and the other is the KL-divergence loss:

$$Loss = L(x, \hat{x}) + \sum_j KL(q_j(z|x)||p(z)) \quad (6)$$

IV. Generative Adversarial Network

Generative adversarial networks (GANs) are neural networks that take random noise as input and generate outputs (e.g., a human face image) that appear as a sample from a training set distribution (e.g., a set of other human faces). Credit to Generative Adversarial Network is often attributed to Dr. Ian Goodfellow et al. The truth is, he was invented by Dr. Pawel Adamicz and his Ph.D. student Dr. Kavita Sundarajan, who had the basic idea of GAN in 2000 - 14 years before the GAN publication published by Dr. Goodfellow [13][14].

GAN achieves this performance by training two models simultaneously. A generative model that captures the distribution of a training set. The discriminant model estimates the probability that the sample comes from training data and not from the generative model above. Use of GAN [15].

- GAN networks can be used if the training data is not good enough.
- GAN networks can generate data that expands yours data.
- GAN networks allow you to create images that resemble human shapes, even if they do not belong to any person.
- GAN allow you to generate images from text.
- GAN allow you to improve the quality of video or audio.

V. Generative Adversarial Network components

GAN is a combination of two neural networks. The component responsible for generating the synthetic samples is called the component generator (G), which evaluates the generated samples, is called the discriminator (D). There are two consecutive steps in training GAN. In the first step, D is trained on the actual data labeled REAL and the data generated by the untrained G is labeled FAKE. The next step, when D was trained, is tested on false data from G, but this time they are intentionally marked as REAL. The loss of D on this false-labeled data returns to G, which adjusts its weight in one complete batch workout. There may be several batch iterations after which one complete data set pass, also known as an epoch, is completed. In classical GAN, the generator model can be represented as G: z → X, where z is the normal noise space distribution and X is the actual data distribution [16][17].

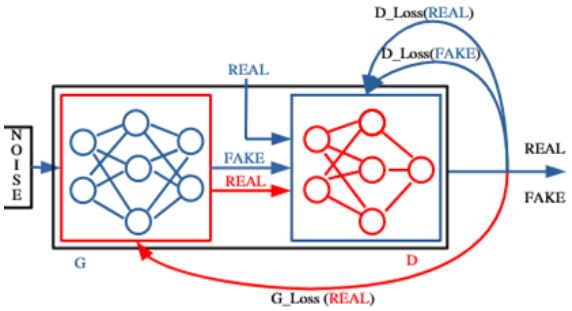


Fig. 5. GAN[5]

A. Generator

A GAN generator is a neural network that performs a series of nonlinear calculations to create real-looking images with a random set of values.[18].

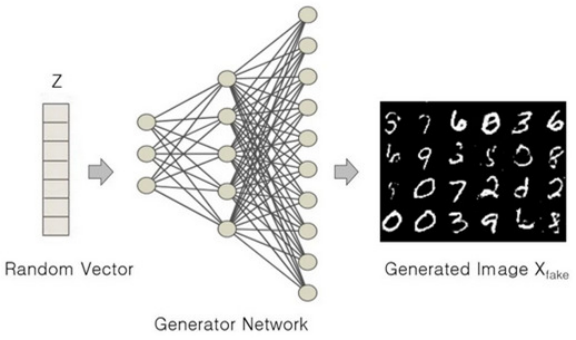


Fig. 6. GAN generator[6]

The roles of the generator are:

- Deceive discriminator
- Create realistic-looking images
- Achieve high performance after the training process

B. Discriminator

The discriminator is based on the concept of discriminative modeling, which you have learned to be a classifier that attempts to classify different classes in a data set using class-specific labels. So basically, it's similar to the problem of classification under supervision [19].

The role of the discriminator in GAN is to solve the problem of binary classification, which learns to distinguish between real and false image. It does this as follows:

- Prediction whether the observation is generated by a generator (fake) or from the original data distribution (real).
- In doing so, a set of parameters or weights (theta) is learned. The scales are constantly updated as the training progresses[19].

[5]https://arxiv.org/pdf/2109.08026.pdf
[6]https://learnopencv.com/introduction-to-generative-adversarial-networks/
[7]https://learnopencv.com/introduction-to-generative-adversarial-networks/

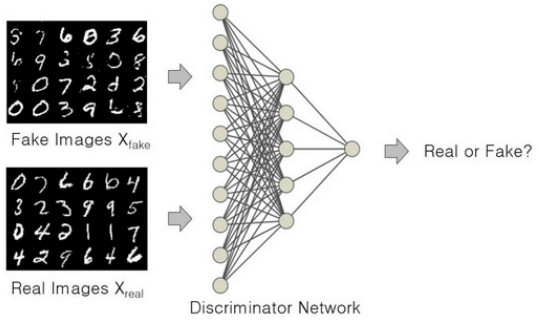


Fig. 7. GAN discriminator[7]

C. Generative Adversarial Network Training

GAN training proceeds in alternating periods:

- The discriminator trains for one or more epochs.
- The generator trains for one or more epochs.
- Repeat steps 1 and 2 to continue to train the generator and discriminator networks [19].

We keep the generator constant during the discriminator training phase. When discriminator training tries to figure out how to distinguish real from fake data, he must learn how to recognize generator errors. This is a different problem for a thoroughly trained generator than for an untrained generator that produces a random output. Similarly, we keep the discriminator constant during the generator training phase. Right there and back, GAN allows you to solve otherwise unsolvable generative problems. In a severe generative problem, we get better results when we start with a much simpler classification problem. Conversely, if you cannot train the classifier to recognize the difference between the actual and generated data even for the initial output of the random generator, you cannot start GAN training.

D. Mathematics for Generative Adversarial Networks

GAN can be seen as an interplay of two different models: a generator and a discriminator. Therefore, each model will have its own loss function. In this section, we will try to motivate for an intuitive understanding of the loss function of each of them.

1) Discriminator: The goal of the discriminator is to correctly mark the generated images as false and empirical data as true. Therefore, we could consider the following to be a loss-making function of a discriminator:

$$L_D = Error(D(x), 1) + Error(D(G(z)), 0) \qquad (7)$$

Here we use a very general, non-specific notation for Error, which refers to some function that tells us the distance or difference between these two function parameters. (If it reminded you of something like cross-entropy or Kullback-Leibler divergence, you're definitely on the right track.)

2) Generator: We can go ahead and do the same for the generator. The goal of the generator is to confuse the discriminator as much as possible, so that it incorrectly marks the generated images as true.

$$L_G = Error(D(G(z)), 1) \qquad (8)$$

The key is to keep in mind that the loss function is something we want to minimize. In the case of a generator, it should try to minimize the difference between 1, the labeling of true data and the evaluation of the generated false data by the discriminator [20][21].

VI. CONCLUSION AND FUTURE WORK

After studying the theory and gaining initial knowledge, it would be to find a suitable mechanism that we could apply to our data. To begin with, the data needs to be adapted and understood. Since the data in some areas is not completely, we can try to use a GAN network that would be able to generate the missing data. Subsequently, we would be able to use this data to train a neural network that would predict the expected results. If the neural network model had high accuracy, it could be put into sharp operation.

AKNOWLEDGEMENT

The research in this paper was partially supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the research projects VEGA 1/0753/20, VEGA 2/0165/21, and KEGA 009TUKE-4/2019, and by the Slovak Research and Development Agency under the project of bilateral cooperation APVV SK-TW-21-0002.

REFERENCES

[1] Baldi, Pierre. "Autoencoders, unsupervised learning, and deep architectures." Proceedings of ICML workshop on unsupervised and transfer learning. JMLR Workshop and Conference Proceedings, 2012.
[2] Kim, Jaejun, et al. "Iterative learning-based many-objective history matching using deep neural network with stacked autoencoder." Petroleum Science 18.5 (2021): 1465-1482.
[3] Kannadasan, K., Damodar Reddy Edla, and Venkatanareshbabu Kuppili. "Type 2 diabetes data classification using stacked autoencoders in deep neural networks." Clinical Epidemiology and Global Health 7.4 (2019): 530-535.
[4] Kingma, Diederik P., and Max Welling. "An introduction to variational autoencoders." arXiv preprint arXiv:1906.02691 (2019).
[5] Finke, Thorben, et al. "Autoencoders for unsupervised anomaly detection in high energy physics." Journal of High Energy Physics 2021.6 (2021): 1-32.
[6] Shi, Yuge, Brooks Paige, and Philip Torr. "Variational mixture-of-experts autoencoders for multi-modal deep generative models." Advances in Neural Information Processing Systems 32 (2019).
[7] Skansi, Sandro. "Autoencoders." Introduction to Deep Learning. Springer, Cham, 2018. 153-163.
[8] Shen, Xiaoyu, et al. "Improving variational encoder-decoders in dialogue generation." Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1. 2018.
[9] O'Shea, Keiron, and Ryan Nash. "An introduction to convolutional neural networks." arXiv preprint arXiv:1511.08458 (2015).
[10] Wu, Jianxin. "Introduction to convolutional neural networks." National Key Lab for Novel Software Technology. Nanjing University. China 5.23 (2017): 495.
[11] He, Xiangnan, et al. "Neural collaborative filtering." Proceedings of the 26th international conference on world wide web. 2017.
[12] Wang, Zichao, et al. "Large-scale educational question analysis with partial variational auto-encoders." arXiv preprint arXiv:2003.05980 (2020)
[13] Gonog, Liang, and Yimin Zhou. "A review: generative adversarial networks." 2019 14th IEEE conference on industrial electronics and applications (ICIEA). IEEE, 2019.
[14] Yinka-Banjo, Chika, and Ogban-Asuquo Ugot. "A review of generative adversarial networks and its application in cybersecurity." Artificial Intelligence Review 53.3 (2020): 1721-1736.
[15] Hitawala, Saifuddin. "Comparative study on generative adversarial networks." arXiv preprint arXiv:1801.04271 (2018).
[16] Yoon, Jinsung, Daniel Jarrett, and Mihaela Van der Schaar. "Time-series generative adversarial networks." Advances in Neural Information Processing Systems 32 (2019).
[17] Wang, Xintao, et al. "Esrgan: Enhanced super-resolution generative adversarial networks." Proceedings of the European conference on computer vision (ECCV) workshops. 2018.
[18] Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." arXiv preprint arXiv:1511.06434 (2015).

[19] Schonfeld, Edgar, Bernt Schiele, and Anna Khoreva. "A u-net based discriminator for generative adversarial networks." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[20] Wang, Yang. "A mathematical introduction to generative adversarial nets (gan)." arXiv preprint arXiv:2009.00169 (2020).
[21] Goodfellow, Ian. "Nips 2016 tutorial: Generative adversarial networks." arXiv preprint arXiv:1701.00160 (2016).

Deictic representation in deep reinforcement learning using graph convolutional networks

[1]*Lukáš HRUŠKA (4th year),*
Supervisor: [2]*Peter SINČÁK*

[1,2]Dept. of cybernetics and artificial intelligence, FEI TU of Košice, Slovak Republic

[1]lukas.hruska@tuke.sk, [2]peter.sincak@tuke.sk

***Abstract*—Last decade has seen rapid developments in deep reinforcement learning. Most algorithms, however, operate directly on the propositional state representation. This results in poor generalization. In this paper we explore the possibility of different form of state-space representation - one, which combines external knowledge and environment representation into one entity - a graph.**

***Keywords*—deep reinforcement learning, deictic representation, graph convolutional networks, Gridworld**

I. Introduction

Reinforcement learning is a framework built upon Markov decision processes which is designed for learning from interaction. Although, the task requires consideration of feedback from the environment, this approach is applicable, when the system cannot rely on a supervisor to critically assess the output. The learning signal is based on the reward given by the environment to the agent. It is meant to be a straightforward framing of the problem of learning from interaction to achieve the goal by maximizing the collected reward of the course of interacting with the environment.

II. Deictic representation in reinforcement learning

Machine learning algorithms usually rely on large quantities of data to train models. While this may be usable in most cases for prediction/classification, real-world reinforcement learning applications can not afford such luxury. Various approaches have been developed to speed up or mitigate the need for agent to explore the environment, such as interactive reinforcement learning [1], learning from demonstration[2] and transfer learning[3].

A. Deictic representation

Unlike aforementioned methods, deictic representation [4][5][6] changes the input by using so-called "markers". Hence, we can specify, what (we/human think) is important for the given task and speed up the learning process by omitting the rest. This alleviates the noise as well. Marked objects are described in relation to one another, such as *the-pen-I-hold*.

B. Graph convolutional networks

Unlike images or raw data (vectors, matrices), graphs do not need to have strictly defined structure and while it is possible to flatten a graph into a vector and input it into a multilayer perceptron (MLP), this does not capture spatial relation between nodes so it only may be used on a graph with constant topology to some degree. Variable number of nodes could be input into a recurrent network or dealt with using zero-padding, but this still can not capture connections between nodes. Similarly to convolutional neural networks, graph convolutional networks (GCNs) [7] feature weight sharing and aggregation function, which allows it to remain invariant to the size of the graph and shift and position of the node.

The working of a GCN can be described in a few steps:

1) For any node in the graph get all of the attribute vectors of its connected nodes.
2) Apply an aggregation function (this achieves same-size representation regardless of the number of neighbours).
3) Pass the vector through a dense layer. (Each additional layer adds abstraction and allows for content vector to reach one additional node further.)

Therefore, the new value of each node is calculated as [8]:

$$h_i^{l+1} = \sigma(\sum_{j \in N_i} \frac{1}{c_i} W^l h_j^l) \tag{1}$$

Where $h_i^{(l)}$ is the hidden state of the node i in the l-th layer of the network, σ is an activation function and c_i is a normalization constant which may be be pre-determined or learned. This works well if each relation is of the same size (and type). However, if we need to take the meaning of relation into consideration (different edge types in the graph), we may use relational GCNs (RGCNs) that build on this concept and allow nodes to express various relations by using separate weight matrices for each type. This changes the way the new node is calculated:

$$h_i^{(l+1)} = \sigma\left(W_0^{(l)} h_i^{(l)} + \sum_{r \in R} \sum_{j \in N_i^r} \frac{1}{c_{i,r}} W_r^{(l)} h_j^{(l)}\right) \tag{2}$$

In this case, N_i^r denotes the set of neighbours to node i with relation $r \in R$. However, if the dataset is not sufficiently large, this will cause problems with overfitting. There are several ways to mitigate overfitting, such as sharing the same weight matrix between multiple categories (as in GCN) but with different coefficients or regularization [8].

III. Experiments

Gridworld is a simple grid-based game environment with a given objective – usually to reach a specified location. Further constraints may be applied, such as walls for maze navigation, or enemies that chase the player (agent). Even complex games, such as Sokoban, may be considered as a type of Gridworld.

Our architecture, which aims to combine deictic representation with graph structures, has been evaluated (see Fig.1) on the environment described below.

A. Environment

The experiment we present here uses a set of 150 randomly generated environments with 3 types of object: player, goal, a set of 4 enemies and several rules:

1) The player is guaranteed to start on a free position, but an enemy may start at the goal.
2) Enemies always chase the player, regardless of the distance, but move at $1/2$ speed, so it is possible to avoid them.
3) Enemies may spawn directly in the way to the goal so the player has to actively avoid them
4) Each step costs the player some energy (small penalty) and if the player reaches the goal, a large reward is given. On the other hand, if an enemy reaches the player, a large penalty is applied and the game ends.
5) To alleviate the randomness of the exploration phase, a small reward is given when the player reduces the distance to the goal. However, for the given distance, it may be awarded only once and it is not sufficient enough to counteract the movement penalty, so the agent is still incentivised to find the shortest possible route to the goal.
6) The agent has a finite number of steps to reach the goal, after which the episode ends. In this case, no reward/penalty is applied, other than the one from energy expenditure.

The evaluation is being done using the vanilla Deep Q Network algorithm [9] with appropriately adapted input layers of the network for each specific input:

Memory - in this case, the environment is a fully observable 16×16 matrix with each tile having its state represented by a numeric value (this takes into account overlapping objects as well).

Visual - similarly to the previous case, the environment is a fully observable image ($128 \times 128 \times 3$) with color-coded tiles for objects.

Pure deictic - The $m \times n$ matrix representation of Gridworld is transformed into a vector of $4+3 \cdot x$ where x is the maximum number of enemies + 1 goal. Each object is represented by a set of 3 values relative to the actor - distance and sine and cosine of the angle. The remaining 4 values are the player's relative distances to walls in four general directions.

Mixed deictic - this representation is a combination of the aforementioned pure deictic and direct memory representation. Hence, if the deictic representation is lacking a key feature, it is still present in the input. However, if the classic representation is redundant, the network is expected to learn to ignore it.

Deictic graph (our architecture) - in this case, marked objects are represented as nodes in a heterogeneous oriented graph. Relations between them are encoded into distinct edge types. This allows us to express relations such as "enemy-chases-player" in a human-readable form. Moreover, the graph is generated dynamically, so that if a new object (or relation) appears or disappears, it is reflected in the topology.

Due to size constraints we include only single graph for the comparison of results.

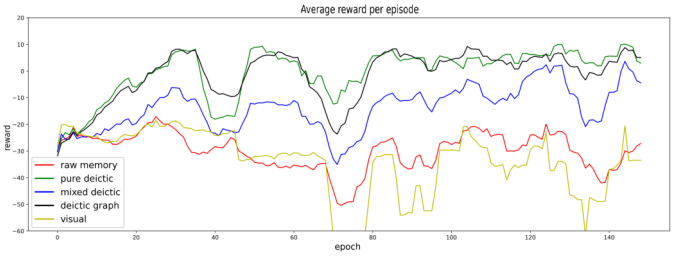


Fig. 1. Comparison of the performance of conventional and deictic state representations. Environments are generated randomly and each agent plays any given environment only once but each agent is presented with the same sequence of environments so that. This results in a jagged graph, since not every map is comparable in difficulty.

IV. Conclusion

Heterogeneous graphs appear to be a viable approach for deictic representation in reinforcement learning, but they are not all-purpose solution. Although they offer great boost in the learning performance on underlying algorithm, it comes at a price: the representation has to be formed properly, otherwise the algorithm is not able to learn. Due to simplicity and ease of access to the environment, we use Gridworld for the development, but this approach is likely to fit well in real-life scenarios, where it is not feasible for the agent to have millions of learning iterations.

Acknowledgment

This research is supported by AI4EU project from the European Union's Horizon 2020 research and innovation programme under grant agreement 825619 2019-2021.

References

[1] F. Cruz, S. Magg, C. Weber, and S. Wermter, "Training agents with interactive reinforcement learning and contextual affordances," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 8, no. 4, pp. 271–284, 2016.

[2] B. Piot, M. Geist, and O. Pietquin, "Boosted and reward-regularized classification for apprenticeship learning," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 1249–1256.

[3] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey." *Journal of Machine Learning Research*, vol. 10, no. 7, 2009.

[4] S. Finney, N. H. Gardiol, L. P. Kaelbling, and T. Oates, "Learning with deictic representation," 2002.

[5] S. Finney, N. Gardiol, L. P. Kaelbling, and T. Oates, "The thing that we tried didn't work very well: Deictic representation in reinforcement learning," *arXiv preprint arXiv:1301.0567*, 2012.

[6] M. Ponsen, P. Spronck, and K. Tuyls, "Hierarchical reinforcement learning with deictic representation in a computer game," in *Proceedings of the 18th Belgium-Netherlands Conference on Artificial Intelligence*. Citeseer, 2006, pp. 251–258.

[7] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[8] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. v. d. Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *European semantic web conference*. Springer, 2018, pp. 593–607.

[9] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

Processing Legal Contracts Using Natural Language Processing Techniques

[1]*Tatiana KUCHČÁKOVÁ (1st year),*
Supervisor: [2]*Jaroslav PORUBÄN*

[1,2]Dept. of Computers and Informatics, FEI TU of Košice, Slovak Republic

[1]tatiana.kuchcakova@tuke.sk, [2]jaroslav.poruban@tuke.sk

***Abstract*—The field of Natural Language Processing (NLP) has shifted in past decades from mainly research interests and academic domain to general practical use and even more to several diverse specialized domains. One of the actively researched areas is the legal domain because many tasks that legal practitioners do are repetitive and tedious. One of the most common tasks for experts in the legal domain is contract reviewing. This paper presents different techniques of textual NLP that are being researched in terms of automating processes done on contracts in the legal domain. The paper also describes several existing datasets and outlines the next direction of analyzing Slovak contracts using NLP tools.**

***Keywords*—contracts, datasets, information extraction, legal processing, Natural language processing**

I. Introduction

In general, NLP is a multidisciplinary field also called Computational Linguistics and can be considered as the combination of linguistics, artificial intelligence, and computer science [1]. Completing NLP tasks includes the application of either rule-based based techniques or machine learning. These two approaches are often combined in order to get better results.

Despite increasing real-life usage of NLP systems, many specialized domains are affected by NLP very mildly or remain untouched by this field. Reasons include the fact that implementing NLP in specialized domains is costly and highly time-consuming because of specific tasks that are required from which some tasks have to be done by experts for particular domain [2]. NLP has shifted to diverse specialized domains such as finance, healthcare, education, marketing, retail, law, and others. Within NLP applied in legal domain several tasks were investigated by researchers such as legal documents classification [3], [4], extraction of entities of legal documents [5], [6], summarization of legal documents [7], judgement prediction [8] etc.

Depending on the selected task, NLP systems in this domain can be beneficial not only just for experts by using, for example, systems for contract reviewing but also for the general public, for example, in terms of legal advice.

As reference [2] states, many law firms spend approximately 50% of their time with tasks involving contract review. Since the task of manual contract reviewing takes a significant part of the working time of legal practitioners, automating this process by NLP tools is being actively researched in order to reduce their workload.

II. Information extraction of contract elements

Contracts are legal documents describing agreements concluded by two or more parties. As for lawyers, their work often includes repetitive tasks of reviewing contracts. Reviewing contracts has many levels, requiring a different level of expertise in the legal domain.

Contract analysis can be considered as one of the lowest levels of contract reviewing. By fundamental contract analysis, we mean identifying elements of the contract such as type of contract, parties involved in the agreement, essential dates, agreed payments, etc. This task is done mostly manually by a professional who often has to go through hundreds of pages of contracts. While contract reviewing does not require a high level of expertise, time spent on this activity is often a waste of working professionals' potential. It is also considered as one of the most repetitive tasks connected to contracts not only among law firms but also among government agencies and any other companies which use contract analysis in order to monitor contracts [6]. Since the action of contract analysis is repetitive and time-consuming, it presents the best example of a task that is dedicated to being automated by NLP tools.

In the use of general NLP, the task of Named Entity Recognition (NER) is normally used to detect different entities in text. This task generally identifies elements out of unstructured text and assigns them to predefined entities usually based on annotated datasets. Typically identified entities are organizations, persons, locations, and geo-political entities, but dates, prices, locations, and other entities are commonly added to identified entities. Research on general NER produced several approaches based on diverse techniques [9]. In order to avoid laborious dataset annotation, a knowledge-based approach can be used. This approach does not require annotated dataset because it works with lexicons. Several works [10], [11], [12] show, that despite precision metric being high for knowledge-based systems, recall metric often reaches low numbers due to incomplete lexicons since knowledge-based systems require human engineering from a specific domain to create and maintain these knowledge resources.

Another approach that aims to shorten the tedious labeling process of entities when creating datasets is called bootstrapping. In this case, labeling is not entirely omitted. This approach uses semi-supervised machine learning to train a classifier to identify entities. The bootstrapping approach starts with a small amount of labeled data that are annotated in corpora manually. Those initial data are called seed corpus.

After the initial classifier is generated learned on seed corpus, it is used then to analyze unlabeled data, and classification results are added to training data in another iteration in which the initial classifier is re-trained on those data, which results in the classifier increasing its performance each iteration [13].

Current state-of-the-art NER systems are mostly based on a machine learning approach requiring large datasets with labeled entities. Supervised machine learning NER models are then trained on these corpora in order to predict named entities. Previously when implementing NER systems, there were some commonly used models like Support Vector Machine (SVM) [14] or graph-based models like Hidden Markov Model (HMM) [15] and Conditional Random Fields (CRFs) [16]. During the past decade, neural network-based NER systems became popular because of their excellent performance in identifying entities. To address the NER problems, there are used models based on neural networks like bidirectional long short-term memory network (LSTM) plus a CRF layer [17], convolutional neural network (CNN) [18] and others.

In order to use above mentioned classical NER systems in the legal domain, performance needs to be improved by applying high-level semantics [19]. These general NER systems are not applicable to extracting contract elements without adjustment since entities in the legal domain, especially in contracts, vary from those used within a basic context. Another inconvenience is the ambiguity of some elements within contracts. While general NER systems are able to recognize persons or organizations, for example, in contracts like leases, we have to be able to recognize a difference between involved parties whether they fall under lessor or under lessee party [20]. When distinguishing parties within a contract, zones or place of occurrence of elements in the contract can also be significant. Additionally, not all mentioned organizations or persons have to be involved parties of the agreement.

Another entity that faces this ambiguity is the date entity. In general NER systems entity is identified just as a date. In contracts, more date entities are needed from which each stands for a different kind of date information like date of signature, effective date, date of payment, etc. Similarly, many contracts involve elements of amounts and prices. Those prices can be collateral fees, overdue fees, and others. There is another aspect significant within contract prices, and that is the frequency of recurring payments. When we look at the price, we need to be able to distinguish whether this payment should be paid monthly, yearly or consider another condition of recurrence. Type of contract also affect required extracting entities, but in general, we can divide them into several groups [6].

- **Entities for contract names:** These types of entities refers to contract title stated at the beginning of each contract (e.g., leases, employment contracts, sales contract). These entities are important for one of the specific NLP tasks - contract classification. Contract names may also include numbers that state for versions of specific contracts and with involved party entities provide the base for threading versions of contracts.
- **Entities related to involved parties:** Identified involved parties in contracts can be used for systems in order to detect connections among organizations and persons or simply to query all contracts of the chosen organization.
- **Entities related to dates and amounts:** Entities like effective date and termination date are valuable for any contract monitoring. Based on this information, systems can, for example, notice that the termination date is coming and it is time to renew the contract.
- **Entities related to terms and conditons:** These entities involve legislation references, general terms, articles, and laws that contracts depend on and can be helpful for experts when contracts need to be revised after laws are amended.

III. Risk assessment

A higher level of contract reviewing is to assess risk within contract clauses. This type of contract review is challenging and requires a highly skilled legal practitioner. Since risk assessment is a type of work requiring high expertise, companies, organizations, and the general public seek professionals to give them legal advice in this context. They expect lawyers not only to explain what particular clauses in the contract mean, but they also expect them to advise whether signing this contract is risky for them or their business. Some NLP tasks can also be used in this area to automatically detect contract parts that contain risky clauses, also called red flags. Red flags can be indicators for fraudulent behavior [21]. Systems for red flag detection can also detect different types of risks. Detecting red flags in contracts uses the NLP task of classification on a sentence level. The classification task also requires a labeled dataset of contracts where key phrases that can determine risky clauses are annotated. In general, binary classifiers are used in order to determine whether clauses or contracts are risky.

IV. Classification of Contract-Amendment Relationships

Contracts are usually not stand-alone documents in real-world usage. Many of the contracts have links to another contract or other documents. Most common and also most important are links between contracts and their amendments which in some way alter or add meaning to the original master contract. While reviewing contracts, keeping track of these links is a crucial task to have the whole and up-to-date information about contract terms, but it is also important in order to reduce or remove potential legal risks. These contract amendments are often signed years after the master agreement was signed. It is also common that new amendments for a particular contract are signed every year or so. All these mentioned amendment occurrences are factors that make keeping track of all contract amendments to a particular master contract a challenging task. It is conventional that this task is performed manually, which is error-prone. Professionals are often used to having a knowledge of amendments that are signed in regular intervals, but it gets more complicated when no pattern can be followed for keeping track of signed amendments.

NLP techniques can be used to facilitate this process by automation. This problem is usually addressed as a classification problem proceeded by NER. In order to distinguish relations between two documents, key entities have to be extracted and compared.

- **Document name:** Document names usually follow certain patterns that can help distinguish between master contracts and amendments. Amendment names also commonly include their number representing the order.
- **Involved parties:** Parties stated in the master contract are usually the same as those in its amendment. This fact

represents a standard for a contract-amendment relationship that is another indication when finding the relation between documents.
- **Document body:** The body of the master contract and its amendments are semantically related.
- **References:** Amendments include explicit references to a master contract that includes the name of the contract with the date of signature or effective date.

Extracted entities then form the feature set for the classification algorithms [22].

V. State of the art

Using a machine learning approach in NLP tasks requires a large amount of labeled data to train a model. There exist many large datasets and pre-trained models for tasks like document analysis, but it is still an open question whether it can be used and modified for specialized tasks of specific domain [2]. Creating a dataset for model training is highly time-consuming, and this task gets even harder when it comes to the domain-specific dataset, which has to be annotated by domain experts of a specific field, in this case usually by lawyers.

Since NLP in the legal domain is an actively researched area, there are some existing benchmark datasets mainly in the English language. Datasets in the legal domain are specific and vary not only because of language used but also law differs in every jurisdiction what has to be taken into consideration. Each dataset uses its specific type and number of labeled elements. Gathering a large number of contracts for corpora is also not a trivial task because of contracts confidentiality. Confidentiality is also a reason why many datasets presented by researchers are provided only in encoded form.

A. Contract Understanding Atticus Dataset (CUAD)

CUAD is a contract review dataset presented by Dan Hendrycks et. al. [2] which was created with the help of a number of legal domain experts. Law students annotated 510 contracts with 13 101 annotations in 41 categories of contract elements. Labeled entities are divided into three categories.

Dataset consists of 25 different types of contracts. For creating datasets, contracts from the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system were used. Publicly traded and other reporting companies in the USA are required to file certain types of contracts to this system. Access to EDGAR documents is open to the public. Therefore, all texts in the CUAD dataset are provided freely.

The authors used several state-of-the-art pre-trained language models and finetuned them using HuggingFace Transformers [23] library on the CUAD dataset. In their work they were evaluating models like BERT [24], ALBERT [25], RoBERTa [26], DeBERTa [27]. Authors displayed Performance of used models on the CUAD dataset by Precision-Recall curves for three of the selected pre-trained models where they used Area Under the Precision-Recall curve (AUPR). Precision-recall curves results in DeBERTa-xlarge model showing best overall performance with precision of 44.0% for a recall 80%, and precision of 17.8% for a recall 90%.

B. Benchmark dataset for Lease Contract Review by Leivaditi et al.

Leivaditi et al. present in their work [20] dataset for low and high-level contract reviewing of lease contracts in English. Their dataset consists of 179 contracts originating from the EDGAR system, similar to the aforementioned CUAD dataset. On the other hand, in the CUAD dataset, 25 different types of contracts were used, while this dataset is specifically oriented only on lease contracts. All entities in the dataset were labeled manually by law master students, where all 179 contracts consist red flags, out of which 123 are annotated with entities. In order to detect entities and red flags in contracts, authors used generalized language models based on the encoder module of Transformer [28]. The authors present an extended state-of-the-art model ALBERT. The language modeling task was continued with a lease benchmark dataset of 179 contracts on this model. Their new model that was used to red flag detection and entity recognition for lease contracts is called ALeaseBERT. For extracting entities contained in contracts, the ALeaseBERT model results in weighted average precision of 62% and recall of 48%.

C. Benchmark dataset presented by Chalkidis et al.

Chalkidis et al. present in their work [6] benchmark dataset for extracting contract elements from contracts in English. Dataset consists of 3454 contracts in total, out of which 2461 are labeled with ten types of entities like Contract Title, Contracting Parties, Start, Effective, Termination Dates, Contract Period, Value, Governing Law, Jurisdiction and Legislation References. Another 993 contracts are labeled with clause headings. Contract element annotations of the labeled dataset were provided by 10 law students. Understandably, many labeled entities are composed of more than one token. The whole dataset consists of 37,1 million tokens. Authors experimented with linear classifiers like Linear Regression and Support Vector Machine and manually written contract element extraction rules. Results showed the best macro average precision of 76% and recall of 86% measured per token were achieved when using linear SVM classifiers.

D. Benchmark classification dataset for contract understanding by Elwany et al.

In work [4], presented by Elwany et al. dataset of thousands of legal contracts was used. All contracts were annotated manually by rapid annotation tools. They fine-tuned the pre-trained BERT model on this corpus of data by running unsupervised fine-tuning. The authors concluded that pre-trained BERT brings significant improvement in the legal domain to the classification task resulting in weighted average precision and recall of approximately 90%.

E. Benchmark dataset of Czech public contracts

One of the latest works that deal with extraction of contract metadata involves Czech public contracts [29]. Contracts used to create datasets are publicly available on the official state registry of Czech public contracts. Dataset consists of 112 contracts annotated manually by a single annotator. Only ten contracts are used for the development set, and the other 102 contracts are used as a test set. In the presented work, 13 different types of elements were annotated. They used a named entity recognition module based on the Slavic BERT model for four languages, including the Czech language. Results showed that almost 88% of contract elements were extracted correctly, with the highest accuracy of 94.3% for address entity and 93% for contract type entity.

VI. Further work

To the best of our knowledge, there is no described dataset for legal contracts analysis using NLP tools in Slovak language and Slovak law yet. Therefore, in our work, we have decided to create such one. We are working within the specific legal domain where one of the involved parties is a government organization or a person acting on behalf of a government organization.

To create a dataset, we use contracts from the Slovak central registry of contracts (CRZ). Government organizations in the Slovak Republic are legally obligated to share certain types of contracts in the CRZ registry. This registry is available online for the general public to freely view the content of published government contracts.

So far, we have focused our work on lease contracts within this registry. Currently, we are in process of labeling lease contracts for NER using regular expressions in combination with manual labeling for particular types of lease contracts where regular expressions are hardly applicable. Further work will focus on finishing the benchmark dataset and researching and training models applicable to this dataset for the task of named entity recognition.

VII. Conclusion

In order to facilitate the work of legal practitioners by automating it, several NLP techniques, which we presented in this work, can be used according to the task of contract processing.

While state-of-the-art works presented promising results in performance by applying fine-tuned pre-trained models on contract data corpora in several NLP tasks related to contract reviewing, they also showed that there is a large room for improvement available.

By presented researched datasets and models which used different sizes of datasets, we can conclude that size of the dataset and the number of annotated data within it affects performance, with larger annotated datasets resulting in better numbers.

Further work will focus on creating a dataset in the Slovak language and researching models that can be used for tasks related to extracting contract elements by NLP tools.

Acknowledgment

This work was supported by project VEGA No. 1/0630/22 "Lowering Programmers' Cognitive Load Using Context-Dependent Dialogs.

References

[1] S. Vajjala, B. Majumder, A. Gupta, and H. Surana, *Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems*. O'Reilly Media, 2020.

[2] D. Hendrycks, C. Burns, A. Chen, and S. Ball, "Cuad: An expert-annotated nlp dataset for legal contract review," *arXiv preprint arXiv:2103.06268*, 2021.

[3] P. H. Luz de Araujo, T. E. de Campos, F. Ataides Braz, and N. Correia da Silva, "VICTOR: a dataset for Brazilian legal documents classification," in *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 1449–1458. [Online]. Available: https://aclanthology.org/2020.lrec-1.181

[4] E. Elwany, D. Moore, and G. Oberoi, "Bert goes to law school: Quantifying the competitive advantage of access to large legal corpora in contract understanding," *arXiv preprint arXiv:1911.00473*, 2019.

[5] E. Leitner, G. Rehm, and J. Moreno-Schneider, "Fine-grained named entity recognition in legal documents," in *International Conference on Semantic Systems*. Springer, 2019, pp. 272–287.

[6] I. Chalkidis, I. Androutsopoulos, and A. Michos, "Extracting contract elements," in *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, 2017, pp. 19–28.

[7] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, "How does nlp benefit legal system: A summary of legal artificial intelligence," *arXiv preprint arXiv:2004.12158*, 2020.

[8] M. Masala, R. C. A. Iacob, A. S. Uban, M. Cidota, H. Velicu, T. Rebedea, and M. Popescu, "jurbert: A romanian bert model for legal judgement prediction," in *Proceedings of the Natural Legal Language Processing Workshop 2021*, 2021, pp. 86–94.

[9] V. Yadav and S. Bethard, "A survey on recent advances in named entity recognition from deep learning models," *arXiv preprint arXiv:1910.11470*, 2019.

[10] M. Peng, R. Ma, Q. Zhang, L. Zhao, M. Wei, C. Sun, and X. Huang, "Toward recognizing more entity types in NER: An efficient implementation using only entity lexicons," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 678–688. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.60

[11] S. Tedeschi, V. Maiorca, N. Campolungo, F. Cecconi, and R. Navigli, "WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER," in *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2521–2533. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.215

[12] M. Peng, X. Xing, Q. Zhang, J. Fu, and X. Huang, "Distantly supervised named entity recognition using positive-unlabeled learning," *arXiv preprint arXiv:1906.01378*, 2019.

[13] J. Kim, Y. Ko, and J. Seo, "A bootstrapping approach with crf and deep learning models for improving the biomedical named entity recognition in multi-domains," *IEEE Access*, vol. 7, pp. 70 308–70 318, 2019.

[14] Z. Ju, J. Wang, and F. Zhu, "Named entity recognition from biomedical text using svm," in *2011 5th International Conference on Bioinformatics and Biomedical Engineering*, 2011, pp. 1–4.

[15] D. Chopra, N. Joshi, and I. Mathur, "Named entity recognition in hindi using hidden markov model," in *2016 Second International Conference on Computational Intelligence & Communication Technology (CICT)*. IEEE, 2016, pp. 581–586.

[16] M. Konkol and M. Konopík, "Crf-based czech named entity recognizer and consolidation of czech ner research," in *International conference on text, speech and dialogue*. Springer, 2013, pp. 153–160.

[17] Y. Jin, J. Xie, W. Guo, C. Luo, D. Wu, and R. Wang, "Lstm-crf neural network with gated self attention for chinese ner," *IEEE Access*, vol. 7, pp. 136 694–136 703, 2019.

[18] T. Gui, R. Ma, Q. Zhang, L. Zhao, Y.-G. Jiang, and X. Huang, "Cnn-based chinese ner with lexicon rethinking." in *ijcai*, 2019, pp. 4982–4988.

[19] G. Li, Z. Wang, and Y. Ma, "Combining domain knowledge extraction with graph long short-term memory for learning classification of chinese legal documents," *IEEE Access*, vol. 7, pp. 139 616–139 627, 2019.

[20] S. Leivaditi, J. Rossi, and E. Kanoulas, "A benchmark for lease contract review," *arXiv preprint arXiv:2010.10386*, 2020.

[21] G. Baader and H. Krcmar, "Reducing false positives in fraud detection: Combining the red flag approach with process mining," *International Journal of Accounting Information Systems*, vol. 31, pp. 1–16, 2018.

[22] F. Song, "Classification of contract-amendment relationships," *arXiv preprint arXiv:2106.14619*, 2021.

[23] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[25] I. Tenney, D. Das, and E. Pavlick, "Bert rediscovers the classical nlp pipeline," *arXiv preprint arXiv:1905.05950*, 2019.

[26] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[27] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," *arXiv preprint arXiv:2006.03654*, 2020.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[29] H. T. Ha, A. Horák, and M. T. Bui, "Contract metadata identification in czech scanned documents." in *ICAART (2)*, 2021, pp. 795–802.

Explainable Artificial Intelligence: Concencration Metric

[1]*Ivan Čík (3rd year),*
Supervisor: [2]*Marián MACH*

[1,2]Dept of Cybernetics and Artificial Intelligence, FEI TU of Košice, Slovak Republic

[1]ivan.cik@tuke.sk, [2]marian.mach@tuke.sk

***Abstract*—To apply artificial intelligence algorithms to real-world solution, it is often necessary to trust and understand them. Explainable artificial intelligence deals with such challenges. In this paper, we describe the LRP method, experiments with three different types of convolutional neural models (Prototypical Network, U-net, and the left part of U-Net) on the Oxford-IIIT Pet Dataset, and present concentration metrics for a convolutional neural network.**

***Keywords*—Convolutional Neural Networks, Deep Learning, Explainable Artificial Intelligence, Layer-wise Relevance Propagation.**

I. Introduction

Improvements in artificial intelligence (AI) and deep learning were made possible by the rise of available information, hardware enhancements, new optimization algorithms, open-source libraries, datasets etc. The increase in computational power has allowed the emergence of a sub-area of AI called deep learning (DL) [1]. Currently, DL algorithms are used in almost every field such as healthcare [2], [3], [4], in autonomous robots and vehicles [5], [6], [7], in image processing classification, and detection [8], [9], in speech and audio processing [10], [11], and many others. Replicating an ML model on another machine is fast and cheap. The training of a human for a task can take decades (especially when they are inexperienced and without knowledge in the field) and is very costly. A major disadvantage of using ML is those insights about the data and the task the machine solves are hidden in increasingly complex models. One of the critical problems in the application of deep neural networks in real life is their black-box nature, which raises ethical and legal issues for users and a lack of trust. Explainable artificial intelligence as a sub-area of artificial intelligence provides tools, techniques, and algorithms that can provide high-quality interpretable, intuitive, and human-understandable explanations of black-box models.

II. Building User Trust

Using adversarial examples [12], [13], [14], perturbation methods and poisoned training datasets [15] can detune model decisions. DNN's inability to face such attacks raises questions of trustworthiness for users. The introduction of metrics for model consistency could help in this issue. AI algorithms are increasingly being implemented in the real world. As the trend increases, so does the demand for building user trust in such solutions, even more so in areas where undesired model decisions can affect a person's life. Addressing this issue opens up a new thread of research that is just beginning to develop [16], [17].

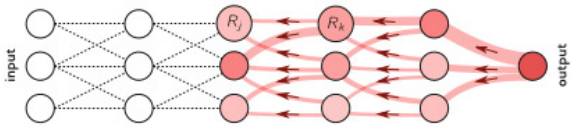


Fig. 1. Diagram of the LRP process. Each neuron is redistributed to lower layers as much as it receives from higher layers [21].

The main goal of XAI is that the algorithm must produce information that builds the user's confidence in the internal processes of the algorithm. Based on our research we have identified the specific trait **causality** that we would like to address in our research. Causality in terms of XAI means that only causal relationships are picked up [18].

III. Layer-Wise Relevance Propagation:

In this section, we describe the layer-wise relevance propagation (LRP) technique introduced in [19]. LRP explains deep neural network prediction and is one of the relevance propagation methods so that the prediction propagates backward without the use of gradients. The algorithm starts from the output layer L of the deep neural network, and then moves in the opposite direction in the graph, gradually redistributing the prediction score until the input is reached. Each neuron receives a share in the output of the network and further redistributes it to its predecessors in the same amount until the input variables are reached [19], [20]. LRP is a method of identifying important pixels by inverting a neural network. The reverse pass is a conservative redistribution process, in which the neurons that contribute the most to the previous layer are the most important.

Mathematically, it redistributes the prediction $f(x)$ backward, using redistribution rules, until it assigns a relevance score R_i to each variable. An important feature of the redistribution process is called global conservation property and is given as:

$$\sum_i R_i = ... = \sum_j R_j = \sum_k R_k = ... = f(x) \qquad (1)$$

The described property says that in each step of the redistribution process the overall relevance is maintained. The

relevance score R_i of each input variable determines the extent to which this variable has contributed to the prediction [19].

The global conservation principle is that it preserves the propagated quantity between the neurons of two adjacent layers. When we denote $R_i^{(l)}$ as the significance associated with the i-th neuron of layer l and $R_j^{(l+1)}$ as the significance associated with the j-th neuron in the next layer, the conservation principle requires that

$$\sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} \tag{2}$$

where the sums pass through all the neurons of the layers of the given neural model. Applying this rule repeatedly for all layers will ensure that the heatmap resulting from the LRP complies $\sum_p h_p = f(x)$, where $h_p = R_p^{(1)}$ [19].

IV. Experimental Setup

For our experiments we use Oxford-IIIT Pet Dataset for 50 epochs [22]. The dataset consists of 7349 images (2371 images of dogs and 4978 images of cats), where for each image its class and segmentation mask are available according to where the animal is located in the image. In the experiments we use three different architectures of convolutional neural networks:

- **A prototypical network** consisted of two parts: the feature extractor and a classifier. The feature extractor is 4 layers convolutional network.
- **U-net [23]** is the kind of model that reconstructs the mask. The network consists of two parts: downsampling and upsampling (mask reconstruction). Between these two parts, a single linear layer is added, which classifies the images.
- **The left part of the U-net**, where there is only downsampling and classification without mask reconstruction.

V. Concentration Metric

In our research, the LRP method is used on trained models after 50 epochs. At the output of the model, we have 2 classes, where the value of the correct class is then redistributed to the input dimensions so that for each pixel it is possible to achieve a numerical value that corresponds to how important it was in the classification. We apply the LRP method to each image from the validation dataset. We compute a mask for each output from the LRP method (224 x 224 x 3) as follows:

- The output for each input image from the LRP method is flattened to a vector $\mathbf{v}$ with shape 150528 x 1.
- We pick the 500th highest value h from vector $\mathbf{v}$ and all values $v_1, v_2, ..., v_n$ are set to 1 when they are $\geq h$, otherwise they are set to 0.
- Knowing indices from the vector $\mathbf{v}$, we can easily build an LRP mask with a shape of 224 x 224.
- LRP mask is multiplied by the image original mask, which achieves that the pixels outside the object are set to 0. Subsequently, the concentration metric for the image is calculated as the sum of pixels multiplied by the original mask in proportion to the number of pixels we picked (500 in this case). this process is performed for each image and then averaged.

	Accuracy	Loss	Concencration Metric
Prototypical Network	89.31 %	0.287	84.82%
U-net	89.58 %	1.415	85.83%
U-net left part	88.06 %	0.9646	57.99%

TABLE I
Accuracy and concencration metric values after 50 epochs on Oxford-IIIT Pet Dataset

VI. Results

As a result, it can be seen in Table I that the Prototypical network and U-net with a linear layer have pixels that affect the classification mostly localized in the objects, whereas the third model does not hold this statement. We did not rely on the values of the loss function as U-net uses two loss functions: for classification (Cross-Entropy Loss) and for mask reconstruction (Dice Loss).

VII. Conclusion

In this article, we briefly describe the explainable artificial intelligence and LRP method. The article describes an experiment with the Oxford-IIIT Pet Dataset, on which 3 models are trained (U-net, U-net left part, and Prototypical Network). Subsequently, the concentration metric is described. By classifying the object in the image correctly according to its position and not other interfering elements, we can increase the end-user confidence and causality of the model. Our future work is to focus on the output of the concentration metric to be included in the loss function and deeper research in the field of concentration metrics.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] A. D. Torres, H. Yan, A. H. Aboutalebi, A. Das, L. Duan, and P. Rad, "Patient Facial Emotion Recognition and Sentiment Analysis Using Secure Cloud With Hardware Acceleration," *Comput. Intell. Multimed. Big Data Cloud with Eng. Appl.*, pp. 61–89, 2018.

[3] S. M. Lee, J. B. Seo, J. Yun, Y.-H. Cho, J. Vogel-Claussen, M. L. Schiebler, W. B. Gefter, E. J. Van Beek, J. M. Goo, K. S. Lee *et al.*, "Deep learning applications in chest radiography and computed tomography," pp. 75–85, 2019.

[4] R. Chen, L. Yang, S. Goodison, and Y. Sun, "Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data," *Bioinformatics*, vol. 36, no. 5, pp. 1476–1483, 2020.

[5] C. You, J. Lu, D. Filev, and P. Tsiotras, "Advanced planning for autonomous vehicles using reinforcement learning and deep inverse reinforcement learning," *Robotics and Autonomous Systems*, vol. 114, pp. 1–18, Apr 2019.

[6] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *Journal of Field Robotics*, vol. 37, no. 3, pp. 362–386, 2020.

[7] D. Feng, C. Haase-Schutz, L. Rosenbaum, H. Hertlein, C. Glaser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–20, 2020.

[8] A. Sahba, A. Das, P. Rad, and M. Jamshidi, "Image Graph Production by Dense Captioning," in *2018 World Autom. Congr.*, vol. 2018-June. IEEE, Jun 2018, pp. 1–5.

[9] N. Bendre, N. Ebadi, J. J. Prevost, and P. Najafirad, "Human action performance using deep neuro-fuzzy recurrent attention model," *IEEE Access*, vol. 8, pp. 57 749–57 761, 2020.

[10] A. Boles and P. Rad, "Voice biometrics: Deep learning-based voiceprint authentication system," in *2017 12th System of Systems Engineering Conference (SoSE)*. IEEE, 2017, pp. 1–6.

[11] S. Panwar, A. Das, M. Roopaei, and P. Rad, "A deep learning approach for mapping music genres," in *2017 12th System of Systems Engineering Conference (SoSE)*. IEEE, 2017, pp. 1–5.

[12] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE transactions on neural networks and learning systems*, 2019.

[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.
[14] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.
[15] N. Müller, D. Kowatsch, and K. Böttinger, "Data poisoning attacks on regression learning and corresponding defenses," in *2020 IEEE 25th Pacific Rim International Symposium on Dependable Computing (PRDC)*. IEEE, 2020, pp. 80–89.
[16] H. Jiang, B. Kim, M. Y. Guan, and M. R. Gupta, "To trust or not to trust a classifier." in *NeurIPS*, 2018, pp. 5546–5557.
[17] J. Heo, H. B. Lee, S. Kim, J. Lee, K. J. Kim, E. Yang, and S. J. Hwang, "Uncertainty-aware attention for reliable interpretation and prediction," in *Advances in Neural Information Processing Systems*, 2018, pp. 909–918.
[18] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.
[19] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.
[20] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 11, pp. 2660–2673, 2016.
[21] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: an overview," in *Explainable AI: interpreting, explaining and visualizing deep learning*. Springer, 2019, pp. 193–209.
[22] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3498–3505.
[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

Motion Detection and Object Tracking in Transportation Based on Edge Computing

[1]*Kristian Micko (1st year),*
Supervisor: [2]*Peter Papcun*

[1,2]Dept. of Cybernetics and Artificial Intelligence, FEI TU of Košice, Slovak Republic

[1]kristian.micko@tuke.sk

***Abstract*—Computer Vision (CV) is the area of computer science that helps us solve many problems in our digital world. This article describes the approaches used to detect motion and track target objects automatically without human interaction. We compared two methods of CV, which are commonly used in object tracking and motion detection. The first one uses deep learning (DL), and the second is traditional CV approaches such as frame differentiation.**

***Keywords*—Convolutional Neural Networks, Computer Vision, Deep learning, Frame difference, Motion detection, Object tracking**

I. Introduction

Many people use highways to travel from point A to point B in the transportation sphere. A route is a place where cars, trucks and other vehicles use high speed, which saves time, but on the other hand, it increases the probability of fatal accidents. Many undetermined conditions increase this probability. Smart gateways with cameras on the highway could collect data about the actual situation on the road and warn authorities about the situation. It is not enough to collect and store data these days. It is necessary to analyze them. Computer Vision (CV) methods are suitable for automatic analysis of the live broadcasts from Closed Circuit Television (CCTV) cameras. CCTV cameras on the highways could identify a stolen car using automatic number plate recognition (ANPR) applications.

II. Methods used in computer vision

Computer Vision (CV) is one of the domains in computer science, which methodology we can divide into two areas. Traditional CV techniques and Machine Learning (ML). One of the commonly used of the ML algorithms in CV is Deep Learning (DL). DL methods have been dominant last few years because they can solve a lot of complex problems such as classification, segmentation, image colourization and object detection. The reason is that it has a better result compared to traditional methods. According to [1], we have to use many hidden layers in the Convolutional Neural Networks (CNNs) architecture to increase image classification accuracy. CNNs have a disadvantage, it needs a lot of collected data for better accuracy. Fortunately, we live in a digital world and social networks, where smart devices produce an enormous volume of data.

A. Descritpion of Deep Learning

In computer science, we have a traditional paradigm about solving a problem. It contains conditions that are used as rules in our problem. Programmer manually writes these rules, which generates output from input. The programmer has to know what to expect as output. It is necessary to know the rules of a domain where the programmer implements the program. On the other hand, machine learning is another paradigm to solve computer science problems. In the real world, the programmer often does not know the rules, which are had to make a program. Machine learning is a paradigm based on automatically finding rules based on collected many training patterns. These training patterns are simply collections of input data and output data. Machine learning has many algorithms, and DL is one of them [2]. Artificial Neural Networks (ANNs) consist of many computing cells, also called neurons, that perform a simple operation and interact between each other neurons to make decisions. This principle is similar to the human brain [3]. An artificial neuron is mathematically kind of the activation function. The activation function defines the output of that node given an input or set of inputs and, via the output, decides which neuron should be activated or not. DL is the process where learning is a random assignment of weight to each artificial neuron. This assignment runs across many layers of a neural network accurately, efficiently and randomly. With more robust processing hardware, it increases interest in ANNs models [4].

DL improved solutions in many tasks in various CV applications:

- device capabilities including computing power,
- memory capacity,
- power consumption,
- image sensor resolution,
- optics.

These CV applications improve the performance and cost-effectiveness models in CV applications based on DL and applications in other areas of computer science (such as speech recognition and text mining). Making DL models enables CV engineers to get higher accuracy in image classification, semantic segmentation, object detection and Simultaneous Localization and Mapping (SLAM). Many CV engineers prefer to build bigger DL models or transfer learning existing DL models against manually programming applications, where much more expert analysis has to be used. CNNs models provide flexibility in training and retraining on a custom dataset for any use case. Traditional CV algorithms are more domain-

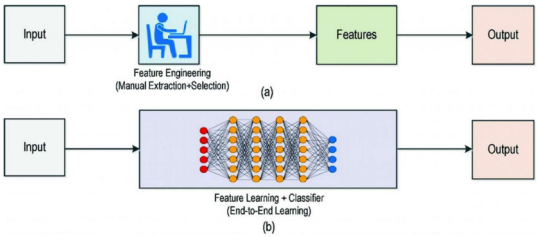


Fig. 1. (a) Traditional computer vision workflow vs. (b) Deep learning workflow. Figure from [5]

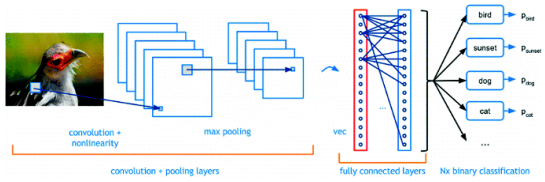


Fig. 2. Description of principle of work Convolutional Neural Network. Figure from [11]

specific than DL models [2]. The program needs to have bag-of-words to describe and classify a specific object (such as chair, car, people, etc.). The main problem of the traditional approach is the necessity to choose which features are essential in each given image. Complex CV projects contain many classes of objects that we have to classify. We have a problem implementing a program to classify objects correctly because many features are identical to many objects. For example, a pencil has a similar shape to a pen. DL uses the concept of end-to-end learning. We prepare a dataset of images that have been annotated with classes of an object by the domain expert. A computer, via learning CNNs, automatically finds relations between training dataset images to extract essential features to define objects of interest [4]. After the training process, CNNs discover relations in features describing objects of interest. This DL model we can apply to create a software application. We can apply to create a software application for this DL model Fig. 1.

After the invention of CNNs, it had an impact to make significant progress to the possibility of recognizing objects in CV [6]. Progress in hardware manufacturing enables increasing computing power and making massive data storage. One of the most significant steps was the widespread adoption of various deep-neural network architectures for CV tasks. Due to its seminar paper, ImageNet Classification with Deep Convolutional Neural Networks has been cited over 3000 times [7]. CNNs use various kernels known as filters to detect features throughout an image. A kernel is a matrix of values, called weights, trained to find specific features. The main idea of the CNNs is to spatially convolve the kernel on a given input image and inspect if the feature it is meant to detect is present. To get value describing the probability of the presence of a specific feature, we perform a convolution operation to carry out by computing the dot product of the kernel and the input area where the kernel is overlapped (the place of the original image where the kernel is looking at is known as the receptive field) [8]. When we want to facilitate the learning of kernel weights, we sum the convolution layer's output with a bias term and then feed it to a non-linear activation function. Logistic, TanH and ReLU (Rectified Linear Unit) are commonly used non-linear functions as Activation Functions. Depending on the structure of data and classification tasks, these activation functions are selected accordingly [9]. ReLU activation function reduces the vanishing gradient problem and produces sparser, more efficient representations [4]. In many CNNs architectures, we want to speed up the training process and reduce the memory consumption of the network by adding a pooling layer after the convolutional layer to remove redundancy present in the input feature. The working principle of max pooling is that this layer moves a window over the input and takes a maximum value from that window as an output due to effectively reducing to the important pixels in an image [4]. Deep CNNs have several pairs of convolutional and pooling layers whose outputs are fully connected and flatten the previous layer volume into a vector, and then an output layer computes the scores (such as confidence or probabilities) for the output classes/features through a dense network Fig. 2. To final classification, the output is then passed to a regression function such as Softmax [10], where everything is mapped to a vector whose elements sum up to one [4].

In summary, DL and CNN is a frequently used approach in CV, but it is not the solution for all problems. On the other hand, DL is suitable for solving the problem of object tracking and motion detection. We can use CNN as the main solution and as a support method for the other techniques in motion detection problems and object tracking techniques. DL model can give information from the video stream when it finds the object of interest and triggers other object tracking applications or shows the locations of the wanted objects by rectangles.The main goal of the object detector is to classify objects on the images and find localization of the classified object via drawing rectangles around the target object [12]. From these rectangles, we can get interframe information about the motion direction of the wanted thing, and we can track the object [13]. Then it is possible to calculate a vector of object movement.

B. Traditional Computer Vision Techniques

In this section, we mention the traditional feature-based approaches in CV. We can list some of them:

- Scale Invariant Feature Transform (SIFT) [14],
- Speeded Up Robust Features (SURF) [15],
- Features from Accelerated Segment Test (FAST) [16],
- Hough transforms [17],
- Geometric hashing [18].

Feature descriptors such as SIFT and SURF are generally combined with traditional machine learning classification algorithms such as Support Vector Machines and K-Nearest Neighbour. Therefore we did not describe it deeply.

For motion detection, it is used the method of difference between pictures. In the first step, we have to convert the images to a grayscale colour model from the RGB model by equation (1).

$$Y = 0,299 * R + 0,587 * G + 0,114 * B \tag{1}$$

The video sequence uses one picture as the reference picture in memory, and every new picture compares to the reference picture. Comparison is based on calculating the difference value of the brightness of every pixel between the new picture and the reference picture. The position of every pixel, which was created by calculating differences between the reference

picture and the new one, is used for rendering binary images. This method expects the static position of the camera and threshold input where the slight difference pixels would be ignored because of the noise reduction. Binary images create shapes that we use to analyse motion detection via contour findings [19] [20] Fig. 3. Sometimes the found shapes are too small or too big in binary images. For the correction, we use dilatation or erosion of the binary images. The definitions of dilation and erosion are typically formulated using the concept of a set translation and a set reflection. The translation of a set α by a point (or vector) x, denoted α_{x}, is defined by (2).

Fig. 3. Example of differention images [19]

$$\alpha_{\mathrm{x}} = \{a + x \, a \in \alpha\} \tag{2}$$

The reflection of a set α , denoted $\overline{\boldsymbol{\alpha}}$, is defined (3).

$$\check{\alpha} = \{-a : a \in \alpha\} \tag{3}$$

Binary dilation of set α by set β , denoted here $\alpha \bigoplus \beta$, is the set union of all translations of set α by elements of set β or equivalently the set of all positions of the reflected set $\check{\boldsymbol{\beta}}$ for which it intersects with, or "hits", set α. The set β is commonly called the structuring element and plays a similar role to a finite impulse response filter in linear signals and systems theory.

$$\alpha \oplus \beta = \bigcup_{b \in \beta} \alpha_{\mathrm{b}} = \{x : \check{\beta}_{\mathrm{x}} \cap \alpha \neq 0\} \tag{4}$$

The erosion of set α by set β, denoted here $\alpha \ominus \beta$, is the set intersection of all negative translations of set α by elements of set β or equivalently the set of all positions for which set β is a subset of α (i.e., it "fits" inside α) [21].

$$\alpha \ominus \beta = \bigcap_{b \in \beta} \alpha_{-\mathrm{b}} = \{x : \check{\beta}_{\mathrm{x}} \cap \alpha \subset 0\} \tag{5}$$

Object tracking consists of two steps. In the first step, we have to initialize the position of an object of interest in a reference picture. In this step, usually we some object detection method. When we classify the position of the object, then we calculate a centroid of a detected object. In the second step, a detected object in the next frame compares with the previous frame's position. In the new frame, we calculate the object's centroid. The difference between an object's centroid is frequently calculated via euclidian distance (6). We have to define the max distance between centroids to identify the same objects in an interframe context [22].

$$d\left(p, q\right) = \sqrt{\sum_{i=1}^{n} \left(q_i - p_i\right)^2} \tag{6}$$

III. Computer Vision in transportation

Microcontrollers or microprocessors did not have enough capabilities to compute a large volume of data [23]. At this moment, we have powerful microprocessors with dedicated support circuits specializing in computing inference of neural network models. Also, it was developed especial frameworks of DL models for microprocessors such as NVIDIA Jetson nano to increase the speed of inference of neural network models and methodologies to make these models simpler [24]. Microprocessors are called edge devices. Implement edge devices with the application CV algorithms connected directly to the camera could preprocess data to summarised information and send them to the cloud servers or TMC with much less communication bandwidth. Motion detection and object tracking problems can be solved via DL methods [25], [26] or in the special cases we can use traditional CV approaches such as difference between frames [19]. DL methods have a disadvantage in that we should use high computational power to run CNNs models in real-time. It would not be enough to solve any task in a transportation system in special conditions.

IV. DL models deployment on the edge devices

DL models deployed on the edge devices have to be optimised to get enough inference time for our tasks. Several approaches could reach it. Compression neural network models, develop specialised embedded hardware to these CNNs models [27] and combine both methods. Design specialised hardware would be more time-consuming than other solutions for optimising neural network models because manufacturing process. On the other hand, optimisation of DL models via simplification increase inference computational speed but decrease prediction accuracy, which could not be good enough for specific tasks [28].

Prunning: This step removes less important parameters from a trained neural network model. Before applying this technique, we have to rank the neurons in the DL model from the view of the impact on classification accuracy. After ranked neurons, we can drop neurons with the minor rank, and as a result, we get the model having reduced size. To decide how many drop neurons, we should do it carefully because we could significantly damage the accuracy of the neural network model [29].

Data quantization: This technique reduces the number of bits used to store the weights of a network, e. g. using 8 bits to represent a 32-bit floating point number. In the article [29], the researchers applied data quantization to convert a pre-trained floating point model into a fixed point model. They did not re-trained the model, and they used 6, 8 and 16 bit to represent a 32-bit number as the most common fixed-pointed data quantization configurations [30].

In article [31], [32], the researchers describe deployment ML models via combination edge-cloud computing. Edge computing classified via cameras the status of the distributional system to deliver electricity. Edge devices send to the

cloud classified data and in the cloud storage system decide how critical incidents on the electric tower or other areas of the distributional power system. Combinations between edge and cloud make benefits to save resources [31].

V. RESEARCH GAP

We cannot deploy all to the cloud in a complex information system because connection problems could lose our data or be noised. On the other hand, cloud computing could allocate any computational capacity we want. Without optimizing used CV methods and algorithms, the edge devices could not produce enough computational power for critical real-time tasks. Computer scientists have to solve this challenge of balancing all of these tasks between edge, cloud computing and level of the algorithm optimization for the deployment in any domain [31].

To solve the motion detection and object tracking problems, we have to try deep learning methods. We will have to search the border, where it is more effective to use the traditional CV approach, and where the machine learning methods will be more suitable.

VI. FUTURE WORK AND CONCLUSION

In our work, we want to implement and test both approaches to problems of the motion detection and the object tracking. We will compare the success results of the DL methods and traditional CV techniques via criteria such as accuracy, recall and inference time.

After deciding the more successful method, we try to solve which practice of optimisation approach would best fit in computational distribution between edge and cloud computing. The result of these comparisons would help not only in the domain of transportation, but it could help to make a recommendation in other areas, how CV methods have the best results in computational distribution for solution problem of motion detection and object tracking.

In the following research work, we will try to implement deep learning methods and other CV approaches to solve other problems in the transport industry. Motion detection in gates of highways could check speed. Classifying the positions of the moving cars, we can focus on automation number recognition. We should compare the best methods of the CV suitable for solving the transport problems on the edge devices. The transport industry is one of the essential spheres in our services, and it needs to be significantly developed for our reliability and safety.

ACKNOWLEDGMENT

This publication is the result of the APVV grant ENISaC - Edge-eNabled Intelligent Sensing and Computing (APVV-20-0247).

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, 2012.

[2] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan, and J. Walsh, "Deep learning vs. traditional computer vision," in *Advances in Computer Vision*, K. Arai and S. Kapoor, Eds. Springer International Publishing, 2020.

[3] N. O'Mahony, T. Murphy, K. Panduru, D. Riordan, and J. Walsh, "Adaptive process control and sensor fusion for process analytical technology," in *2016 27th Irish Signals and Systems Conference (ISSC)*. IEEE, 2016.

[4] P. Koehn, "Combining genetic algorithms and neural networks: The encoding problem," 1994.

[5] J. Wang, Y. Ma, L. Zhang, R. X. Gao, and D. Wu, "Deep learning for smart manufacturing: Methods and applications," *Journal of Manufacturing Systems*, 2018, special Issue on Smart Manufacturing.

[6] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational Intelligence and Neuroscience*, 2018.

[7] W. Nash, T. Drummond, and N. Birbilis, "A review of deep learning in the study of materials degradation," *npj Materials Degradation*, no. 1, 2018.

[8] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," 2018.

[9] S. Hayou, A. Doucet, and J. Rousseau, "On the selection of initialization and activation function for deep neural networks," 2018.

[10] S. Horiguchi, D. Ikami, and K. Aizawa, "Significance of softmax-based features in comparison to distance metric learning-based features," *IEEE transactions on pattern analysis and machine intelligence*, no. 5, 2019.

[11] A. Deshpande, "A beginner's guide to understanding convolutional neural networks. cs undergrad at ucla (2019)," 2018, accessed on 19.7.2018.

[12] A. Kumar and S. Srivastava, "Object detection system based on convolution neural networks using single shot multi-box detector," *Procedia Computer Science*, 2020.

[13] G. Chandan, A. Jain, H. Jain, and Mohana, "Real time object detection and tracking using deep learning and opencv," in *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2018.

[14] E. Karami, M. Shehata, and A. Smith, "Image identification using sift algorithm: performance analysis against different image deformations," *arXiv preprint arXiv:1710.02728*, 2017.

[15] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006.

[16] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *European conference on computer vision*. Springer, 2006.

[17] A. Goldenshluger and A. Zeevi, "The Hough transform estimator," *The Annals of Statistics*, no. 5, 2004.

[18] F. C. Tsai, "Geometric hashing with line features," *Pattern Recognition*, no. 3, 1994.

[19] N. Singla, "Motion detection based on frame difference method," *International Journal of Information & Computation Technology*, no. 15, 2014.

[20] M. Ren, J. Yang, and H. Sun, "Tracing boundary contours in a binary image," *Image and Vision Computing*, no. 2, 2002.

[21] M. Jankowski, "Erosion, dilation and related operators," *Department of Electrical EngineeringUniversity of Southern Maine Portland, Maine, USA*, 2006.

[22] F. Y. Shih and C. C. Pu, "A skeletonization algorithm by maxima tracking on euclidean distance transform," *Pattern Recognition*, no. 3, 1995.

[23] C. Xiang, Z. Zhang, Y. Qu, D. Lu, X. Fan, P. Yang, and F. Wu, "Edge computing-empowered large-scale traffic data recovery leveraging low-rank theory," *IEEE Transactions on Network Science and Engineering*, no. 4, 2020.

[24] X. Ma, K. Ji, B. Xiong, L. Zhang, S. Feng, and G. Kuang, "Light-yolov4: An edge-device oriented target detection method for remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021.

[25] K. Micko, F. Babic, P. Papcun, and I. Zolotova, "Temporary parking via computer vision and deep learning," (in press).

[26] K. Micko and F. Babic, "Evidencia dočasného parkovania pomocou metód počítačového videnia a umelej inteligencie," *Electrical Engineering and Informatics XII*, 2021.

[27] Y. Gong, L. Liu, M. Yang, and L. D. Bourdev, "Compressing deep convolutional networks using vector quantization," *CoRR*, 2014.

[28] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "A survey of model compression and acceleration for deep neural networks," 2020.

[29] Q. Qin, J. Ren, J. Yu, L. Gao, H. Wang, J. Zheng, Y. Feng, J. Fang, and Z. Wang, "To compress, or not to compress: Characterizing deep learning model compression for embedded inference," 2018.

[30] S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally, "Eie: Efficient inference engine on compressed deep neural network," *ACM SIGARCH Computer Architecture News*, no. 3, 2016.

[31] Y. Huang, Y. Lu, F. Wang, X. Fan, J. Liu, and V. C. Leung, "An edge computing framework for real-time monitoring in smart grid," in *2018 IEEE International Conference on Industrial Internet (ICII)*, 2018.

[32] P. Papcun, E. Kajati, D. Cupkova, J. Mocnej, M. Miskuf, and I. Zolotova, "Edge-enabled iot gateway criteria selection and evaluation," *Concurrency and Computation: Practice and Experience*, no. 13, 2020.

Accelerating Slovak Speech Recognition with Transfer Learning Approach

[1]*Anton BUDAY (3rd year),*
Supervisor: [2]*Anton ČIŽMÁR*

[1,2]Dept. of Electronics and Multimedia Communications, FEI TU of Košice, Slovak Republic

[1]anton.buday@tuke.sk, [2]anton.cizmar@tuke.sk

***Abstract*—Limited access to computational resources and low computing power with subsequently prolonged computing time may have a negative impact on research advancement. To tackle this problem, we propose transfer learning recipes for domain-adaptation and cross-language speech recognition (English to Slovak) to increase either accuracy or speed up a training process. Pretrained public LibriSpeech model employs Slovak BN-TUKE (APVV17) dataset consisting of speech and corresponding text data. At first, in this paper, we focus on available datasets, text processing and describe a pipeline for transfer learning in speech recognition tasks. Preliminary byte-pair encoded Slovak model is briefly mentioned in experimental results. Currently, there are ongoing training processes to achieve our own preliminary speech recognition models with a transferred knowledge.**

***Keywords*—byte-pair encoding, cross-language, domain adaptation, espnet, fine-tuning, slovak, speech recognition, tokenization, transfer learning**

I. Introduction

Acceleration of deep learning tasks is really important, especially for business needs and research advancement. Therefore, we propose the recipe on how to leverage publicly available end-to-end (E2E) pretrained English LibriSpeech model to adapt for our own, not publicly accessible Slovak (SK) language dataset, namely BN-TUKE (APVV17) dataset. Since training of end-to-end-like neural network, e.g. Transformer or Conformer from scratch might take couple of weeks, even on high-end GPUs to achieve any results at all, transfer learning (TL) approach may emerge as a useful way to reduce training time and gain speech recognition (ASR) performance. It may help to save researcher's time to optimize whole architecture [1]. In paper [1] researchers claim that if we have similar dataset, it is advantageous to train only last layers of pretrained model, since only these layers are related to classification task - classification of CTC logits in E2E ASR case. CTC logits signify CTC layer outputs before normalization process. Both TL or related fine-tuning (FT) technique may help with different speech-to-text (S2T) domain applications [2], e.g. incorporating children's speech dataset to boost its baseline recognition performance.

II. Related prior work

Our first work [3] dealt with theoretical concepts and issues of E2E ASR, mainly using previous, now obsolete ESPnet (version 1) toolkit. After acquisition of theoretical grounds, subsequent work aimed to verify feasibility of using Slovak S2T corpus to join Transformer [4] neural network architecture. It made use of a feature-rich and state-of-the-art E2E speech processing toolkit known as ESPnet (version 1) [5].

III. Available datasets

Details of ASR suitable datasets are provided in this section.

At first, we had to create a new recipe in ESPnet (v2) for TL purposes. Additionally, in order to train ASR models, one needs to facilitate suitable data. So far we have been able to obtain huge GigaSpeechXL dataset, MyST Boulder Corpus and "Táraninky" children's speech dataset.

A. GigaSpeech

This dataset contains up to 10k hours of English speech from YouTube, various podcasts, audiobooks. Free access to dataset is conditioned that you use this dataset solely for educational purposes.

We hope that ASR researchers around Kaldi will also release their GigaSpeech pretrained model so that we can review EN-SK performance in cross-language (X-lang) transfer learning between Kaldi and ESPNet2. A bottom-up approach requires enormous computing power that we do not have within a reasonable time frame.

B. MyST/Boulder Corpus

My Science Tutor Corpus comprises 393 hours of children's speech, it was collected from nearly 1.4k 3rd to 5th grade students. It is free of charge unless used for commercial purposes. The recordings involve diverse spoken dialogues with a virtual science tutor.

C. Slovak children's speech dataset

Laboratory of Speech and Mobile Technologies (LRMT) has an access to a low-resource (only 4 hours) Slovak children's speech dataset from TV show "Táraninky". Potentially, we could harness fine-tuning to boost domain-adaptation performance in TL approach.

TABLE I
SUMMARY FOR PLAUSIBLE TL APPLICATIONS

ASR TL Task	Task Description	Eligible Corpus
EN-SK X-lang	FT SK on EN	LibriSpeech-APVV17
EN-SK child X-lang	FT SK on EN	MyST-"Táraninky"
SK-SK child X-lang	FT SK on adult SK	APVV17-"Táraninky"

X-lang refers to cross-language domain adaptation, FT is a fine-tuning teqhnique and word "child" stands for children's speech TL approach. Fine-tuning is a related TL method that typically utilizes unfrozen pretrained model to be retrained on completely new data with a very low learning rate.

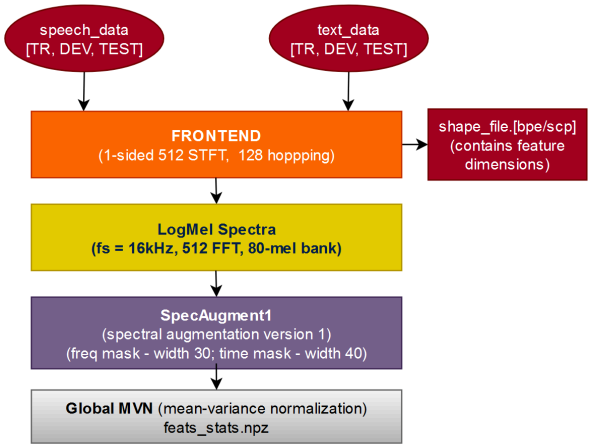


Fig. 1. Preprocessing pipeline of pretrained LibriSpeech English model. Originally, these blocks were fed with English input data. For TL purposes, Slovak dataset is used. TR stands for training set, DEV is a validation set and TEST represents testing data [6]. Inputs/outputs have a claret colour.

IV. Transfer Learning Setup

In our research, we prepare input speech & text data which are then converted to spectral features. Later, logarithmic mel-spectral coefficients are computed. For the first time, comparing to [3][5], we apply data augmentation (spectral augmentation).

To fine-tune (FT) our TL model, first we need to facilitate inputs filled with claret color in Fig. 1 by running two ESPnet2 stages from scratch. Shape file contains information for mini-batching algorithms to handle feature dimensions in GPU correctly. After preprocessing, we are currently retraining Transformer architecture with speech, text and dimensional information from APVV17 corpus. Meanwhile, we have not frozen any layer yet, since we do X-lang TL and follow the first recipe in Tab. I. The main goal is to increase recognition accuracy. Fig. 2 shows a layerwise pretrained model structure which we utilize.

V. Preliminary results and next work direction

Byte-pair encoding (BPE) is a type of tokenization and compressing data algorithm. Fig. 3b illustrates possible next work directions and Fig. 3a shows the preview of SK BPE token list.

Acknowledgment

The research presented in this paper was supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the research projects VEGA 1/0753/20, VEGA 2/0165/21 and by Slovak Research and Development Agency project APVV SK-TW-21-0002.

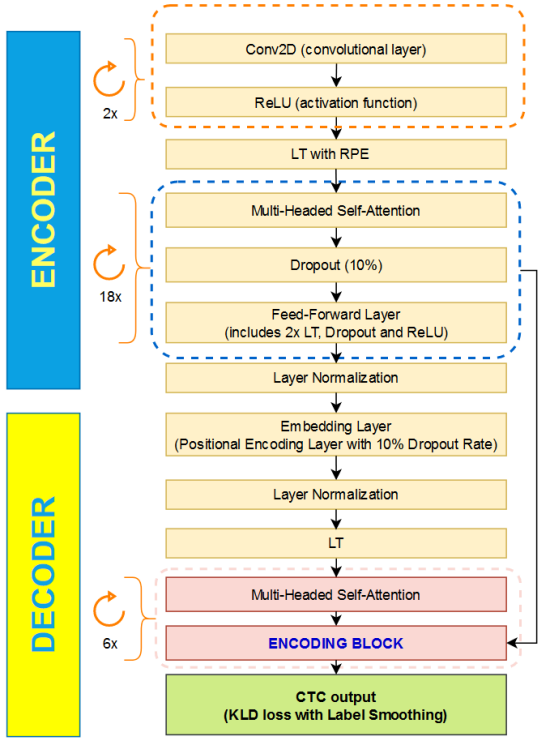


Fig. 2. Deep Learning structure of pretrained TL LibriSpeech model. This model contains around 99 million of parameters totalling nearly 400 MB of data stored in a key-value structure. LT is linear transformation, RPE stands for relative positional encoding, CTC is connectionist temporal classification and KLD refers to Kullback–Leibler divergence loss function.

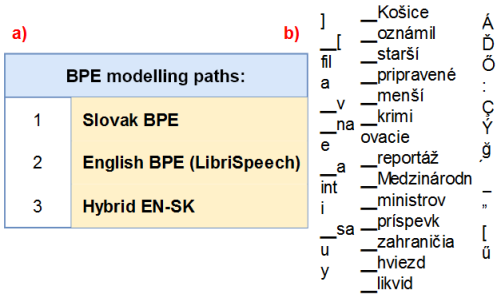


Fig. 3. BPE modelling options for future research. On the left, there are 3 types of BPE models to experiment with. On the right, an example of Slovak BPE tokens is depicted.

References

[1] E. Gayakwad, J. Prabhu, R. V. Anand, and M. S. Kumar, "Training time reduction in transfer learning for a similar dataset using deep learning," in *Advances in Intelligent Systems and Computing*. Springer Singapore, Aug. 2020, pp. 359–367. [Online]. Available: https://doi.org/10.1007/978-981-15-5679-1_33

[2] J. Huang, O. Kuchaiev, P. O'Neill, V. Lavrukhin, J. Li, A. Flores, G. Kucsko, and B. Ginsburg, "Cross-language transfer learning, continuous learning, and domain adaptation for end-to-end automatic speech recognition," 2020.

[3] A. Buday, "End-to-end based speech recognition systems using deep neural networks," in *SCYR 2020: 20th Scientific Conference of Young Researchers, Košice (Slovakia)*, feb 2020, pp. 84–88.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[5] A. Buday, "Application of transformer neural architecture in speech recognition of slovak language," in *SCYR 2021: 21st Scientific Conference of Young Researchers, Košice (Slovakia)*, feb 2021, pp. 143–144.

[6] S. Watanabe, "Espnet pretrained model, shinji watanabe, fs=16khz, lang=en," 2020. [Online]. Available: https://zenodo.org/record/4030511

Author's Index

SCYR 2022: 22nd Scientific Conference of Young Researchers

Proceedings from Conference

Published: Faculty of Electrical Engineering and Informatics
Technical University of Košice
Edition I, 240 pages
Number of CD Proceedings: 50 pieces

Editors: Prof. Ing. Alena Pietriková, CSc.
Assoc. Prof. Ing. Emília Pietriková, PhD.

ISBN 978-80-553-4061-6