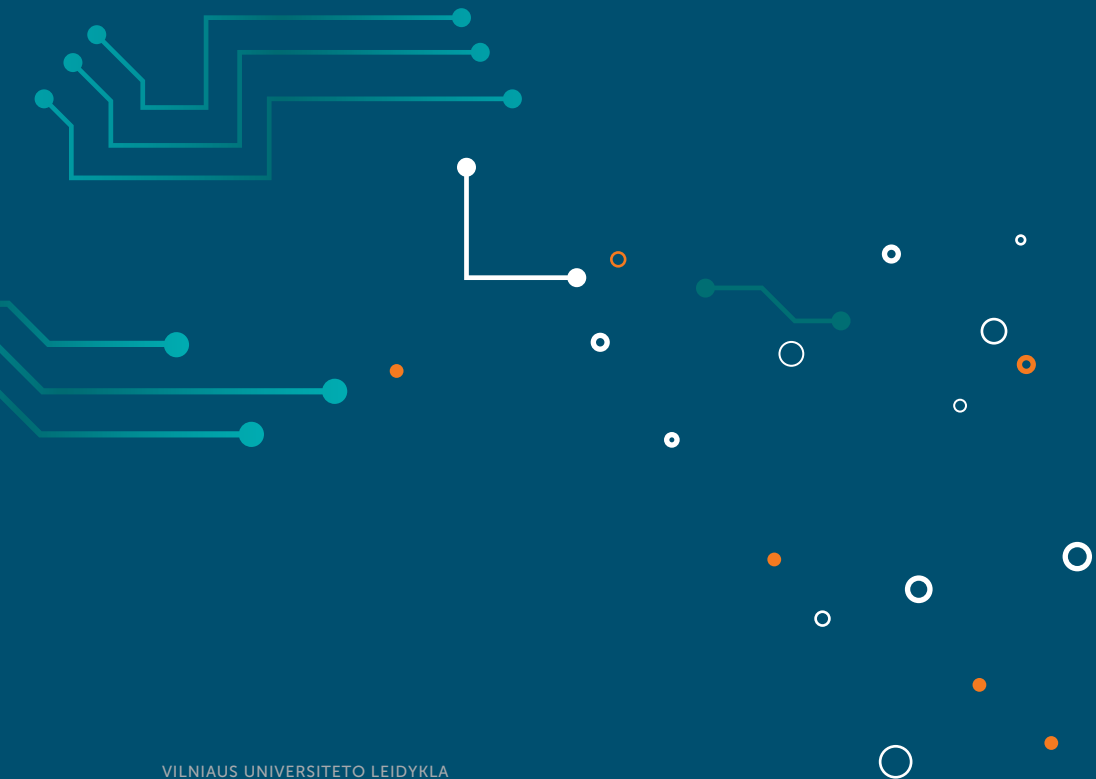




Konferencijos „Lietuvos magistrantų informatikos ir IT tyrimai“ darbai



2025 m.





eISSN 2783-784X

Konferencijos „Lietuvos magistrantų informatikos ir IT tyrimai“ darbai

2025 m. gegužės 13 d.



VILNIAUS
UNIVERSITETO
LEIDYKLA

2025

Programinis komitetas:

Dr. Jolita Bernatavičienė
Prof. habil. dr. Gintautas Dzemyda
Doc. dr. Gražina Korvel
Prof. dr. Olga Kurasova
Doc. dr. Algirdas Lančinskas
Doc. dr. Viktor Medvedev
Doc. dr. Martynas Sabaliauskas
Prof. dr. Julius Žilinskas

eISSN 2783-784X

<https://doi.org/10.15388/LMITT.2025>

Autorių teisės © Autoriai, 2025. Publikavo Vilniaus universiteto leidykla.

Tai yra atvirosios prieigos žurnalas. Žurnalas platinamas vadovaujantis *Kūrybinių bendrijų licencija – Priskyrimas (CC BY)*, kuri leidžia laisvai ir neribotai naudoti, kaip mano esant reikalinga, be kita ko, kopijuoti, daryti pakeitimus ir kurti išvestinius kūrinius, naudoti komerciniais ir nekomerciniais tikslais nurodant informacijos šaltinį ir autorių.

Įžanga

Konferencija „Lietuvos magistrantų informatikos ir IT tyrimai“ – tai jau šeštą kartą vykstantis renginys, skirtas pristatyti naujausius studentų pasiekimus informatikos ir IT srityse. Šio renginio tikslas – ugdyti studentų mokslinių darbų rengimo bei rezultatų pristatymo įgūdžius, supažindinti su kitų studentų atliekamais tyrimais, paskatinti studentus domėtis moksline veikla. Renginys subūrė studentus iš kelių Lietuvos aukštųjų mokyklų. Šiomet norinčių dalyvauti konferencijoje buvo beveik dvigubai daugiau nei anksčiau metais. Konferencijoje savo pranešimus skaitė studentai iš Kauno technologijos universiteto, Vilniaus Gedimino technikos universiteto, Vilniaus kolegijos ir Vilniaus universiteto. Konferencijoje aptariamos įvairiausios temos, nuo programų sistemų kūrimo iki dirbtinio intelekto, ir tai suteikia puikią galimybę dalyviams pagilinti savo žinias, keistis idėjomis bei užmegzti vertingus profesinius ryšius.

Konferenciją organizuoja Lietuvos mokslų akademija (LMA) kartu su Vilniaus universitetu. LMA – tai įstaiga, jungianti žymiausius Lietuvos ir savo veikla su Lietuva susijusius mokslininkus. Ji yra nepriklausoma Lietuvos Respublikos Seimo, Vyriausybės ir jai pavaldžių institucijų ekspertė ir patarėja mokslo bei studijų, kultūros, socialinės raidos, ūkio, gamtosaugos, sveikatos apsaugos, technologijų bei kitais klausimais. LMA įgyvendina šalies mokslui ir eksperimentinei plėtrai reikšmingus ES struktūrinių fondų projektus, rengia šalies bei tarptautines mokslines konferencijas, užsienio mokslininkų seminarus, mokslininkų susitikimus, akademinius skaitymus, parodas. Tikime, kad dalyvavimas šioje LMA kuruojamoje konferencijoje paskatins magistrantus ir kitus studentus tęsti mokslinę veiklą ir pabaigus studijas.

Konferencijos darbuose publikuoti recenzuoti studentų parengti moksliniai straipsniai. Tai dažniausiai pirmosios mokslinės publikacijos, tačiau tikimės, kad ateityje virs į straipsnius prestižiniuose mokslo žurnaluose. To norėtume palinkėti konferencijos dalyviams.

Organizatoriai

dr. Jolita Bernatavičienė

prof. habil. dr. Gintautas Dzemyda

prof. dr. Olga Kurasova

doc. dr. Viktor Medvedev

prof. dr. Julius Žilinskas

Contents

<i>Ieva Andrijauskaitė</i> . Neurotechnology and <i>neural rights</i>	7
<i>Indrė Blagnytė</i> . Multidimensional Visualization of Maternal Health Data	16
<i>Juozapas Rokas Čypas, Viktor Medvedev, Juozas Dautartas</i> . Kenkėjiškų programų aptikimo gerinimas taikant kelių klasių gerybinės programinės įrangos analizę	24
<i>Miglė Gervytė, Tadas Žvirblis</i> . Pointwise Hypothesis Testing of Biomedical Near-Infrared Spectroscopy Signals	28
<i>Mantas Gričius</i> . Širdies plakimo garso įrašų (fonokardiogramų) klasifikavimui skirtų metodų tyrimas	37
<i>Mahammad Ismayilov, Dalia Čalnerytė</i> . Integration of Image Decomposition Methods and CNN for Image Classification	45
<i>Eugenijus Gabrielius Ivanauskas, Saulius Jokubaitis</i> . Vertės pokyčio rizikos vertinimas taikant vienmatį režimų pasikeitimo MS–GJR–GARCH kopulos modelį	55
<i>Nora Jasiukėnaitė, Rūta Levulienė</i> . Vietinių turistų skaičiaus Lietuvos apgyvendinimo įstaigose prognozavimas naudojant „Google Trends“ duomenis.	64
<i>Martynas Jokubaitis, Igoris Belovas</i> . Vilniaus universiteto priėmimo rodiklių analizė ir prognozavimas (XXI a., I ketvirtis)	70
<i>Dominykas Kaminskas</i> . Debesų šalinimas iš ortografinių nuotraukų naudojant giliuosius generatyvinius tinklus	76
<i>Atėnė Kasperavičiūtė, Laura Ringienė</i> . Verslo valdymo sistema – sprendimas efektyvesniam ikimokyklinio ugdymo procesų valdymui	84
<i>Oskaras Klimašauskas, Gintautas Dzemyda</i> . Skatinamojo mokymosi taikymas automobiliui apmokyti sėkmingai pravažiuoti duotu maršrutu	93

<i>Gintarė Košubienė, Saulius Masteika.</i> Implementing EVM-Based Self-Sovereign Identity to Meet European Digital Identity Compliance for Decentralized Finance	99
<i>Glauco Lima, Ernestas Filatovas, Marco Marcozzi, Remigijus Paulavičius.</i> A Review of Quantum-Based Diffusion Models in Generative AI	109
<i>Mantas Mačiūnas, Viktor Skorniakov.</i> Decoding An Automobile's Technical Specification From Its Identification Number	121
<i>Alona Makeieva, Svetlana Kubilinskienė.</i> Virtual Reality Simulation Game for Economics Students: Usability Evaluation	130
<i>Matas Lukšys, Viktor Medvedev.</i> Požymių konvertavimo į vaizdus metodų palyginimas kenkėjiškų programų aptikimo efektyvumui gerinti	142
<i>Arvydas Norgėla, Jurgita Markevičiūtė.</i> Studentų nubyrėjimo Vilniaus universitete vertinimas taikant mašininio mokymo metodus	152
<i>Andrius Paulauskas.</i> Didelio masto orlaivių sistemos koordinavimas pasitelkiant būrio metaforą	160
<i>Irmantas Pilypas, Irma Šileikienė.</i> Atsitiktinio miško modelio taikymas užpildant trūkstamas ekonominių rodiklių reikšmes NUTS 2 lygmeniu	167
<i>Guoda Puslytė, Rūta Levulienė.</i> Samdomojo darbo pajamų praleistų reikšmių įrašymas taikant daugiareikšmį įrašymą	182
<i>Simas Rimašauskas, Igoris Belovas.</i> Recognising the contents in digitised financial documents.	187
<i>Dovydas Šablevičius, Asta Slotkienė.</i> Comparative Evaluation of Speech-to-Text Models for Lithuanian Transcription: Effects of Audio Formats and Recording Environments	197
<i>Kamilė Samusiovaitė, Domas Nemanius, Kristina Lapin.</i> Naudotojo apsaugos nuo duomenų viliojimo atakų įrankių palyginimas	209
<i>Evelina Vaitkevičiūtė, Gražina Korvel.</i> Propagandos technikų lingvistinių požymių lyginamoji analizė taikant natūralios kalbos apdorojimo metodus.	218

Matas Valatka, Jolita Bernatavičienė. Neuroniniai tinklais grįstų triukšmo šalinimo EKG signale metodų eksperimentinis tyrimas.	228
Rugilė Vasaitytė. Laiko eilutes transformuojančių į vaizdus, siekiant taikyti CNN klasifikavimo uždaviniui spręsti, metodų tyrimas	238
Dominykas Vilčinskas, Rūta Levulienė. Samdomųjų darbuotojų darbo užmokesčio prognozavimas taikant mašininio mokymosi metodus.	245
Lukas Voveris, Jolita Bernatavičienė. Saulės elektrinių matavimų klasifikavimas remiantis NLP ir mašininio mokymusi.	250
Anton Zagzin, Pijus Zlatkus, Vasilij Savin, Skaidymo metodų vertinimas skirtingiems duomenų bazių tipams	260
Airidas Žaliauskas, Viktor Medvedev. Multimodaliųjų modelių taikymas vaizdų antraščių generavimui lietuvių kalba	271
Dovydas Marius Zapkus, Asta Slotkienė. Quality Evaluation of Large Language Models Generated Unit Tests: Influence of Structured Output	281
Eimantė Žemytė, Potencialių pardavimų baigties prognozavimas naudojant mašininio mokymosi modelius	289
Aidas Žygas, Gražina Korvel. Muzikos garso šaltinių atskyrimo giliojo mokymosi modelio SCNet apmokymas skirtingais duomenų rinkiniais	297

Neurotechnology and *neural rights*

Ieva Andrijauskaitė

Vilniaus universitetas Teisės fakultetas,
Saulėtekio al. 9, Vilnius, 10221 Vilniaus m. sav.
ieva.andrijauskaite@tf.stud.vu.lt

Abstract. Neurotechnology, which includes a range of devices and procedures that interact directly with the human nervous system, has become a pioneering field with the potential to revolutionize medicine, communication and human development. The rapid progress and constant development poses complex serious ethical and legal challenges. Especially, with regard to human rights such as mental privacy and cognitive freedom. The paper explores the definition and examples of neurotechnologies and its applications, their impact on cognitive behavior and emotions, the human rights they affect, and the emerging concept of *neural rights*. Existing systems are analyzed through a legal lens and the need to balance between innovation and ethical imperatives is sought.

Keywords: neurotechnology, human rights, mental privacy, ethics, neurolaw.

1 Introduction

Neurotechnology encompasses a variety of methods and devices designed to create a direct interface between technical systems and the nervous system. These systems, which include electrodes, computers and advanced prosthetic devices, perform two main functions: they record neural signals and translate them into operational commands for external systems, or they actively modulate the brain's activity through electrical or optical stimulation [1]. They have applications in medicine [2], neuroscience [3], the development of artificial intelligence [4], and even in everyday technologies such as wearable brain monitoring devices. From deep brain stimulation to direct Brain-Computer Interfaces – neurotechnologies are opening up new possibilities for treating neurological disorders, improving people's quality of life and even enhancing cognitive abilities. However, technological advances are also accompanied by major ethical, legal and social challenges that require comprehensive regulation and proper control of their ethical use.

The emergence of increasingly advanced neurotechnologies raises fundamental questions about data protection, informed consent, risks of discrimination and even possible fraudulent practices. One of the main issues is mental privacy. If a person's brain activity data is made available without his or her consent, there is a risk that it could be used for commercial, political or even criminal purposes. In addition to that, the impact on cognitive freedom is also relevant: will there be a future possibility of manipulating people's thoughts, attitudes or behavior and if so - what will be the legal safeguards?

The development of neurotechnologies is not only relevant in a scientific and/or medical context, but also in the wider public discourse. The scientific community is approaching this topic through the prism of neuroethics and neurolaw. Neuroethics research, as developed by authors such as Rafael Yuste, Marcello Ienca and Nita Farahany, examines the impact of neurotechnologies on personal autonomy, mental privacy and cognitive freedom. In the field of neurolaw, attention has been paid to legal instruments that can protect human rights in the context of neurotechnologies such as the International Covenant on Civil and Political Rights or The Universal Declaration of Human Rights. However, the concept of "neurorights" is not (yet) incorporated into international legal norms although there has already been a debate on the regulation of neurotechnologies in the EU [5].

The aim of this study is to undertake a comprehensive analysis of existing legal frameworks, to assess their relationship to the protection of human rights, and to assess the extent to which these rights are effectively guaranteed in the context of neurotechnology. Particular attention is paid to the most sensitive and vulnerable areas where existing legal mechanisms may be lacking or in need of considerable improvement. By exploring the interaction between advances in neurotechnology and fundamental human rights, this study seeks to identify potential legal gaps and propose solutions that ensure a balanced approach that promotes both innovation and security. It focuses on the most sensitive areas. In the light of these concerns, the study seeks to answer the key question: how can human rights protection be ensured in the context of the rapid development of neuro-technologies, and what legal mechanisms can most effectively strike a balance between innovation and security? In order to achieve this goal, the paper formulates a number of key objectives. First, it defines the concept of neurotechnologies and provides examples of their application

in order to highlight their impact on different areas of society. Secondly, the impact of neurotechnologies on human behavior and emotions is analyzed, assessing the potential risks and benefits. Thirdly, it examines whether existing human rights safeguards are sufficient to ensure human integrity, freedom of thought and personal autonomy. Fourthly, it presents possible avenues for legal solutions, identifies the areas most in need of improvement and compares existing legal frameworks to determine their effectiveness. The study is based on a literature analysis, comparative and legal research methods.

2 Defining neurotechnology

Neurotechnology is a multidisciplinary field that integrates neuroscience, biomedical engineering, computer science, and artificial intelligence to develop tools and methodologies that interact directly with the nervous system. Neurotechnology refers to devices, systems, and procedures — encompassing both hardware and software — that directly access, monitor, analyze, predict or modulate the human nervous system in order to understand, influence, restore, or anticipate its structure, activity, function, or intentions (speech). Neurotechnology combines elements of neuroscience, engineering, and computing [6].

In recent years, applications have expanded considerably, from medical applications to everyday use. Functional applications of these technologies are particularly obvious and practically relevant such as:

Brain-computer interfaces (BCIs): these are devices that allow direct communication between the brain and external devices, enabling individuals to control computers or prostheses using neural signals. For example, BCIs have been developed to help paralysed people control digital interfaces or robotic limbs. A recent example is a 64-year-old ALS patient who, using a brain implant developed by Synchron, is able to control Amazon Alexa with his thoughts, thus regaining his lost independence [7].

Deep Brain Stimulation (DBS): This is a medical procedure that involves implanting electrodes in specific areas of the brain to treat movement disorders such as Parkinson's disease and, more recently, psychiatric conditions such as depression. DBS modulates neural activity to relieve symptoms. The reviewed studies show preliminary evidence toward fine-tuning DBS surgery to help improve depressive symptoms that are highly

associated with a debilitating disease, thus, providing hope for improving quality of life [8].

Neuroimaging technologies: Techniques such as functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) allow the visualization and monitoring of brain activity. These tools are essential for both clinical diagnosis and research on brain function. For example, EEG is used for real-time monitoring of brain waves to help diagnose epilepsy and other neurological conditions [9].

3 Impact on human behavior

We are already intimately connected to our machines. Researchers at Google calculated this year [Aut. Note – in 2017] that the average user touches their phone nearly one million times annually [10]. This close relationship inevitably has an impact on overall human development. The aforementioned technologies can improve concentration, memory and decision-making processes by targeting specific brain areas associated with higher cognitive functions. They also allow for enhanced information processing and even the creation of direct mechanisms for translating thoughts into actions, which can lead to faster learning and improved cognitive skills.

However, such long-term external modulation of brain activity can lead to neuroadaptive changes that can affect a person's independent thinking, decision-making autonomy, or even lead to unwanted side-effects such as headaches, epileptic seizures, or even destabilization of the emotional state [11]. In addition, data security is at stake: if neurotechnologies collect and analyze brain activity data, there is a risk that this data could be used illegally, violating mental privacy rights [12]. For instance, brain implants like DBS are vulnerable to attack by third parties who want to exert malicious control over the users' brain activity. This risk of modification of a person's brain activity through unauthorized use of neurodevices by third parties, also called as "brainjacking". "Brainjacking" can lead to several harmful consequences, such as: the unauthorized extraction of neural information, violating a person's right to mental privacy. As well as interference with neural implants, such as interruption of stimulation, drainage of battery power, tissue damage or impairment of motor functions, thus violating the right to mental integrity [13].

4 Neural rights

Article 15 of the International Covenant on Economic, Social and Cultural Rights stipulates right of everyone to *enjoy the benefits of scientific progress and its applications*. However, the technological advances are thus having an increasing impact on human rights, with both positive and negative consequences, especially in the areas of mental privacy, cognitive freedom, mental integrity and protection against discrimination. The technologies, capable of interacting directly with the human brain, have the potential to improve health, cognitive abilities and overall quality of life, as mentioned above, but at the same time they raise existential questions about the autonomy of the individual and the right to the integrity of his or her mental processes. Insufficiently regulated development of neurotechnologies can open the way to abuse with potentially irreversible consequences.

Mental Privacy is becoming one of the most vulnerable aspects due to the ability to monitor and interpret brain activity in real time. Traditional data protection measures have been designed to protect information that is consciously shared by the individual, but neurotechnologies are changing this paradigm as they are able to extract unconsciously generated information that even the individual may not be fully aware of, and in some cases even unaware of at all. This problem becomes even more serious when the technologies are used on a mass scale, for example in the workplace or in educational institutions, where employers or school administrations can monitor the cognitive activity of employees and students to assess their productivity, interactions with each other or even their psychological state. Such practices can violate not only the right to mental privacy, but also human dignity, by turning internal thought processes into readable data that can be used without explicit consent [14]. The right to privacy itself is unarguably a fundamental human right, recognized in article 12 of the Universal Declaration of Human Rights stating that: *no one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation. Everyone has the right to the protection of the law against such interference or attacks*, as well as, article 17 of the International Covenant on Civil and Political Rights (ICCPR) and in many other international and regional human rights instruments. According to the United Nations Human Rights Council (HRC), *privacy presumes that individuals should have an area of autonomous development, interaction, and liberty. People should have a “private sphere” with or without interaction with*

others, free from state intervention as well as excessive unsolicited interventions from other uninvited individuals [15]. That gives a glimpse of the legal security to a human private sphere of any kind, including brain. Article 17 ICCPR covers to all interferences with a person's privacy, regardless of whether they emerge from state officials or natural or legal persons. However, the right to mental privacy has not yet been explicitly recognized as a specific human right. And the privacy to any mental states seems to have only implicit protection under distinctive human rights and freedoms, such as, the right to privacy, the right to freedom of thought, and the right to freedom of expression. Nevertheless, the right to mental privacy is a specific right through the lens of neurotechnology and ought to be a potential candidate for a right that is given a particularly high priority.

The right to cognitive liberty could also be referred to as the right to "mental self-determination" [16]. Which refers to the individual's right to control his or her thoughts, decisions and cognitive abilities without external intervention, is also a subject of debate in the context of neurotechnology. *If the right to mental privacy may help protect the mind from external access and inspection, the principle of cognitive liberty has been invoked to protect mental states from external influence and interference* [17]. While technologies can help treat neurological disorders and improve certain cognitive abilities, they can also be used to modify attitudes, behaviors or even patterns of beliefs. In this view, the right to cognitive liberty encompasses a broad spectrum of freedoms and rights such as the *freedom of thought and rumination, the right to self-access and self-alteration, and to consent to or refuse changes to our brains and our mental experiences* [18].

If the power to subtly manipulate people's thinking through neurotechnology were to be acquired by the state or private organizations, it could pose a serious threat to democratic values and individual autonomy. For example, one can imagine situations in which political forces use techniques to influence voter behavior or corporations that apply cognitive assistance to certain employees to make them more productive but lose the ability to independently assess their working conditions and express critical thinking. According to article 9 of the European Convention on Human Rights *Everyone has the right to freedom of thought, conscience and religion; this right includes freedom to change his religion or belief and freedom*. This right has both an internal and an external dimension. The internal aspect – freedom of thought, conscience and religion - is absolute and cannot be

legally restricted. While in contrast, the external dimension - the right to express one's beliefs - can be restricted in certain circumstances. The nature of the internal dimension ensures that no one can interfere with individuals' most private thoughts, whether by forcing them to change their beliefs or by using methods to expose their private thoughts. As the drafters of Article 9 of the European Convention on Human Rights pointed out, this protects individuals not only against state-imposed confessions, but also against forced examinations, judicial practices, or anything that undermines intellectual autonomy and conscience [18].

While the right to mental integrity is explicitly recognized in established human rights frameworks, its precise scope and limitations remain undefined. Its application in the context of neurotechnology therefore remains unclear. However, certain foundational principles have begun to emerge in the European legal area in relation to mental integrity, which provide valuable insights into the ongoing discourse on neurotechnologies and their potential consequences. For example, case law [19] relating to Article 8 of the European Convention on Human Rights emphasizes that mental health is an essential component of private life, which is linked to mental integrity. Furthermore, the European Court of Human Rights has confirmed [20] that mental integrity covers not only physical integrity but also many aspects of a person's identity, including gender identity, sexual orientation, the right to a name and the right to control one's image. That leads also to the importance of non-discrimination as every person has the right to have his physical, mental, and moral integrity respected.

Advances in neurotechnology and artificial intelligence are transforming understanding of the human mind by blurring the boundaries between cognitive processes and technological intervention. These developments are not only redefining the scope of our mental life, but also raising fundamental questions about the relevance of existing human rights protection. As the analysis show, there is a growing recognition that mental privacy, mental integrity and cognitive liberty are fundamental moral principles that need to be addressed more clearly by a legal framework that is more responsive to these technological developments.

5 Conclusion

Neurotechnology is emerging as one of the fastest growing areas of science and technology, offering revolutionary solutions for treating neurological

disorders, enhancing cognitive abilities, and creating human-machine interfaces. The potential of these technologies is enormous: not only can they make everyday life easier, but they can also open up new horizons in the fields of artificial intelligence, analysis of brain activity and even cognitive behavior of a human. However, these advances are accompanied by fundamental ethical and legal issues related to the protection of human rights.

One of the most important challenges is mental privacy, as modern technologies are able to directly collect, analyze and interpret brain activity data. This raises the risk that information obtained without an individual's consent may be used for commercial, political or even criminal purposes, making traditional data protection measures inadequate in this area. Cognitive freedom, meaning, the right to form and express one's thoughts without control, is also an increasingly important issue, as it can be used not only for therapeutic purposes, but also to adjust people's behaviors, attitudes and even decisions without people knowing it. If not properly regulated, these technologies could become a new tool for governments or corporations to manipulate the public mind.

In addition, mental integrity, which includes the right to the integrity of one's own neurological condition, is threatened by possible cyber-attacks on brain implants and the vulnerability of neurotechnological devices. Technological advances open the way not only to brain stimulation, but also to potentially forced interventions that can alter a person's identity or even will. To add, important area is protection against discrimination, as the development of neuro-technologies can reinforce social exclusion dividing people by their race, sexuality or brain capacities.

Implementing more precise and technologically relevant rights would lead to a more efficient legal system. This would allow more effective access to justice in cases of violation of rights, strengthening both individual autonomy and public confidence in technological progress, while maintaining the necessary balance between innovation and security.

All in all, many challenges come with the latest innovations - for ethics, morality and human rights. These include not only data protection, informed consent mechanisms or fairness in access to technology, but also the broader discourse on the direction in which human-technology interactions should evolve in the future. And they should definitely do so in a human-centric way.

References

- [1] Oliver Muller, Stefan Rotter Neurotechnology: Current Developments and Ethical Issues Front. Syst. Neurosci., 13 December 2017 Volume 11 - 2017 | <https://doi.org/10.3389/fnsys.2017.00093>
- [2] Bhidayasiri Roongroj The grand challenge at the frontiers of neurotechnology and its emerging clinical applications Front. Neurol. (2024) <https://doi.org/10.3389/fneur.2024.1314477>
- [3] Andrea Cometa, Jacopo Carpaneto et. al. Clinical neuroscience and neurotechnology: An amazing symbiosis iScience, Volume 25, Issue 10, 105124 (2022) <https://doi.org/10.1016/j.isci.2022.105124>
- [4] Berger S., Rossi F. AI and Neurotechnology: Learning from AI Ethics to Address an Expanded Ethics Landscape Communications of the ACM Vol. 66 No. 3 (2023) <https://cacm.acm.org/research/ai-and-neurotechnology/>
- [5] Neurotechnology and neurorights - Privacy's last frontier (2023) <https://www.europarl.europa.eu/thinktank/de/events/details/neurotechnology-and-neurorights-privacy-/20231019WKS05721>
- [6] UNESCO's recommendation on the Ethics of Neurotechnology <https://unesdoc.unesco.org/ark:/48223/pf0000391074>
- [7] Shapiro L. Man controls Alexa with thoughts via brain-computer interface *ALS News today* (September 19, 2024) <https://alsnewstoday.com/news/man-controls-alexa-thoughts-brain-computer-interface/>
- [8] N. Simay Gökbayrak, Irene Piryatinsky, Rebecca A. Gavett, Omar J. Ahmed Mixed effects of deep brain stimulation on depressive symptomatology in Parkinson's disease: a review of randomized clinical trials Front. Neurol. (2024)
- [9] <https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/electroencephalogram-eeeg>
- [10] Yuste, R., Goering, S., Arcas, B. et al. Four ethical priorities for neurotechnologies and AI Nature 551, 159–163 (2017). <https://doi.org/10.1038/551159a>
- [11] Ienca, M., Andorno, R. Towards new human rights in the age of neuroscience and neurotechnology. Life Sci Soc Policy 13, 5 (2017). <https://doi.org/10.1186/s40504-017-0050-1>
- [12] Yuste, R., Goering, S., Arcas, B. et al. *supra note 10*
- [13] Pycroft L, Boccard SG, Owen SLF, Stein JF, Fitzgerald JJ, Green AL, Aziz TZ. Brainjacking: Implant Security Issues in Invasive Neuromodulation World Neurosurg. 2016; 92: 454–62
- [14] Ligthart S, Ienca M, Meynen G, et al. Minding Rights: Mapping Ethical and Legal Foundations of 'Neurorights.' Cambridge Quarterly of Healthcare Ethics. 2023;32(4):461–481. <https://doi.org/10.1017/S0963180123000245>
- [15] (A/HRC/39/29: The right to privacy in the digital age - Report of the United Nations High Commissioner for Human Rights, 3 August 2018, para 5
- [16] Ienca, M., Andorno, R., *supra note 11*
- [17] Ligthart S, Ienca M, Meynen G, et al., *supra note 14*
- [18] Ligthart S, Kooijmans T, Douglas T, Meynen G. Closed-Loop Brain Devices in Offender Rehabilitation: Autonomy, Human Rights, and Accountability. Cambridge Quarterly of Healthcare Ethics. 2021;30(4):669-680. <https://doi.org/10.1017/S0963180121000141>
- [19] ECtHR 26 November 2009, appl.no. 25282/06 (Dolenc/Croatia)
- [20] ECtHR (GC) 29 March 2016, appl.no. 56925/08 (Bédat/Switzerland)

Multidimensional Visualization of Maternal Health Data

Indrė Blagnytė

Vilnius University, Faculty of Mathematics and Informatics,
Naugarduko g. 24, Vilnius, Lithuania
indre.blagnyte@mif.stud.vu.lt

Abstract. Visualizing multidimensional health data poses challenges in selecting methods that effectively reveal patterns and separations. This study evaluates five visualization techniques for maternal health risk data: scatter plot matrix, parallel coordinates, RadViz, principal component analysis (PCA), and multidimensional scaling (MDS). Both standardized and normalized data are used to assess group separation effectiveness. Direct visualization methods and PCA show limited separation, especially for medium-risk. MDS with Manhattan distance and standardized data provides the best separation. Results show that the visualization method determines the ideal scaling approach, with no single technique universally optimal for multivariate health data.

Keywords: Multidimensional Visualization, scaling, direct visualisation, PCA, MDS.

1. Introduction

Data visualisation can be a useful tool for public health specialists and researchers to support decision making; however, choosing the best visualization technique can be tricky [1][2]. Health data tends to be complex, and visualising it in understandable ways has applications not only for diagnostic purposes but also for public communication of health information [3]. Health data tends to be multivariate, however, most common visualisation techniques are only meant for two or three-dimensional data [3][4]. For this specific multidimensional data, visualization techniques are necessary [4][5].

A different obstacle is choosing the best scaling method for health data [6]. A 2021 study by M. Ahsan explored the effect of six different scaling methods on Machine learning models using health data. The scaling methods affected the model accuracy, however, no one method was universally best.

In this study, the data chosen was maternal health risk data that has been collected from hospitals and clinics in rural areas of Bangladesh [7].

It was first gathered for the purpose of diabetes research, but later applied for evaluating pregnancy risk. The goal of this analysis is to find the best type of visualization method for showing group separation of this kind of health data. As well as exploring the differences between standardization and normalization for this visualization task, and finding the preferred one.

Five visualization methods were chosen for the visualization task. Three direct visualization methods: scatter plot matrix, parallel coordinates, and RadViz, and two dimensionality reduction techniques: principal component analysis (PCA) and multidimensional scaling (MDS). The scatter plot matrix visualizes all possible pairwise combinations of features as scatter plots [5]. For the parallel coordinates plot, each feature is represented by a parallel axis, and each n-dimensional point is represented by a polyline crossing each of the n axes at the appropriate feature value [5][8]. RadViz generates nonlinear mappings of high-dimensional data onto a plane by modelling a physical spring system where the variables create anchor points [9]. PCA linearly transforms high-dimensional data so that most of the variance is conserved in the first few components, allowing for dimensionality reduction by elimination of the last few components [5][10]. MDS, when used for dimensionality reduction, uses a pairwise distance matrix and tries to find low-dimensional points so that distances between the points in the low-dimensional space are as close to the original proximities in the matrix [5][11]. This uses nonlinear transformations, and the type of distances chosen affects the results of the analysis [11][12].

2. Dataset

The chosen dataset was Maternal Health Risk data from the UCI Machine Learning Repository. The data set has 1013 instances, 7 total features, 6 of which are numerical and 1 categorical. The categorical variable represents predicted risk level during pregnancy, the possible values being: low-risk, mid-risk, and high-risk. The numerical variables are as follows: age, systolic or upper blood pressure, diastolic or lower blood pressure, blood glucose levels, body temperature, and resting heart rate. All analyses were done with both normalized (to range [0,1]) and standardized (mean 0, standard deviation 1) data, and results compared. The visualisation was judged on group separation by the categorical variable, where the best group separation was determined visually based on two criteria: the visible distinctness between clusters representing different risk groups, and the reduced overlap area among groups compared to other methods.

3. Direct visualization and comparison to Random Forest.

Three direct visualization methods were attempted: scatter plot matrix, parallel coordinates plot, and RadViz plot. The standardized and normalized data plots were almost identical; only the standardized data plots were included. Random Forest analysis was also performed to compare the variable importance with the significance of variables in the direct visualization methods. Random Forest showed that blood sugar had the highest importance for predicting risk level, with all others lagging behind, and body temperature showing the lowest importance (Fig. 1).

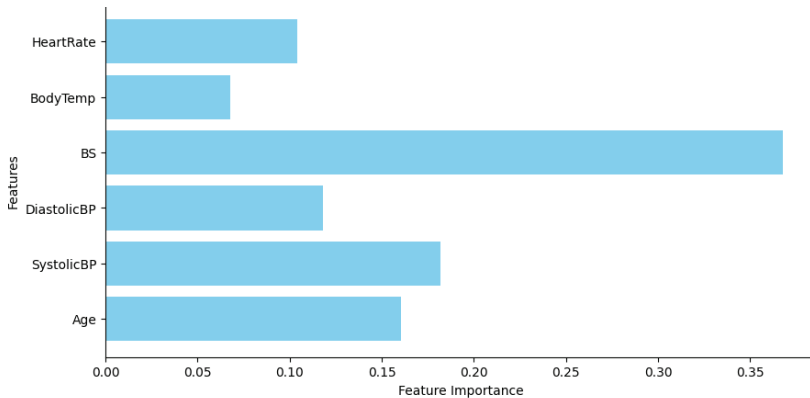


Figure 1. Feature importance using the Random Forest algorithm. The highest importance is shown by blood sugar. Generated using Python 3.11.9.

3.1 Scatter plot matrix

From the individual feature graphs, we can see low-risk and medium-risk groups following a similar distribution and high-risk groups differing from these (Fig. 2). The exceptions to this are blood sugar, where all the distributions differ, and body temperature, where high and medium-risk are more similar. For the pairwise plots, there can be seen some separation between high-risk and low-risk in all pairs with blood sugar, especially in the blood sugar and age pairwise plot. This is consistent with blood sugar showing the highest importance for Random Forest. However, there is still a lot of separation, and the medium-risk group is not separated from the other groups.

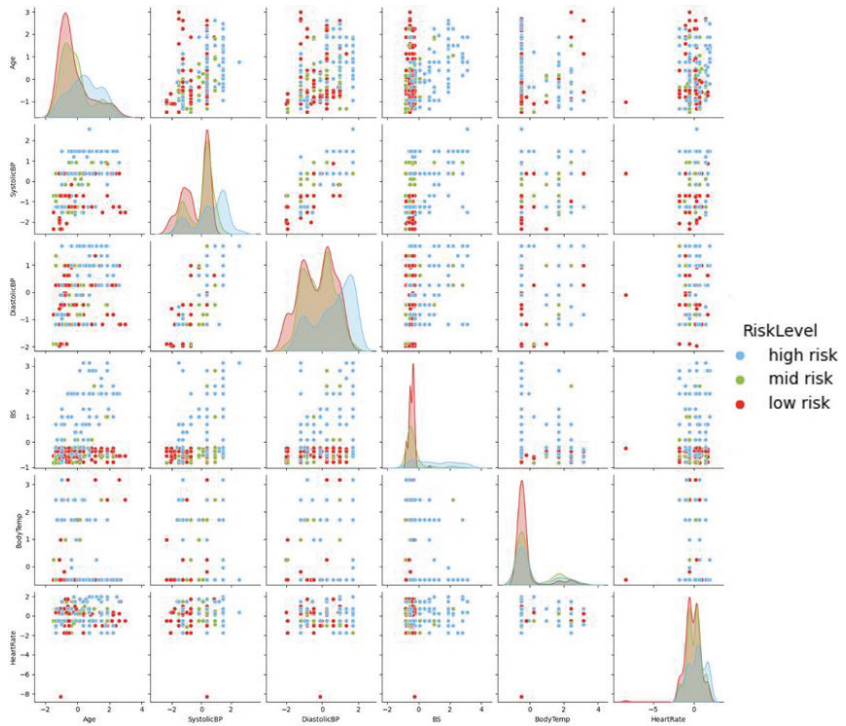


Figure 2. Scatter plot matrix of the standardized Maternal Risk data. The clearest separation between groups can be seen in the blood sugar versus age plot, where high-risk cases tend to form a distinct cluster away from low and medium-risk groups. However, medium and low-risk groups still show significant overlap across most feature combinations. Generated using Python 3.11.9.

3.2 Parallel coordinates plot

The ranges of the groups in the parallel coordinates plot (Fig. 3) overlap significantly, so the groups do not have separation. The mean lines, however, show some separation for all features except body temperature, where the high-risk and mid-risk lines overlap, which might explain the low importance of this variable in Random Forest. Overall, high-risk shows the best separation from the other groups.

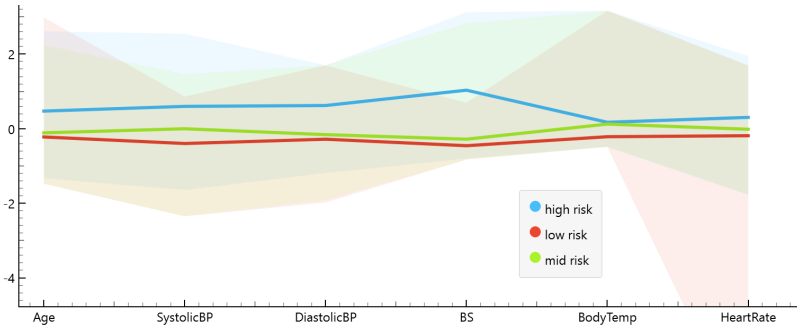


Figure 3. Parallel coordinates plot of the standardized Maternal Health Risk data, with the individual lines hidden and only range and mean shown. The high-risk group's mean line consistently deviates from the low and medium-risk groups, especially for blood sugar and systolic blood pressure. Body temperature, however, shows substantial overlap between all groups. Generated using Orange 3.38.1.

3.3 RadViz plot

RadViz plot (Fig. 4) shows some separation between high-risk and low-risk. However, there is still a good amount of overlap. The medium-risk group overlaps heavily with both high-risk and low-risk. Interestingly, the groups seem to separate out mostly by heart rate, which showed quite small importance in Random Forest, and not blood sugar, which showed the largest importance.

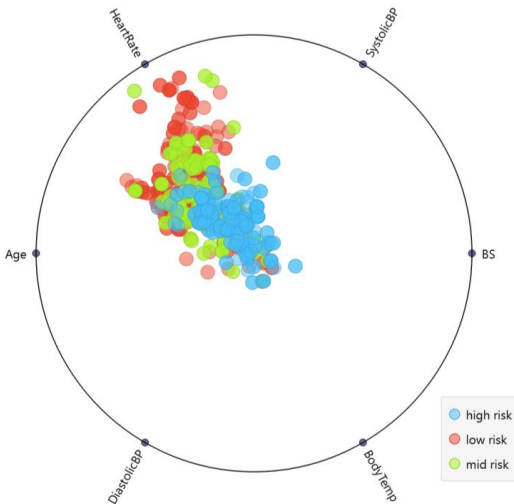


Figure 4. RadViz plot of the standardized Maternal Health Risk data. Some separation can be seen between high-risk and low-risk groups, mainly along the heart rate anchor. Generated using Orange 3.38.1.

4. Principal Component Analysis (PCA)

For the standardized data, the variance explained by the first two principal components was 62.53%, and for normalized data, 70.24%. Both PCA plots showed a lot of overlap (Fig. 5A, B). The most separated group was the high-risk group, with medium and low-risk groups being deeply overlapped. Of the two versions, the normalized data was slightly preferable because of the similar amount of separation and higher explained variance. RadViz plot arguably showed better overall group separation than PCA visualization, though not significantly, and therefore, PCA was not necessarily optimal for visualizing this dataset.

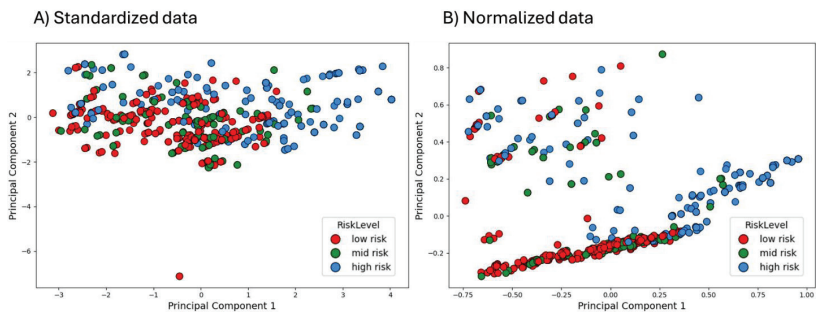


Figure 5. Scatter plot for the first two principal components: A) using standardized data, B) using normalized data. Although the high-risk group forms a somewhat distinct cluster, the medium and low-risk groups overlap heavily along the first two principal components. Generated using Python 3.11.9.

5. Multidimensional Scaling (MDS)

Multidimensional scaling was attempted using standardized and normalized data, using both Euclidean and Manhattan distances. Visualizations using Euclidean distances did not show any better separation than PCA. Using Manhattan distances with the normalized data showed significantly better separation of the high-risk group from others when compared to both PCA and direct visualization methods (Fig. 6A). However, the plots showed poor separation of the low and medium-risk groups. For the standardized data, the separation for the high-risk was not as good, however, the overall separation between all three groups was better (Fig. 6 B). Of all visualization methods, MDS using Manhattan distances and standardized data showed

the best group separation between all groups. Better than PCA and better than all direct visualization methods.

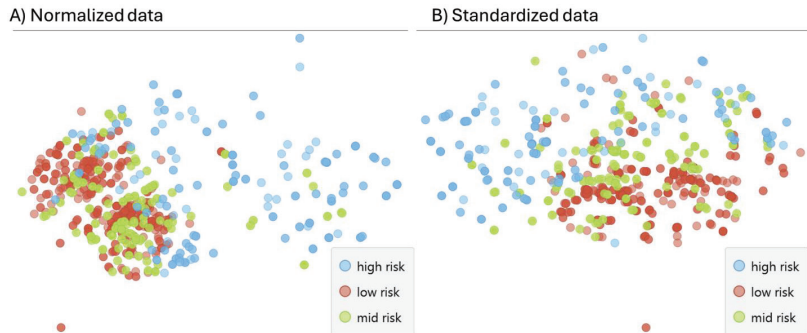


Figure 6. MDS plot using Manhattan distances and PCA initial solution: A) done with normalized data, B) done with standardized data. High-risk instances form a distinct cluster, while medium and low-risk groups overlap. Standardised data show improved low and medium-risk group separation. Generated using Orange 3.38.1.

6. Conclusions

This study explored visualization methods for maternal health risk data and the impact of standardization versus normalization. To objectively assess the quality of group separation, visual inspection was combined with consideration of group compactness and distinctness from other groups. All direct visualisation methods showed limited separation, with medium-risk being most overlapped. Choosing between normalization and standardization did not impact this separation. PCA was also not optimal for showing group separation, with normalization being preferred due to higher explained variance of the first two components. MDS using Manhattan distance and standardized data showed the clearest differentiation between low, medium, and high-risk groups, with less visual overlap than PCA, RadViz, or direct visualization techniques. Thus, based on these qualitative criteria and consistent visual patterns across repetitions, MDS was judged the most effective method. In terms of Normalization vs standardization, neither was preferred for all methods, and therefore, while standardization was optimal for the best method, the conclusion was that the method of visualization dictates the scaling method just as much as the nature of the data itself.

References

- [1] Park, S., Bekemeier, B., Flaxman, A., & Schultz, M. (2021). Impact of data visualization on decision-making and its implications for public health practice: a systematic literature review. *Informatics for Health and Social Care*, 47(2), 175–193.
- [2] Austin, R. R., Mathiason, M. A., & Monsen, K. A. (2022). Using data visualization to detect patterns in whole-person health data. *Research in Nursing & Health*, 45, 466–476.
- [3] O'Connor, S., Waite, M., Duce, D., O'Donnell, A., & Ronquillo, C. (2020). Data visualization in health care: The Florence effect. *Journal of Advanced Nursing*.
- [4] Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1998). An investigation of methods for visualising highly multivariate datasets. *Case Studies of Visualization in the Social Sciences*, 43.
- [5] Dzemyda, G., Kurasova, O., Žilinskas, J. (2012) Multidimensional Data Visualization: Methods and Applications. Springer Optimization and Its Applications. Springer New York.
- [6] Ahsan, M. M., Mahmud, M. A. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. *Technologies*, 9(3), 52.
- [7] Ahmed, M., Kashem, M.A., Rahman, M., & Khatun, S. (2020). Review and Analysis of Risk Factor of Maternal Health in Remote Area Using the Internet of Things (IoT). Lecture Notes in Electrical Engineering.
- [8] Moustafa, R.E. (2011), Parallel coordinate and parallel coordinate density plots. *WIREs Comp Stat*, 3: 134-148.
- [9] Rubio-Sánchez, M., Raya, L., Diaz, F., & Sanchez, A. (2015). A comparative study between radviz and star coordinates. *IEEE transactions on visualization and computer graphics*, 22(1), 619-628.
- [10] Ringnér, M. (2008). What is principal component analysis?. *Nature biotechnology*, 26(3), 303-304.
- [11] Buja, A., Swayne, D. F., Littman, M. L., Dean, N., Hofmann, H., & Chen, L. (2008). Data Visualization With Multidimensional Scaling. *Journal of Computational and Graphical Statistics*, 17(2), 444–472.
- [12] Muller, S. T. (2022). *Distance, similarity, and multidimensional scaling*

Kenkėjiškų programų aptikimo gerinimas taikant kelių klasių gerybinės programinės įrangos analizę

Juozapas Rokas Čypas, Viktor Medvedev, Juozas Dautartas

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Duomenų mokslo ir skaitmeninių technologijų institutas,
Akademijos g. 4, LT-08412, Vilnius, Lietuva
juozapas.cypas@mif.stud.vu.lt

Santrauka. Šiame tyrime siūloma metodika, apimanti gerybinės ir kenkėjiškos programinės įrangos kategorizavimą, siekiant pagerinti kenkėjiškų programų aptikimo tikslumą. Siekiama išanalizuoti, kaip skirtingų programų požymiai padeda atskirti kenkėjišką programinę įrangą nuo gerybinės. Metodika grindžiama statinės analizės duomenimis ir šiuolaikiškais duomenų apdorojimo bei vizualizavimo metodais.

Raktiniai žodžiai: kibernetinis saugumas, gerybinė programinė įranga, kenkėjiškos programinės įrangos aptikimas, programinės įrangos klasifikavimas, gilieji neuroniniai tinklai, mašininis mokymasis.

1 Įvadas

Šiandieninėje skaitmeninėje erdvėje kibernetinio saugumo svarba yra kaip niekad aktuali. Kibernetinių atakų metu kasmet padaroma vis didesnė žala, kuri pastaruoju metu savo dydžiu užima trečios didžiausios pasaulio ekonomikos vietą po JAV ir Kinijos [1]. Į globalius karinius konfliktus įsitraukusios šalys skiria milžiniškas lėšas kibernetiniam karui. Kenksminga programinė įranga (angl. *malware*) technologinio vystymosi akivaizdoje sparčiai tobulėja, tad kibernetinių atakų atpažinimas ir prevencija yra itin svarbus iššūkis. Praktikoje naudojami mašininio mokymosi metodai norint išmokyti modelius atpažinti ir užkirsti kelią kibernetinėms grėsmėms įrenginiuose ar tinkluose.

2 Kenkėjiškų programų analizės metodai

Kenkėjiškos programos, kurios dažniau šiomis dienomis vadinamos virusais, yra analizuojamos dviem metodikom: statine ir dinamine analize. Stati-

nės analizės metu tyrėjai gilinasi į tokias failo savybes kaip sekcijų kiekis, failų tipas, kompiliatorius, esami žodžių junginiai failė (angl. *strings*), *Windows API* kvietimai ir pan. Šios analizės metu failas nėra paleidžiamas, siekiama patikrinti jo išorines savybes jo neatidarant, nėra gaunama informacija apie įtartinę failo funkcionalumą. Nors šis analizės būdas turi savo privalumų, jį gana lengva apeiti modifikuojant failo savybes. Todėl siekiant susidaryti pilnesnį vaizdą apie failo funkcionalumą, taip pat yra taikoma dinaminė analizė, kurios metu failas yra paleidžiamas izoliuotoje aplinkoje (angl. *sandbox*) siekiant išnagrinėti paleidžiamo proceso savybes ir veiksmus. Šios analizės metu galime pastebėti tokius dalykus kaip sisteminius kvietimus (angl. *system calls*), į kokius IP adresus bando kreiptis procesas ir pan. Tačiau siekiant atpažinti dar nematytas virusų grėsmes dažnai yra naudojami mašininio mokymosi arba giliojo mokymosi metodai, kuriems apmokyti naudojami statinės ir dinaminės analizės metu surinkti duomenys. Tačiau duomenų rinkiniai, kurie naudojami šių modelių mokymui, neretai būna nesubalansuoti ir kenksmingos programinės įrangos duomenų imtis būna gerokai didesnė nei gerybinės (nekenksmingos, nekenkėjiškos) programinės įrangos (angl. *goodware*, *benign*). Kenkėjiškos programinės įrangos kūrėjai dažnai taiko naujas metodikas bei keičia savo įrankių funkcionalumą, todėl duomenų rinkiniai nespėja pakankamai dažnai atsinaujinti ir prisitaikyti prie pokyčių. Tuo tarpu, nekenksmingos programinės įrangos kūrėjai esminio funkcionalumo nekeičia, nes jų tikslai nėra išvengti aptikimo ar apgauti anti-virusines sistemas. Todėl šio tyrimo metu bus nagrinėjami nekenksmingos programinės įrangos bruožai, suskirstant nekenksmingą programinę įrangą į kelias dažniausiai pasitaikančias grupes, aprėpiančias dažniausiai naudojamą funkcionalumą. Tyrimo metu taip pat bus bandoma palyginti, kaip šių programų bruožai atsiskiria nuo kenksmingos programinės įrangos. Šis tyrimas yra grindžiamas statinės failų analizės metodu, nes šis metodas įprastai yra pirmas antivirusinių programų atliekamas žingsnis norint nustatyti, ar failas yra pavojingas. Statinė analizė duoda greitius rezultatus, tačiau ją galima apeiti manipuliuojant failo metaduomenimis. Dinaminė analizė, nors ir tikslesnė, reikalauja *sandbox* aplinkos bei daugiau resursų, be to, ne visus failus gali pavykti atidaryti *sandbox* aplinkoje.

3 Duomenų rinkinio paruošimas

Paprastai dauguma virusų klasifikavimo modelių yra mokomi remiantis viešai prieinamais duomenų rinkiniais, tokiais kaip EMBER [2] ar SOREL-20M [3].

Šiuose rinkiniuose gerybiniai failai traktuojami kaip viena klasė, toks supaprastinimas lemia klaidingai teigiamus (angl. *false positive*) rezultatus, kai virusai struktūriškai panašūs į tam tikrą nekenksmingos programinės įrangos grupę. Šiame tyrime siūlome naują požiūrį į kenkėjiškų programų aptikimą, analizuojant ne tik kenkėjiškas programas, bet ir įvairias gerybinės programinės įrangos kategorijas.

Šiame tyrime bus pasiūlytas duomenų rinkinio paruošimo metodas, pagrįstas nekenksmingos programinės įrangos kategorizavimu. Siūlomos kategorijos apima ofiso programas, sisteminius įrankius, žaidimus, vaizdo redagavimo programas ir kt., ir vieną kenkėjiškų programų kategoriją. Naudojant pasiūlytą metodą galima geriau suprasti gerybinės programinės įrangos ypatumus ir identifikuoti anomalijas, kurios gali sufleruoti kenkėjišką elgesį. Naudojant statinę analizę, iš kiekvieno programos failo galima išgauti daugiau nei 2000 požymių įskaitant failo antraštės (angl. *header*) informaciją, sekcijų pasiskirstymą, importuojamas funkcijas, simbolių eilutes, baitų pasiskirstymą. Toks didelis požymių skaičius kelia iššūkių tiek skaičiavimų efektyvumui, tiek modelių generalizacijai apdorojant nematytus duomenis. Siekiant sumažinti duomenų dimensiją ir išlaikyti svarbiausią informaciją, planuojama taikyti įvairius dimensijos mažinimo metodus, tokius kaip pagrindinių komponentų analizė (PCA), t-SNE, UMAP ir autoenkoderiai. Dimensijos mažinimas leidžia ne tik sumažinti duomenų požymių skaičių, bet ir vizualizuoti skirtingų programų klasių pasiskirstymą, leidžiantį pastebėti panašumus tarp tam tikrų kenkėjiškų ir gerybinių programų kategorijų.

4 Tyrimo metodika

Šio tyrimo metu bus tikrinama hipotezė ar pasiūlyta programų kategorizavimo strategija gali pagerinti kenkėjiškų programų aptikimo tikslumą. Yra tyrimų, kurie rodo, kad kai kurių kenkėjiškų programų struktūra bei atliekamos funkcijos, norėdamos išvengti aptikimo, siekia imituoti tam tikrų gerybinių programų požymius [4]. Tokie pastebėjimai pabrėžia būtinybę gilinti gerybinės programinės įrangos analizę ir įtraukti ją į kenkėjiškų programų aptikimo modelių mokymą.

Tyrimo metu bus siekiama apmokyti dirbtiniu intelektu pagrįstus modelius, pavyzdžiui, dirbtinį neuroninį tinklą. Kitas būdas yra pamėginti statinės analizės metu išgautus požymius transformuoti į vaizdus (pavyzdžiui, panaudojant GASF, GADF, MTF, GAFMAT metodus [5]) arba kenkėjiškos pro-

graminės įrangos dvejetainį kodą transformuojant į vaizdą [6]. Tokiais būdais gauti vaizdai gali būti naudojami apmokyti konvoliucinį neuroninį tinklą atskirti kenkėjišką įrangą nuo gerybinės.

Padėka: Finansavimą skyrė Lietuvos mokslo taryba (LMTLT), sutarties Nr. S-MIP-24-116.

Literatūra

- [1] Morgan, S. (žiūrėta 2025 m. balandžio 10 d.). Cybercrime to cost the world 8 trillion annually in 2023. <https://cybersecurityventures.com/cybercrime-to-cost-the-world-8-trillion-annually-in-2023/>
- [2] Roth, P., & Anderson, H. S. (2018). EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models. <https://doi.org/10.48550/arXiv.1804.04637>
- [3] Harang, R., & Rudd, E. M. (2020). SOREL-20M: A Large Scale Benchmark Dataset for Malicious PE Detection. <https://doi.org/10.48550/arXiv.2012.07634>
- [4] Yin, H., Lou, B., & Reiher, P. (2023). A Method for Summarizing and Classifying Evasive Malware. (p. 455-470). New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3607199.3607207>
- [5] Budžys, A., Kurasova, O., & Medvedev, V. (2024 m. rugpjūčio 29 d.). Deep Learning-Based Authentication for Insider Threat Detection in Critical Infrastructure. *Artificial Intelligence Review*, 57(10), 1-35. <https://doi.org/10.1007/s10462-024-10893-1>
- [6] Meenpal, T., & Kumar, N. (2019). Texture-Based Malware Family Classification. 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), (p. 1-6). <https://doi.org/10.1109/ICCCNT45670.2019.8944659>

Pointwise Hypothesis Testing of Biomedical Near-Infrared Spectroscopy Signals

Miglė Gervytė, Tadas Žvirblis

Vilnius University, Faculty of Mathematics and Informatics,
Naugarduko g. 24, LT-03225 Vilnius
migle.gervyte@gmail.com, tadas.zvirblis@mf.vu.lt

Abstract. This study uses a pointwise statistical approach to analyze Near-Infrared Spectroscopy (NIRS) signals in preterm infants with and without Patent Ductus Arteriosus (PDA). Three oxygenation signals— SpO_2 , $\text{rSO}_2\text{-1}$ (cerebral), and $\text{rSO}_2\text{-2}$ (renal)—were analyzed across no-PDA, PDA, and hsPDA groups. Smoothed signals were tested using pointwise ANOVA and Tukey HSD to detect significant group differences. Results showed distinct patterns in $\text{rSO}_2\text{-1}$ and $\text{rSO}_2\text{-2}$, with the hsPDA group standing out in $\text{rSO}_2\text{-1}$ and the no-PDA group in $\text{rSO}_2\text{-2}$, demonstrating the value of this method in biomedical signal analysis.

Keywords: Biomedical Near-Infrared Spectroscopy Signals, Pointwise Data Analysis, Functional Data Analysis, Pointwise Hypothesis Testing, Pointwise ANOVA, Pointwise post-hoc Tukey's HSD Test

1 Introduction

Analyzing biomedical signals like Near-Infrared Spectroscopy (NIRS) is challenging due to their continuous nature, high dimensionality, and clinical noise. NIRS, a non-invasive method for monitoring tissue oxygenation [1], is particularly relevant in the context of Patent Ductus Arteriosus (PDA)—a condition where a fetal blood vessel fails to close after birth, potentially leading to serious complications [2]. Since PDA affects tissue oxygenation, NIRS data can offer valuable diagnostic insights.

However, most existing studies simplify NIRS signals to discrete averages, often missing time-based patterns essential for clinical interpretation. This study aims to address that gap by smoothing noisy NIRS signals and applying pointwise hypothesis testing to detect statistically significant differences in oxygenation patterns between preterm infants with different PDA statuses.

2 Related Works

Previous studies on NIRS in neonatology have mainly compared average oxygenation values between PDA groups. Van der Laan et al. used non-parametric tests and found no significant differences [3], while others, including Schwarz and Navikienė, applied ANOVA-based methods and identified some group-level differences, particularly involving the hsPDA group [4, 5]. However, these approaches often rely on summary statistics, missing time-specific patterns in the signals.

Functional Data Analysis (FDA) has been proposed as a more suitable framework for continuous biomedical data, with applications like functional PCA and canonical correlation offering richer temporal insights [6]. Still, its use in clinical research remains limited. This study builds on that gap by using pointwise testing on smoothed NIRS signals to capture interpretable, time-localized differences between groups.

3 Dataset and Methodology

Dataset

Study was carried out in the tertiary-level neonatal intensive care unit of the Neonatology Center, Vilnius University Hospital Santaros Klinikos, from 2017 November to 2020 June. The study was approved by the Vilnius Regional Biomedical Research Ethics Committee (No.158200-17-940-446, issued on 2017 September 12th). And registered at clinicaltrials.com (reg. No NCT04295395). Informed parental consent was obtained before enrolment.

Infants included in the study were very-low-birth-weight (<1500 g), born at <32 weeks gestation, and at least 72 hours old. NIRS measurements were taken using the NONIN SenSmart X-100 system with neonatal sensors (8004CB-NA, EQUANOX™). Sensors were placed on the forehead for cerebral and on the lower back for renal oxygenation. Recordings were continuous over 12 hours, with brief repositioning every 3 hours to prevent skin irritation.

Using NIRS technology, 3 signals had been measured:

- Cerebral oxygenation levels (rSO 2-1);
- Renal oxygenation levels (rSO 2-2);
- Blood oxygenation (SpO 2).

Table 1. Newborn groups description

Group	Description	Number of patients
no-PDA	Newborns with closed ductus arteriosus therefore without PDA	63
PDA	Newborns with hemodynamically insignificant patent ductus arteriosus thus without treatment	41
hsPDA	Newborns with hemodynamically significant patent ductus arteriosus thus with pharmacological treatment	20

Data Preparation

Firstly data had to be prepared. The dataset consisted of continuously collected medical data from preterm newborns, which inevitably included missing values (NAs) due to factors such as signal interruptions or technical issues. To address these missing values, a two-step approach was applied:

- signals with 30% or more NA values for a specific patient were excluded
- signals with less than 30% missing data were imputed using interpolation – for NAs within the signal linear interpolation was used, for NAs at the beginning of the signal the first observed value was carried backward to fill in missing values, for NAs at the end of the signal the last observed value was carried forward to fill in missing values.

Moreover, any oxygenation values lower than 20 were replaced with 20, ensuring that the data aligned with the physiological expectations.

Functional Data Representation and Smoothing

Since NIRS signals were collected every 1 minute over a 12-hour period, Functional Data Analysis (FDA) framework was adopted to transform these high-frequency, discrete observations into smooth, continuous functions. This approach enables more robust statistical analysis by preserving temporal dependencies and reducing noise. For this, B-spline basis functions were used to represent each discrete signal as continuous curve. B-spline basis functions are particularly suited for modeling non-periodic physiological data due to its flexibility and computational efficiency [7]. Each discrete signal is represented using B-spline basis functions:

$$x(t) \approx \sum_{m=1}^M c_m \phi_m(t), \quad (1)$$

where $\phi_m(t)$ are the B-spline basis functions, c_m are the corresponding coefficients, and M is the number of basis functions. The choice of M controls the smoothness of the function: higher values provide flexibility, while lower enforce smoother trends.

To determine the optimal number of basis functions and the degrees of smoothness, Generalized Cross-Validation (GCV) was used:

$$GCV = \frac{n \cdot SSE}{(n - df)^2}, \quad (2)$$

where SSE is the sum of squared errors, n is the number of observations, and df is the effective degrees of freedom. Then the elbow method together with `optim.basis` function from `fda.usc` library [8] were used to determine the optimal number of basis functions.

Hypothesis testing

Pointwise ANOVA

To detect localized differences in NIRS signals across groups, pointwise Analysis of Variance (ANOVA) was applied. Although it does not analyze curves as a whole, it provides valuable insights into the differences between data groups and the statistical significance of those differences [9]. This method checks the null hypothesis that all groups have the same mean value at each time t against the alternative hypothesis that at least one pair of data groups has different mean values at time t . Null and alternative hypothesis can be expressed as:

$$\begin{aligned} H_0 &: \mu_1(t) = \dots = \mu_k(t) \\ H_1 &: \exists i, j \text{ such that } \mu_i(t) \neq \mu_j(t) \end{aligned} \quad (3)$$

To test hypothesis at each time, the total sum of squares (SST) is calculated by summing between group sum of squares (SSB) and the within-group sum of squares (SSW), where:

$$SSB(t) = \sum_{j=1}^k n_j \left(\bar{Y}_j(t) - \bar{Y}(t) \right)^2, \quad (4)$$

where $\bar{Y}_j(t) = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}(t)$ is the mean for group j at the time t and $SSW(t) = \sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij}(t) - \bar{Y}_j(t))^2$.

The F-statistic is then calculated at every time t as:

$$F(t) = \frac{MSB(t)}{MSW(t)}, \quad (5)$$

where $MSB(t) = SSB(t)/(k - 1)$ is the mean square between groups, $MSW(t) = SSW(t)/(N - k)$ is the mean square within groups.

This means that the F-statistic compares the ratio of variance explained by group differences to the variance due to random error. By comparing $F(t)$ to the F-distribution, the p -value and significance of the difference at each time t are determined.

Pointwise Post Hoc Analysis – Tukey HSD Test

To investigate pairwise differences between groups, Tukey's Honest Significant Difference (HSD) test is used as a post hoc analysis. This pointwise test evaluates whether the mean values of two groups differ significantly at each time point t . The difference is defined as:

$$\Delta_{k_1, k_2}(t) = |\bar{Y}_{k_1}(t) - \bar{Y}_{k_2}(t)|, \quad (6)$$

Where $\bar{Y}_{k_1}(t)$ and $\bar{Y}_{k_2}(t)$ are the mean values of groups k_1 and k_2 at the time t , respectively.

Tukey HSD test calculates the critical value for pairwise comparisons using the studentized range distribution. This calculation incorporates the number of groups, the total number of observations, and the mean square within groups (obtained from the ANOVA calculation). The test identifies statistically significant differences between group pairs at each time point, providing a more granular view of group-level differences over time.

4 Results and Discussion

NIRS signals were smoothed using B-spline basis functions to reduce noise and prepare the data for statistical analysis. The optimal number of basis functions was selected for each signal and group using Generalized Cross-Validation (GCV) and validated with the Elbow method.

The number of basis functions varied depending on the signal complexity: hsPDA groups required fewer functions (12–16), while no-PDA groups showed more variability, requiring up to 57. This reflects differences in signal structure across groups.

Figures 1–3 show raw (colored) and smoothed (black) curves for each group and signal: SpO₂, rSO₂-2, and rSO₂-1. The smoothing clearly reduced measurement noise while preserving key physiological trends.

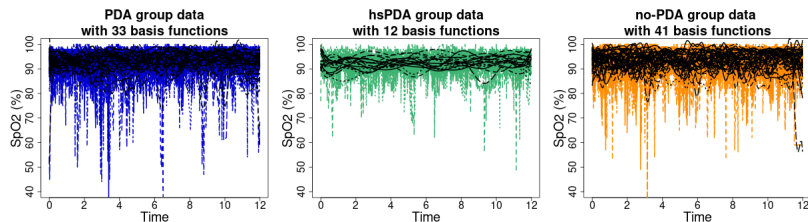


Fig. 1. Raw SpO₂ (blue, green and orange) signals and smoothed curves (black) in newborn groups (PDA, hsPDA, no-PDA): X-axis represents time (hours), and Y-axis represents blood oxygenation levels

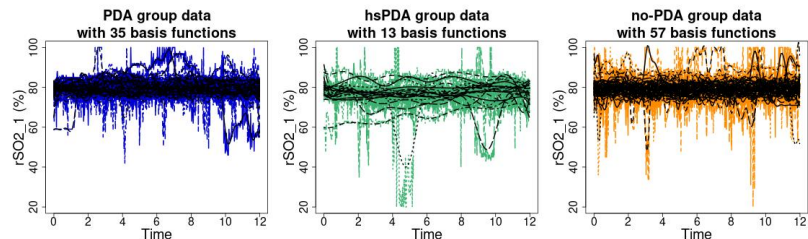


Fig. 2. Raw rSO₂-1 (blue, green and orange) signals and smoothed curves (black) in newborn groups (PDA, hsPDA, no-PDA): X-axis represents time (hours), and Y-axis represents blood oxygenation levels

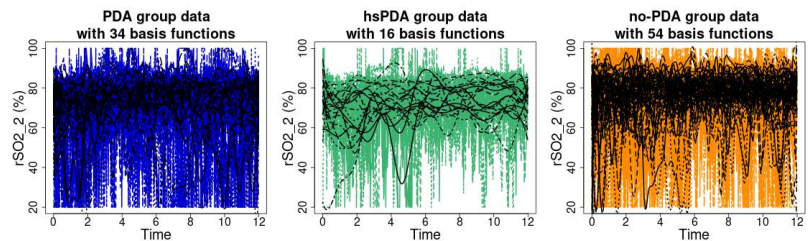


Fig. 3. Raw rSO₂-2 (blue, green and orange) signals and smoothed curves (black) in newborn groups (PDA, hsPDA, no-PDA): X-axis represents time (hours), and Y-axis represents blood oxygenation levels

To identify statistically significant differences in oxygenation signals between newborn groups, pointwise ANOVA and Tukey HSD post hoc tests were applied to the smoothed signals.

As shown in Figure 4, pointwise ANOVA for SpO₂ (blood oxygenation) revealed mostly non-significant differences across groups, with *p*-values rarely falling below the 0.05 threshold. The Tukey HSD plots (Fig. 5) confirm this finding, with no sustained significant pairwise differences. This suggests that arterial oxygen saturation remains relatively stable across PDA statuses.

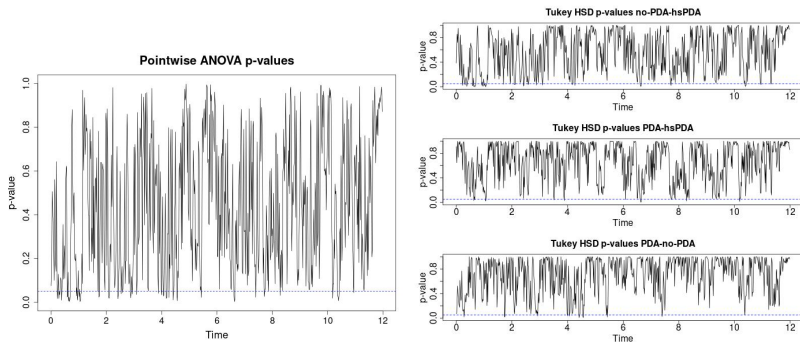


Fig. 4. Pointwise ANOVA *p*-values for rSO₂-1 and pointwise Tukey HSD *p*-values for SpO₂ signals across time

Figure 5 presents pointwise ANOVA results for rSO₂-1 (cerebral oxygenation), with more time periods with significant differences compared to SpO₂ signal. The time period with the most significant differences is between 2 and 6 hours, with additional peaks before 2 hour and more frequent peaks between 6 and 10 hours, while the only period without significant *p*-values is from 10 to 12 hours. The post-hoc test revealed that PDA and no-PDA groups show only a few isolated peaks with statistically significant differences, while hsPDA with no-PDA and hsPDA with PDA pairs show statistically significant differences for longer and continuous time periods (Fig. 5). It shows that the hsPDA group, when compared to the other two groups, has the most significant differences over time.

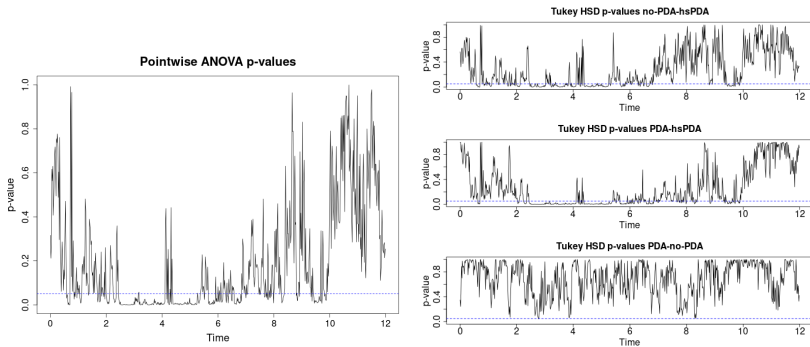


Fig. 5. Pointwise ANOVA p-values for rSO_2-1 and pointwise Tukey HSD p-values for rSO_2-1 signals across time

Figure 6 shows that pointwise ANOVA identified extended time regions with significant group differences in rSO_2-2 (renal oxygenation), particularly between hours 1 and 3, around 6th hour and from 8 to 12 hours, with additional peaks at other times. The Tukey HSD post-hoc test shows that the PDA and no-PDA pair shows the most significant differences, the hsPDA and no-PDA group also exhibit many periods with significant differences, while in contrast, the hsPDA and PDA pair shows the fewest significant differences (Fig. 6). From Tukey HSD tests, it is evident that the no-PDA group differs the most from the other two groups, as it has continuous periods of significant differences with both PDA and hsPDA groups.

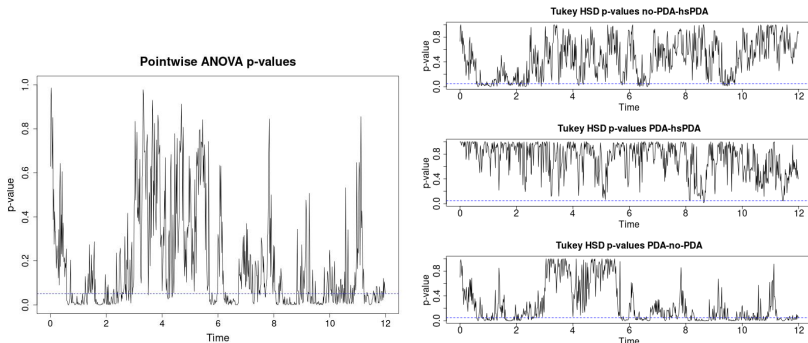


Fig. 6. Pointwise ANOVA p-values for rSO_2-2 and pointwise Tukey HSD p-values for rSO_2-2 signals across time

5 Conclusions

This study applied a pointwise statistical approach to smoothed NIRS signals to investigate group differences in oxygenation patterns among preterm infants with and without PDA. The results showed that while SpO_2 remained largely similar across groups, significant differences were observed in cerebral and renal oxygenation. The hsPDA group showed distinct patterns in $\text{rSO}_2\text{-1}$, while the no-PDA group stood out in $\text{rSO}_2\text{-2}$.

These findings demonstrate that pointwise analysis can reveal time-specific physiological differences that are often missed by traditional summary-based methods. This approach offers a simple yet powerful framework for analyzing biomedical signals and may support more nuanced interpretations in clinical research.

References

- [1] Jöbsis, F. F. (1977). Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science*, 198(4323), 1264–1267.
- [2] Dice, J. E., & Bhatia, J. (2007). Patent ductus arteriosus: an overview. *The Journal of Pediatric Pharmacology and Therapeutics*, 12(3), 138–146.
- [3] van der Laan, M. E., Roofthoof, M. T., Fries, M. W., Berger, R. M., Schat, T. E., van Zoonen, A. G., Tanis, J. C., Bos, A. F., & Kooi, E. M. (2016). A hemodynamically significant patent ductus arteriosus does not affect cerebral or renal tissue oxygenation in preterm infants. *Neonatology*, 110(2), 141–147.
- [4] Schwarz, C. E., Preusche, A., Wolf, M., Poets, C. F., & Franz, A. R. (2018). Prospective observational study on assessing the hemodynamic relevance of patent ductus arteriosus with frequency domain near-infrared spectroscopy. *BMC Pediatrics*, 18, 1–7
- [5] Navikienė, J., Viršilė, E., Vankevičienė, R., Liubšys, A., & Jankauskienė, A. (2021). Brain and renal oxygenation measured by NIRS related to patent ductus arteriosus in preterm infants: a prospective observational study. *BMC Pediatrics*, 21(1), 559.
- [6] Barati, Z., Zakeri, I., & Pourrezaei, K. (2013). Functional data analysis view of functional near-infrared spectroscopy data. *Journal of Biomedical Optics*, 18(11), 117007.
- [7] Ullah, S., & Finch, C. F. (2013). Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology*, 13, 1–12.
- [8] Febrero-Bande, M., & Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: The R package *fda.usc*. *Journal of Statistical Software*, 51(4), 1–28. <https://www.jstatsoft.org/v51/i04/>
- [9] Zhang, J. (2014). Analysis of variance for functional data. *Monographs on Statistics and Applied Probability*, 127, 127.

Širdies plakimo garso įrašų (fonokardiogramų) klasifikavimui skirtų metodų tyrimas

Mantas Gričius

Kauno technologijos universitetas,
K. Donelaičio g. 73, LT-44249 Kaunas
mantas.gricius@ktu.edu

Santrauka. Pasaulio sveikatos organizacijos duomenimis, širdies ir kraujagyslių ligos yra pagrindinė mirties priežastis pasaulyje. Dirbtinio intelekto klasifikavimo algoritmai gali padėti aptikti širdies ligas ankstyvoje stadijoje. Šio darbo tikslas – sukurti algoritmą fonokardiogramų klasifikavimui (tipinis / netipinis signalas) ir ištirti duomenų įtaką klasifikavimo rezultatams. Viena pagrindinių problemų šios srities sprendimuose yra ribotas apmokymo duomenų kiekis. Dažniausiai šiuose tyrimuose naudojamas duomenų rinkinys yra „2016 CinC/PhysioNet iššūkiu“ duomenų bazė. Tačiau šiame duomenų rinkinyje egzistuoja duomenų disbalansas – tipinių garso signalų yra beveik 4 kartus daugiau nei netipinių. Todėl tyrime išbandomi 3 skirtingi klasifikavimo modeliai (atraminių vektorių klasifikatorius (SVM), atsitiktiniai miškai, konvoliucinis neuroninis tinklas (CNN)) ir palyginama, kaip duomenų disbalansas paveikia klasifikavimo tikslumą. SVM klasifikavo netipinius signalus prasčiausiai (jautrumas – 0,6683), o geriausiai klasifikavo CNN modelis (MAcc – 0,8470). Duomenų disbalansui pašalinti buvo atliekama netipinių signalų augmentacija ir sukurti sintetiniai duomenys, kurie išlygino įrašų kiekį kiekvienai klasei. Pašalinus duomenų disbalansą, visi modeliai klasifikavo netipinius garso signalus ženkliai tiksliau, SVM pasiekė net 20% geresnį jautrumo įvertį. O tipinių signalų klasifikavimo tikslumas suprastėjo maždaug 5%.

Raktiniai žodžiai: CNN, GAN, širdies garso įrašai, klasifikavimas, dirbtinių neuronų tinklas, mašininis mokymas.

1 Įvadas

Pasaulio sveikatos organizacijos duomenimis, širdies ir kraujagyslių ligos yra pagrindinė mirties priežastis pasaulyje [1]. Širdies vožtuvų ligos (angl. valvular heart diseases, VHD) yra viena iš pagrindinių širdies ir kraujagyslių ligų priežasčių, dėl kurių visame pasaulyje yra didelis mirtingumas [2].

Ankstyvas tokių problemų diagnozavimas gali padėti sumažinti mirtingumą sustabdant ligos vystymąsi.

Šio darbo tikslas – sukurti algoritmą fonokardiogramų (angl. phonocardiogram, PCG) klasifikavimui (tipinis / netipinis signalas) ir ištirti duomenų įtaką klasifikavimo rezultatams. Sukurto algoritmo ir egzistuojančių sprendimų rezultatų vientisumui užtikrinti, naudojamas „2016 CinC/PhysioNet iššūkiu“ duomenų rinkinys [3]. Tikslui pasiekti atliekami eksperimentai naudojant skirtingus mašininio mokymosi (angl. machine learning, ML) modelius. „2016 CinC/PhysioNet iššūkiu“ duomenų rinkinyje širdies plakimo garso įrašai yra suskirstyti į tipinius ir netipinius, tačiau tipinių fonokardiogramų yra gerokai daugiau (2575) nei netipinių (665). Toks duomenų disbalansas sukuria klasifikavimo modelio šališkumą tipiniams garso įrašams. Dėl šios priežasties naudojami papildomi garsų augmentacijos metodai, tokie kaip Gauso triukšmas (angl. Gaussian noise), laiko tempimas (angl. time stretch), tono keitimas (angl. pitch shift), garso keitimas (angl. shift) bei generatyvusis priešininkų tinklas (angl. generative adversarial network, GAN) sintetiniams duomenims kurti. Tai turėtų padėti pašalinti šališkumą tipiniams signalams, efektyviau apmokyti modelius ir gauti tikslesnius klasifikavimo rezultatus.

2 Egzistuojančių fonokardiogramų klasifikavimo sprendimų ir literatūros analizė

Literatūroje matoma tendencija, jog pasirinktas duomenų rinkinys „Physionet Challenge 2016“ iš ties gausiai paplitęs šios srities moksliniuose darbuose [2, 4, 5, 6, 7] ir kadangi duomenų rinkinys turi labai didelę įtaką klasifikavimo tikslumui, naudodami tą patį duomenų rinkinį galėsime geriau palyginti gautus rezultatus. Požymių išskyrimui naudojami labai įvairūs ir skirtingi metodai. Krishnan et al., [4], 2020 ir Alkhodari et al., [2], 2021 naudojo 3-jų sluoksnių vienmatį konvoliucinį neuroninį tinklą (angl. convolutional neural network, CNN); Shuvo et al., [6], 2021 taip pat naudojo CNN, tačiau 4-ių sluoksnių; tuo tarpu Zabihi et al., [5], 2016 išskaičiavo požymius pasitelkiant 5 skirtingus garso signalų apdorojimo metodus; o Demir et al., [8], 2019 pasitelkė net iš anksto apmokytus giliuosius neuroninius tinklus (angl. pretrained deep neural networks, DNN) VGG 16, VGG 19 ir AlexNet. Analizėje matome, jog dažniausiai naudojamas požymių išskyrimo metodas buvo kelių sluoksnių konvoliucinis neuroninis tinklas. Tyrimų rezultatams tarpusavyje palyginti naudojamas balansuotas tikslumas (angl. bal-

anced accuracy, MAcc). Tyrimai taip pat naudojo skirtingus būdus ir pačiam fonokardiogramų klasifikavimui. Krishnan et al., [4], 2020 ir Zabihi et al., [5], 2016 pasitelkė tiesioginio sklaidimo neuroninius tinklus (angl. feed-forward neural networks, FNN) ir atitinkamai pasiekė 0,8574 ir 0,8590 MAcc. Alk-hodari et al., [2], 2021 naudojo dvikryptį ilgalaikės trumpalaikės atminties (angl. bi-directional long short-term memory, BiLSTM), rekurentinį neuro-ninį tinklą (angl. recurrent neural network, RNN) ir pasiekė 0,8731 MAcc. O geriausią rezultatą pasiekti pavyko Chen et al., [7], 2020 – 0,9391 MAcc. Iš atliktos analizės galime matyti pagrindines širdies plakimo garso įrašų klasifikavimo darbų tendencijas: dauguma darbų modelio apmokymui, va-lidacijai ir testavimui naudoja „Physionet Challenge 2016“ duomenų rinkinį; gan dažnai pasirenkama naudoti konvoliucinius neuroninius tinklus poži-ymių išskyrimui; klasifikavimui dažnai naudojami įvairūs dirbtinių neuronų tinklai (FNN, RNN, CNN). Taip pat verta atsižvelgti ir į esamų sprendimų silpnąsias savybes ir didžiausius iššūkius. Chen et al., [9], 2021 atliko iš-samią fonokardiogramų klasifikavimo darbų analizę ir aptiko pagrindines problemas ir iššūkius:

- Ribotas duomenų kiekis apmokymui;
- Apmokymo efektyvumas ir sparta;
- Giliesiems dirbtiniams neuroniniams tinklams apmokyti reikalingi atminties ir skaičiavimo ištekliai.

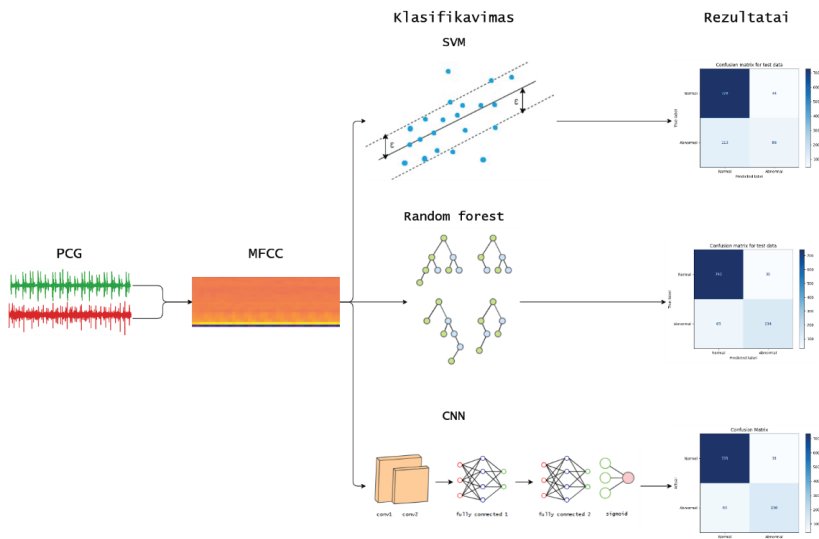
Atsižvelgiant į šiuos iššūkius galime orientuotis į sritis, kurias galima patobulinti ir galimai gauti geresnius rezultatus arba atrasti kaip palengvinti patį klasifikavimo procesą.

3 Eksperimentų eiga

Prieš pradėdant eksperimentus vienas labai svarbus žingsnis yra tinkamai atskirti ir apdoroti duomenis. Kadangi duomenų kiekis nėra labai didelis (3240 įrašai), norint atlikti tikslesnį duomenų testavimą pasirinkta skirti 30% duomenų (972 įrašai) testavimui ir 70% duomenų (2268 įrašai) mode-lio apmokymui. Eksperimentams svarbu ne tik duomenų kiekis, tačiau ir jų kokybė. Peržvelgus duomenis buvo pastebėtas duomenų disbalansas. Iš 2268 įrašų tik 457 įrašai apmokymo imtyje buvo netipiniai, tai reiškia, kad tipinių širdies plakimo garso įrašų apmokymo imtyje yra beveik 4 kartus daugiau nei netipinių. Tai daugumai modelių sukuria šališkumą ir didesnę

tikimybę, kad garsai bus atpažinti kaip tipiniai, kas gali nulemti labai prastą netipinių fonokardiogramų klasifikavimą. Toks duomenų disbalansas suteikia progą palyginti kaip skirtingi modeliai susitvarko su šia problema ir kaip šios problemos pašalinimas įtakoja tiek tipinių, tiek netipinių širdies plakimo garsų klasifikavimą. Požymių išskyrimas yra pirmas žingsnis šios srities tyrimų eksperimentuose. Šiame darbe požymių išskyrimui naudojami Mel cepstrum koeficientai (angl. Mel Frequency Cepstrum Coefficients, MFCC). Pasirinkta naudoti pirmuosius 20 MFCC koeficientų, nes tolimesni koeficientai gali lemti modelio persimokymą. Duomenų rinkinyje garsai yra skirtingos trukmės (nuo 5 iki 120 sekundžių) ir prieš požymių išskyrimą nėra niekaip apkarpmi ar apdorojami siekiant išlaikyti kuo daugiau informacijos. Dėl skirtingos trukmės įrašų, gaunamas MFCC kadrų skaičius kiekvienam koeficientui skiriasi priklausomai nuo garso trukmės. Daugumai modelių naudoti skirtingų matmenų įvesties duomenis nėra priimtina, todėl modelių apmokymui ir testavimui naudojami kiekvieno MFCC koeficiento kadrų vidurkiai, taigi kiekvienas duomenų įvesties įrašas turi lygiai 20 savybių. Eksperimentams pasirinkti naudoti 3 skirtingi modeliai (žr. 1 pav.): atraminių vektorių klasifikatorius (angl. support vector machine, SVM) pasirinktas kaip pirminis modelis, dėl populiarumo šios srities tyrimuose; atsitiktinių sprendimų miškų (angl. random forest) modelis pasirinktas, kadangi dėl savo prigimties ir veikimo būdo gali nesukurti šališkumo ir geriau atlikti klasifikavimą esant duomenų disbalansui; 2-jų sluoksnių CNN modelis dėl literatūroje pasiektų rezultatų. Pasirinkus kelis skirtingus modelius galėsime palyginti, kaip duomenų disbalansas įtakoja modelių klasifikavimo tikslumą. Kiekvienas modelis taip pat turi parametrus, nuo kurių priklauso klasifikavimo tikslumas. SVM modeliui pasirinkta naudoti radialinės bazinės funkcijos branduolį (angl. Radial Basis Function kernel, RBF), dėl gebėjimo geriau prisitaikyti ir klasifikuoti netiesinius duomenis. Atsitiktinių miškų algoritmui pasirinkta naudoti 50 sprendimų medžių siekiant užtikrinti pusiausvyrą tarp modelio tikslumo ir skaičiavimų efektyvumo. Minimalus įrašų skaičius medžio padalijimui pasirinktas numatytasis parametras – 2 įrašai, tai užtikrina, kad medžiai būtų pakankamai detalūs ir užfiksuotų sunkiai aptinkamas savybes. Minimalus įrašų skaičius lapuose (galutinėse medžių viršūnėse) pasirinktas taip pat 2 įrašai, siekiant šiek tiek sumažinti persimokymo (angl. overfitting) riziką, tačiau vis dar išlaikant gan aukštą medžio detalumą. Kadangi atliekama binarinė klasifikacija (šir-

dies garsai klasifikuojami į tipinius ar netipinius) CNN pasirinkta naudoti binarinės kryžminės entropijos (angl. Binary Cross Entropy, BCE) nuostolių funkciją. Kiti CNN parametrai reikalavo šiek tiek eksperimentavimo. Išbandžius keletą skirtingų optimizatorių, pasirinktas Adafactor optimizatorius ir 0.01 mokymosi greitis. Pasirinkus parametrus ir atlikus eksperimentus, tolimesniu žingsniu, siekiama įvertinti, kaip modelio rezultatai kinta naudojant subalansuotą duomenų rinkinį, kuriame abiejų klasių įrašų skaičius yra vienodas ar bent jau panašus. Tam reikia arba sumažinti tipinių garso įrašų kiekį, arba padidinti netipinių įrašų kiekį. Kadangi duomenų kiekis nėra labai didelis, jį sumažinus būtų sudėtinga sėkmingai apmokyti modelį, taigi reikia padidinti netipinių įrašų kiekį. Šiam tikslui pasiekti atliekama netipinių garso įrašų augmentacija pasitelkiant Gauso triukšmą, laiko tempimą, tono keitimą, garso keitimą. Gauso triukšmas suteikia papildomą triukšmą garso įrašams ir imituoja realistinę aplinką (pvz., kūno judesiai, elektroninio stetoskopo triukšmas). Tai padeda modeliui geriau prisitaikyti prie triukšmingų signalų ir tiksliau atpažinti netipinių širdies garsų savybes. Laiko tempimas padeda imituoti įvairius netipinius širdies garsus su skirtingais širdies ritmais (lėtesnis – bradikardija, greitesnis – tachikardija). Tono keitimas simuliuoja skirtingų pacientų anatomiją ir stetoskopo padėtį – dėl kūno masės ar padėties širdies garsai gali turėti aukštesnį arba žemesnį toną, tokių garsų įvairovė leidžia modeliui mažiau priklausyti nuo įvairių dažnių ir labiau atkreipti dėmesį į garsų laiką ir struktūrą (pvz. užesius, papildomus garsus). Garso keitimas perstumia garso signalą laike, todėl modelis neprisiriša prie garsų vietos įrašė. Tai padeda modeliui atpažinti reikšmingas savybes bet kurioje įrašo vietoje ir būti nepriklausomam nuo įrašo trukmės. Pasitelkiant šiuos 4 augmentacijos metodus ir juos atsitiktiniu būdu kombinuojant, kiekvienam netipiniam garso įrašui sukuriama po dar 3 sintetinius garso įrašus bei taip subalansuojamas apmokymo duomenų rinkinys. Prijungus sintetinius duomenis prie apmokymo duomenų, gaunamas duomenų rinkinys turi 1811 tipinių širdies plakimo garso įrašų ir 1828 netipinių. Tuomet naudojant papildytą duomenų rinkinį vykdoma ta pati 1 pav. pavaizduota veiksmų seka. Atlikus klasifikavimą kiekvienu atveju apskaičiuojama klaidų matrica, kurios pagalba galima apskaičiuoti ir kitas modelio vertinimo metrikas (tikslumą, jautrumą, F įvertį, specifiškumą).



1 pav. Eksperimentų eigos schema.

4 Rezultatai

Atliktų eksperimentų rezultatai atvaizduoti 1 ir 2 lentelėse. Modelio tikslas tyrime yra suklasifikuoti gautą garso signalą kaip tipinį arba netipinį širdies plakimą. Iš loginės ir probleminės svarbos, svarbiau, kad būtų teisingai klasifikuojami netipiniai širdies plakimo garso signalai, nes šiuo atveju būtų galima įspėti pacientą, kad jam ištis gresia pavojus ir reikia kardiologo konsultacijos. Todėl apžvelgiant rezultatus daugiausiai skirsime dėmesio jautrumo įverčiui, kuris reprezentuoja kiek netipinių širdies plakimo garsų buvo suklasifikuota teisingai. 1 lentelėje galime matyti kokią įtaką turi modelio pasirinkimas apmokant modelį klasifikavimui, kai duomenų rinkinys yra nesubalansuotas ir dominuoja vienos klasės įrašai. Matome, jog SVM modelis labiausiai nukentėjo nuo duomenų disbalanso ir klasifikavo tipinius signalus itin gerai (specifiškumas – 0,9777), o netipinius itin blogai (jautrumas – 0,6683). Nors ir perėjus prie atsitiktinių miškų modelio rezultatai neturėjo ženklų pokyčių, galime pastebėti, jog modelio prigimtis leido šiam modeliui ne taip stipriai prisirišti prie dominuojančios klasės ir klasifikuoti netipinius signalus maždaug 1,5% tiksliau. Na o geriausius rezultatus pasiekti pavyko naudojant CNN modelį: MAcc – 0,8470, jautrumas - 0,7307. Taip pat galime

pastebėti, kad kai modeliai tiksliau klasifikuoja netipinius signalus, tipinių signalų klasifikavimo tikslumas šiek tiek suprastėja (žr. 1 lentelės specifiškumo stulpelį).

1 lentelė. Rezultatai naudojant originalius duomenis (be augmentacijos).

MODELIS	TIKSLUMAS	F1 ĮVERTIS	SPECIFIŠKUMAS	JAUTRUMAS	MAcc
SVM	0,9115	0,7637	0,9777	0,6683	0,8230
RANDOM FOREST	0,9104	0,7654	0,9725	0,6826	0,8276
CNN	0,9135	0,7835	0,9633	0,7307	0,8470

2 lentelėje pateikti modelių klasifikavimo rezultatai kai apmokymo imtyje naudojami sintetiniai duomenys išlyginti įrašų kiekį kiekvienai klasei ir panaikinamas duomenų disbalansas. Galima pastebėti, jog subalansavus duomenų rinkinį visi modeliai ženkliai tiksliau klasifikuoja netipinius garsus. SVM jautrumas pagerėjo net beveik 20% ir kadangi atsitiktinių miškų modelis jau nebeturi pranašumo dėl duomenų disbalanso, SVM pranoksta atsitiktinių miškų modelį klasifikuojant ir netipinius širdies plakimo garso signalus. Tačiau taip pat matomas skirtumas ir 2 lentelės specifiškumo stulpelyje. Nors pokytis ne toks ženklus, pašalinus duomenų disbalansą, modeliai maždaug 5% prasčiau klasifikuoja tipines fonokardiogramas.

2 lentelė. Rezultatai naudojant duomenų rinkinį papildytą sintetiniais duomenimis.

MODELIS	TIKSLUMAS	F1 ĮVERTIS	SPECIFIŠKUMAS	JAUTRUMAS	MAcc
SVM	0,9002	0,7810	0,9188	0,8317	0,8753
RANDOM FOREST	0,8909	0,7644	0,9083	0,8269	0,8676
CNN	0,9135	0,8090	0,9293	0,8557	0,8925

5 Išvados

Nors ir „2016 CinC/PhysioNet iššūkiu“ duomenų rinkinys sudarytas iš mokliškai patvirtintų ir tikrai kokybiškų duomenų, deja, tipinių garso signalų kiekis yra gerokai didesnis nei netipinių, kas sukuria duomenų disbalansą. Taigi ribotas duomenų kiekis apmokymui vis dar lieka viena iš esminių problemų šio tipo tyrimuose. Šiame darbe buvo įvertinta duomenų disbalanso įtaka skirtingiems dirbtinio intelekto klasifikavimo metodams ir kaip pasikeičia tokių dirbtinio intelekto modelių klasifikavimo tikslumas pašalinus

duomenų disbalanso problemą. Duomenų disbalansas didžiausią įtaką turėjo SVM klasifikavimo modeliui, kuris pasiekė tik 0,6683 jautrumą, atsitiktinių medžių algoritmas klasifikavo netipinius širdies plakimo garso signalus maždaug 1,5% tiksliau (jautrumas – 0,6826), o CNN klasifikavo net 9% tiksliau (jautrumas – 0,7307). Duomenų disbalansui pašalinti buvo atlikta augmentacija netipiniams garsams ir sukurta sintetinių duomenų, įrašų kiekiui kiekvienai klasei išlyginti. Naudojant duomenų rinkinį su sintetiniais duomenimis visi 3 naujai apmokyti modeliai klasifikavo netipinius širdies plakimo garso signalus ženkliai tiksliau. SVM klasifikavo netipinius signalus net 20% tiksliau ir pranašok atsitiktinių miškų algoritmą. Tipinius garso signalus, deja, visi modeliai maždaug 5% procentais klasifikavo prasčiau. Taigi, nors ir tipinių signalų klasifikavimas šiek tiek suprastėjo, netipinių garso signalų klasifikavimas tapo ženkliai tikslesnis ir tai parodo kokią įtaką duomenys turi dirbtinio intelekto klasifikavimo algoritams ir net sintetiniai duomenys gali turėti stiprų poveikį siekiant atpažinti širdies plakimo garsus tiksliau.

Literatūra

- [1] Cardiovascular diseases (CVDs). URL: <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-cvds>
- [2] ALKHODARI, M. - FRAIWAN, L. Convolutional and recurrent neural networks for the detection of valvular heart diseases in phonocardiogram recordings. In *Computer Methods and Programs in Biomedicine*. 2021. Vol. 200, p. 105940. doi: 10.1016/j.cmpb.2021.105940
- [3] CLIFFORD, G.D. ir kt. Recent advances in heart sound analysis. In *Physiological Measurement*. 2017. Vol. 38, no. 8, p. E10. doi: 10.1088/1361-6579/aa7ec8
- [4] KRISHNAN, P.T. ir kt. Automated heart sound classification system from unsegmented phonocardiogram (PCG) using deep neural network. In *Physical and Engineering Sciences in Medicine*. 2020. Vol. 43, no. 2, p. 505–515. doi: 10.1007/s13246-020-00851-w
- [5] ZABIHI, M. ir kt. Heart sound anomaly and quality detection using ensemble of neural networks without segmentation. In *2016 Computing in Cardiology Conference (CinC)*. 2016. p. 613–616.
- [6] SHUVO, S.B. ir kt. CardioXNet: A Novel Lightweight Deep Learning Framework for Cardiovascular Disease Classification Using Heart Sound Recordings. In *IEEE Access*. 2021. Vol. 9, p. 36955–36967. doi: 10.1109/ACCESS.2021.3063129
- [7] CHEN, Y. ir kt. Classification of heart sounds based on the combination of the modified frequency wavelet transform and convolutional neural network. In *Medical & Biological Engineering & Computing*. 2020. Vol. 58, no. 9, p. 2039–2047. doi: 10.1007/s11517-020-02218-5
- [8] DEMIR, F. ir kt. Towards the classification of heart sounds based on convolutional deep neural network. In *Health Information Science and Systems*. 2019. Vol. 7, no. 1, p. 16. doi: 10.1007/s13755-019-0078-0
- [9] CHEN, W. ir kt. Deep Learning Methods for Heart Sounds Classification: A Systematic Review. In *Entropy*. 2021. Vol. 23, no. 6, p. 667. doi: 10.3390/e23060667

Integration of Image Decomposition Methods and CNN for Image Classification

Mahammad Ismayilov, Dalia Čalnerytė

Kaunas University of Technology,
Studentų g. 50, Kaunas, Lithuania
mahammad.ismayilov@ktu.edu, dalia.calneryte@ktu.lt

Abstract. This study explores integrating Haar wavelet decomposition techniques with convolutional neural networks for image classification on the MNIST dataset. The research demonstrates that without losing significant accuracy by applying the 1-level, 2-level, and 3-level decomposition techniques, the model can reduce the dimensionality and the number of parameters required by the convolutional neural network model. During the training, the 1-level Haar CNN results achieved optimal performance, demonstrating competitive accuracy and computational efficiency compared to the baseline CNN model. This approach highlights the potential of wavelet decomposition techniques to enhance CNN performance with limited computational resources.

Keywords: Convolutional Neural Network (CNN), Haar Wavelet Decomposition, Image Classification, Image Decomposition, MNIST Dataset.

1 Introduction

Image classification is one of the main types of image processing tasks, employed in the stages of diagnosing diseases [1], automatic defect identification [2,3], surveillance systems [4], and other problems. A hybrid deep neural network model was proposed to classify different eczema types [1], a convolutional neural network (CNN) model was implemented to identify flaws in the real-time food packaging control [2], and random forest model was employed in real-time home textile fabric defect inspection system [3]. Image classification problem can be defined as assigning a class for an image based on its content. Artificial intelligence (AI) methods like CNN enabled to automate processes and therefore reduce the human power needed to perform the repetitive tasks. It also enabled transferring experts' knowledge and making the decision processes easier.

One of the main problems in the image processing tasks is the collection of the dataset that is large enough to train the model. Since most of the AI

methods are based on the idea that models learn the features relevant for the classes from the raw data, the training procedure usually requires large number of images in the dataset to achieve practically acceptable results. Moreover, complex datasets that demonstrate high diversity in the features representing the objects from the same class lead to the need for training large models and this results in the need for large computational and memory resources. Conventional image processing technologies, such as Fourier transformation or discrete wavelet decomposition, can be applied to highlight representative features in the image and speed up the training process of the classification model. Such integration of conventional image preprocessing techniques and AI methods ensures the sustainable pipeline for image classification that balances computational resources and accuracy.

In this paper, the methodology to integrate conventional image decomposition methods with convolutional neural networks is presented. It investigates the impact of multi-level decomposition levels as a preprocessing step to reduce input dimensionality and the number of parameters. The approach is experimentally demonstrated with the MNIST (Modified National Institute of Standards and Technology) dataset, which is a subset of a larger NIST handwritten digits dataset from 0 to 9 with 60,000 examples of the training set and 10,000 examples of the test set, where each image is 28x28 pixels in size [5].

2 Related Work

Various integrations of wavelet decomposition and CNN were presented in research related to image classification. It is possible to simply use wavelet decomposition as feature extraction method and use the obtained features in machine learning model for classification. For example, Haar wavelet decomposition features were applied to monitor meat quality and classify it as fresh, frozen, and rotten based on its texture [6]. The property of the Haar decomposition to extract features of repetitive structures was employed to develop a real-time home textile fabric defect inspection machine [3]. After several modifications, such as brightness compensation and Gaussian blur filter, the results of Haar wavelet transform were used as input in Random Forest Classifier. The developed system demonstrated high accuracy and short inference time [3].

One of the most popular approaches of wavelet decomposition and CNN combination is wavelet pooling in CNN [7]. Although it demonstrated good performance in texture classification and image annotation tasks compared to the conventional CNN models with much higher number of parameters, the wavelet pooling is a computationally expensive procedure and can be difficult to implement in practical tasks [8].

Another approach is based on image feature extraction using wavelet decomposition and combining the obtained results with output of different CNN levels in various ways. The wavelet transformation was applied to decompose hyperspectral images of three tea types (black, green, and yellow) and used as input for a method based on a lightweight CNN and support vector machine [9]. The CNN-enhanced multi-level Haar wavelet features fusion network was proposed to alleviate the issue that ordinary CNN learns mainly spatial characteristics and do not take into account the spectral features [10]. Spectral features extracted using Haar transforms were combined with conventional CNN output at different layers in the fire surveillance system [4]. The proposed model resulted in lower computational cost and reduced number of false alarms compared to the ordinary CNN. The wavelet function was incorporated into CNN blocks as activation function in the model developed to detect breast cancer [11]. The wavelet decomposition was used to replace the early convolution layers in the deep learning model design to classify medical images [12]. It was demonstrated that fixed feature extraction method reduced the number of parameters the neural network needs to learn and therefore speeds up the training process. In addition, the models with wavelet decomposition were more stable compared to the one that used convolutions, thus, it leads to wider application possibilities in the data-limited domains [12]. The issue related to data insufficiency to train the deep CNN model for hyperspectral polarimetric synthetic aperture radar (PolSAR) imagery classification stimulated incorporating Haar wavelets as an effective feature extraction technique for a three-branch deep CNN model in order to improve accuracy and mitigate noise [13].

3 Methodology

The proposed method, shown in Figure 1 below, first applies decomposition to the images. Then, the decomposition results are stacked into the array for multi-channel input in the CNN model, which afterward is trained and evaluated.

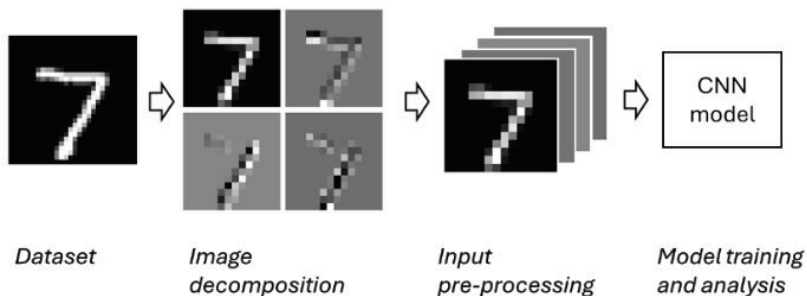


Figure 1. The proposed method

3.1 Image Decomposition

The Haar wavelet transform was initially introduced for signal processing, and it was called the 1D wavelet transform. This wavelet transform is based on decomposing every signal into approximation and detail coefficients, where the Haar wavelet function is crucial for this decomposition as it defines how the signal is split into detail and approximation components [14]. Mathematically, the function can be written as:

$$\psi(t) = \begin{cases} 1, & 0 \leq t < \frac{1}{2} \\ -1, & \frac{1}{2} \leq t < 1 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Here $\psi(t)$ is the Haar wavelet function, often called the “mother wavelet”, and t is an independent variable, which is typically referred to as position in image processing.

The Haar wavelet transform can be extended to 2D for image decomposition, which can be computed using the 1D Haar wavelet decompositions. The multi-level decomposition technique is a continuation of the non-standard approach for two-level decomposition and more, as the standard method does not inherently define multi-level processing.

From the visualization aspect, Figure 2 illustrates the image decomposition to sub-bands, arranging approximate coefficient (A), horizontal (H), vertical (D), and diagonal (D) components, and so on. The approximation captures the low-frequency part, representing the overall structure of the

image, and the horizontal component contains high-frequency information along the horizontal direction, while the vertical component contains high-frequency information along the vertical direction. Additionally, the diagonal component captures high-frequency variations in both directions.

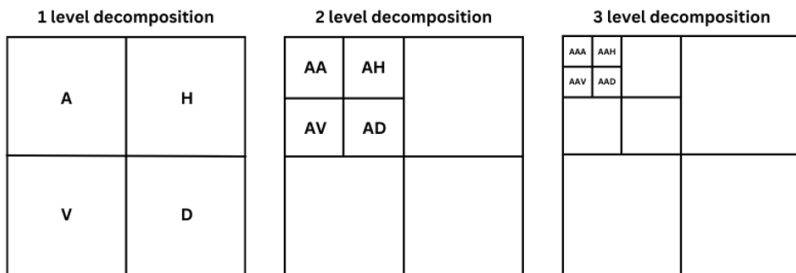


Figure 2. Visualization of the decomposition levels in Haar wavelet transform

Afterward, the 2-level decomposition applies the Haar transform again but only to the approximation sub-band from the first level, which generates a further smoothed approximation (AA) and produces horizontal, vertical, and diagonal details of the approximations (AH, AV, AD). Then, the process repeats the same on the AA sub-band from the 2-level, which creates the third-level approximation (AAA) and horizontal, vertical, and diagonal details of the second-level approximation (AAH, AAV, AAD).

4 Model Evaluation

The convolutional neural network model, integrated with different decomposition levels of the Haar wavelet transform, is defined by the following layers, as shown in Figure 3 below. The main difference between the 1-level, 2-level, and 3-level decomposition is that the CNN model has different number of channels depending on the Haar decomposition level, resulting in different feature resolutions and abstraction levels.

This CNN architecture has been chosen because the MNIST dataset is relatively simple, and even simple models can achieve good accuracy. Moreover, the simplicity of the model is ideal for clearly demonstrating the impact of the Haar wavelet transform.

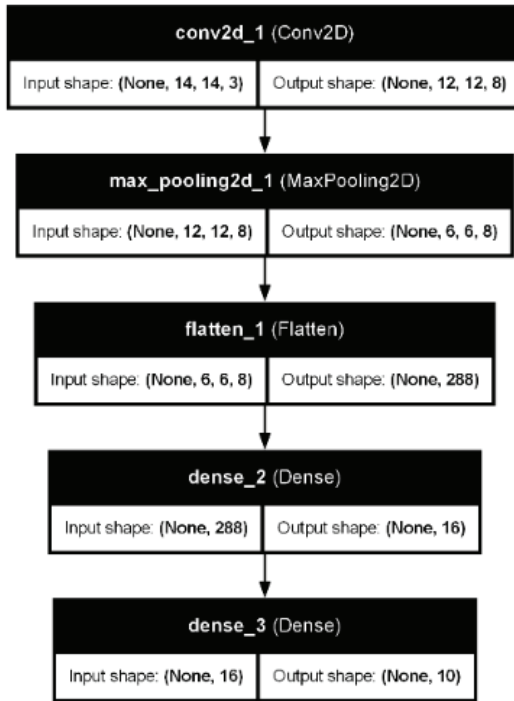


Figure 3. The architecture of the CNN 3-level decomposition model

CNN models with 3 different Haar decomposition levels were compared to a baseline CNN trained on MNIST images containing monochrome 70,000 handwritten digits with 60,000 training and 10,000 test images. Both models share a similar CNN architecture, differing primarily in the input dimensions due to the Haar preprocessing step. After training the models, benchmark results were provided to compare the number of parameters, accuracy, loss, and training time.

Accuracy measures how often the model correctly predicts the outcome:

$$Accuracy = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\hat{y}_i = y_i) \quad (2)$$

where y_i represents the true label of the i^{th} sample in the dataset, \hat{y}_i are the predicted labels produced by the model and $n_{samples}$ is the total number

of samples in the dataset. The function $1(\hat{y}_i = y_i)$ returns 1 if the predicted label matches the actual label.

Categorical Cross-entropy is used as the loss function, which is suitable for multi-class classification where the labels are integer-encoded. For multiclass, the cross-entropy loss formula is shown in equation 3:

$$Loss = - \sum_{i=1}^C y_i * \log(\hat{y}_i) \quad (3)$$

where y_i is the true label from the one-hot encoded target vector, \hat{y}_i is the predicted probability for class i and C is the number of classes.

5 Results

The models experimented with MNIST dataset images and normalized to the [0, 1] range by dividing pixel values by 255 and trained for 5 epochs due to the dataset's simplicity. A Haar wavelet transform was applied at three decomposition levels, initially reducing the image resolutions to 14×14, 7×7, and 4×4 pixels, and the 1-level and 2-level decompositions were resized back to 14×14 pixels for uniform multi-channel CNN input, as shown in Figure 4 below.

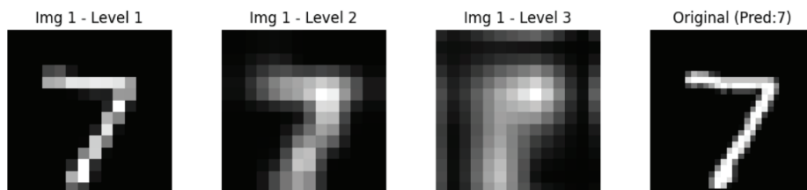


Figure 4. Decomposition level results of CNN model

The CNN models were trained using the TensorFlow library on a laptop equipped with the 12th Gen Intel Core i7-12700H processor, 16 GB of RAM, and an NVIDIA GeForce RTX 3060 GPU with 6 GB video memory.

In general, all models illustrated excellent performance on the MNIST dataset, which is shown in Table 1 below. Nevertheless, comparing the wavelet decomposition levels with the CNN Baseline model, Haar CNN 2-level and Haar CNN 1-level achieved the highest accuracy with 97.50% and 97.45% of results. Moreover, Haar CNN 1-level gained the lowest test loss of 0.0845 and required fewer parameters, indicating slightly better generalization than other models.

Table 1. Benchmark results of Baseline CNN and Haar CNN models

Model	Training Time (s)	Parameters	Test Loss	Test Accuracy
Baseline CNN	9.96	21898	0.0919	0.9725
Haar CNN 1-level	10.87	4874	0.0845	0.9745
Haar CNN 2-level	8.76	4946	0.0893	0.9750
Haar CNN 3-level	8.47	5018	0.1162	0.9651

The unified plots in Figure 5 with training and validation accuracy/loss illustrate that no major overfitting was observed. However, comparing all Haar CNN models, the Baseline CNN model exhibits slightly faster convergence and final accuracy. The loss comparison plot shows a steady decrease in loss for all models, with Haar CNN Level 1 achieving the lowest final loss among the Haar models.

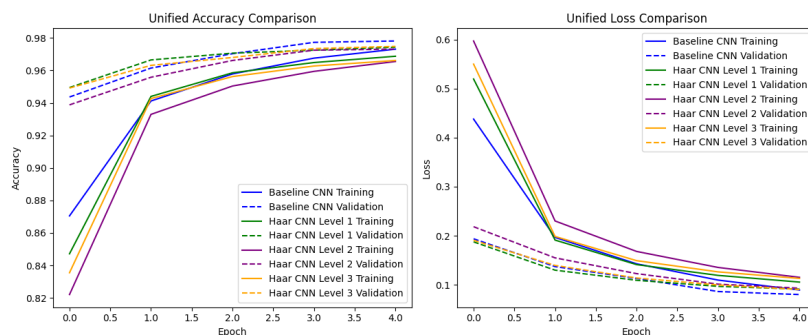


Figure 5. Accuracy and loss comparison results

The Confusion Matrix results in Figure 6 below showed that the Baseline CNN achieves minimal misclassifications for digits like „0“, „1“, „3“, „6“, „7“, and „9“; however, compared with the Haar CNNs, the 1-level decomposition model presents improved accuracy for specific digits such as „5“ and „8“, indicating that the wavelet decomposition technique captures edge and shape variations more effectively. Although Haar CNN 2 and 3 level models offer slight adjustments, they introduce misclassifications in digits like „6“ and „4“, making them less consistent.

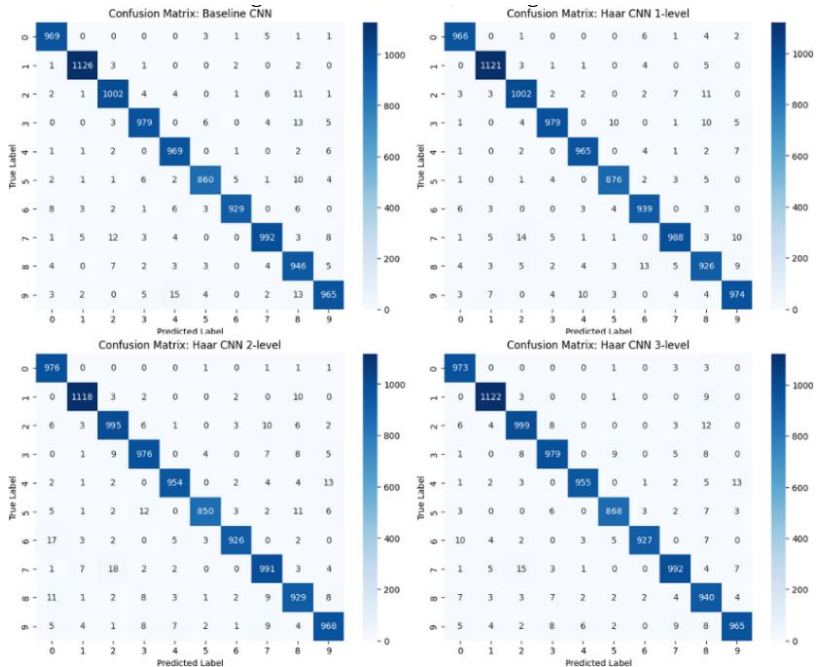


Figure 6. Confusion matrix results

6 Conclusions

To sum up, combining convolutional neural network and Haar wavelet transform has shown that the decomposition method can successfully reduce the dimensionality needed by CNN models. With the lowest test loss and competitive accuracy compared to the baseline CNN model, the Haar CNN with 1-level decomposition shows the best possible balance between accuracy, computational efficiency, and model complexity out of all the assessed models. However, this research was focused on approximation coefficients, omitting the detail coefficients, which contain valuable directional and edge-specific information. Future research will integrate detail coefficients to leverage the Haar decomposition's capabilities fully and will look at methods that methodically lower the number of factors needed at each level of decomposition to improve efficiency and model generalization, which is particularly advantageous in settings with limited resources.

References

- [1] Rasheed, A., Umar, A. I., Shirazi, S. H., Khan, Z., Nawaz, S., & Shahzad, M. (2022). Automatic eczema classification in clinical images based on hybrid deep neural network. *Computers in Biology and Medicine*, 147, 105807. <https://doi.org/10.1016/j.compbiomed.2022.105807>
- [2] Medus, L. D., Saban, M., Francés-Villora, J. V., Batailler-Mompeán, M., & Rosado-Muñoz, A. (2021). Hyperspectral image classification using CNN: Application to industrial food packaging. *Food Control*, 125, 107962. <https://doi.org/10.1016/j.foodcont.2021.107962>
- [3] Barman, J., Wu, H.-C., & Kuo, C.-F. J. (2022). Development of a real-time home textile fabric defect inspection machine system for the textile industry. *Textile Research Journal*, 92(23–24), 004051752211114. <https://doi.org/10.1177/00405175221111477>
- [4] Huang, L., Liu, G., Wang, Y., Yuan, H., & Chen, T. (2022). Fire detection in video surveillances using convolutional neural networks and wavelet transform. *Engineering Applications of Artificial Intelligence*, 110, 104737. <https://doi.org/10.1016/j.engappai.2022.104737>
- [5] Deng, L. (2012). The MNIST database of handwritten digit images for machine learning research [Dataset]. *Papers with Code*. <https://paperswithcode.com/dataset/mnist>
- [6] Kiswanto, K., Hadiyanto, H., & Sedyono, E. (2024). Meat Texture Image Classification Using the Haar Wavelet Approach and a Gray-Level Co-Occurrence Matrix. *Applied System Innovation*, 7(3), 49. <https://doi.org/10.3390/asi7030049>
- [7] Fujieda, S., Takayama, K., & Hachisuka, T. (2018). Wavelet convolutional neural networks. <https://doi.org/10.48550/arXiv.1805.08620>.
- [8] Zhao, L., & Zhang, Z. (2024). A improved pooling method for convolutional neural networks. *Scientific Reports*, 14, 1589. <https://doi.org/10.1038/s41598-024-51258-6>
- [9] Cui, Q., Yang, B., Liu, B., Li, Y., & Ning, J. (2022). Tea category identification using wavelet signal reconstruction of hyperspectral imagery and machine learning. *Agriculture*, 12(8), 1085. <https://doi.org/10.3390/agriculture12081085>
- [10] Guo, W., Xu, G., Liu, B., & Wang, Y. (2022). Hyperspectral image classification using CNN-enhanced multi-level Haar wavelet features fusion network. *IEEE Geoscience and Remote Sensing Letters*, 19, 6008805. <https://doi.org/10.1109/LGRS.2022.3167535>
- [11] Oyelade, O. N., & Ezugwu, A. E. (2022). A novel wavelet decomposition and transformation convolutional neural network with data augmentation for breast cancer detection using digital mammogram. *Scientific Reports*, 12, 5913. <https://doi.org/10.1038/s41598-022-09983-z>
- [12] Masquelin, A. H., Cheney, N., Kinsey, C. M., & Bates, J. H. T. (2021). Wavelet decomposition facilitates training on small datasets for medical image classification by deep learning. *Histochemistry and Cell Biology*, 155(2), 309–317. <https://doi.org/10.1007/s00418-020-01961-y>
- [13] Jamali, A., Mahdianpari, M., Mohammadimanes, F., Bhattacharya, A., & Homayouni, S. (2022). PolSAR image classification based on deep convolutional neural networks using wavelet transformation. *IEEE Geoscience and Remote Sensing Letters*, 19, 4510105. <https://doi.org/10.1109/LGRS.2022.3185118>
- [14] Talukder, K. H., & Harada, K. (2007). Haar wavelet based approach for image compression and quality assessment of compressed image. *IAENG International Journal of Applied Mathematics*, 36(1), 1–9. <https://doi.org/10.48550/arXiv.1010.4084>

Vertės pokyčio rizikos vertinimas taikant vienmatį režimų pasikeitimo MS–GJR–GARCH kopulos modelį

Eugenijus Gabrielius Ivanauskas, Saulius Jokubaitis

Vilniaus Universitetas, Matematikos ir informatikos fakultetas,
Naugarduko g. 24, Vilnius
gabrielius.ivanauskas@mif.stud.vu.lt, saulius.jokubaitis@mif.vu.lt

Santrauka. Vertės pokyčio rizikos rodiklis yra vienas dažniausiai naudojamų pinigų finansų įstaigų sektoriuje galimam nuostoliui vertinti. Šiame darbe pastarasis rodiklis yra vertinamas taikant vienmatį režimų pasikeitimo MS–GJR–GARCH kopulos modelį, kuris modeliuoja ne tik volatilumą bei jo pokytį laike, bet ir atsižvelgia į priklausomybės struktūrą. Nagrinėjant du portfelius prieita prie išvadų, kad labiau diversifikuotam krepšeliui tiksliausiai riziką vertino dviejų režimų MS–GJR–GARCH modelis su asimetriniu Studento–t pasiskirstymu standartizuotoms liekanoms bei *Joe* kopula, o labiau volatiliam portfeliui – identiškas modelis volatilumui, bet su *Gumbel* kopula. Rezultatais akcentuojama kopulų svarba rizikos vertinime, išryškintas kopulų dinamikos modeliavimo poreikis.

Raktiniai žodžiai: vertės pokyčio rizika, režimų pasikeitimas, MS–GJR–GARCH, kopula.

1 Įvadas

Finansų valdyme rizika yra apibrėžiama kaip galimybė, kad grąža iš investicijų arba kitų sandorių bus mažesnė už prognozuotą, o blogiausiu atveju – neigiama [1]. Rizika, prarandant finansus pinigų finansų įstaigų sektoriuje, yra vienas esminių rodiklių, ką šios turi vertinti nuolat. Tai leidžia laiku identifikuoti galimas grėsmes ir imtis prevencinių priemonių, siekiant apsaugoti savo kapitalą ir (arba) klientų lėšas. Nuo tikslaus rizikos vertinimo gali priklausyti finansų paskirstymas, strategija bei jos valdymas, pinigų finansų įstaigų atsparumas šokams kriziniu laikotarpiu, o visa tai formuoja įmonės patikimumą investuotojų akyse. Siekiant prisidėti prie globalaus finansinio stabilumo užtikrinimo ir gairių apibrėžimo 1974 metais buvo įkurtas Bazelio komitetas. Pastarasis tarnavo ir tebetarnauja kaip forumas arba platforma reguliariam bendradarbiavimui pinigų finansų įstaigų priežiūros reikalams [2].

Vertės pokyčio rizika (VPR, angl. *Value-at-risk, VaR*) yra vienas pagrindinių rodiklių vertinant rinkos riziką. Jis yra apibrėžiamas kaip maksimalus galimas nuostolis tam tikru laikotarpiu, esant pasirinktam reikšmingumo lygmeniui. Nors VPR koncepcijos ištakos siekia XX amžiaus pradžią, šio rodiklio vertinimas buvo pradėtas tik pastarojo amžiaus pabaigoje. Tą paskatino 1987 metais įvykęs vertybinių popierių rinkos žlugimas po kurio prasidėjo minėtojo rodiklio skaičiavimo tobulinimai, naujų algoritmų paieška [3]. Vienas dažnesnių metodų vertinant minėtą rodiklį yra volatilumo modeliai. Tačiau dėl to, kad skirtingais laikotarpiais rinkos volatilumas nėra vienodas, praktikoje įprasta nagrinėti modelius, kurie efektyviai tai įvertina – dviejų režimų ARCH modelis [4], trijų režimų daugiamačio GARCH kopulos modelis [5]. Taip pat VPR vertinimui naudojami ir skirtingi kopulų įtraukiantys modeliai, kurie skirti portfelio aktyvų priklausomybės struktūrai įvertinti [6, 7, 8].

Šio straipsnio tikslas – plėsti VPR rodiklio vertinimą bei analizę, skirtingo diversifikavimo krepšeliams rasti tiksliausių modelį VPR vertinimui. Modelis turi būti konstruojamas atsižvelgiant į režimų pasikeitimą, modeliuojant volatilumą MS-GJR-GARCH modeliu bei priklausomybės struktūrą vertinant kopulomis. Režimų pasikeitimo modelio naudojimas grindžiamas jo efektyvumu praktikoje [4, 5, 8], o pasirinktas volatilumo modelis – demonstruojamu tikslumu bei lankstumu (asimetriškumas) [8, 9, 10]. Atradus tinkamą junginį, svarbu apžvelgti modelio savybes bei pagrįsti jo tinkamumą teoriškai.

2 Metodologija

Šiame straipsnyje konstruojami du portfeliai: pirmajame pasirinkta naudoti tris indeksus, atliepiančius Jungtinių Amerikos Valstijų bei Europos rinkas – S&P 500 (indeksas, jungiantis JAV 500 didžiausių įmonių), FTSE 100 (atspindi Londono biržoje prekiaujamų 100 didžiausių įmonių) bei EURO STOXX 50 (indeksas, apimantis 50 didžiausių bendrovių iš euro zonos). Antras krepšelis sudarytas iš trijų pasirinktų akcijų: *Walt Disney*, *Intel* ir *JP Morgan*. Duomenys yra dieninio dažnumo, juose naudojamos dienos uždarymo kainos. Visas duomenų rinkinys yra nuo 2007–03–30 iki 2025–03–16, iš viso 4346 stebiniai.

Antrojo portfelio rizika yra didesnė, grąža gali priklausyti tiek nuo lokalių (šalies/sectoriaus) šokų, tiek nuo ekonominės bei geografinės aplinkos

dinamikos (verslo/kreditų ciklų), kompanijų reputacijos bei vartotojų pasitikėjimo ir pan. Tuo tarpu diversifikuotas portfelis – mažesnės rizikos, atsparesnis lokaliems šokams. Bendros ekonominės, socialinės arba geografinės tendencijos yra svarbios, tačiau indeksai apima ir neigiamas žinias atsveriančius (balansuojančius) aktyvus. Todėl pastarasis portfelis plačiai nagrinėjamas ir naudojamas tiek institucinių, tiek smulkiųjų investuotojų.

Kasdienės grąžos apskaičiuojamos pagal (1) formulę:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}, \quad (1)$$

čia P_t yra aktyvo uždarymo kaina. Daroma prielaida, kad R_t yra stacionarus procesas su vidurkiu μ ir standartiniu nuokrypiu σ . Volatilumas yra modeliuojamas taikant režimų pasikeitimo vienmatį GJR-GARCH modelį, kuriam neigiami šokai turi didesnę poveikį už teigiamus:

$$h_{k,t} = \alpha_{0,k} + (\alpha_{1,k} + \alpha_{2,k}I\{y_{t-1} < 0\})y_{t-1}^2 + \beta_k h_{k,t-1}, \quad (2)$$

čia k yra esamo režimo reikšmė (1, 2 arba 3, priklausomai nuo nagrinėjamo modelio), parametrai $\alpha_{0,k} > 0$, $\alpha_{1,k} > 0$, $\alpha_{2,k} > 0$, $\beta_k \geq 0$, $I\{\cdot\}$ yra indikatoriaus funkcija, kuri įgyja reikšmes 1 arba 0. Tada režimų pasikeitimo modelis apibrėžiamas su (3) formule:

$$y_t | (s_t = k, I_{t-1}) \sim D(0, h_{k,t}, \epsilon_k), \quad (3)$$

čia s_t yra Markovo grandinė, įgyjanti reikšmes $\{1, \dots, K\}$, I_{t-1} yra informacijos aibė iki laiko momento $t - 1$, $D(\cdot)$ yra tolydusis skirstinys su vidurkiu 0, sąlygine dispersija $h_{k,t}$ bei papildomu formos parametrų vektoriumi ϵ_k . Režimai nustatomi taikant Markovo grandinės *Monte Carlo* metodą, kurio parametrai naudojami kaip pradinis taškas galutiniam didžiausio tikėtimumo optimizavimo algoritmui. Naudojant standartizuotas įvertinto modelio liekanas, jų jungtinis skirstinys vertinamas parenkant kopulos funkciją:

$$F(u_1, \dots, u_n) = C(F_1(u_1), \dots, F_n(u_n)). \quad (4)$$

Tinkamas kopulos įvertinimas reikalauja adekvačių F_n skirstinių, kurie leidžia įvertinti teisingą $C(\cdot)$ funkciją. Tyrime nagrinėjamos šios kopulos: eliptinė kopulų šeima (normalioji kopula, T kopula), Archimedo (*Clayton*, *Gumbel*, *Frank*, *Joe*), *Vine*. Skirtingomis kopulų šeimomis siekiama įvertinti įvairių formų skirstinių uodegų priklausomybes (pagrindinis priklausomybės matas yra *Kendall's tau* koeficientas), o labiausiai duomenims tinkama kopula nau-

dojama sintetinių stebėjimų generavimui ir grąžų simuliacijai vieną dieną į priekį:

$$\hat{R}_{t+1} = F^{-1}\{\mathbf{s}_t; \boldsymbol{\theta}_t\} \times \hat{h}_{t+1}, \quad (5)$$

čia $F^{-1}\{\cdot\}$ yra marginaliojo skirstinio atvirkštinė funkcija, \mathbf{s}_t yra kopulos generuotos sintetinių reikšmių vektorius, $\boldsymbol{\theta}_t$ – marginaliojo skirstinio parametrų vektorius, \hat{h}_{t+1} – prognozuotas volatilumas vieną dieną į priekį. Pagal (5) formulę įvertinus portfelio grąžų skirstinį, yra apskaičiuojamas VPR rodiklis:

$$\widehat{VPR}_\alpha = \hat{F}_{\hat{R}_{t+1}}^{-1}(\alpha), \quad (6)$$

čia $\hat{F}_{\hat{R}_{t+1}}^{-1}(\alpha)$ empirinis α lygmens kvantilis. Galiausiai gautas kvantilis yra lyginamas su tikromis to laikotarpio grąžomis. Jeigu tikroji grąža yra mažesnė už apskaičiuotą rodiklį, tai laikoma modelio pažeidimu. Visi skaičiavimai yra taikomi slenkančiam langui, kurio ilgis yra visada 3846 dienų, o skaičiavimai kartojami 500 kartų kiekvienai naujai dienai. Siekiama, kad pažeidimai statistiškai reikšmingai nesiskirtų nuo rizikos lygmens ir būtų nepriklausomi. VPR modelio adekvatumas vertinamas taikant *Kupiec* ir *Christoffersen* testus. Siekiant įvertinti ir suprasti straipsnyje pasiūlytos metodologijos naudą, atliekamas rezultatų palyginimas su BEKK modeliu gautais VPR įverčiais [13]. Skirtingai nuo aptartos metodologijos, BEKK leidžia modeliuoti portfelį kaip aktyvų sistemą, atsižvelgiančią į inovacijų kovariacinę struktūrą.

3 Rezultatai

Straipsnyje nagrinėjamos grąžos – stacionarios bei turinčios kintantį volatilumą (*Augmented Dickey-Fuller*, Lagranžo daugiklio testai). Nustatyta, kad visų aktyvų grąžas tiksliausiai nusako dviejų režimų MS-GJR-GARCH modeliai¹. Dviejų režimų modelis išskyrė aukštos bei žemos sklaidos periodus (režimus) su atitinkamomis parametrizacijomis. Tinkamas režimo nustatymas yra ypač svarbus prognozuojant VPR, klaida šioje vietoje lemtų rizikos pervertinimą arba nuvertinimą. Potencialiems ateities tyrimams verta būtų ištirti, ar grąžų tarpusavio priklausomybė jautri režimų pokyčiams. Atlikus modelio liekanų analizę nustatyta, kad statistiškai reikšmingai tiksliausias buvo asimetrinis Studento-t (toliau – T) skirstinys (remiantis Kolmogorovo–Smirnovu testu). Alternatyviai, taikyta neparimetrinė skirstinių modeliavi-

¹ Tik šiuo atveju iš liekanų buvo pašalinti visi ARCH efektai.

mo metodika, tačiau dėl tiksliai įvertintos parametrinės skirstinio formos neparametriniai metodai reikšmingo indėlio nepridėjo, todėl visiems žemiau aprašytiems rezultatams taikytas asimetrinis T skirstinys.

Kaip etaloninis (angl. *benchmark*) modelis pasirinktas MS–GJR–GARCH (kopulų modeliavimas neatliekamas), t.y. standartizuotoms liekanoms pritaikomas asimetrinis T skirstinys ir simuliuojama grąža vieną dieną į priekį (pagrindiniai rezultatai pateikiami 1 lentelėje). Rezultatai parodė, kad tokio modelio rizika, pirmojo krepšelio indeksų grąžų duomenims, esant 5 % reikšmingumo lygmeniui yra nepakankamai įvertinta – fiksuota 7,2 % pažeidimų. Tokia išvada grindžiama *Kupiec* testu, kuris rodo statistiškai reikšmingą skirtumą nuo pasirinkto lygmens. Viena iš to priežasčių – vienmačiai volatilumo modeliai neatsižvelgia į duomenų tarpusavio priklausomybę.

Papildant analizę daugiamačiu GARCH (BEKK) modeliu pastarasis demonstruoja statistiškai reikšmingai tikslų 5,8 % pažeidimų rezultatą, kurie taip pat yra nepriklausomi. Modelis tiesiogiai skirstinių uodegų nevertina, tačiau indeksų grąžų krepšeliui tokio modelio pakanka tinkamam rizikos vertinimui. Gauti rezultatai yra tikėtini modeliuojant plačiai diversifikuotus portfelius – vienetiniai kelių akcijų šokai atsveriami bendros dinamikos.

Modeliuojant priklausomybę su skirtingomis kopulų šeimomis ir įvertinus VPR pažeidimų procentą, daroma išvada, kad dauguma jų pervertina riziką, o pažeidimų procentas neviršija 1,2 %. Nagrinėjamų kopulų šeimos dažniausiai pasižymi pernelyg centruotomis kairiosiomis uodegomis. Tinkama buvo *Joe* kopula, kuri parodė 4,8 % pažeidimų – tai reikšmingai geresnis rezultatas negu etaloninių modelių. Kartu tai iliustruoja pasiūlytos metodologijos naudą ir pranašumą. Remiantis *Kupiec* ir *Christoffersen* testais daroma išvada, kad rodiklis statistiškai reikšmingai nesiskiria nuo pasirinkto reikšmingumo lygmens, o pažeidimai yra nepriklausomi.

Joe kopulos tinkamumas indikuoja stiprią dešinės (atitinkamai silpną kairiosios) uodegos priklausomybę. Tą vaizduoja *Kendall's tau*, kurio vidutinė reikšmė slenkančio lango metu yra 0,33 (gana silpna priklausomybė) bei 1 pav. esantys *Joe* kopulos kontūro ir empirinio tankio grafikai. Jie nubrėžti prieš tai pasirinkus slenkančio lango vidurinės datos (2024–02–28) standartizuotas S&P500 ir FTSE100 indeksų liekanas, joms pritaikius asimetrinio T skirstinio marginalųjį skirstinį bei transformuojant į reikšmes priklausančias intervalui (0;1). Iš empirinio grafiko pastebima sunkesnė dešinioji uodega, kuri ir būdinga pasirinktai kopulai. Rezultatai leidžia daryti išvadą, kad pasirinkto portfelio diversifikacija lemia atsparumą neigiamiems šokams,

tačiau neapriboja portfelio vertės augimo potencialo. Tikėtina, kad suveikia investavimo į apsaugines (angl. defensive) akcijas mechanizmai, kurie mažina riziką, taip išlaikydami stabilų ir ganėtinai pastovų augimą [12].

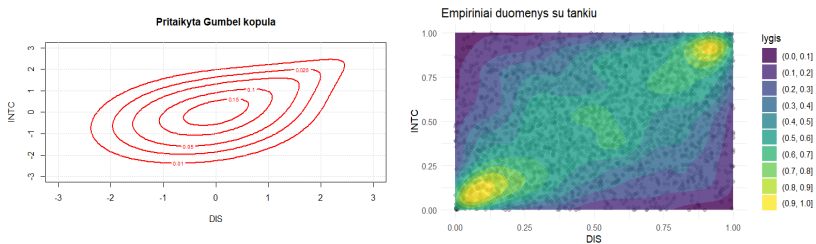
Analizuojant antrąjį (rizikingą) portfelį gaunami šie rezultatai: dviejų režimų MS-GJR-GARCH modelis yra statistiškai reikšmingai tinkamiausias. Kaip ir indeksų gražų atveju adekvačiausias marginalusis skirstinys liekanoms yra asimetrinis T. Pakartojus tuos pačius žingsnius tolimesnėje analizėje darytos panašios išvados (pagrindinius rezultatus žr. 1 lentelėje), jog modeliuojant tik volatilumą be priklausomybės struktūros modelis stipriai nuvertina riziką, o mažiau diversifikuoto krepšelio atveju VPR pažeidimo procentas padidėja iki 8,2 %. Priešingai nei indeksų atveju, BEKK riziką irgi nuvertina. Tikėtina, kad nagrinėjamų akcijų priklausomybės struktūra yra sudėtingesnė, todėl BEKK specifikacijos neužtenka. Darant eksperimentą su skirtingomis kopulomis, pastebėta, kad *Joe* kopula pažeidimai sudaro 6 %, tačiau tai pagal *Kupiec* testą, statistiškai reikšmingai nesiskiria nuo nagrinėjamo lygmens. Pastebėtina, kad visais atvejais pažeidimų procentas yra didesnis nei indeksų gražų atveju, o toks padidėjimas lemia modelių tinkamumą (*Kupiec*, *Christoffersen* testai). Visgi tiksliausią rizikos vertinimą pademonstravo *Gumbel* kopula (su vidutiniu *Kendall's tau* 0,28), modeliuojanti dešinės uodegos priklausomybę, su 4,7 % pažeidimų, kurie yra statistiškai reikšmingai nepriklausomi. Šios kopulos tinkamumą iliustruoja 2 pav. apmokytos *Gumbel* kopulos kontūro bei empirinio tankio grafikai, kurie buvo nubrėžti analogiškai kaip ir 1 pav. atveju. Parinktos *Gumbel* kopulos struktūra leidžia daryti išvadą, kad teigiamos modeliuojamo portfelio naujienos yra stipriai priklausomos, greičiausiai lemiamos ekonominės bei vartojimo aplinkos augimo. Neigiamos naujienos yra susijusios silpniau, tikėtina, nulėmtos konkrečių įmonės sprendimų ar verslo plėtros nesėkmių. Tai rodo tam tikrą lygio diversifikaciją, tačiau optimaliam rizikos valdymui rekomenduotina akcijų krepšelių plėsti, nagrinėjant daugiau nei 3 akcijas.

Abiejų krepšelių rezultatai rodo, kad vienos kopulos vertinimas visiems duomenims greičiausiai yra mažiau efektyvus, nes esant skirtingiems volatilumo laikotarpiams gali būti naudingos vis kitokios kopulų savybės, kitaip tariant – būtų naudingos dinaminės kopulos. Nors slenkantis langas leidžia tam tikrą dinamiką, reikia atkreipti dėmesį, kad šiuo atveju turima omenyje galimybė keistis kopulos modeliui priklausomai nuo rinkos elgesio. Kopulų parinkimo lankstumo potencialų pranašumą akcentavo ir [5, 8, 11] autorių straipsniai. Laike kintančių kopulų poreikį galima būtų grįsti tuo, kad pasi-

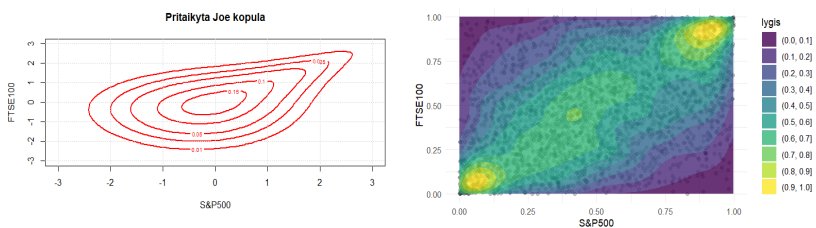
rinktas slenkantis langas turėjo labiau volatilių laikotarpių (COVID-19, karo Ukrainoje pradžia), nei nagrinėtas testuojamas periodas. Nors šiuos periodus siekiame atskirti režimais, verta akcentuoti ir galimą laikiną priklausomybių formos pokytį. Šios rizikos pervertinimas mažiau pastebimas labiau volatiliame akcijų gražų krepšelyje greičiausiai dėl jo didesnės rizikos. Be to, ARCH efektų nebuvimas standartizuotose liekanose bei jų statistiškai reikšmingas tinkamumas asimetriniam T skirstiniui patvirtina [5] straipsnio išvada, kad neretai teisingos kopulos parinkimas yra svarbesnis už volatilumo modelio parinktį. Tam, kad modelis pasiektų norimą tikslumą būtina atrasti tinkamą sąryšį, kuris paaiškintų priklausomumo struktūrą.

1 lentelė. Pagrindinių modelių rezultatų suvestinė

Modelis	Indeksų gražų krepšelis		Akcijų gražų krepšelis	
	Pažeidimai	Tinkamumas (Kupiec, Christoffersen)	Pažeidimai	Tinkamumas (Kupiec, Christoffersen)
MS-GJR-GARCH + normalioji kop.	0,6 %	Rizika pervertinama	3,7 %	Modelis tinkamas
MS-GJR-GARCH + Studento-t kop.	0,6 %	Rizika pervertinama	3,7 %	Modelis tinkamas
MS-GJR-GARCH + Clayton kop.	0,6 %	Rizika pervertinama	3,3 %	Modelis tinkamas
MS-GJR-GARCH + Gumbel kop.	1,2 %	Rizika pervertinama	4,7 %	Modelis tinkamas (tiksliausias)
MS-GJR-GARCH + Frank kop.	0,8 %	Rizika pervertinama	3,9 %	Modelis tinkamas
MS-GJR-GARCH + Joe kop.	4,8 %	Modelis tinkamas (tiksliausias)	6,0 %	Modelis tinkamas
MS-GJR-GARCH + Vine kopulos	0,6 %	Rizika pervertinama	3,7 %	Modelis tinkamas
MS-GJR-GARCH	7,2 %	Rizika nuvertinama	8,2 %	Rizika nuvertinama
BEKK	5,8 %	Modelis tinkamas	7,4 %	Rizika nuvertinama



1 pav. Gumbel kopulos kontūro ir empirinis tankio grafikai pasirinktiems indeksams 2024 m. vasario 28 dienai.



2 pav. Joe kopulos kontūro ir empirinis tankio grafikai pasirinktiems indeksams 2024 m. vasario 28 dienai.

4 Išvados

Atlikus vertės pokyčio rizikos analizę dviems skirtingai volatiliems krepšeliams, daroma išvada, kad pasiūlytas MS–GJR–GARCH bei kopulų metodikos apjungimas leidžia reikšmingai pagerinti VPR modeliavimo tikslumą lyginant su priklausomybės struktūros nemodeliuojančiais vienmačiu MS–GJR–GARCH bei daugiamačiu BEKK modeliais. Kopulų vertinimas leidžia lanksčiai modeliuoti aktyvų tarpusavio priklausomybės formą. Tinkamai išskirti aukštos bei mažos grąžų sklaidos režimai, kartu su tinkamai nustatyta priklausomybe, užtikrina ne tik tikslų rizikos valdymą, bet ir leidžia suprasti bei paaiškinti nagrinėjamo portfelio struktūrą. Tačiau esant pakankamai dideliai imčiai vienos kopulos pritaikymas gali būti nepakankamas. Nors volatilitumo modelis bei marginalusis duomenų pasiskirstymas gali rodyti gerą pritaikymą prie duomenų, rinkoje įvykę stipresni sukrėtimai gali reikalauti kitokios priklausomumo struktūros. Papildžius tyrimo metodiką įtraukiant dinaminį kopulos vertinimą, stebimi struktūros pokyčiai galėtų būti siejami su pasauliniais įvykiais. Be to, priklausomai nuo portfelio struktūros, natū-

ralu tikėtis, kad viena kopulos šeima gali būti pernelyg ribojanti – skirtingos aktyvų poros gali turėti skirtingas priklausomybės. Todėl kitas natūralus metodikos plėtinys – įtraukti *Vine* kopulų modeliavimą siekiant maksimalaus lankstumo.

Literatūra

- [1] Investing Basics: Risk. Financial Industry Regulatory Authority (FINRA): <https://www.finra.org/investors/investing/investing-basics/risk>. Žiūrėta: 2025 03 25.
- [2] Bank of International Settlements (BIS), History of the Basel Committee: <https://www.bis.org/bcbs/history.htm>. Žiūrėta: 2025 03 25.
- [3] Holton, G. A. (2013). Value-at-Risk: Theory and Practice, Second Edition, published online at <https://www.value-at-risk.net>. Žiūrėta: 2025 03 25.
- [4] Gau, Y., Tang, W. (2004). Forecasting value-at-risk using the markov-switching arch model. *Econometric Society 2004 Far Eastern Meetings*, 715.
- [5] Fülle, M. J., Herwartz, H. (2024). Predicting tail risks by a Markov switching MGARCH model with varying copula regimes. *Journal of Forecasting*, 43(6), 2163–2186.
- [6] Byun, K., Song, S. (2021). Value at Risk of portfolios using copulas. *CSAM 2021*, 28, 59–79.
- [7] Huang, J., Lee, K., Liang, H., Lin, W. (2009). Estimating value at risk of portfolio by conditional copula-GARCH method. *Mathematics and Economics*, 45(3), 315–324.
- [8] Sampid, M. G., Hasim, H. M., Dai, H. (2018) Refining value-at-risk estimates using a Bayesian Markov-switching GJR-GARCH copula-EVT model. *PLoS ONE*, 13(6).
- [9] Bob, N.K. (2013). Value at risk estimation A GARCH-EVT-copula approach. *Stockholms universitet, Matematiska institutionen*, 1–41.
- [10] Nugroho, D. B., Kurniawati, D., Panjaitan, L. P., Kholil, Z., Susanto, B., Sasongko, L. R. (2019). Empirical performance of GARCH, GARCH-M, GJR-GARCH and log-GARCH models for returns volatility. *J. Phys.: Conf. Ser.* 1307 012003.
- [11] Berger T. (2013), Forecasting value-at-risk using time varying copulas and EVT return distributions. *International Economics*, 133, 93–106.
- [12] Bajaj Finance, Defensive Stocks: <https://www.bajajfinserv.in/defensive-stocks>. Žiūrėta: 2025 03 25.
- [13] Fülle M. J., Lange A., Hafner C. M., Herwartz H. (2024). BEKKs: An R Package for Estimation of Conditional Volatility of Multivariate Time Series. *Journal of Statistical Software*, 111(4), 1–34.

Vietinių turistų skaičiaus Lietuvos apgyvendinimo įstaigose prognozavimas naudojant „Google Trends“ duomenis

Nora Jasiukėnaitė, Rūta Levulienė

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Taikomosios matematikos institutas Naugarduko g. 24, Vilnius
nora.jasiukenaite@mif.stud.vu.lt

Santrauka. Šiame darbe į laiko eilučių prognozavimo modelius – sezoninį ARIMA modelį su egzogeniniais kintamaisiais (SARIMAX) ir atraminių vektorių regresiją (AVR) – įtraukiami internetinės paieškos sistemos „Google“ duomenys, siekiant pagerinti modelių tikslumą vietinių turistų skaičiaus Lietuvos apgyvendinimo įstaigose prognozavimui. Nustatyta, jog modeliai, sudaryti naudojant „Google Trends“ duomenis, pasižymi mažesnėmis paklaidomis.

Raktiniai žodžiai: turizmas, Google Trends, laiko eilutės, SARIMAX, AVR.

1 Įvadas

Trumpalaikio apgyvendinimo paslaugos yra vienas svarbiausių aspektų, padedančių užtikrinti turizmo sektoriaus plėtrą – apgyvendinimo paslaugų prieinamumas ir kokybė gali lemti keliautojų sprendimus dėl kelionės maršruto bei trukmės. Be to, „Keliauk Lietuvoje“ duomenimis, 2023 m. Lietuvos piliečiai sudarė net 66 % visų turistų Lietuvoje [1]. Taigi, svarbu užtikrinti tikslingą apgyvendinimo paslaugų plėtrą bei gerinimą, o būsimų vietos turistų srautų prognozė gali būti naudojama kaip įrankis šiam tikslui.

Turistų srautų prognozavimui naudojami įvairūs metodai – tiek klasikiniai laiko eilučių [2], [3], tiek mašininio mokymosi bei dirbtinio intelekto modeliai [4], [5]. Pastarųjų metų tyrimai rodo, jog internetinių paieškos sistemų, kaip „Google“ ir „Baidu“, užklausų populiarumo duomenų įtraukimas į šiuos modelius gali žymiai pagerinti jų tikslumą [6], [7], [8]. Tačiau nėra ištirta, kaip ši metodika veikia Lietuvos kontekste, naudojant paieškos užklausas lietuvių kalba. Šio darbo tikslas – nustatyti, ar „Google Trends“ duomenų įtraukimas į klasikinį laiko eilučių bei mašininio mokymosi modelius pagerina vietinių turistų skaičiaus Lietuvos apgyvendinimo įstaigose prognozavimo tikslumą.

2 Duomenys

Tyrime naudojami viešai prieinami Valstybės duomenų agentūros mėnesiniai turistų skaičiaus Lietuvos apgyvendinimo įstaigose 2012-2024 m. duomenys [9]. Taip pat įtraukiami 2012-2024 m. mėnesiniai „Google Trends“ [10] duomenys užklausoms „apartamentai“, „viešbutis“, „Vilnius“, „Kaunas“ ir „lankytinos vietos“, kuriuose pateikiamas santykinis užklausos populiarumas pagal bendrą paieškų geografinėje vietovėje skaičių ir atitinkamą laikotarpį. [10] Naudotinos užklausos atrinktos iš didesnės aibės pagal koreliacijos koeficiento su turistų skaičiumi reikšmę.

Dėl situacijos specifiškumo ir įtakos prognozavimo tikslumui, COVID-19 pandemijos laikotarpio (2020 m. sausis - 2021 m. liepa) duomenys buvo laikomi išskirtimis bei koreguoti padalinus laiko eilutę pagal mėnesį ir naudojant tiesinę interpoliaciją reikšmių įrašymui.

Laiko eilutė buvo padalinta į mokymo ir testavimo aibes – paskutinių turimų, t. y. 2024, metų duomenys naudojami modelių tikslumui įvertinti.

3 Metodai ir rezultatai

Tyrime laiko eilučių prognozavimui taikyti du modeliai – sezoninis ARIMA modelis su egzogeniniais kintamaisiais (SARIMAX) [11] bei atraminių vektorių regresija (AVR) su Gauso branduoliu [12].

ARIMA yra plačiausiai taikomas klasikinis laiko eilučių analizės ir prognozavimo modelis. SARIMAX yra šio modelio atmaina, išplečianti jo galimybes įtraukiant sezoniškumo duomenyse išskyrimą – taip leidžianti modelį taikyti nestacionariems duomenims – ir naudojant išorinius kintamuosius, kurie suteikia papildomos informacijos ir pagerina modelio tikslumą [11]. SARIMAX modeliui būtina nustatyti parametrus: autoregresijos (p), integravimo (d) ir slenkančio vidurkio (q) bei sezoninių autoregresijos (P), integravimo (D) ir slenkančio vidurkio (Q) komponentų reikšmes. Parametras p nustato autoregresinių lagų skaičių, d keičia diferencijavimo skirtumų skaičių, o q nurodo slenkančio vidurkio lango dydį. Parametrai P , Q ir D yra anksčiau minėtų parametrų atitikmenys, skirti apdoroti sezoniškumu pasižyminčius duomenis [11]. Taip pat nustatomas ir stebėjimų skaičius sezone s – mėnesiniams duomenims šio parametro reikšmė lygi 12.

AVR yra prižiūrimo (angl. *supervised*) mašininio mokymosi modelio – atraminių vektorių klasifikatoriaus (AVK) – atmaina, pritaikyta regresijos uždaviniams spręsti. AVR transformuoja mokymo duomenis į aukštesnės dimensijos erdvę, kurioje duomenims yra taikomas tiesinis modelis [13]. Taikant atsitiktinių vektorių regresijos modelį, nustatomi modelio sudėtingumo pa-

parametrai – C , kuris nustato santykį tarp modelio sudėtingumo ir tikslumo, ir ϵ , kuriuo kontroliuojama leistina paklaida [13].

Šie modeliai buvo pasirinkti dėl jų populiarumo sprendžiamos problemos kontekste [14], galimybės įtraukti egzogeninius kintamuosius bei tinkamumo mažoms duomenų aibėms. Tyrimo metu buvo išbandyti ir sudėtingesni mašininio mokymosi modeliai (atsitiktinių miškų regresija [15], *XGBoost* (angl. *Extreme Gradient Boosting*) [16] ir *LightGBM* (angl. *Light Gradient Boosting Machine*) [17]), tačiau dėl riboto duomenų kiekio (turimoje duomenų aibėje yra 156 stebėjimai, pritaikius diferencijavimą ir įtraukus kintamųjų lagus skaičius sumažėjo iki 131), šių modelių veiksmingumas nebuvo optimalus ir buvo susidurta su persimokymo problema, tad jie nėra įtraukiami.

Abiejų modelių atveju išbandytos trys skirtingos kintamųjų kombinacijos: bazinis modelis (naudojant tik istorinius turistų skaičiaus duomenis); įtraukus visų užklausų „Google Trends“ duomenis; įtraukus 3 užklausų, pasižyminčių didžiausiu koreliacijos koeficientu su turistų skaičiumi, „Google Trends“ duomenis.

Siekiant padidinti modelio tikslumą ir patenkinti modelio prielaidas, SARIMAX modelių skaitiniai egzogeniniai kintamieji („Google Trends“ duomenys) paversti į stacionarius – panaikintas sezoniskumas bei tendencija naudojant diferencijavimą. Modelių parametrai buvo parinkti minimizuojant Akaičės informacinį kriterijų (AIC).

AVR modeliai sudaryti naudojant nestacionarius duomenis dėl šio modelio galimybių išskirti sudėtingus, netiesinius ryšius, sezoniskumą ir tendencijas duomenyse [13]. Modelių hiperparametrai parinkti naudojant paiešką gardelėje ir slenkančio lango kryžminę patikrą [18].

Modelių palyginimui naudota vidutinė absoliutinė procentinė paklaida (MAPE), vidutinė absoliutinė paklaida (MAE) bei šaknis iš vidutinės kvadratinės paklaidos (RMSE). Šios metrikos skaičiuojamos pagal formules:

$$MAPE = \frac{1}{N} \sum_{t=1}^N \left| \frac{\hat{Y}_t - Y_t}{Y_t} \right|,$$

$$MAE = \frac{1}{N} \sum_{t=1}^N |\hat{Y}_t - Y_t|,$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N (\hat{Y}_t - Y_t)^2}.$$

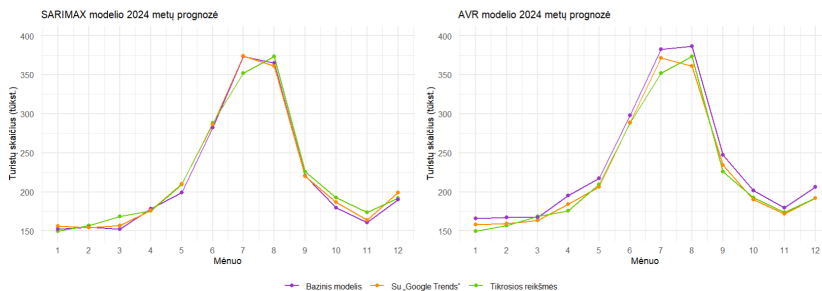
Čia N – stebėjimų skaičius testavimo aibėje, \hat{Y}_t – t -ojo stebėjimo prognozuota reikšmė, Y_t – t -ojo stebėjimo tikroji reikšmė.

1 lentelėje pateiktos bazinio bei geriausio modelio su „Google Trends“ duomenimis tikslumo metrikos testinėje aibėje. Rezultatai rodo, jog tiek SARIMAX, tiek AVR modelio prognozavimo tikslumas padidėja įtraukus „Google Trends“ duomenis, tačiau AVR modelio atveju šis skirtumas yra ryškesnis. Taip pat rezultatai atskleidžia, jog AVR modelis, naudojantis „Google Trends“ duomenis, yra tiksliausias – jo absoliutinė procentinė paklaida siekia vos 2,66 %.

1 lentelė. Modelių tikslumo metrikos. Skliausteliuose nurodytas procentinis pokytis, lyginant su baziniu modeliu.

Modelis		MAPE	MAE	RMSE
SARIMAX	Bazinis	0,0394	8616	10534
	Su „Google Trends“	0,0318 (-19,3 %)	7183 (-16,6 %)	9370 (-11 %)
AVR	Bazinis	0,0614	13298	15324
	Su „Google Trends“	0,0266 (-56,7 %)	6111 (-54 %)	8154 (-46,8 %)

1 paveiksle pateikti bazinių ir „Google Trends“ duomenis naudojančių modelių prognozių ir tikrųjų turistų skaičiaus reikšmių testinėje aibėje grafikai. Grafikai patvirtina modelių tikslumo metrikų rezultatus – tiek SARIMAX, tiek AVR modelio prognozė yra arčiau tikrųjų reikšmių, kai įtraukiami „Google Trends“ duomenys. AVR atveju šis skirtumas ryškesnis – bazinis modelis žymiai pervertina turistų skaičių.



1 pav. Modelių prognozės 2024 metams.

4 Išvados

Tyrimo metu gauti rezultatai atskleidė, jog „Google Trends“ duomenys žymiai pagerina vietinių turistų skaičiaus Lietuvos apgyvendinimo įstaigose prognozavimo tikslumą. Didžiausias tikslumas pasiektas naudojant mašininio mokymosi algoritmą – atraminių vektorių regresijos modelį su internetinės paieškos užklausų populiarumo informacija. Tyrimas gali būti plečiamas įtraukiant sudėtingesnius mašininio mokymosi arba jungtinius algoritmus bei didinant internetinės paieškos užklausų aibę.

Literatūra

- [1] VšĮ „Keliauk Lietuvoje“ (2024). Turizmo sektoriaus apžvalga. Prieiga: https://lithuania.travel/other_files/Turizmo%20sektoriaus%20ap%C5%BEvalga_Keliauk%20Lietuvoje_2024.pdf (žiūrėta 2025-04-23).
- [2] Lim, C., McAleer, M. (2001). Forecasting tourist arrivals. *Annals of Tourism Research*, 28(4), 965-977.
- [3] Cho, V. (2001). Tourism forecasting and its relationship with leading economic indicators. *Journal of Hospitality & Tourism Research*, 25(4), 399-420.
- [4] Claveria, O., Monte, E., Torra, S. (2015). Tourism Demand Forecasting with Neural Network Models: Different Ways of Treating Information. *International Journal of Tourism Research*, 17(5), 492-500.
- [5] Sun, S., Du, Z., Zhang, C., Wang, S. (2022). Improving multi-step ahead tourism demand forecasting: A strategy-driven approach. *Expert Systems with Applications*, 210, 118465.
- [6] Höpken, W., Eberle, T., Fuchs, M. et al (2019). Google Trends data for analysing tourists' online search behaviour and improving demand forecasting: the case of Åre, Sweden. *Information Technology & Tourism*, 21, 45-62.
- [7] Yang, X., Pan, B., Evans, J. A., Benfu, L. V. (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management*, 46, 386-397.
- [8] Bokelmann, B., Lessmann, S. (2019). Spurious patterns in Google Trends data-An analysis of the effects on tourism demand forecasting in Germany. *Tourism Management*, 75, 1-12.
- [9] Valstybės duomenų agentūra (2025). Turistų skaičius apgyvendinimo įstaigose. Prieiga: <https://osp.stat.gov.lt/lt/statistiniu-rodikliu-analize?hash=21ac96a1-7229-4a43-bdbb-c1ae9175226f##> (žiūrėta 2025-04-23).
- [10] Rogers, S. (2016). What is Google Trends data — and what does it mean? Google News Lab. Prieiga: <https://medium.com/google-news-lab/what-is-google-trends-data-and-what-does-it-mean-b48f07342ee8> (žiūrėta 2025-04-23).
- [11] Alharbi, F. R., Csala, D. (2022). A Seasonal Autoregressive Integrated Moving Average with Exogenous Factors (SARIMAX) Forecasting Model-Based Time Series Approach. *Inventions*, 7(4), 94.
- [12] Ojemakinde, B. T. (2006). Support Vector Regression for Non-Stationary Time Series. *Magistro darbas. University of Tennessee*, 28-42.

- [13] Crone, S. F., Guajardo, J., Weber, R. (2006). A study on the ability of Support Vector Regression and Neural Networks to Forecast Basic Time Series Patterns. *Artificial Intelligence in Theory and Practice, IFIP 19th World Computer Congress*, 149-158.
- [14] Song, H., Qiu, R. T. R., Park, J. (2019). A review of research on tourism demand forecasting: Launching the Annals of Tourism Research Curated Collection on tourism demand forecasting. *Annals of Tourism Research*, 75, 338-362.
- [15] Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5-32.
- [16] Chen, T., Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.
- [17] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, 30.
- [18] Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16(4), 437-450.

Vilniaus universiteto priėmimo rodiklių analizė ir prognozavimas (XXI a., I ketvirtis)

Martynas Jokubaitis¹, Igoris Belovas²

¹ Vilniaus universitetas, Šiaulių akademija,
Vytauto g. 84, LT-76352, Šiauliai
martynas.jokubaitis@sa.stud.vu.lt

² Vilniaus universitetas, Matematikos ir Informatikos fakultetas,
Duomenų mokslo ir skaitmeninių technologijų institutas,
Akademijos g. 4, LT-08412, Vilnius
igoris.belovas@mif.vu.lt

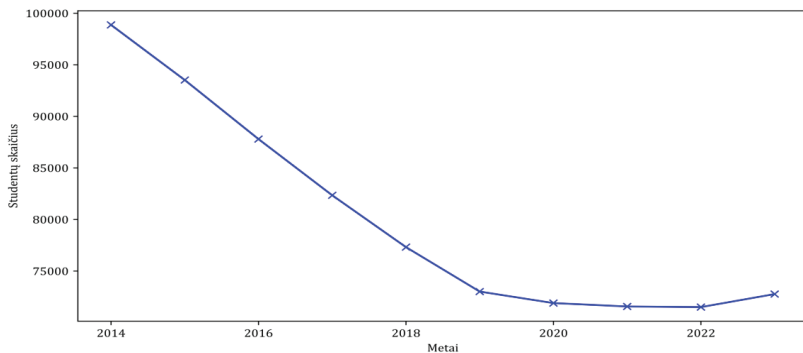
Santrauka. Aukštasis mokslas yra vienas iš svarbesnių veiksnių, teikiančių ne tik asmeninę naudą, bet ir prisidedančių prie geresnės, sveikesnės visuomenės kūrimo. Lietuvoje aukštojo mokslo prieinamumui išaugus, šalis tapo viena labiausiai išsilavinusių visoje Europos Sąjungoje. Ženklus universitetų studentų skaičius mažėjimas 2021–2022 metais stabilizavosi ir pradėjo rodyti teigiamą dinamiką. Vilniaus universitete, kurį pasirenka didžiausia stojančiųjų į universitetus dalis (2023–2024 metais ji sudarė 42–43 %), nuo 2023 metų priimtųjų studijuoti skaičius tik augo ir 2024 metais Vilniaus universitete studijas pradėjo didžiausias per 2019–2024 metų laikotarpį pirmakursių skaičius. Darbe nagrinėjami Vilniaus universiteto XXI amžiaus pirmo ketvirčio priėmimo rodikliai. Galimų ateities tendencijų įvertinimui bei analizei yra taikomas ARIMA modelis. Prognozės rezultatai rodo, kad stojančiųjų rodiklis išliks stabilus. Taip pat darbe nustatoma, jog ARIMA modelis tiksliau įvertina stojančiųjų skaičių už anksčiau naudotą AR modelį.

Raktiniai žodžiai: ARIMA, RMSE, MAE, prognozavimas, studentų skaičius, aukštasis mokslas, priėmimo rodikliai.

1 Įvadas

Smarkiai iki 2021–2022 metų kritęs Lietuvos universitetų studentų skaičius paskutiniiais metais ne tik stabilizavosi, bet ir pradėjo rodyti augimo tendencijas [1] (žr. 1 pav.). Tokioje netikėtai besikeičiančioje aplinkoje yra svarbu gebėti prisitaikyti prie naujų aplinkybių. Studentų skaičiaus prognozavimo galimybes nagrinėjančiuose moksliniuose darbuose pabrėžiama, kad pro-

gnozės padeda ne tik numatyti stojimų tendencijas, bet ir priimti duomenimis pagrįstus strateginius sprendimus dėl resursų paskirstymo, gerinti tiek mokymo, tiek mokymosi procesus.



1 pav. Lietuvos universitetų studentų skaičiaus dinamika

Lavilles ir Arcilla [2], tirdami Mindanao universiteto stojančiųjų rodiklius, siekiant pagerinti universiteto gebėjimą tiksliau prognozuoti priėmimą ir tobulinti kursų pasiūlą, nustatė, kad laiko eilučių metodai, tokie kaip slenkančiųjų vidurkių, viengubo ir dvigubo eksponentinio glodinimo, pasižymėjo didesniu tikslumu nei anksčiau universiteto naudotas *naive* metodas, kai sekančio žingsnio prognozė yra praėjusio žingsnio reikšmė. Bousnguar ir kt. [3], nagrinėdami priėmimo statistiką, taikė ARIMA, eksponentinio glodinimo, neryškių laiko eilučių (*angl.* Fuzzy Time Series, FTS) ir neuroninių tinklų ilgos trumpalaikės atminties (*angl.* Long Short-Term Memory, LSTM) modelius. Savo darbe jiems pavyko nustatyti, kad esant turimam 18 metų duomenų rinkiniui, tiksliausius rezultatus pateikė FTS metodas. Išsamus stojančiųjų skaičiaus vertinimo tyrimas, siekiantis užtikrinti tvarų edukacinės sistemos vystymą, buvo atliktas Kinijoje [4]. Prognozavimui buvo naudota polinominė regresija, dispersinė analizė ir Holto eksponentinis glodinimas, kuris pasiekė tikslesnius rezultatus.

Lietuvoje studentų skaičiaus mažėjimo problemą nagrinėjo Šimkevičienė, kuri savo magistro baigiamajame darbe atliko Lietuvos didžiųjų miestų aukštųjų mokyklų priėmimo rezultatų lyginamąją statistinę analizę 2008–2021 metų laikotarpiu ir nustatė studentų skaičiaus mažėjimo tendencijas. Verta pažymėti, kad autorės tyrimas apsiribojo duomenimis iki 2021 m., o naujausi duomenys rodo bendrą Lietuvos universitetų studentų skaičiaus

augimo tendenciją. Savo tyrimuose autorė studentų skaičiaus laiko eilutėms analizuoti taikė autoregresinį vieno parametro $AR(p)$ modelį [5]. Šio darbo tikslas – žengti sekantį žingsnį ir pritaikyti autoregresinį slenkančiųjų vidurkių ARIMA modelį Vilniaus universiteto priėmimo rodikliams.

2 Duomenys ir metodas

Darbe naudojami Vilniaus universiteto priimtų studentų ir numatytų vietų skaičiaus duomenys 2000-2024 metų laikotarpiu. Apžvelgus atliktus tyrimus [2,3,4,5], reikia pastebėti, kad aukščiau minėtų darbų autoriai neatsižvelgė į nustatyto plano dydį ir vertino tik bendrus priimtų studentų skaičius. Šiame darbe studentų skaičiaus įvertinimui yra naudojamas priimtų studentų ir numatytų vietų santykis, taip atsižvelgiant ir į Universiteto išsikeltus metinius priėmimo tikslus.

Tyrimo yra taikomas ARIMA (angl. *Autoregressive Integrated Moving Average*) metodas, žymimas kaip $ARIMA(p,d,c)$, kur p – autoregresinio proceso eilė, d – stacionarių duomenų eilė, c – slankiojo vidurkio proceso eilė [3]. Modelio tikslumui įvertinti naudojami vidutinės absoliučios paklaidos MAE (angl. *Mean Absolute Error*) ir šaknies vidurkio kvadratinės paklaidos RMSE (angl. *Root Mean Square Error*) rodikliai. $ARIMA(p, d, c)$ modelio parametrai yra apskaičiuojami tinklelio paieškos (angl. *grid search*) metodu, kai p, d, c reikšmės kinta nuo 0 iki 2 ir minėtomis paklaidomis remiantis parenkamas geriausias rezultatus rodantis modelis [3].

ARIMA modelis apskaičiuotas naudojant 20 metų laikotarpį (nuo 2000 iki 2019 metų). Sekantiems 5 metams (nuo 2020 iki 2024 metų) modelis yra testuojamas ir vertinamos paklaidos; toliau yra pateikiamos prognozės 5 metams į priekį.

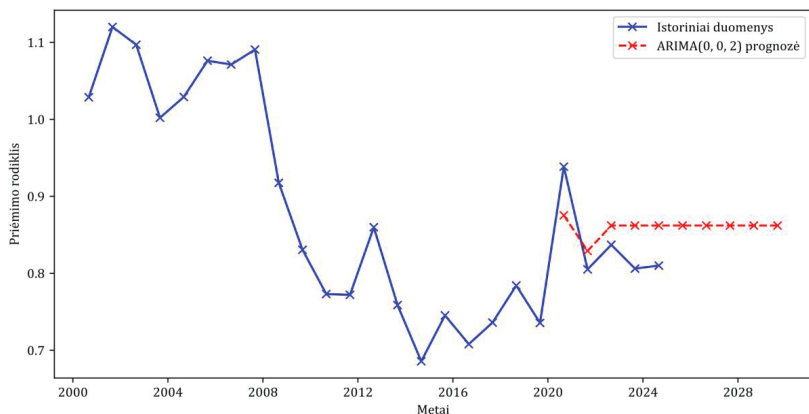
3 Rezultatai

Analizuojant MAE ir RMSE paklaidas, pateiktas 1 lentelėje, matoma, kad ARIMA modelis, kurio autoregresinio proceso p ir stacionarių duomenų d eilės lygios 0 ir slankiojo vidurkio proceso eilė c lygi 2, $ARIMA(0, 0, 2)$, pasiekia geriausias rezultatus iš visų naudotų modelių. Ties šiais parametrais tiek vidutinės absoliučios paklaidos MAE, tiek ir šaknies vidurkio kvadratinės paklaidos RMSE reikšmės yra mažiausios.

1 lentelė. Skirtingų parametų ARIMA modelių paklaidų rezultatai.

ARIMA(p,d,c)	MAE	RMSE
ARIMA(0, 0, 0)	0,066	0,068
ARIMA(0, 0, 1)	0,074	0,081
ARIMA(0, 0, 2)	0,044	0,047
ARIMA(0, 1, 0)	0,104	0,116
ARIMA(0, 1, 1)	0,110	0,121
ARIMA(0, 1, 2)	0,106	0,120
ARIMA(0, 2, 0)	0,249	0,254
ARIMA(0, 2, 1)	0,150	0,155
ARIMA(0, 2, 2)	0,173	0,177
ARIMA(1, 0, 0)	0,102	0,114
ARIMA(1, 0, 1)	0,103	0,115
ARIMA(1, 0, 2)	0,155	0,161
ARIMA(1, 1, 0)	0,107	0,118
ARIMA(1, 1, 1)	0,106	0,117
ARIMA(1, 1, 2)	0,098	0,111
ARIMA(1, 2, 0)	0,174	0,178
ARIMA(1, 2, 1)	0,150	0,155
ARIMA(1, 2, 2)	0,156	0,160
ARIMA(2, 0, 0)	0,101	0,114
ARIMA(2, 0, 1)	0,093	0,109
ARIMA(2, 0, 2)	0,151	0,157
ARIMA(2, 1, 0)	0,110	0,123
ARIMA(2, 1, 1)	0,100	0,115
ARIMA(2, 1, 2)	0,102	0,114
ARIMA(2, 2, 0)	0,111	0,122
ARIMA(2, 2, 1)	0,176	0,181
ARIMA(2, 2, 2)	0,166	0,171

2 pav. pateiktas grafikas, rodantis istorinius analizuoto rodiklio duomenis bei naudoto modelio reikšmes. Iš grafiko matyti, kad modelis gana tiksliai nuspėja 2020, 2021 ir 2022 metų rodiklio pokyčius testuoto laikotarpio metu. Ir pateikia prognozes 5 metams į priekį, kur prognozuojamas rodiklis nemažėja. Tai atitinka anksčiau aptartą studentų skaičiaus mažėjimo tendencijų stabilizavimąsi.



2 pav. Istoriniai duomenys ir ARIMA(0, 0, 2) prognozė

2 lentelėje pateikti nuo 2020 iki 2024 metų testavimo periodo istoriniai duomenys bei modelio prognozė, kartu su prognozuojamomis reikšmėmis sekantiems 5 metams. Apibendrinant, galime teigti, kad yra nustatyta, jog Vilniaus universiteto studentų priėmimo rodiklį tiksliausiai nusako ARIMA(0, 0, 2) parametų modelis.

2 lentelė. ARIMA(0, 0, 2) modelio prognozė

Data	Istoriniai duomenys	ARIMA(0, 0 2) prognozė
9/1/2020	0,939	0,875
9/1/2021	0,805	0,829
9/1/2022	0,837	0,862
9/1/2023	0,806	0,862
9/1/2024	0,810	0,862
9/1/2025		0,862
9/1/2026		0,862
9/1/2027		0,862
9/1/2028		0,862
9/1/2029		0,862

4 Išvados ir aptarimas

Šimkevičienė savo tyrime naudojo vieno parametro AR(p) modelius bei nustatė, kad studentų skaičių geriausiai aprašo AR(3) modelis. Šiame darbe pritaikius AR(p) modelius, kai p parametras kinta nuo 1 iki 3, analizuojamiems duomenims, jų prognozių paklaidos pateiktos 3 lentelėje. Kaip matyti iš lentelės, visi šie modeliai stipriai atsilieka nuo šiame darbe nustatyto ARIMA(0, 0, 2) modelio. Jo prognozių paklaidos yra daugiau nei du kartus mažesnės nei bet kurio pritaikyto AR(p) modelio, todėl jį galima laikyti geresniu pasirinkimu.

3 lentelė. AR(p) modelių paklaidų rezultatai

Modelis	MAE	RMSE
AR(1)	0,102	0,114
AR(2)	0,101	0,114
AR(3)	0,121	0,132
ARIMA(0, 0, 2)	0,044	0,047

Nors darbe yra akcentuojamas ARIMA modelis (kaip sekantis žingsnis po anksčiau naudoto AR modelio), tęsiant tyrimus ateityje, planuojama įtraukti į palyginimą kitus užsienio autorių naudotus metodus: eksponentinio glodinimo, neryškių laiko eilučių (angl. *Fuzzy Time Series, FTS*) ir neuroninių tinklų ilgąs trumpalaikės atminties (angl. *Long Short-Term Memory, LSTM*) modelius.

Literatūra

- [1] Oficialiosios statistikos portalas. URL: <https://osp.stat.gov.lt/statistiniu-rodikliu-analize?hash=4decceb8e-1383-46e9-a86f-ecee7485f558#/>
- [2] Lailles, R. Q., Arcilla M. J. B. (2012). Enrollment Forecasting for School Management System. *International Journal of Modeling and Optimization*, 2(5), 563-566. <https://doi.org/10.7763/IJMO.2012.V2.183>
- [3] Bousnguar, H., Najdi, L., Battou, A. (2022). Forecasting approaches in a higher education setting. *Education and Information Technologies*, 27, 1993-2011. <https://doi.org/10.1007/s10639-021-10684-z>
- [4] Li, X. (2023). Sequence Model and Prediction for Sustainable Enrollments in Chinese Universities. *Sustainability*, 15, 214. <https://doi.org/10.3390/su15010214>
- [5] Šimkevičienė, I., (2023). Studentų priėmimo į Lietuvos didžiųjų miestų aukštąsias mokyklas lyginamoji statistinė analizė. *Magistro baigiamasis darbas*. URL: <https://talpykla.elaba.lt/elaba-fedora/objects/elaba:192828115/datastreams/MAIN/content>

Debesų šalinimas iš ortografinių nuotraukų naudojant giliuosius generatyvinius tinklus

Dominykas Kaminskas

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Didlaukio g. 47, LT-08303 Vilnius
dominykas.kaminskas@mif.stud.vu.lt

Santrauka. Debesys daro neigiamą poveikį Žemės stebėjimui ir palydovų užfiksuojamiems vaizdams. Šiame darbe analizuojami metodai, skirti debesų pašalinimui iš RGB (regimosios šviesos) tipo nuotraukų. Tyrimo metu 3 giliaisiais generatyviais tinklais paremti trukdžių šalinimo metodai – *CloudGAN* [1], *LaMa* [2] ir *RePaint* [3] – pritaikyti debesuotam Vilniaus miesto ortografinių (palydovinių) nuotraukų rinkiniui. Šių modelių sugeneruotos nuotraukos panaudotos trejetų neuroninio tinklo apmokymui. Šalinimo metodų efektyvumui nustatyti apskaičiuotos nuotraukų panašumo metrikos. Geriausiai pasirodė *LaMa* modelis, pasižymėjęs savo greitaveika, siūlomu funkcionalumu ir aukštais tyrimo rezultatais.

Raktiniai žodžiai: gilusis mokymas, generatyviniai tinklai, difuziniai modeliai, debesų šalinimas, palydovinės nuotraukos.

1 Įvadas

Debesys bet kuriuo laiko momentu dengia apie du trečdalius (67 %) Žemės paviršiaus [4]. Daugiau nei 50 % palydovų užfiksuotų vaizdų padengti debesų masyvais. Tai daro neigiamą poveikį Žemės stebėjimui iš kosmoso bei surinkamai duomenų kokybei. Šiai problemai spręsti atliekami įvairūs tyrimai, siekiantys sukurti debesų pašalinimo iš palydovinių nuotraukų metodus.

Tradiciniai debesų šalinimo algoritmai iš palydovinių nuotraukų grindžiami papildoma informacija – laiko eilučių [5] arba spektriniais [6] duomenimis. Tačiau papildomi duomenys gali būti nepasiekiami arba išvis neegzistuoti.

Gilieji neuroniniai tinklai per pastaruosius keletą dešimtmečių padarė nepaprastą pažangą [7]. Generatyvinio dirbtinio intelekto pagalba įmanoma kurti realistiškus vaizdus ir generuoti objektus ant nuotraukų paviršiaus.

Šiame darbe lyginami 3 giliaji generatyvinių tinklų metodai, pasižymėję savo efektyvumu šalinant trukdžius iš nuotraukų bei generuojant naujus vaizdus [8]. Atlikus literatūros analizę nuspręsta pritaikyti 2 bendro pobūdžio generatyvinius trukdžių šalinimo algoritmus – *LaMa* ir *RePaint* – bei išbandyti 1 debesų šalinimo užduočiai skirtą generatyvinį neuroninį tinklą – *CloudGAN*. Darbo tikslas – nustatyti efektyviausią metodą debesų šalinimo iš palydovinių nuotraukų užduočiai spręsti. Modelių efektyvumas palygintas naudojantis trejetų neuroniniu tinklu, gebančiu apskaičiuoti nuotraukų panašumo metrikas.

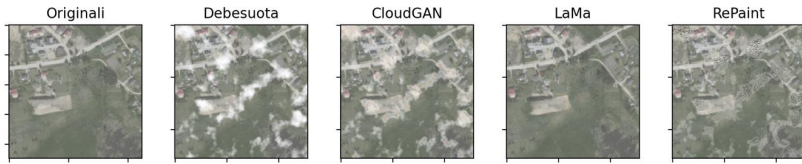
2 Gilieji generatyviniai tinklai

Generatyvinis dirbtinis intelektas yra vienas didžiausių 21 a. kompiuterių mokslo proveržių, pakeitęs žmonių profesinį, asmeninį ir socialinį gyvenimus [23]. Generatyviniai modeliai geba kurti tekstą, garsą ar net vaizdą, todėl gali būti pritaikyti duomenų analizei, nuotraukų generavimui ar kodo optimizacijai [9]. Šie modeliai taip pat gali būti panaudoti trukdžių šalinimui iš vaizdų (angl. *image inpainting*) - remiantis kontekstine nuotraukos informacija pageidaujamoje zonoje generuojamos naujos pikselių reikšmės.

Tyrimo metu išbandyti 3 generatyviniai modeliai:

- *CloudGAN* - generatyvinių besivaržančių tinklų (angl. *Generative Adversarial Network*) modelis, skirtas atpažinti ir pašalinti debesų masyvus iš palydovinių RGB tipo nuotraukų [1]. Išbandyta originali autorių implementacija, siūlanti patogias konsolines komandas programos vykdymui.
- *LaMa* - generatyvinių besivaržančių tinklų (GAN) modelis, leidžiantis pašalinti trukdžius bei įterpti pageidaujamus objektus nuotraukose [2]. Šiame darbe išbandytas *Python* aplinkos paketas "*simple-lama-inpainting*" [10], leidžiantis greitai instaliuoti programinę įrangą ir įvertinti jos efektyvumą.
- *RePaint* - difuzinis dirbtinio intelekto modelis [11], skirtas vaizdų atkūrimo nuotraukose užduotims atlikti [3]. Panaudota kompanijos *Hugging Face* algoritmo implementacija iš "*Diffusers*" bibliotekos [12].

Trukdžių šalinimo algoritmų rezultatą galima vizualiai palyginti 1 paveikslėlyje.



1 pav. Pateiktos originalios ir debesuotos nuotraukos bei skirtingų trukdžių šalinimo algoritmų (*CloudGAN*, *LaMa*, *RePaint*) išvestys.

Visi šiame darbe išbandyti generatyviniai modeliai surašyti 1 lentelėje. Pateiktas vykdymo laikas, skirtas vienai nuotraukai apdoroti. Galima pastebėti, jog *CloudGAN* ir *LaMa* modeliai savo našumu smarkiai lenkia *RePaint* algoritmą. Taip yra todėl, nes difuziniai modeliai paremti iteratyviu veikimo principu, susidedančiu iš nykstamojo proceso (angl. *forward process*) ir grįžtamojo proceso (angl. *backward process*) [11]. Difuzinė architektūra reikalauja daugiau operacijų skaičiaus nei generatyviniai besivaržančių tinklų (GAN) modeliai tai pačiai užduočiai įvykdyti.

1 lentelė. Modelių palyginimas.

Modelis	Tipas	Vykdymo laikas, sek	Debesų aptikimas	Nuotraukos dydis, px
<i>CloudGAN</i>	GAN	0,8	Taip	256 x 256
<i>LaMa</i>	GAN	0,1	Ne	Be apribojimų
<i>RePaint</i>	Difuzinis	11	Ne	256 x 256

CloudGAN yra vienintelis šiame darbe aptariamas modelis, gebantis automatiškai aptikti debesis – generuoti debesų kaukes (angl. *cloud masks*). Nenurodžius debesų sluoksnio, algoritmas bandys pats atpažinti debesų teritorijas ir jas pašalinti, o nuotraukos zonas, kurias modifikavo, pažymės sugeneruotame debesų sluoksnyje. Tačiau šis funkcionalumas tyrimo metu pasirodė neefektyvus – spalvos iškreipiamos, paveikslėlis tampa neatpažintamas (sugadintas). Todėl galima teigti, jog efektyviai pašalinti debesis be trukdžių sluoksnio (angl. *inpaint mask*) su darbe aprašomais metodais neįmanoma.

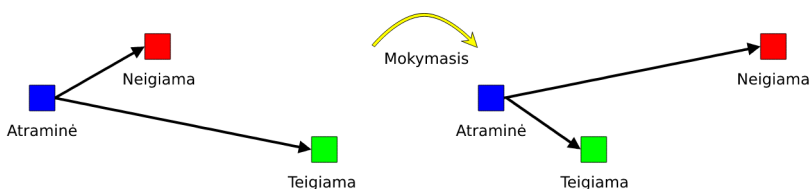
Vienintelis šiame darbe išbandytas modelis, įvesčiai priimantis įvairaus dydžio nuotraukas (netaikantis apribojimų), yra *LaMa*. Nors tinklas buvo apmokytas tik su 256 x 256 px rezoliucijos vaizdais, algoritmas sėkmingai

pritaikomas ir didesnės rezoliucijos nuotraukoms. Autorių teigimu generatyvinis tinklas sėkmingai veikia ir su nuotraukų rezoliucijomis, siekiančiomis 2000 x 2000 px. Likę modeliai (*CloudGAN* ir *RePaint*) įvesčiai priima tik standartinių matmenų vaizdus, todėl visame darbe nuspręsta taikyti vieną rezoliuciją (256 x 256 px).

3 Trejetų neuroninis tinklas

Nuotraukų panašumo metrikos naudojamos algoritmų efektyvumui įvertinti [13]. Šiame darbe pritaikytas gilusis konvoliucinis tinklas palydovinių vaizdų panašumui nustatyti (toliau – trejetų neuroninis tinklas). Sėkmingam nuotraukų panašumo įvertinimui pasitelkta trejetų nuostolių funkcija (angl. *triplet loss*) [14].

Trejetų neuroninio tinklo mokymui skirtą duomenų rinkinį visada sudarys nuotraukų trejetai: atraminė nuotrauka (angl. *anchor image*), teigiama nuotrauka (angl. *positive image*) ir neigiama nuotrauka (angl. *negative image*). Tikslas - mažinti atstumą tarp atraminės-teigiamos nuotraukų poros tuo pačiu metu didinant atskirtį tarp atraminės ir neigiamos nuotraukų (2 paveikslėlis). Šių atstumų skirtumas (toliau - vidutinis skirtumas) nurodo tinklo efektyvumą skirti panašias nuotraukas nuo nepanašių.



2 pav. Trejetų neuroninio tinklo mokymo metu siekiamas tikslas.

Ortografinių nuotraukų panašumo požymiams nustatyti buvo pasirinktos 5 neuroninių tinklų architektūros: *MobileNet* [15], *EfficientNet* [16], *EfficientNetV2* [17], *VGG-16* [18] ir *ResNet* [19], bei jų įvairios konfigūracijos. Vilniaus universiteto mokslininkai savo darbe nustatė, jog naudojant *VGG-16* architektūrą kartu su trejetų nuostolių funkcija galima pasiekti iki 99 % tikslumą [20], todėl šis modelis laikytinas standartu.

4 Darbo eiga

Toliau aprašoma detali tyrimo eiga:

1. Iš Vilniaus miesto ortografinių žemėlapių nuotraukų [21] buvo sudarytas duomenų rinkinys, susidedantis iš daugiau kaip 20 tūkst. nuotraukų. Dirbtinio intelekto modeliui apmokyti pasirinkta 60 % visų nuotraukų, o validacijai ir testavimui skirta po 20 %.
2. Duomenų rinkiniui buvo pritaikytas dirbtinių debesų generavimo įrankis *Satellite Cloud Generator* [22]. Tai parametrizuota *Python* programa, leidžianti ant nuotraukų paviršiaus generuoti realistiškai atrodančius debesis.
3. Debesuotoms nuotraukoms pritaikyti trukdžių šalinimo algoritmai (*CloudGAN, LaMa, RePaint*) debesims panaikinti.
4. 5 kartus atskirai apmokytas trejetų neuroninis tinklas trukdžių šalinimo metodų efektyvumui įvertinti pritaikius:
 - a. Originalų duomenų rinkinį;
 - b. Debesuotų nuotraukų rinkinį;
 - c. *CloudGAN* pašalintų debesų nuotraukų rinkinį;
 - d. *LaMa* pašalintų debesų nuotraukų rinkinį;
 - e. *RePaint* pašalintų debesų nuotraukų rinkinį.
5. Apskaičiuotos nuotraukų panašumo metrikos (vidutinis skirtumas ir tikslumas) bei palyginti rezultatai.

5 Rezultatai

Kiekvieną kartą tinklui baigus mokymosi procesą buvo vykdomi šie eksperimentai:

- Greitaveikos eksperimentas – skaičiuojamas laikas, per kurį neuroninis tinklas ortografines nuotraukas paverčia požymių žemėlapiais (angl. *feature maps*).
- Vidutinio skirtumo eksperimentas - trejetų neuroninio tinklo tikslo funkcijos reikšmė pateikus testavimui skirtą nuotraukų rinkinį (mokymo metu nematyti vaizdai).
- Tikslumo eksperimentas – tinklo gebėjimas pateikus nuotrauką rasti panašias iš didelės krūvos vietovės vaizdų.

Generatyvinio modelio efektyvumas šalinant debesis iš ortografinių nuotraukų vertinamas pagal tai, kiek tinklas eksperimentų metu prano-

ko debesuotų nuotraukų rinkinio pasiektus rezultatus ir kiek atsiliko nuo originalių nuotraukų (angl. *ground truth*). Trejetų neuroninio tinklo VGG-16 architektūros eksperimentų rezultatai pateikti 2 lentelėje. Paryškintos tos trukdžių šalinimo algoritmų reikšmės, kurios pranoko debesuotų nuotraukų rinkinio pademonstruotus rezultatus.

2 lentelė. Trejetų neuroninio tinklo VGG-16 architektūros pasiekti rezultatai.

Nuotraukų rinkinys	Laikas, sek ↓	Vidutinis skirtumas ↑	Tikslumas, % ↑
Originali	13	0,22	99,18
Debesuota	21	0,17	88,62
<i>CloudGAN</i>	17	0,15	87,16
<i>LaMa</i>	13	0,21	90,46
<i>RePaint</i>	13	0,15	73,22

Debesų daroma įtaka nuotraukų panašumo nustatymui – vidutiniškai apie 5-10 % kritęs tikslumas. Sėkmingai pasirodęs *LaMa* modelis pranoko debesuotų nuotraukų rinkinį visose panašumo metrikose ir pasiekė beveik 99 % tikslumą su *ResNet* architektūra, todėl yra efektyviausias išbandytas algoritmas šiame darbe šalinant debesis iš ortografinių nuotraukų. Šio modelio sugeneruotos nuotraukos beveik visais atvejais pademonstravo apie 2-5 % didesnį tikslumą už debesuotus vaizdus. Ne toks efektyvus buvo *CloudGAN* modelis, tyrimų metu pranokęs debesuotų duomenų rinkinį tik 7 iš 18 kartų (apie 40 %). Prasčiausiai pasirodė *RePaint* modelis, kurio pašalintų nuotraukų rinkinys eksperimentų metu vos kelis kartus pagerino debesuotų nuotraukų panašumo metrikas. Žinant, jog bandymų metu šio modelio vykdymo laikas buvo lėčiausias, būtų galima teigti, jog *RePaint* nėra efektyvus algoritmas šalinant debesis iš ortografinių nuotraukų.

6 Išvados

Debesų poveikis palydoviniam Žemės stebėjimui ar nuotoliniams moksliniams tyrimams – prarasta dalis surenkamos informacijos. Tradiciniai metodai duomenų atkūrimui pagrįsti papildomais informacijos šaltiniais - laiko eilučių ar spektriniais. Tačiau papildomi duomenys ne visada egzistuoja. Todėl šiame darbe palyginti generatyvinių tinklų modeliai, gebantys pašalinti dalį debesų iš ortografinių nuotraukų. Tyrimų metu geriausiai pasirodė *LaMa* algoritmas, kurio sugeneruotų nuotraukų rinkinio rezultatai panašu-

mo metrikomis (vidutiniu skirtumu ir tikslumu) aplenkė debesuotų nuotraukų rinkinio rezultatus. Prasčiausiai pasirodė *RePaint* algoritmas, nepasižymėjęs nei vykdymo laiku, nei panašumo metrikomis. Taip pat šis tyrimas parodė, jog nėra efektyvaus būdo pašalinti debesų iš RGB tipo palydovinių nuotraukų neturint debesų kaukės (angl. *cloud mask*).

7 Padėka

Padėka skirta Vilniaus universiteto Matematikos ir informatikos fakulteto Informacinių technologijų atviros prieigos centrui (ITAPC) už suteiktus HPC išteklius šio darbo skaičiavimams atlikti. Eksperimentai buvo vykdomi paskirstytų skaičiavimų tinkle (PST) - tai 2021 metais pristatytas naujasis universiteto superkompiuteris, kurio bendras teorinis našumas yra ne mažesnis nei 0,5 Pflops DP.

Literatūra

- [1] Kluiver, S., Schonenberg, J. (2022). CloudGAN: Detecting and Removing Clouds from RGB-images using Image Inpainting.
- [2] Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., ... & Lempitsky, V. (2022). Resolution-robust large mask inpainting with fourier convolutions. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 2149-2159).
- [3] Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., & Van Gool, L. (2022). Repaint: Inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11461-11471).
- [4] King, M. D., Platnick, S., Menzel, W. P., Ackerman, S. A., & Hubanks, P. A. (2013). Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua satellites. *IEEE transactions on geoscience and remote sensing*, 51(7), 3826-3852.
- [5] Alonso-Sarria, F., Valdivieso-Ros, C., & Gomariz-Castillo, F. (2024). Imagery Time Series Cloud Removal and Classification Using Long Short Term Memory Neural Networks. *Remote Sensing*, 16(12), 2150. <https://doi.org/10.3390/rs16122150>
- [6] Hasan, C., Horne, R., Mauw, S., & Mizera, A. (2022). Cloud removal from satellite imagery using multispectral edge-filtered conditional generative adversarial networks. *International Journal of Remote Sensing*, 43(5), 1881–1893. <https://doi.org/10.1080/01431161.2022.2048915>
- [7] Liu, J., Kong, X., Xia, F., Bai, X., Wang, L., Qing, Q., & Lee, I. (2018). Artificial intelligence in the 21st century. *Ieee Access*, 6, 34403-34421.
- [8] Liu, Z. (2024). Inpainting Ortho-Rectified Satellite Images. *Photogrammetry and Remote Sensing Institute of Geodesy and Photogrammetry, ETH Zürich*.
- [9] Gupta, R., Nair, K., Mishra, M., Ibrahim, B., & Bhardwaj, S. (2024). Adoption and impacts of generative artificial intelligence: Theoretical underpinnings and research agenda. *International Journal of Information Management Data Insights*, 4(1), 100232.

- [10] Simple LaMa Inpainting. (2022). <https://github.com/enesmsahin/simple-lama-inpainting>
- [11] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33, 6840-6851.
- [12] Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Nair, D., Paul, S., Berman, W., Xu, Y., Liu, S., & Wolf, T. (2022). Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers>
- [13] Mitchell, H.B. (2010). Image Similarity Measures. In: *Image Fusion*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-11216-4_14
- [14] F. Schroff, D. Kalenichenko, J. Philbin. Facenet: A unified embedding for face recognition and clustering. (2015). *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 815–823).
- [15] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
- [16] Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105-6114). PMLR.
- [17] Tan, M., & Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In *International conference on machine learning* (pp. 10096-10106). PMLR.
- [18] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition.
- [19] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
- [20] Valaitis, V., Marcinkevičius, V., & Jurevičius, R. (2020). Learning aerial image similarity using triplet networks. In *Numerical computations: theory and algorithms: international conference on numerical computations: theory and algorithms NUMTA 2019. Part 2 / edited by: Sergeyev, YD; Kvasov, DE*. https://doi.org/10.1007/978-3-030-40616-5_15
- [21] V. Ž. ūkio duomenų centras. (2025). Lietuvos erdvinės informacijos portalas. <https://www.geoportal.lt/geoportal/>
- [22] Czerkawski, M., Atkinson, R., Michie, C., & Tachtatzis, C. (2023). SatelliteCloudGenerator: Controllable Cloud and Shadow Synthesis for Multi-Spectral Optical Satellite Images. *Remote Sensing*, 15(17), 4138. <https://doi.org/10.3390/rs15174138>
- [23] Banh, L., & Strobel, G. (2023). Generative artificial intelligence. *Electronic Markets*, 33(1), 63.

Verslo valdymo sistema – sprendimas efektyvesniam ikimokyklinio ugdymo procesų valdymui

Atėnė Kasperavičiūtė, Laura Ringienė

Vilniaus universitetas, Matematikos ir Informatikos fakultetas,
Akademijos g. 4, LT-08412 Vilnius
atene.kasperavičiūtė@mif.stud.vu.lt, laura.ringiene@mif.vu.lt

Santrauka. Ikimokyklinio ugdymo įstaigos atlieka svarbų vaidmenį vaikų socialinėje ir akademinėje raidoje, tačiau jų veiklos valdymas neretai susiduria su administraciniais ir organizaciniais iššūkiais. Skaitmenizavimas tampa esminiu įrankiu, padedančiu optimizuoti procesus, palengvinti administracinį darbą ir gerinti ugdymo kokybę. Nepaisant technologinių galimybių, švietimo įstaigos dažnai naudoja nesuderintas sistemas, kurios apsunkina duomenų valdymą ir neefektyviai išnaudoja darbuotojų laiką. Todėl atsiranda poreikis vieningai sistemai, kuri užtikrintų sklandų duomenų mainymąsi ir veiklos valdymą. Šio darbo tikslas – išanalizuoti skaitmenizavimo poveikį ikimokyklinio ugdymo įstaigoms, įvertinti dabartinių informacinių technologijų taikymo efektyvumą ir pasiūlyti vieningos sistemos galimą sprendimą.

Raktiniai žodžiai: ikimokyklinio ugdymo įstaiga, verslo valdymo sistema, veiklos valdymo sistema, skaitmenizavimas, „Microsoft Dynamics 365 Business Central“

1 Įvadas

Ikimokyklinis ugdymas Lietuvoje yra svarbi švietimo sistemos dalis, kuri padeda vaikams ugdyti pagrindines kompetencijas, skatina socialinę adaptaciją ir mažina mokymosi sunkumų bei socialinės atskirties riziką [1] [2]. Šiuolaikinės ikimokyklinės įstaigos atlieka dvejopą funkciją – jos bendradarbiauja su šeima, siekdamos užtikrinti nuoseklų vaiko ugdymą darželyje, ir padeda šeimai, ugdančiai vaiką. Siekdamos suderinti ugdymo procesą su šeimos poreikiais, šios įstaigos suteikia galimybę tėvams derinti darbą, vaikų ugdymą ir priežiūrą, siūlydamos lanksčias paslaugas, pavyzdžiui, ugdymo metodikų pasirinkimą, ilgesnį darželio darbo laiką ar vaikų priežiūrą savaitgaliais. Tokiu būdu ikimokyklinės įstaigos prisideda prie vaikų ugdymo ir šeimų gerovės užtikrinimo [3].

Pasak D. Vaičiūnaitės ir V. Brazdeikio [4] atliktos švietimo problemos analizės, šiuolaikinės švietimo įstaigos, siekdamos optimizuoti savo veiklos procesus, naudoja įvairias taikomas programas, kurios palengvina procesų valdymą ir komunikaciją. Kadangi šioms įstaigoms suteikiama galimybė informacines sistemas pasirinkti pagal savo poreikius, jose dažnai vyrauja pavienės, nesuderintos sistemos, kurios nėra tarpusavyje integruotos ir nesikeičia duomenimis, nors jose dažniausiai operuojama ta pačia informacija. Ši praktika ne tik apsunkina darbo procesus, bet ir neefektyviai išnaudoja darbuotojų laiką.

Atsižvelgiant į šiuos iššūkius, atsiranda poreikis vieningos veiklos valdymo sistemos, kuri galėtų apimti visus ikimokyklinės įstaigos veiklos procesus bei efektyviau atspindėti organizacijos veiklą. Todėl šio darbo tikslas – išanalizuoti skaitmenizavimo poveikį ikimokyklinio ugdymo įstaigoms, įvertinti dabartinių informacinių technologijų taikymo efektyvumą ir pasiūlyti vieningos sistemos sprendimą, kuris optimizuotų administracinius procesus ir gerintų veiklos efektyvumą ir kokybę.

2 Skaitmenizavimo svarba

Verslo ir veiklos procesai organizacijose yra tarpusavyje susiję, todėl net mažiausi trikdžiai gali paveikti jų efektyvumą. Šiuolaikinės ikimokyklinio ugdymo įstaigos susiduria su poreikiu ne tik organizuoti vaikų ugdymą, bet ir valdyti daugybę informacijos apdorojimo, administracinių ir veiklos procesų. Todėl skaitmenizavimas tampa pagrindiniu įrankiu modernizuojant veiklos valdymą.

Skaitmenizavimo reikšmė ikimokyklinio ugdymo įstaigoms atskleidžiama R.N. Fitria, D. Haryadi ir R. Bustan atliktame tyrime [5], kuriame patvirtinama, kad mokymosi priemonių informacinė sistema (SIPP) reikšmingai palengvina pedagogų darbą, automatizuodama duomenų suvedimą ir ataskaitų rengimo procesą. Palyginus auklėtojų naudojamus kasdienius įrankius „Microsoft Word“ ir „Microsoft Excel“ su autorių sukurta sistema SIPP, pastebima, kad naudojant informacinę sistemą, ataskaitos parengiamos efektyviau: „labai greitai“ ir „greitai“ parengiančių ataskaitų apklaustųjų dalis yra didesnė (71,16 % su SIPP, 57,69 % su kasdieniais įrankiais), o lėtai parengiančių ataskaitas procentas reikšmingai sumažėjo (1,92 % su SIPP, 3,85 % su kasdieniais įrankiais). Taip pat tyrėjų rezultatai parodė, kad vartotojams naudotis SIPP sudarant ataskaitas buvo paprasta ir patogiu (nei vienas dalyvis nenurodė šio proceso kaip „sunkaus“ ar „labai sunkaus“). Šie tyrimo duomenys rodo, kad SIPP padeda sumažinti darbo krūvį ir popierinių dokumentų naudojimą, taip

pat padeda išvengti klaidų pildant pasikartojančius laukus ataskaitose dėl realizuoto automatinio laukų užpildymo, taip prisidedant prie geresnės ikimokyklinio ugdymo kokybės bei efektyvesnio proceso valdymo.

Dar viename autorių R. Rohita ir N. R. Hidayat atliktame tyrime [6] vertinamas SIPP veiksmingumas ir efektyvumas, padedant pedagogams supaprastinti pamokų planavimo procesą. Tyrime veiksmingumas buvo nustatomas pagal naudojimo patogumą, patikimumą, pritaikomumą, reakcijos laiką ir rezultatų kokybę. 50,8 % respondentų sutiko, kad SIPP padėjo pasiekti pamokų planavimo tikslus, o sistema buvo itin gerai įvertinta pagal prieinamumą (58,6 %), pritaikomumą (54,5 %) ir patikimumą (55,9 %). Efektyvumas buvo matuojamas pagal laiko, sąnaudų ir pastangų taupymą bei popieriaus naudojimo mažinimą. Pastebėta, kad 65,6 % respondentų nurodė greitesnį pamokų planavimą naudojant SIPP, o 50,8 % teigė, jog sumažėjo išlaidos (pvz., popieriaus, rašalo, įrišimo sąnaudų) dėl mažesnio spausdinimo poreikio. Be to, 58,6 % darbuotojų pritarė, kad SIPP padėjo sukurti popieriaus nereikalaujančią darbo aplinką, leidžiančią efektyviau saugoti ir pasiekti duomenis. Taigi, SIPP pasitvirtino kaip vertinga priemonė, optimizuojanti pamokų planavimą, mažinanti mokytojų darbo krūvį ir gerinanti planavimo tikslumą.

3 Ikimokyklinio ugdymo įstaigų informacinės sistemos

Lietuvos ikimokyklinių ugdymo įstaigos neturi vieningos, bendros informacinės sistemos visai darželio veiklai valdyti. Išnagrinėjus ikimokyklinių įstaigų finansinių ataskaitų rinkinių aiškinamuosius raštus buvo nustatyta, jog šiose organizacijose dažniausiai dominuoja du sistemų tipai - viena informacinė sistema naudojama ikimokyklinės ugdymo įstaigos veiklos valdymui ir administravimui, o kita vaikų veiklų ir lankomumo organizavimui, valdymui ir administravimui. Su vaiku susijusias veiklas atlikti dažniausiai naudojami elektroniniai dienynai, o kitoms įstaigos veikloms, kaip sutarčių sudarymas, ataskaitų generavimas, statistikų išvedimas, tėvų, vaikų, personalo informacijos rinkimui ir pan. naudojamos verslo, buhalterinės, dokumentų ar finansų valdymo sistemos.

Palyginus keturias sistemas (žr. 1 lentelę), kurias naudoja Lietuvos darželiai: „Labbis“, „Avily“, „Debetas“ ir „Biudžetas VS“ buvo nustatyta, jog visos sistemos turi bazinius buhalterinius apskaitos modulius, bet pagrindinis jų trūkumas, jog nėra vaikų veiklų administravimo funkcijų (kaip pavyzdžiui, lankomumo žymėjimas), todėl ikimokyklinėms įstaigoms reikia atskiros sistemos, kaip pavyzdžiui „Mūsų darželis“ ar „ELIIS“. Ši fragmentacija

apsunkina duomenų valdymą, sukelia darbo dubliavimą ir mažina veiklos efektyvumą. Vieningos sistemos diegimas galėtų iš esmės pagerinti procesų standartizavimą ir administracinio darbo našumą.

1 lentelė. Sistemų palyginimas

Sistema Kriterijus	„Biudžetas VŠ“	„Debetas“	„Labbis“	„Avilys“
Sistemos vieningumas	Sistema veikia modulinio principu, bet jį veikia tik vieningoje sistemoje.	Bazinę sistemą sudaro 7 atskiri moduliai, veikiantys kaip atskiros sistemos.	Visi moduliai dirba su viena duomenų baze. „KONTO“ integruotas kartu su „LOGAS“ ir „BONUS“ modeliais.	Visas sistemos funkcionalumas yra vienoje platformoje.
Duomenų saugojimas ir archyvavimas	+	+	+	+
Užduočių valdymas	-	-	+	+
Vartotojų vaidmenų funkcionalumas	+	+	+	+
Ataskaitų rengimas	+	+	+	+
Sutarčių sudarymas	+	-	+	+
Duomenų analizė ir verslo įžvalgos	-	-	+	-
Dokumentų rengimas pagal šablonus	+	+	+	+
Sistemos komponentų integracija	SODRA Valstybės išdas, VMI, Bankų sistemos, Viešųjų pirkimų sistemos	-	VMI, SODRA, „Oracle BI“, Mokinių registras, „Microsoft Office“	„Microsoft Office“, pasirinktos finansų, personalo valdymo sistemos „MS Active directory“
Kaina	Vidutinė (prieinama)	Ekonomiška, puikiai tinka mažoms įmonėms	Aukšta (pilnos VVS versijos brangios)	Vidutinė, priklauso nuo pasirinktų modulių
Debesų kompiuterijos palaikymas	+	-	+	+
Vaikų lankomumo žymėjimas	-	-	-	-

Šaltinis: lentelė sudaryta autorės pagal šaltinius [7], [8], [9], [10].

4 Verslo valdymo sistemos

Atlikta ikimokyklinio ugdymo įstaigų sistemų analizė atskleidė poreikį vieningai sistemai, kuria naudotųsi visi įstaigos darbuotojai. Tai sistema, kuri efektyviai apjungtų vaikų ir organizacijos veiklų administravimą bei organizavimą. Optimalus šių įstaigų veiklų valdymo sprendimas – verslo valdymo sistema (toliau – VVS).

VVS – tai integruota informacinė sistema, skirta pagrindinių verslo procesų valdymui realiuoju laiku, naudojant bendrinamus duomenis ir informaciją [11]. Sistemos kūrimui pasirinktas VVS sprendimas, nes jis suteikia šiuos pagrindinius privalumus: sklandų skyrių ir darbuotojų bendradarbiavimą, efektyvų ataskaitų valdymą, centralizuotą duomenų bazę, automatizuotus kasdieninius procesus, vieningą ir lengvai išmokstamą vartotojo sąsają.

G. Balkytės ir J. Jurevičienės [12] straipsnyje apie kompiuterizuotos apskaitos programų pasirinkimą atliktas tyrimas atskleidė informacinių sistemų teikiamą naudą įstaigų darbuotojams. Tyrimo rezultatai rodo, kad dauguma respondentų pritarė, jog kompiuterizuotos programos pagerino informacijos prieinamumą (60 sutiko, 2 nesutiko), padidino duomenų kontrolę ir analizės galimybes (67 sutiko, 2 nesutiko), užtikrino duomenų patikimumą (61 sutiko, 1 nesutiko), sumažino darbo ir duomenų dubliavimą (55 sutiko, 3 nesutiko) bei padidino apskaitos darbo efektyvumą (71 sutiko, 3 nesutiko). Šie duomenys patvirtina, kad kompiuterizuotos apskaitos sistemos yra reikšmingas įrankis organizacijų veikloje, padedantis optimizuoti procesus, gerinti duomenų valdymą ir didinti darbo efektyvumą.

VVS analizei pasirinktos trys pasaulyje plačiai naudojamos sistemos: „Microsoft Dynamics 365 Business Central“, „Oracle NetSuite“ ir „SAP Business One“. Žemiau esančiose lentelėse (žr. 2 – 5 lenteles) verslo valdymo sistemos palyginamos pagal tris kriterijus: integracija, plėtimas, vartotojo sąsaja, licencijavimo modeliai ir diegimo kaštai. Šie aspektai gali turėti esminės įtakos organizacijų sprendimui rinktis vieną ar kitą sistemą.

Atlikus verslo valdymo sistemų analizę, ikimokyklinio ugdymo įstaigų informacinei sistemai pasirinkta „Microsoft Dynamics 365 Business Central“ (toliau - „Business Central“), nes ji išsiskiria lengva integracija, pritaikomumu ir intuityvia vartotojo sąsaja. Aiški ir prieinama kainodara leidžia efektyviai planuoti biudžetą, nes sistema pasižymi aiškiai apibrėžtu licencijavimo modeliu. Lyginant su konkurentais, „Business Central“ yra ekonomiškiausias pasirinkimas diegimo kaštu atžvilgiu, kas yra didelis privalumas mažoms ir vidutinėms įstaigoms, siekiančioms optimizuoti savo veiklą neperžengiant biudžeto ribų. „Business Central“ užtikrina efektyvią integraciją – ji sklandžiai

veikia su „Microsoft“ produktais bei trečiųjų šalių sistemomis, leidžiant sujungti visus verslo procesus į vientisą duomenų valdymo sistemą. Ši VVS lengvai išplečiama ir pritaikoma – funkcionalumą galima plėsti tiek programiniu kodu, tiek papildomomis programomis ar plėtiniais, kurie gali būti pritaikyti konkrečioms organizacijos poreikiams. O intuityvi vartotojo sąsaja yra labai panaši į „Office 365“, todėl darbuotojai gali lengvai perprasti sistemą ir greitai pradėti ja naudotis, be sudėtingų mokymų. Dėl šių priežasčių „Business Central“ buvo pripažinta tinkamiausia sistema, kuri užtikrins optimalią veiklos valdymo kokybę ir efektyvumą.

2 lentelė. Sistemų palyginimas pagal integraciją

Sistema	Integracija
„Microsoft Dynamics 365 Business Central“	Sistema labai lengvai integruojama su „Microsoft“ produktais, tokiais kaip „Office 365“, „Azure“ bei „Outlook“. „Microsoft AppSource“ patalpintos programos ir plėtiniai leidžia praplėsti sprendimo funkcionalumą. Taip pat su partnerių pagalba galima sukurti integracijas su trečiųjų šalių sistemomis.
„Oracle NetSuite“	Sistema turi iš karto sukurtas integracijas su „Google“, „Oracle“ ir „Salesforce“ programomis. Taip pat siūlomas produktų rinkinys, kuris efektyviai veikia, jei atitinka įmonės poreikius. Tačiau kartais tarp įmonės poreikių ir siūlomo funkcionalumo atsiranda spragų, kurias gali būti sudėtinga pašalinti naudojant tik pagrindinį rinkinį.
„SAP Business One“	„SAP Store“ siūlo įvairias programas, leidžiančias praplėsti sistemos funkcionalumą. Taip pat galima integruoti trečiųjų šalių programas, tačiau kiekvienai integracijai reikalingas atskiras įrankis.

Šaltinis: lentelė sudaryta autorės pagal šaltinius [13], [14], [15]

3 lentelė. Sistemų palyginimas pagal plėtimą

Sistema	Plėtimas
„Microsoft Dynamics 365 Business Central“	Galima lengvai ir maksimaliai išnaudoti plėtimo galimybę ir pritaikyti jį prie specifinių verslo poreikių, išnaudojant jo lankstumą įvairiose srityse. Daugeliui modifikacijų kodavimo nereikia – naudotojai gali greitai integruoti papildomas funkcijas, pritaikyti sąsają bei modulius pagal savo darbo pobūdį. Naudojant „Visual Studio Code“ galima atlikti sudėtingus vartotojo sąsajos patobulimus ir funkcionalumo praplėtumus.
„Oracle NetSuite“	Sistema sukurta remiantis geriausiomis verslo valdymo praktikomis, todėl siekiama, kad įmonės kuo labiau pasikliautų standartiniu funkcionalumu ir nenaudotų individualizuotų pritaikymų.
„SAP Business One“	Lengvai pritaikomas ir modifikuojamas pagal kiekvieno vartotojo poreikius, nesukeliant sudėtingų techninių iššūkių. Sistema leidžia atlikti įvairias modifikacijas ir plėtinius tiesiogiai darbo vietose, suteikiant vartotojams daugiau galimybių personalizuoti jų patirtį.

Šaltinis: lentelė sudaryta autorės pagal šaltinius [13], [14], [16]

4 lentelė. Sistemų palyginimas pagal naudotojo sąsaja

Sistema	Naudotojo sąsaja
„Microsoft Dynamics 365 Business Central“	Intuityvi, lengvai valdoma sistema, vartotojo sąsaja primenanti vartotojams jau žinomus „Microsoft Office“ įrankius, tokius kaip „Outlook“, „Word“ ir „Excel“. Dėl šių ypatybių darbuotojai greičiau išmoksta dirbti su sistema. Personalizuotas valdymo skydelis leidžia vartotojams matyti tik svarbiausią informaciją. Nors sąsaja paprasta, pažangesnės funkcijos gali pasirodyti sudėtingos nepatyrusiems vartotojams.
„Oracle NetSuite“	Moderni, švari ir prisitaikanti sąsaja, leidžianti vartotojams matyti duomenis pagal rolę. Vartotojo sąsaja gana sudėtinga gali reikalauti laiko ir mokymų. Sudėtingesnis duomenų valdymas naujokams.
„SAP Business One“	Pritaikomas skydelis pagal vartotojo rolę. Labiau orientuotas į patyrusius vartotojus – pradantiems sistema gali būti sudėtinga. Ne tokia intuityvi kaip kitos modernios VVS sistemos. Sąsaja gali atrodyti „senamadiška“ lyginant su „Business Central“ ar „NetSuite“.

Šaltinis: lentelė sudaryta autorės pagal šaltinius [13], [14], [16]

5 lentelė. Sistemų palyginimas pagal kainas

Sistema	Licencijavimo modelis	Diegimo kaštai
„Microsoft Dynamics 365 Business Central“	Vienam vartotojui: bazinė prenumerata – 70 dolerių (~ 65 Eur) mėnesiui; premium – 100 dolerių (~ 92 Eur) mėnesiui.	Sistemos įdiegimas gali kainuoti nuo 3 000 dolerių (~ 2 771 Eur) iki 10 000 dolerių (~ 9 238 Eur), priklausomai nuo projekto apimties.
„Oracle NetSuite“	Bazinė licencija mėnesiui - 999 doleriai (~ 923 Eur) ir papildomai po 99 dolerius (92 Eur) kiekvienam vartotojui.	Diegimo kaštai svyruoja nuo 25 000 dolerių (~ 23 095 Eur) iki 100 000 dolerių (~ 92 379 Eur) ir daugiau, priklausomai nuo verslo dydžio ir sudėtingumo.
„SAP Business One“	Profesionali licencija – 94 dolerių (~ 87 Eur) mėnesiui, ribota licencija – 54 dolerių (~ 50 Eur) mėnesiui.	Priklausomai nuo diegimo tipo: debesų ar vietinis sprendimas, sistemos diegimo kaina gali svyruoti nuo 30 000 dolerių (~ 27 714 Eur) iki 100 000 dolerių (~ 92 379 Eur).

Šaltinis: lentelė sudaryta autorės pagal šaltinius [13], [14], [16]

Pastaba: Suma Eurais nurodyta pagal lb.lt 2025 m. kovo 25 d. duomenis.

5 Planuojamos verslo valdymo sistemos funkcionalumas

„Microsoft Dynamics 365 Business Central“ pagalba realizavus ikimokyklinio ugdymo verslo valdymo sistemą, šių įstaigų administracija turės vieningą, lengvai prieinamą erdvę visiems svarbiausiems procesams – nuo vaikų re-

gistracijos ir lankomumo sekimo iki finansų valdymo bei personalo administravimo. Pedagogai galės greitai pasiekti reikiamą informaciją apie vaikų pažangą, lengvai planuoti ugdymo veiklas ir bendrauti su tėvais per integruotą platformą.

Visi duomenys bus saugomi vienoje vietoje, todėl nebereikės papildomai siųsti dokumentų ar keistis informacija rankiniu būdu. Automatizuoti duomenų srautai užtikrins, kad buhalteriai visuomet dirbs su naujausia informacija, nes duomenys, įvesti auklėtojo ar administratoriaus, bus iškart matomi sistemoje.

Naujoji sistema reikšmingai sumažins laiko sąnaudas ir klaidų tikimybę, nes sumažės darbo ir duomenų dubliavimo apimtis. Tai taip pat prisidės prie darbo efektyvumo didėjimo, nes darbuotojai galės skirti daugiau dėmesio pagrindinėms užduotims, o automatizuotas duomenų apdorojimas užtikrins didesnę informacijos patikimumą ir tikslumą.

Tėvams ši sistema taps patogiu informacijos centru – jie galės matyti svarbiausius pranešimus, stebėti vaiko kasdienę veiklą, sužinoti apie renginius ar maitinimo tvarkaraščius be papildomų pastangų. Viskas bus lengvai pasiekama tiek kompiuteryje, tiek mobiliajame įrenginyje.

Galiausiai, sujungus visus šiuos elementus, visa ikimokyklinio ugdymo bendruomenė veiks darniai ir efektyviai. Administracinis krūvis sumažės, procesai taps skaidresni, o bendravimas tarp visų suinteresuotųjų – sklandesnis. Tai ne tik supaprastins kasdienybę, bet ir leis daugiau dėmesio skirti svarbiausiam dalykui – vaikų ugdymui ir gerovei.

6 Išvados

Ikimokyklinio ugdymo įstaigos atlieka svarbų vaidmenį vaikų socialinėje ir akademinėje raidoje, tačiau jų veiklos valdymas dažnai susiduria su nesuderintomis informacinėmis sistemomis, kurios apsunkina ugdymo procesus, informacijos valdymą ir sukelia administracinius iššūkius. Technologijų diegimas šiose įstaigose didina veiklos efektyvumą, palengvina darbuotojų atliekamą užduotį ir optimizuoja administracinius procesus, todėl skaitmenizavimas tampa neišvengiamu sprendimu siekiant užtikrinti sklandų duomenų valdymą. Tyrimo rezultatai atskleidė, kad dauguma šiuo metu Lietuvos ikimokyklinio ugdymo įstaigų naudojamų sistemų yra fragmentuotos, turi pasikartojančias funkcijas ir neužtikrina efektyvių duomenų mainų tarp skirtingų veiklos procesų. Jose egzistuoja baziniai buhalterinės apskaitos moduliai, tačiau trūksta vaikų veiklų administravimo funkcijų, kurios yra būtinos sklandžiam

ugdymo proceso valdymui. Verslo valdymo sistema (VVS) yra optimalus sprendimas, leidžiantis efektyviai integruoti įstaigos ir vaikų veiklų valdymą, administravimą ir organizavimą. Tokios sistemos diegimas ne tik sumažintų darbo dubliavimą, bet ir pagerintų procesų integraciją, palengvindamas darbuotojų kasdienes užduotis. Vieninga ir centralizuota informacinė sistema padėtų didinti darbo efektyvumą, gerinti duomenų valdymą ir užtikrinti aukštesnę ugdymo kokybę.

Literatūra

- [1] Malinauskienė, D. (2020). Ikimokyklinio ugdymo pokyčiai: Europos ir Lietuvos kontekstas. *Švietimas: politika, vadyba, kokybė*, 12(1), 4-7.
- [2] Stašionytė, A. (2010). *Lietuvos ikimokyklinių įstaigų sistemas plėtros politikos analizė* (Magistro darbas, Vilniaus universitetas). Kaunas: Lithuanian Academic Libraries Network LABT.
- [3] Litvinienė, J. (2002). *Šeima – vaiko ugdymo institucija*. Klaipėdos universiteto leidykla.
- [4] Vaičiūnaitė, D. ir Brazdeikis, V. (2016). Mokyklos informacinė sistema kaip mokyklos valdymo įrankis, *Švietimo problemos analizė*, 12(62).
- [5] Fitria, N., Haryadi, D. ir Bustan, R. (2021). The Utilization of Learning Device Information System (SIPP) as a Child Development Report Preparation Tool (RAPOT) in Kindergarten, In *International Conference on Educational Sciences and Teacher Profession (ICETeP 2020)*, 403-410.
- [6] Rohita, R. ir Hidayat, N.R. (2023). The Effectiveness of Using Learning Device Information Systems in Preparing Learning Plans, *Jurnal Pendidikan Usia Dini*, 17(1), 155 – 172.
- [7] Balkytė, G. ir Jurevičienė, J. (2021). Kompiuterizuotos apskaitos programų pasirinkimo kriterijai, *Studijos-Verslas-Visuomenė: dabartis ir ateities įžvalgos*, (VI), 4-17.
- [8] Nevda. (n.d.). Finansų valdymas ir apskaita. Nevda.lt. <https://nevda.lt/finansu-valdymas-ir-apskaita/> (Žiūrėta: 2025-03-12).
- [9] Debetas. (n.d.). Debetas. Debetas.lt. <https://debetas.lt/produktai/> (Žiūrėta: 2025-03-12).
- [10] Labbis. (n.d.). Labbis. Labbis.lt. <https://labbis.lt/paslaugos/> (Žiūrėta: 2025-03-12).
- [11] Avilyš. (n.d.). Avilyš. Dvsavilyš.lt. <https://www.dvsavilyš.lt/#kas-yra-avilyš> (Žiūrėta: 2025-03-12).
- [12] Oracle. (n.d.). What is ERP?. Oracle.com. <https://www.oracle.com/erp/what-is-erp/> (Žiūrėta: 2025-03-12).
- [13] Syvantis. (n.d.). Dynamics 365 Business Central and Oracle NetSuite. Syvantis.com. <https://www.syvantis.com/business-central-vs-netsuite-comparison> (Žiūrėta: 2025-03-12).
- [14] Triangle. (n.d.). Dynamics 365 Business Central vs SAP Business One. Triangle.com. <https://www.triangle.es/minicapsulas-en/dynamics-365-business-central-vs-sap-business-one/> (Žiūrėta: 2025-03-12).
- [15] Cargo. (n.d.). Microsoft Dynamics 365 Business Central Guide. Cargo.com. <https://cargas.com/software/microsoft/dynamics-365-business-central/pricing/> (Žiūrėta: 2025-03-12).
- [16] Atkins, D. (2024). NetSuite vs Dynamics 365 Business Central: What's Best for You?. Encore.com. <https://www.encorebusiness.com/blog/netsuite-vs-dynamics-365-business-central> (Žiūrėta: 2025-03-12).

Skatinamojo mokymosi taikymas automobiliui apmokyti sėkmingai pravažiuoti duotu maršrutu

Oskaras Klimašauskas, Gintautas Dzemyda

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Duomenų mokslo ir skaitmeninių technologijų institutas,
Akademijos g. 4, LT-08412 Vilnius
oskaras.klimasauskas@mif.vu.lt

Santrauka. Šiame straipsnyje analizuojamas gilaus skatinamojo mokymosi (DRL) pritaikymas autonominiam automobilio valdymui simuliuotoje 2D lenktynių trasoje. Tyrimo metu lyginami du skirtingi įvesties duomenų tipai: spindulių pagrindu veikianti sistema, naudojanti atstumo iki trasos kraštų matavimus ir automobilio būsenos duomenis, bei vaizdo pagrindu veikianti sistema, apdorojanti aplink automobilių esančius vaizdus. Agentų mokymui naudojamas DDQN algoritmas, o jų veiklos efektyvumas vertinamas pagal nuvažiuotą atstumą. Eksperimentai atliekami su trimis skirtingomis mokymosi strategijomis: pradedant nuo trasos pradžios, pradedant atsitiktinėse trasos vietose ir mokant agentą dviejose skirtingose trasose. Tyrimo tikslas – nustatyti, kaip skirtingi įvesties duomenų tipai ir mokymosi strategijos veikia agentų mokymosi efektyvumą.

Raktiniai žodžiai: Mašininis mokymasis, Gilusis mokymasis, Skatinamasis mokymasis, Maršruto paieška, Demonstracinė aplinka.

1 Įvadas

Skatinamasis mokymasis yra galingas įrankis, leidžiantis agentams mokytis priimti optimalius sprendimus dinamiškose aplinkose. Tačiau skatinamojo mokymosi algoritmų efektyvumas labai priklauso nuo įvesties duomenų ir mokymosi parametrų. Šiame straipsnyje nagrinėjame gilaus skatinamojo mokymosi pritaikymą autonominio automobilio valdymui simuliuotoje aplinkoje. Pagrindinis dėmesys skiriamas dviejų skirtingų įvesties duomenų tipų – spindulių ir vaizdo – palyginimui. Spindulių pagrindu veikianti sistema teikia tikslus atstumo matavimus ir automobilio būsenos duomenis, o vaizdo pagrindu veikianti sistema apdoroja vizualinę informaciją iš aplinkos. Siekiame nustatyti, kurie įvesties duomenų tipai ir mokymosi strategijos leidžia agentams greičiausiai išmokti įveikti lenktynių trasą, pasiekti aukštą va-

žiavimo greitį ir efektyviausiai apibendrinti įgytas žinias naujoje, nematytoje trasoje.

Tyrimai [1] rodo, kad vizualinė informacija gali pasiekti panašų efektyvumą kaip ir struktūruoti duomenys, tačiau reikia:

- Ilgiau mokytis (dėl sudėtingesnio apdorojimo),
- Praleidinėti kadrus [2],
- Taikyti papildomą normalizaciją.

Tai rodo, jog įvesties pasirinkimas nėra tik techninis sprendimas – jis fundamentaliai keičia agento mokymosi eigą.

Be įvesties, agento veikimo efektyvumą taip pat reikšmingai lemia tai, kaip yra suformuota mokymosi aplinka. Mokymas pradedant nuo pastovios padėties leidžia greitai prisitaikyti prie konkretaus scenarijaus, tačiau agentas gali pernelyg užtrukti tyrinėdamas jau perprastą sekciją. Priešingai, atsitiktinių pradžios taškų arba skirtingų žemėlapių strategijos skatina aktyvesnį aplinkos pažinimą ir suteikiant įvairesnių mokymosi duomenų. Tokiu būdu agentui nereikia palaukti, kol išmoksta sekciją, kad gautų duomenų įvairovės. Tokie metodai kaip domeno atsitiktinės atrankos [3] ar mokymasis pagal mokymo programą [4] taip pat remiasi aplinkos įvairinimo idėja, siekiant išspręsti tyrinėjimo problemą. Tyrimai rodo, kad net nedidelis aplinkos sąlygų varijavimas ženkliai pagerina RL sistemų perkėlimo gebėjimus.

2 Metodologija

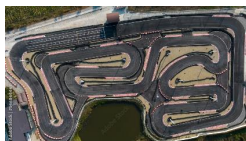
2.1 Aplinkos aprašymas

Eksperimentai buvo atlikti naudojant imituotą 2D juodai baltą lenktynių trasos aplinką, paremta 1 pav. [5] trasos forma, kurioje agentas (virtualus automobilis) pradeda kiekvieną epizodą trasos pradžioje kaip matome 2 pav. Agentas turi penkis veiksmų pasirinkimus, važiuoti pirmyn, atgal, sukti į kairę, sukti į dešinę, nieko nedaryt. Jis yra apdovanojamas proporcingai nuvažiuotam atstumui ir baudžiamas už neveikimą link tikslo, 3 pav. vaizduoja didėjantį atstumą nuo starto būsenos. Susidūrimas su trasos ribas žyminčia balta siena yra griežtai baudžiamas, epizodas nutraukiamas. Epizodas taip pat nutraukiamas po 5000 pasirinktų agento veiksmų. Tyrimo metu buvo įgyvendintos trys mokymosi strategijos:

- Bazinis mokymas: Agentas kiekvieną epizodą pradeda toje pačioje pradinėje pozicijoje.
- Mokymas su atsitiktiniais startais: Agentas kiekvieną epizodą pradeda atsitiktinai parinktoje trasos vietoje.

- Mokymas dviejose trasose: Agentas treniruojamas pakaitomis dviejose skirtingose lenktynių trasose, pradedant nuo fiksuotų pradinių pozicijų kiekvienoje trasoje.

Simuliacija kiekvienai strategijai truko 2000 epochų.



1 pav. Pavyzdinė trasa



2 pav. Imituota juodai balta trasa

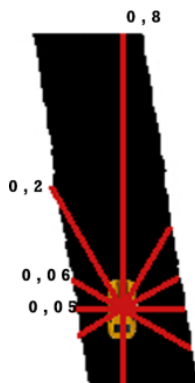


3 pav. Atstumų žemėlapis

2.2 Įvestys

Spinduliais grįsta įvestis

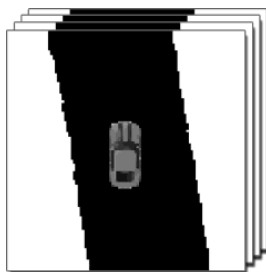
Šis įvesties metodas naudoja tiesiogiai iš simuliacinės aplinkos gaunamus duomenis. Agentui buvo pateikiami dešimt atstumo iki artimiausių trasos ribų matavimų, atliktų skirtingais kampais aplink automobilį, 4 pav vaizduoja jų pasiskirstymą. Be to, kaip įvesties savybės buvo naudojami ir vidiniai automobilio būsenos rodikliai: greitis, atbulinis greitis, teigiamas šoninis poslinkis (dreifas į dešinę) ir neigiamas šoninis poslinkis (dreifas į kairę). Taigi, spinduliais grįstos įvesties vektoriaus dydis buvo 14.



4 pav. Atstumą matuojantys spinduliai

Vaizdu grįsta įvestis

Vaizdu grįstas įvesties metodas rėmėsi vizualine informacija, gaunama tiesiogiai iš simuliacinės aplinkos, nereikalaujant tiesioginių simuliacijos būsenos duomenų. Agentas gavo aplink automobilį esančio vaizdo kadrus. Pradinis 125x125 pikselių spalvotas vaizdas buvo transformuojamas: pasukamas taip, kad automobilis visada būtų orientuotas vertikaliai į viršų, sumažinamas iki 50x50 pikselių ir konvertuojamas į nespalvotą formatą. Siekiant įvertinti automobilio greitį ir pagreitį, agento įvestį sudarė keturi tokie vienas po kito einantys vaizdo kadrai, jų vizualizaciją matome 5 pav. Vaizdu grįstos įvesties tenzorius dydis buvo [4, 50, 50]. Dėl kompiuterinių resursų apribojimų, buvo fiksuojamas tik kas ketvirtas simuliacijos žingsnio vaizdo kadras, todėl vaizdu grįstas metodas gavo mažesnę mokymosi duomenų kiekį nei spindulių metodas.



5 pav. Keturi vaizdo kadrai

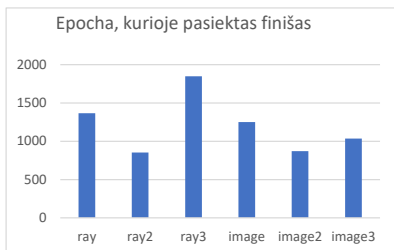
2.3 Algoritmai

Mokymui buvo naudojamas patobulintas Q-mokymosi algoritmas – Dvigubas Gilus Q-Tinklas (DDQN) [6]. Visuose eksperimentuose tarp skirtingų įvesties metodų buvo naudojami identiški hiperparametrai. Tačiau, atsižvelgiant į tai, kad vaizdo įvestis apima vizualinius duomenis, vaizdo informacijos apdorojimui buvo integruotas konvoliucinių neuronų tinklų (CNN) sluoksnis [7].

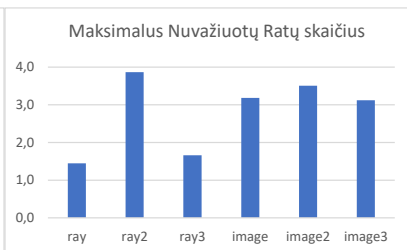
3 Rezultatai

Lentelėse „ray“ nurodo spindulių tipo duomenis, o „image“ – vaizdo. Skaičius 2 nurodo jog tai antroji strategija su atsitiktiniais startais, skaičius 3 – dviejų trasų. Eksperimento rezultatai rodo, kad atsitiktinių pradinių pozicijų naudojimas (ray2, image2) pagerino maksimalų pasiektų ratų skaičių vienoje epochoje, kaip matome 6 pav. ir žiūrint į 7 pav. sumažino epochų skaičių,

reikalingą trasai įveikti. Ray2 agentas pasiekė 3,9 rato per vieną epochą, o image2 – 3,5 rato, palyginti su atitinkamai 1,4 ir 3,2 rato fiksuotoje starto pozicijoje. Be to, šie modeliai pademonstravo geresnį apibendrinimą nematytoje trasoje, ką rodo 8 pav., ypač image2, kuris sugebėjo pilnai apvažiuoti prieš tai nematytą trasą, o ray2 agentas – 73 % trasos. Šie rezultatai rodo, kad atsitiktinės pradinės pozicijos skatina agentą įgyti bendresnius vairavimo įgūdžius ir geriau prisitaikyti prie nežinomų sąlygų.

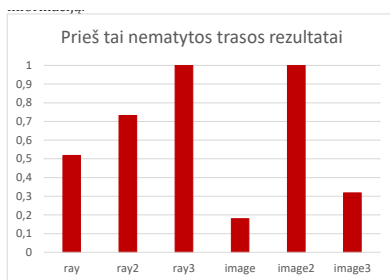


6 pav. Metodų palyginimas pagal finišo pasiekimą



7 pav. Metodų palyginimas pagal nuvažiuotus ratus epochoje

Mokymas dviejose trasose taip pat labiau pagerino apibendrinimą, ypač ray3 agentui, kuris taip pat sugebėjo visiškai įveikti nematytą trasą. Tačiau šis metodas padidino mokymo trukmę: ray3 agentui prirėikė 1848 epochų, o image3 – 1036 epochų, kad pasiektų vienos trasos įveikimą. Vaizdu pagrįsti modeliai buvo stabilesni treniruotėse, tačiau jų našumas nematytoje trasoje buvo mažesnis nei spindulių pagrindu veikiančių modelių. Viena iš galimų priežasčių – dėl prieš tai minėto kadru praleidimo, vaizdo tipo agentai gavo mažiau mokymosi duomenų, kas galėjo neigiamai paveikti jų gebėjimą apibendrinti informaciją.



8 pav. Nematytos trasos nuvažiuota dalis

4 Išvados

Remiantis šiame tyrime atlikta spindulių ir vaizdo tipo įvesčių metodų analize autonominio automobilio valdymo uždavinyje, galima daryti šias išvadas:

1. Vaizdu grįstas metodas, ypač derinant su atsitiktinių pradinių pozicijų strategija, pasižymėjo gebėjimu apibendrinti įgytas žinias ir sėkmingai įveikė visiškai nematytą lenktynių trasą. Šis pasiekimas yra reikšmingas, atsižvelgiant į tai, kad agentas mokymosi metu neturėjo tiesioginės prieigos prie vidinių automobilio būsenos duomenų, tokių kaip greitis ar poslinkis, o veikė tik apdorodamas vizualinę informaciją iš aplinkos ir taip pat gavo mažiau mokymosi duomenų dėl kompiuterio resursų apribojimų.
2. Spinduliais grįstas metodas, ypač naudojant atsitiktinių pradinių pozicijų strategiją, parodė žymiai geresnį našumą treniruočių metu ir padidino agento gebėjimą apibendrinti įgytas žinias, lyginant su fiksuota pradine pozicija. Vis dėlto, treniravimo strategija, apimanti mokymąsi dviejose skirtingose lenktynių trasose (ray3), užtikrino dar geresnį apibendrinimą ir leido agentui sėkmingai įveikti nematytą trasą, nors tokiai strategijai pareikalavo ilgesnio mokymosi laiko apvažiuoti vieną trasą kurioje mokinosi.

Tyrimas atskleidė, kad tinkamas įvesties duomenų tipo pasirinkimas ir mokymosi strategijos pritaikymas turi didelę įtaką gilaus skatinamojo mokymosi agentų našumui autonominio valdymo.

Literatūra

- [1] A. Zhang, R. McAllister, R. Calandra, Y. Gal ir S. Levine, Learning invariant representations for reinforcement learning without reconstruction, 2021 International Conference on Learning Representations.
- [2] M. V., K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra ir A. Riedmiller, Playing Atari with Deep Reinforcement Learning, arXiv:1312.5602 2013.
- [3] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba ir P. Abbeel, Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World, 2017. IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, 2017.
- [4] Y. Bengio, J. Louradour, R. Collobert ir J. Weston, Curriculum learning, 2009 ICML ,09: Proceedings of the 26th Annual International Conference on Machine Learning.
- [5] Adobe Stock, [Tinkle]. Available: <https://stock.adobe.com/lt/images/aerial-top-view-race-kart-track-track-for-auto-racing-top-view-car-race-asphalt-and-curve-grand-prix-street-circuit-aerial-view-asphalt-race-track/566864463>. [Kreiptasi 06 04 2025].
- [6] H. v. Hasselt, A. Guez ir D. Silver, Deep Reinforcement Learning with Double Q-learning, 2015 Proceedings of the AAAI conference on artificial intelligence. Vol. 30. No. 1.
- [7] A. Krizhevsky, I. Sutskever ir G. E. Hinton, ImageNet Classification with Deep Convolutional, Curran Associates, Inc., 2012, pp. 1097-1105).

Implementing EVM-Based Self-Sovereign Identity to Meet European Digital Identity Compliance for Decentralized Finance

Gintarė Košubienė¹ [0009-0001-8548-3253] and
Saulius Masteika² [0000-0002-1770-670X]

^{1,2} Vilnius University, Muitines 8, LT-44280 Kaunas, Lithuania, EU
gintare.kosubiene@knf.vu.lt, saulius.masteika@knf.vu.lt

Abstract. This paper presents an implementation of a Self-Sovereign Identity (SSI) framework using Ethereum-based standards to meet the technical requirements of the European Digital Identity (EUDI) Architecture Reference Framework (ARF). By leveraging ERC-734/ERC-735 standards, the proposed eSSI system enables decentralized key management, verifiable claims, and on-chain auditability. A case study on the Sepolia testnet demonstrates functional alignment with EUDI goals, while highlighting the need for enhanced privacy mechanisms such as zero-knowledge proofs for full compliance.

Keywords: digital identity, SSI, ERC-734/ERC-735, EUDI ARF

1 Introduction

Digital identity is rapidly becoming a foundational layer of the modern digital economy, enabling access to services, legal interactions, and regulatory compliance. As identity fraud, synthetic identities, and document forgery escalate—particularly with the rise of AI-generated deepfakes and falsified credentials—the urgency for secure, privacy-preserving, and verifiable identity systems has never been greater [1].

In response to these challenges, the European Union has been actively developing a common digital identity framework for a decade [2]. The culmination of these efforts is the European Digital Identity (EUDI) initiative, alongside its Architecture Reference Framework (ARF), which lays out technical design principles such as selective disclosure, minimal data use, cross-border interoperability, and robust auditability [3].

Meanwhile, innovation in the private sector is advancing at a significantly faster pace than the development of public infrastructure for EUDI such as the European Blockchain Services Infrastructure (EBSI). Public blockchains,

particularly those using the Ethereum Virtual Machine (EVM), have gained traction as flexible platforms for SSI, offering programmable smart contracts and broad ecosystem adoption [5]. Ethereum-based projects such as uPort pioneered on-chain identity management with decentralized key control and claim issuance, but also revealed challenges in privacy and scalability [22]. More recent solutions, like Privado ID (formerly Polygon ID), leverage zero-knowledge proofs to enhance privacy-preserving verifiability, further demonstrating the evolving capabilities of public chains in supporting regulatory-compliant digital identity [23]. Despite these advancements, none of the existing solutions have been explicitly designed to comply with the specific technical and regulatory requirements set out in the EUDI Architecture Reference Framework (ARF). This forms the basis of our central hypothesis.

Hypothesis: An EVM-compatible digital identity framework, built on SSI principles, can fulfill the requirements of the EUDI ARF when implemented for decentralized finance.

To investigate this, the paper reviews the principles of SSI (Section 2), compares key Ethereum identity standards and their alignment with EUDI (Section 3), and presents a case study of an implementation on the Ethereum Sepolia testnet (Section 4), demonstrating how open blockchain infrastructure can bridge the gap between SSI principles and EUDI requirements.

2 Self-Sovereign Identity (SSI) Principles and Their Technological Foundations

Self-Sovereign Identity (SSI) defines a decentralized and user-centric approach to digital identity management, where individuals retain full control over their credentials and personal data [6]. The conceptual foundation of SSI is based on ten core principles articulated by Christopher Allen—one of the pioneers of decentralized identity [6]. These principles (see Table 1) address both the ideological imperative of user autonomy and the technical challenge of implementing identity systems without central authority.

Each SSI principle can be mapped to a set of enabling technologies such as Decentralized Identifiers (DIDs), Verifiable Credentials (VCs), etc. The mapping approach combined conceptual analysis of Allen's SSI principles

[6] with a review of enabling technologies and relevant standards EUDI ARF [5], W3C [7], OpenID [8] and others [9-15]. This ensured that the resulting mappings reflect both SSI's technical vision and the compliance demands of European digital identity initiatives.

Table 1. SSI Principles and Their Technological Foundation.

SSI principle	Description	Technologies	Standards by EUDI
Existence	Users must be able to exist in the digital world, without the need for a third party.	Decentralized Identifiers (DIDs)	W3C DID [7]
Control	Users must control their identity and how it is used, shared, or hidden.	Consent management, Selective Disclosure	OpenID4VP [8], W3C VC [7]
Access	Users must have full access to their identity data and claims, including records that indicate any changes associated with their identity.	Digital Wallets, Identity Agents	EUDI Wallet
Transparency	Identity systems and their algorithms must be open, auditable, and understandable.	Open-Source Frameworks, Auditable Smart Contracts	ESSIF [9]
Persistence	Identities should be long-lasting, ideally as long as the user wants. But users must also be able to delete them when desired.	Blockchain-based Registries, PKI, Revocation Mechanisms	DKMS [10]
Portability	Identities must be transferable between systems and platforms.	Interoperable Wallets, W3C-compliant Formats	W3C VC [7]
Interoperability	Identities should work across platforms and borders, ensuring usability in diverse, global digital environments.	DIDComm [11], JSON-LD [12], OIDC [13]	DIF [11], EUDI ARF Interfaces [5]
Consent	Users must give informed consent before any identity data is shared.	Zero-Knowledge Proofs	OpenID4VC [8]
Minimization	Only the minimum necessary data should be shared.	Selective Disclosure, Zero Knowledge Proofs	ISO/IEC 27551 [14], W3C VC [7]
Protection	Identities must be protected against tampering and misuse.	Encryption, DKMS, Biometric Factors	ETSI TS 119 312 [15]

The EUDI concept closely aligns with the core principles of Self-Sovereign Identity (SSI), as outlined in the technical standards of the EUDI Architecture Reference Framework (ARF) (see Table 1, last column). The following section explores relevant Ethereum Request for Comment (ERC) standards that serve as the technical bases for implementing SSI in a way that ensures compatibility with EUDI compliance.

3 Comparative Analysis of ERC Standards for EUDI Compliance

Ethereum offers a modular and extensible environment for implementing decentralized identity systems. Several ERC standards have been proposed and partially adopted to enable various components of Self-Sovereign Identity (SSI), including identity creation, key management, credential issuance, and revocation. This section presents a comparative evaluation of both established and emerging ERC standards based on their support for SSI features and alignment with the standards supported by European Digital Identity (EUDI).

ERC-725 / ERC-734 / ERC-735 [16-17] form a foundational trio for on-chain identity management: ERC-725 defines a proxy smart contract that acts as a digital identity controlled by one or more keys, ERC-734 manages the associated keys with varying purposes (management, action, etc.), ERC-735 enables storing verifiable claims (or attestations) about the identity, issued by third parties. Together, these standards support user control, on-chain auditability, and integration with trusted issuers. However, privacy limitations arise when claims are publicly accessible, and additional off-chain data handling is needed to meet GDPR and ARF privacy standards.

EIP-1056 (did:ethr) [18] introduces a lightweight registry-based Decentralized Identifier (DID) method fully aligned with W3C DID specifications. It enables DID ownership and control using Ethereum addresses, key rotation and delegation without requiring full contract deployment, off-chain DID resolution using standard DID documents. EIP-1056 supports SSI's portability and persistence goals while keeping minimal identity data on-chain, aligning well with the EUDI ARF's emphasis on privacy and interoperability.

ERC-780 [19] offers a global claims registry allowing third parties to issue attestations to any Ethereum address. It is efficient for public claims but

lacks selective disclosure mechanisms. Consequently, it is less suitable for privacy-sensitive credentials unless used with zero-knowledge or off-chain validation layers.

ERC-1484 [20] aggregates multiple Ethereum addresses under a single digital identity. This model facilitates identity portability and unified credential management, especially across different applications or key pairs. However, its adoption remains limited, and integration with broader standards such as DIDs and VCs is still evolving.

ERC-1812 [21] is an emerging standard for off-chain verifiable credentials that includes an on-chain revocation registry. Key features include: credential issuance using EIP-712 typed data, support for off-chain storage and selective disclosure, lightweight on-chain revocation checks via credential hashes. This architecture closely aligns with EUDI ARF's privacy and minimal disclosure principles while ensuring verifiability and auditability. It is especially suitable for systems requiring compliance with data protection regulations.

Table 2. Comparison of ERC Standards Alignment with SSI Principles and EUDI Requirements.

Standard	SSI Feature Support	EUDI Alignment
ERC-725/734/735	High (identity, key management, claims)	Strong (requires off-chain privacy support)
EIP-1056	High (DIDs, key delegation)	Strong (W3C DID compatibility)
ERC-780	Moderate (claims registry)	Weak (no selective disclosure)
ERC-1484	Moderate (identity aggregation)	Moderate (experimental)
ERC-1812	Strong (VCs, off-chain privacy)	Strong (selective disclosure, ZK)

Based on the comparison in Table 2, EIP-1056 and ERC-1812 offer the most comprehensive support for SSI implementation in compliance with the EUDI. They enable decentralized key management, privacy-respecting credential issuance, and lightweight revocation mechanisms. Meanwhile, ERC-725/734/735 serve as a robust and flexible foundation for identity prototyping, particularly where on-chain execution, traceability, and compliance testing are required. Their smart contract-native structure makes them ideal for early-stage experimentation and integration with decentralized applications.

4 eSSI Implementation Based on ERC-734/ERC-735

Our European Self-Sovereign Identity (eSSI) system has been implemented using the ERC-734 and ERC-735 standards, chosen for their high SSI feature support and strong compatibility with EUDI framework (see Table 2). A high-level overview of the smart contracts' architecture is described in Figure 1.

As illustrated in Figure 1, the *eSSI*¹ smart contract functions as a deployment and management hub for digital identities and their associated wallets. It contains mappings to track the relationships between each identity and wallet, while also maintaining registries of trusted service providers (issuers) and the set of supported claim topics.

The *eIdentity*² smart contract, based on ERC-734 and ERC-735, represents the core component for managing decentralized identities. It provides interfaces for identity owners to manage authentication keys and maintain claim records. The *eClaimIssuer*³ smart contract ensures that claims associated with an identity can be cryptographically validated and revoked. Claims—issued by external parties—capture information such as identity attributes or status, and are verifiable via digital signatures. Leveraging ERC-735, claims can be updated or revoked as needed, supporting dynamic identity states. For instance, a previously valid claim can be invalidated upon document expiration or regulatory changes, such as the wallet being added to a sanctions list.

By integrating these components, the system provides a scalable and secure digital identity infrastructure. It supports essential SSI features such as on-chain claim revocation, interoperability with decentralized applications (via claim topics), and potential compliance with the EUDI ARF.

5 Scenario Based Simulation

To validate the feasibility and scalability of our eSSI implementation, smart contracts were deployed on the Ethereum Sepolia Testnet, chosen for its high transaction volume and broad potential user base. Over 2,000 digital identities were programmatically deployed, and various claim issuance and revocation scenarios were tested—particularly involving sanction-based claim topics.

¹ <https://sepolia.etherscan.io/address/0xE2a385125BD3D3D62DAB37702984D517B9153b9c#code>

² <https://sepolia.etherscan.io/address/0x75B94C3393D48cf750687B8e532C1Fea28b2013#code>

³ <https://sepolia.etherscan.io/address/0x5bbf8f095312cf2cf54f8f5e7c9c035b22640ded#code>

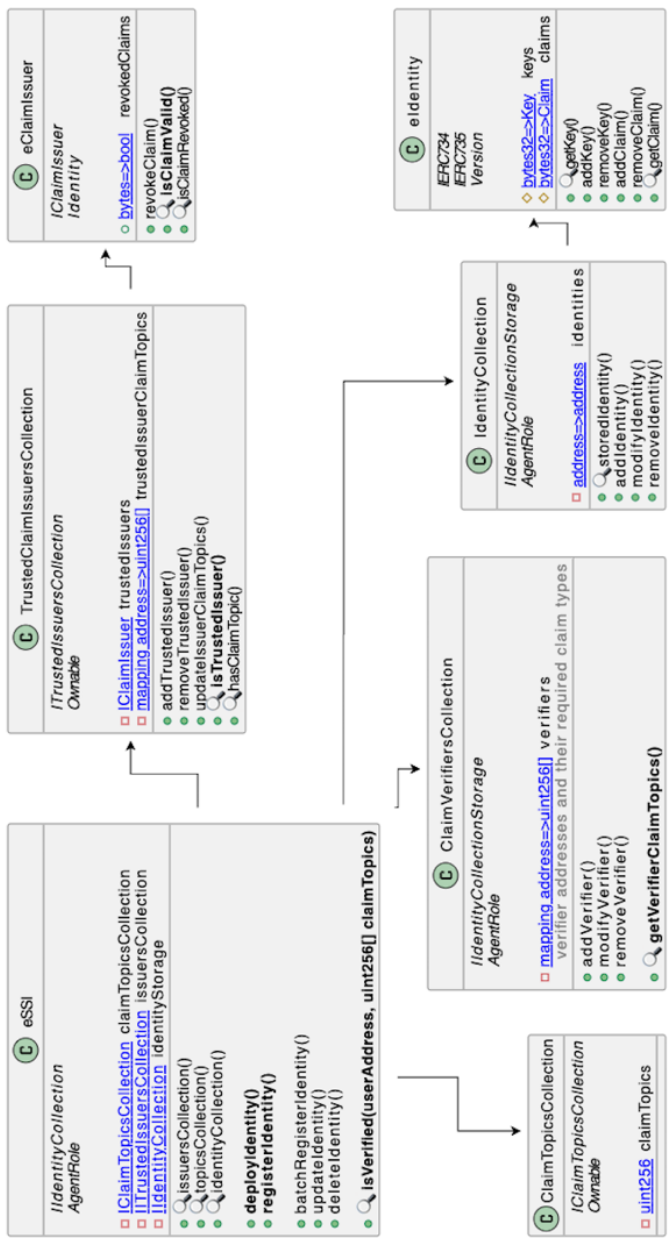


Figure 1. High-Level Smart Contracts Architecture of eSSI.

The simulation demonstrated that an ERC-734/735-based identity system can functionally align with both the foundational principles of SSI and the technical EUDI requirements. Identity owners were able to manage keys independently, and trusted issuers could add or revoke claims in real time. The system correctly manages claim changes enabling dynamic enforcement of trust relationships. For example, when an identity received a claim indicating a regulatory issue, that claim could be revoked, rendering the identity invalid for specific operations. The use of structured claim topics played a key role in facilitating interoperability, allowing decentralized applications to understand and process claims in a standardized way.

However, one major limitation observed during the simulation was the lack of privacy-preserving mechanisms. Since all claim data and interactions were recorded on-chain, the system does not yet address privacy requirements such as those outlined in GDPR or in the EUDI ARF itself. To advance toward production readiness, future versions must incorporate technologies like Zero-Knowledge Proofs (ZKPs) and off-chain data anchoring strategies to ensure selective disclosure and user consent-driven data sharing.

6 Conclusion

The proposed eSSI solution, based on ERC-734/ERC-735 standards, confirms that Ethereum-based SSI implementations can fulfill key EUDI ARF requirements in decentralized finance contexts. This identity system enables decentralized key management, verifiable claims, and on-chain auditability. However, further enhancements—especially privacy-preserving features—are necessary to reach production-readiness. This work outlines a viable, regulation-aware blueprint for digital identity systems built on public blockchain infrastructure.

Acknowledgments. The publication was prepared during the implementation of the project “Implementation of R&D activities, creating APV products by Deverium, UAB” (Project No. 02-020-K-0034). The project is co-financed by the European Union. Findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Deverium, UAB.

References

- [1] Zhang, C.J., Gill A.Q., Liu, B., Anwar, M.J. (2025), AI-based Identity Fraud Detection: A Systematic Review, <https://doi.org/10.48550/arXiv.2501.09239> .
- [2] Official Journal of the European Union, (2014), Regulation (EU) No 910/2014 of the European Parliament and of the Council of 23 July 2014 on electronic identification and trust services for electronic transactions in the internal market and repealing Directive 1999/93/EC, <http://data.europa.eu/eli/reg/2014/910/oj>.
- [3] European Commission, (2021), Architecture and Reference Framework, <https://eu-digital-identity-wallet.github.io/eudi-doc-architecture-and-reference-framework/1.8.0/architecture-and-reference-framework-main/>, last accessed 2025-04-07.
- [4] Biedermann, B., Scerri, M., Kozlova, V., Ellul, J. (2023), A Systematization of Knowledge: Connecting European Digital Identities with Web3. <https://arxiv.org/pdf/2409.19032>.
- [5] Biedermann, B., Scerri, M., Kozlova, V., Ellul, J. (2025), Aggregating Digital Identities through Bridging. An Integration of Open Authentication Protocols for Web3 Identifiers. <http://dx.doi.org/10.48550/arXiv.2501.13770>.
- [6] Allen, C. (2016), The Path to Self-Sovereign Identity, <https://www.lifewithalacrity.com/article/the-path-to-self-sovereign-identity/>, last accessed 2025-04-07.
- [7] W3C, Verifiable Credentials Data Model v1.1, (2022), <https://www.w3.org/TR/vc-data-model/>, last accessed 2025-04-07.
- [8] OpenID, (2022), Whitepaper of OpenID for Verifiable Credentials, https://openid.net/wordpress-content/uploads/2022/06/OIDF-Whitepaper_OpenID-for-Verifiable-Credentials-V2_2022-06-23.pdf , last accessed 2025-04-07.
- [9] The European Self-Sovereign Identity Framework Lab, (2019), SSI Standards Overview, <https://tno-ssi-lab.github.io/standardisation-overview/>, last accessed 2025-04-07.
- [10] Reed, D, Law, J., Hardman, D., Lodder, M., (2019), DKMS (Decentralized Key Management System) Design and Architecture V4, <https://github.com/hyperledger/indy-hipe/tree/49fc d78883d38babe9c95a4e1d150969797cfa2/design/dkms>, last accessed 2025-04-07.
- [11] Identity Foundation, DIDComm Messaging v2.x <https://identity.foundation/didcomm-messaging/spec/>, last accessed 2025-04-07.
- [12] W3C JSON-LD Working Group, JSON for Linking Data, <https://json-ld.org/>, last accessed 2025-04-07.
- [13] OpenID, How OpenID Connect (OIDC) Works, <https://openid.net/developers/how-connect-works/>, last accessed 2025-04-07.
- [14] The International Organization for Standardization and the International Electrotechnical Commission, (2021), ISO/IEC 27551: Information security, cybersecurity and privacy protection — Requirements for attribute-based unlinkable entity authentication, <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:27551:ed-1:v1:en>
- [15] Electronic Signatures and Infrastructures Technical Committee, (2023) Cryptographic Suites, https://www.etsi.org/deliver/etsi_ts/119300_119399/119312/01.04.03_60/ts_119312v010403p.pdf
- [16] ERC-725 Identity Standard, <https://erc725alliance.org/>, last accessed 2025-04-07.
- [17] Frozeman, F.G., (2017) ERC: Key Manager #734, <https://github.com/ethereum/EIPs/issues/734>, last accessed 2025-04-07.
- [18] Ethereum Improvement Proposals, (2018), ERC-1056: Ethereum Lightweight Identity, <https://eips.ethereum.org/EIPS/eip-1056> , last accessed 2025-04-07.

- [19] Tortensson, J., (2017), ERC: Ethereum Claims Registry, <https://github.com/ethereum/EIPs/issues/780> , last accessed 2025-04-07.
- [20] Ethereum Improvement Proposals, (2018), ERC-1484: Digital Identity Aggregator <https://eips.ethereum.org/EIPS/eip-1484>, last accessed 2025-04-07.
- [21] Ethereum Improvement Proposals, (2019), ERC-1812: Ethereum Verifiable Claims, <https://eips.ethereum.org/EIPS/eip-1812>, last accessed 2025-04-07.
- [22] Goel, A., Rahulamathavan, Y., (2024), A Comparative Survey of Centralised and Decentralised Identity Management Systems: Analysing Scalability, Security, and Feasibility, <https://doi.org/10.3390/fi17010001>.
- [23] Kaleido documentation, Privacy & Anonymity, <https://docs.kaleido.io/kaleido-platform/full-stack/privacy/#privacy-solutions>, last accessed 2025-04-07.

A Review of Quantum-Based Diffusion Models in Generative AI

Glauco Lima¹, Ernestas Filatovas², Marco Marcozzi² and Remigijus Paulavičius²

¹ Laboratory of Bioinformatics and Computational Biology, Federal University of ABC, Av. dos Estados, 5001 - Bangú, Santo André, Brazil

² Institute of Data Science and Digital Technologies, Vilnius University, Akademijos str. 4, LT-08412 Vilnius, Lithuania

glauco.endrigo@hotmail.com,

(ernestas.filatovas, marco.marcozzi, remigijus.paulavicius}@mif.vu.lt

Abstract. In recent years, the application of generative AI in several areas has been increasing. Concurrently, quantum computing has been advancing at an accelerated pace, unlocking new possibilities across various fields. This article provides an overview of the integration of quantum computing with generative AI, focusing on diffusion model techniques. We explore use cases documented in recent literature, illustrating how quantum computing techniques, when combined with diffusion models, are being leveraged to drive innovation.

Keywords: Quantum machine learning; Quantum diffusion model; Quantum Computing; Generative AI.

1 Introduction

Diffusion models have become key drivers of progress in generative Artificial Intelligence (AI) [1]. Their capabilities to produce diverse and high-quality synthetic data have led to widespread adoption. Platforms such as AlphaFold3 [2] and RFDiffusion [3] have achieved significant success using this technology in the field of generative biology. In generative imaging, DALL-E 2/3 from OpenAI [4], [5] and Stable Diffusion from Stability AI are good examples [6].

However, traditional diffusion models demand significant computational resources and encounter challenges in scalability and speed. These considerable resource demands present a major bottleneck in the broad deployment of diffusion models, as highlighted in recent research that outlines the high computational costs and the difficulties associated with efficiently training these models [7].

Recent computational challenges have spurred a growing interest in integrating quantum computing with diffusion models. This surge is paralleled by the rapid development of quantum technologies, particularly the advent of Noisy Intermediate-Scale Quantum (NISQ) devices [8]. NISQ processors are built on various quantum physical systems, each leveraging unique strategies. A widely adopted approach in quantum computing relies on superconducting circuits, utilizing qubits designed to operate at cryogenic temperatures [9]. These systems harness Josephson junctions to enable quantum coherence and gate operations. Additionally, superconducting circuits can be employed within quantum annealers, where optimization problems are solved through energy minimization processes [10]. An alternative paradigm is based on photonic quantum computing, where quantum information is encoded within the properties of light, such as polarization and phase. They can operate at room temperature and enable high-speed data transmission [11]. Another class of quantum computing platforms operates by directly manipulating individual atomic systems. Trapped ion-based architectures utilize electromagnetic fields to confine charged atomic species, leveraging their internal states for qubit encoding and employing laser pulses for quantum gate implementation [12]. Similarly, neutral atom-based computing relies on optical tweezers to trap Rydberg atoms [13]. Another type of quantum computing uses topologically protected Majorana-based qubits, and leverages particles called Majorana fermions to store and process quantum information [14].

The primary motivation for utilizing quantum computing lies in its ability to harness the unique properties of quantum mechanical systems, such as superposition, interference and entanglement. These features have the potential to enable the execution of computational tasks that would either be impossible or significantly more challenging on a classical supercomputer [15]. To the best of our knowledge, there is currently no survey on Quantum Diffusion Models (QDMs). This paper aims to fill that gap by reviewing and analyzing recent advances in the field, with a focus on comparing various QDM variants and their underlying architectures.

The remainder of this paper is organized as follows. Section 2 outlines the fundamental principles underlying diffusion models, key quantum encoding techniques and parameterized quantum circuits. Section 3 presents a review of QDMs, highlighting methodological innovations. Finally, Section 4 concludes the paper by summarizing the main findings.

2 Background

2.1 Diffusion Models

This section introduces the core architecture of Diffusion Models (DMs) and two important techniques used in Quantum Machine Learning (QML) that are key to these models.

Diffusion models are a class of generative models gaining traction for their robust performance in data synthesis, particularly in image generation. They operate through a two-phase process rooted in statistical physics. As shown in Figure 1, a data sample undergoes a forward diffusion process, where noise is incrementally added over multiple steps, transforming it into a near-random state. This systematic degradation is intuitive: the data “spreads out” until its original structure is lost. A neural network is then trained to reverse this process, learning to denoise the sample step by step, reconstructing it from noise into a coherent output matching the target distribution [16].

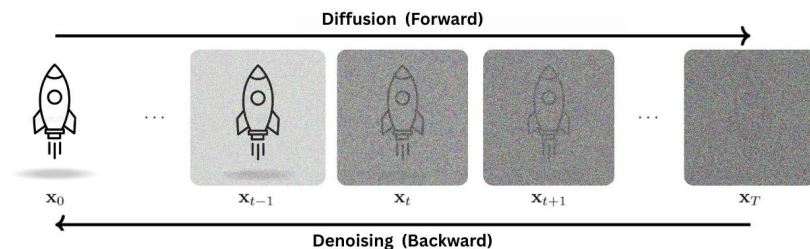


Fig. 1 Illustration of the forward diffusion (top arrow) and backward denoising (bottom arrow) processes in a diffusion model. Starting from the clean image x_0 (left), noise is progressively added to obtain increasingly corrupted versions $\{x_1, x_2, \dots, x_T\}$ (right). During sampling (reverse diffusion), the model iteratively denoises the noisy image x_T back into a clean reconstruction x_0 .

2.2 Types of Encoding

In QMLs, classical data must be encoded into quantum states to enable quantum processing. Various encoding methods have been developed for this purpose [17]. Here we review the ones that have been used in QDMs.

Amplitude encoding, frequently used in QDMs, takes advantage of quantum superposition to efficiently represent high-dimensional data. The corresponding quantum state for a classical data point $x = (x_1, x_2, \dots, x_N)$ is given by:

$$|x\rangle = \frac{1}{\sqrt{\sum_{k=1}^N x_k^2}} \sum_{k=1}^N x_k |i_k\rangle$$

where $|i_k\rangle$ denotes the computational basis state. A key benefit here is that an N -dimensional vector can be encoded using only $\log_2 N$ qubits, which is exponentially more efficient than the classical representation [18].

Another approach, known as *quantum embedding*, has been explored for QDM research [19]. Quantum embeddings leverage quantum computers to map classical data into a high-dimensional Hilbert space. A quantum feature map, implemented as a quantum circuit $\Phi(x, \theta)$, transforms an input x into a quantum state $\Phi(x, \theta)|0\dots 0\rangle$ where θ are trainable parameters. These parameters are optimized—often using classical techniques like gradient descent—to maximize the separation between quantum states corresponding to different classes, measured via metrics such as the Hilbert-Schmidt distance [19].

Additionally, *angle encoding* embeds a classical N -dimensional vector $x = (x_1, x_2, \dots, x_N)$ into an N -qubit product state via the encoded state:

$$|x\rangle = \bigotimes_{k=1}^N R(x_k)|0\rangle$$

where R is typically a single qubit rotation (e.g., R_y). In practice, one first normalizes each feature x_k to $[0,1]$, then sets $\theta_k = 2 \arcsin(\sqrt{x_k})$ and applies $R_y(\theta_k)$ to qubit k , so that measuring each qubit recovers statistics tied to the original data [18].

2.3 Parameterized Quantum Circuits

Parameterized Quantum Circuits (PQCs) are essential to hybrid quantum-classical machine learning approaches [20]. They involve the following steps: (1) state preparation, typically initializing qubits in $|0\rangle^{\otimes n}$; (2) application of parameterized unitary operations (e.g., variational circuits); (3) measurement of the quantum state; and (4) classical optimization, where measurement outcomes are used to update the circuit parameters iteratively. This framework allows exploring high-dimensional Hilbert spaces to learn complex data patterns. However, it faces challenges such as barren plateaus, NISQ noise, and trade-offs between expressibility and trainability [20].

3 Quantum Generative Diffusion Models

This section examines various QDMs, with Table 1 summarizing the typical variants and their characteristics.

In the study [21], the authors propose a quantum adaptation of generative diffusion models, replacing traditional neural networks with PQCs. Two key variations were explored: a latent model and a conditioned model. The latent model employs a classical autoencoder to encode data into a lower-dimensional space before quantum processing, enabling the use of smaller PQCs and improving sample quality. Furthermore, the QDM was adapted into a conditioned version by increasing the Hilbert space dimension with additional qubits to encode labels, allowing for the generation of samples based on specific input conditions. According to the authors, increasing the number of measurements initially improves the quality of generated samples by introducing nonlinearity. One drawback they mention is that beyond a certain threshold, increasing measurements worsens performance by reducing sample variability. Excessive measurements reset qubits to a fixed state, erasing the initial noise information.

The work [22] collects measurement samples from multi-qubit states and then trains a variational diffusion model that progressively corrupts those samples with noise and employs a denoising network of residual and attention blocks to reverse the process. A key highlight of the approach is that it achieves high fidelity while using significantly less memory than RNNs and transformers—for learning the distribution of a 2-qubit quantum state. One limitation, however, is that all experiments are restricted to W and GHZ-type distributions.

The study [23] introduces a novel quantum generative model. This model transforms a target quantum state into a completely mixed state—a state where all possible quantum states are equally probable—through a non-unitary forward process, which increases entropy by not preserving state purity. This transformation is driven by a depolarization channel, a noise model that probabilistically replaces the state with the completely mixed state to simulate realistic quantum noise. The trainable backward process efficiently reconstructs the original state using parameter sharing, which reduces the number of parameters by reusing them across the model, and partial trace operations, which trace out auxiliary subsystems to maintain

non-unitarity while simplifying computations. Timestep embedding enhances this process by integrating information about the diffusion stage into the quantum state, guiding denoising. A pro of the work is its timestep embedding technique, which boosts performance by adaptively distributing quantum states on the Bloch sphere, enhancing temporal information learning over traditional qubit encoding. On the other hand, a con is the high gate complexity, $O(N^2)$, of the denoising circuit.

One notable work is the proposal [24], which introduces three distinct quantum approaches that leverage noise as a beneficial resource in generative modeling. In the first approach, the Classical-Quantum Generative Diffusion Model (CQGDM), the forward diffusion is executed using classical methods. However, the denoising phase is achieved by a Quantum Neural Network (QNN) that can be realized as a fully PQC or as a hybrid quantum-classical network. In the second approach, known as the Quantum-Classical Generative Diffusion Model (QCGDM), the diffusion process itself is performed in a quantum framework by subjecting quantum data to noise via quantum channels—such as depolarizing channels or dynamics modeled by Stochastic Schrödinger Equations—which gradually transform an initial quantum state into a maximally mixed state. The backward denoising is then handled by classical neural networks. The third approach, the Quantum-Quantum Generative Diffusion Model (QQGDM), fully embeds both the diffusion and denoising phases within a quantum domain. The forward process employs quantum noise channels to degrade an initial quantum state into a completely mixed state. The reverse process is carried out using PQCs that incorporate interactions with ancillary qubit systems that act as an environment and are then traced out. This fully quantum implementation enables the exploration and manipulation of complex quantum probability distributions. A highlight is that QQGDM generates non-classical probability distributions, achieving a high average quantum fidelity of 0.997 ± 0.013 in simulations for reconstructing one-qubit states. However, a drawback of their methodology is that its reliance on timestep-specific PQCs without time embedding complicates training, and the use of depolarizing channels limits noise diversity, risking barren plateaus and reduced scalability for broader quantum data distributions.

A thesis [25] explores the foundational principles of QDM using PQCs, highlighting key advancements in the field. This research introduces four

model variants: a base model defining the core architecture, a temporal model incorporating timestep embedding to encode temporal information, a conditional model enabling targeted image generation via label embedding and a hybrid model combining both embedding techniques. Additionally, this thesis presents a model capable of generating full-color images—an achievement that, to the best of our knowledge, is the first of its kind. A constraint is the reliance on single-qubit embedding for timestep and label information, which restricts the model's ability to scale to tasks requiring diverse or numerous labels, as the limited angular range reduces label distinguishability.

The authors of the work [26] proposed two quantum hybrid diffusion models for image synthesis by integrating variational quantum circuits (VQCs) into classical U-Net architectures. The first model, Quantum Vertex U-Net (QVU-Net), replaces ResNet convolutional layers at the U-Net's vertex with quantum layers. The second model, Quantvolutional U-Net (QuantvU-Net), incorporates quantum layers in the encoder for feature extraction. The PQCs are strategically integrated into the U-Net architecture at points where image dimensions are reduced, such as the vertex or second encoder level. This approach minimizes qubit requirements. However, as the authors themselves acknowledge, increasing the number of quantum circuits significantly raises computational demands, which slows training and inference times and creates scalability challenges for larger datasets or more complex models.

The work [27] introduces three novel quantum diffusion based algorithms, Label Guided Generation Inference (LGGI), Label Guided Denoising Inference (LGDI) and Label Guided Noise Addition Inference (LGNAI) to address few-shot learning challenges. The authors leverage QDMs to enhance data generation and inference under limited training samples. LGGI generates synthetic data to augment training for quantum neural networks (QNNs), while LGNAI and LGDI guide noise addition and removal during diffusion/denoising stages using label information. They strategically perform amplitude encoding on classical features and angle encoding on labels during training. They point out that if too many diffusion steps are applied, the original information may degrade excessively into noise, causing the denoising process to overemphasize the label and reconstruct a generic class prototype.

Table 1. Quantum Diffusion Model Studies and Their Contributions

Study	Contribution	Platform (qubits used)	Dataset
Cacioppo et al.(2023)	Proposes QDM with PQCs, introducing latent and conditioned variants for improved sample quality and conditional generation.	PennyLane, 27q - IBM Hanoi (8q full model, 3q latent, 7q conditioned)	Quantum Simulator: MNIST digits {0,1} (16×16), Latent MNIST digits 0-9 (28×28). Quantum Hardware: Reduced latent MNIST digits {0,1} (dimension 4).
Wang et al., (2023)	Uses forward noise-corruption and a parallel reverse denoising network of residual and attention blocks.	Classical system	Measurement outcomes from multi-qubit quantum states.
Chen and Zhao (2024)	Non-unitary forward process (depolarization channel) with parameter-shared, timestep-embedded backward reconstruction.	Tensorcircuit framework (1-8q)	Random quantum states
Parigi et al. (2024)	Classical-Quantum, Quantum-Classical, and Quantum-Quantum Generative DMs using quantum noise, PQCs, and classical NNs.	PennyLane; CQGDM: 4q, QCGDM: 1q, QQGDM: 2q	CQGDM: 1,000 points uniformly in [-1,1]. QCGDM/QQGDM: Random 1-qubit pure states.
Kivijervi (2024)	Explores foundational QDM variants (base, temporal, conditional, hybrid) achieving full-color image generation.	PennyLane (6q base, 7q temporal/conditional, 8q hybrid)	MNIST: 1,024 images resized to 16×16
De Falco et al. (2024)	Improved image quality, faster convergence, fewer parameters via transfer learning.	PennyLane/Flax (12q for VQC, 4q final channel)	MNIST and Fashion MNIST (28×28, 60,000 images)
Wang et al (2024)	Few-shot quantum diffusion algorithms: LGGI, LGNAI, LGDI.	IBM Almaden	MNIST (28×28), Digits MNIST (8×8), Fashion MNIST (28×28)
Shah and Vatsa (2025)	Pairwise Bell-state entangling to reduce qubits required.	PennyLane (6q/8q/10q for 8×8/16×16/32×32 images)	MNIST (28×28, 70,000 images); CIFAR-10 (32×32, 60,000 images). Tested at 8×8, 16×16, 32×32.
Wang et al. (2025)	Adapts DPM ODEs with Carleman linearization for quantum implementation.	Classical system	ImageNet-100 (128×128, 256×256, 512×512) in latent spaces (16×16×4, 32×32×4, 64×64×4)
Han and Patel (2025)	Quantum noise in generative diffusion models enhancing randomness.	127q - IBM Naxca	MNIST: 5,000 images

The study [28] presents an advancement over earlier methods, enabling the processing of higher-dimensional images with intricate pixel structures, even on platforms with limited qubits. They first flatten the 2D image data into a 1D vector, normalize it, and then apply amplitude encoding. Following this, the quantum circuit implements pairwise Bell-state preparation: Hadamard gates are applied to the first half of the qubits (excluding any ancilla qubits) to create superposition, and CNOT gates are used to entangle each qubit from the first half (control qubits from 0 to $n/2 - 1$) with a corresponding qubit in the second half (indices $n/2$ to $n - 1$). This entanglement establishes strong correlations, and this allows the subsequent PQC to operate on a reduced set of qubits while still accessing information from the entire input state. A highlight is that unlike hybrid models, which rely on classical autoencoders, their work reduces parameter count directly within the quantum circuit. A limitation, however, is their work's high computational time compared to classical models; the paper notes an exponential increase during classical simulations of the quantum circuits.

Recent research [29] has transformed the mathematical equations underlying Denoising Diffusion Probabilistic Models (DPMs), particularly the ordinary differential equations (ODEs), into a form suitable for processing by quantum computers through a technique known as "Carleman linearization". This work focuses on two key approaches: DPM-solver-k, which utilizes precise mathematical derivatives to approximate the model's behavior, and UniPC, which employs measurements at various points to estimate the model's evolution. The authors provide theoretical proof that their quantum algorithms would perform efficiently on future fault-tolerant quantum computers. Therefore, the authors take explicit 'out-of-the-box steps' to establish a new avenue for demonstrating quantum computing's utility in machine learning tasks. A critical downside, however, is that the truncation used in their method introduces approximation errors.

A recent work [30] introduces an approach involving three main steps. First, they encode each image into a quantum circuit by applying PCA followed by an angle embedding, transforming the image data into a format suitable for quantum processing. Next, they introduce noise using quantum gates: each qubit receives a rotation via an R_x gate to add noise, and then additional paired rotations that cancel out net movement but compound the noise effect. Finally, after running the circuit on quantum hardware, they measure the qubits to obtain marginal probability distributions,

reverse the angle embedding, and apply an inverse PCA to reconstruct the noisy image for training the diffusion model. Ingeniously, they repurposed quantum decoherence as a source of true physics-driven noise, bypassing classical pseudo-randomness. A downside of their method is that PCA and angle embedding discard spatial details during encoding, which degrades reconstruction fidelity.

Overall, the works on QDMs reviewed here operate within a modest qubit range of 1 to 16. A central trend across these studies is the optimization of qubit usage to overcome the limitations of current quantum hardware. This is achieved through techniques such as encoding data into lower-dimensional latent spaces prior to quantum processing [21], employing pairwise Bell-state preparation to exploit entanglement for efficient information access with fewer qubits [28], and resizing images to conserve computational resources [25].

Another prominent trend is the strategic use of quantum noise, stemming from inherent quantum fluctuations. Studies like [24] and [30] highlight how noise, introduced via quantum channels (e.g., depolarization channels) or gate-based rotations, facilitates the creation of complex, entanglement-driven probability distributions that are computationally infeasible to replicate classically. Predominantly, PQC are employed to reverse the diffusion process [21], [24], [25], [28]. Additionally, there is a clear trajectory toward integrating temporal and conditional information into QDMs, with timestep embedding [23], [25] and label-based conditioning [21], [25] enabling more precise control over the generative process. These trends suggest a field moving toward resource-efficient, noise-augmented, and hybrid models capable of tackling increasingly sophisticated generative tasks.

4 Conclusions

Quantum generative diffusion models draw inspiration from classical diffusion models, which have recently demonstrated state-of-the-art performance in a variety of generative tasks. In this work, we have reviewed recent advances in quantum diffusion models, focusing on aspects such as model architectures. As highlighted in the introduction, diffusion models are pivotal across diverse domains, making their enhancement through quantum computing crucial to overcoming classical limitations in scalability

and efficiency. This review, the first survey in the field of QDMs up to this date, consolidates advancements and provides foundational insights to propel the discipline forward.

Future research should prioritize rigorous evaluations of QDM architectures and methodologies. One suggested avenue for exploration involves leveraging quantum diffusion models in the context of graph-structured data. Such data, with their inherent relational complexity, could rigorously evaluate QDMs' ability to model interconnected systems.

References

- [1] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.
- [2] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.
- [3] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.
- [4] Gary Marcus, Ernest Davis, and Scott Aaronson. A very preliminary analysis of dall-e 2. arXiv preprint arXiv:2204.13807, 2022.
- [5] Mohamad-Hani Temsah, Abdullah N Alhuzaimi, Mohammed Almansour, Fadi Al-jamaan, Khalid Alhasan, Munirah A Batarfi, Ibraheem Altamimi, Amani Alharbi, Adel Abdulaziz Alsuhaibani, Leena Alwakeel, et al. Art or artifact: evaluating the accuracy, appeal, and educational value of ai-generated imagery in dall-e 3 for illustrating congenital heart diseases. *Journal of Medical Systems*, 48(1):54, 2024.
- [6] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion video synthesis with stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22680–22690, 2023.
- [7] Tianshuo Xu, Peng Mi, Ruilin Wang, and Yingcong Chen. Towards faster training of diffusion models: An inspiration of a consistency phenomenon. arXiv preprint arXiv:2404.07946, 2024.
- [8] Kishor Bharti, Alba Cervera-Lierta, Thi Ha Kyaw, Tobias Haug, Sumner Alperin-Lea, Abhinav Anand, Matthias Degroote, Hermanni Heimonen, Jakob S Kottmann, Tim Menke, et al. Noisy intermediate-scale quantum algorithms. *Reviews of Modern Physics*, 94(1):015004, 2022.
- [9] Wei-Yang Liu, Dong-Ning Zheng, and Shi-Ping Zhao. Superconducting quantum bits. *Chinese Physics B*, 27(2):027401, 2018.
- [10] Sheir Yarkoni, Elena Raponi, Thomas Bäck, and Sebastian Schmitt. Quantum annealing for industry applications: Introduction and review. *Reports on Progress in Physics*, 85(10):104001, 2022.

- [11] Shuntaro Takeda and Akira Furusawa. Toward large-scale fault-tolerant universal photonic quantum computing. *APL Photonics*, 4(6), 2019.
- [12] James D Siverns and Qudsia Quraishi. Ion trap architectures and new directions. *Quantum Information Processing*, 16:1–42, 2017.
- [13] Karen Wintersperger, Florian Dommert, Thomas Ehmer, Andrey HOURSANOV, Johannes Klepsch, Wolfgang Mauere, Georg Reuber, Thomas Strohm, Ming Yin, and Sebastian Luber. Neutral atom quantum computing hardware: performance and end-user perspective. *EPJ Quantum Technology*, 10(1):32, 2023.
- [14] David Aasen, Morteza Aghaee, Zulfi Alam, Mariusz Andrzejczuk, Andrey Antipov, Mikhail Astafev, Lukas Avilovas, Amin Barzegar, Bela Bauer, Jonathan Becker, et al. Roadmap to fault tolerant quantum computation using topological qubit arrays. *arXiv preprint arXiv:2502.12252*, 2025.
- [15] Yazhen Wang. *Quantum computation and quantum information*. 2012.
- [16] Paul A Geroski. Models of technology diffusion. *Research policy*, 29(4-5):603–625, 2000.
- [17] Deepak Ranga, Aryan Rana, Sunil Prajapat, Pankaj Kumar, Kranti Kumar, and Athanasios V Vasilakos. Quantum machine learning: Exploring the role of data encoding techniques, challenges, and future directions. *Mathematics*, 12(21):3318, 2024.
- [18] Tuan A Ngo, Tuyen Nguyen, and Truong Cong Thang. A survey of recent advances in quantum generative adversarial networks. *Electronics*, 12(4):856, 2023.
- [19] Seth Lloyd, Maria Schuld, Aroosa Ijaz, Josh Izaac, and Nathan Killoran. Quantum embeddings for machine learning. *arXiv preprint arXiv:2001.03622*, 2020.
- [20] Daniel T Chang. Parameterized quantum circuits with quantum kernels for machine learning: A hybrid quantum-classical approach. *arXiv preprint arXiv:2209.14449*, 2022.
- [21] Andrea Cacioppo, Lorenzo Colantonio, Simone Bordoni, and Stefano Giagu. Quantum diffusion models. *arXiv preprint arXiv:2311.15444*, 2023.
- [22] Yong Wang, Shuming Cheng, Li Li, and Jie Chen. Learning quantum distributions with variational diffusion models. *IFAC-PapersOnLine*, 56(2):5888–5893, 2023.
- [23] Chuangtao Chen and Qinglin Zhao. Quantum generative diffusion model. *arXiv e-prints*, pages arXiv–2401, 2024.
- [24] Marco Parigi, Stefano Martina, and Filippo Caruso. Quantum-noise-driven generative diffusion models. *Advanced Quantum Technologies*, page 2300401, 2024.
- [25] Nikolai Theien Kivijervi. Quantum diffusion model. *Master's thesis*, 2024.
- [26] Francesca De Falco, Andrea Ceschini, Alessandro Sebastianelli, Bertrand Le Saux, and Massimo Panella. Quantum hybrid diffusion models for image synthesis. *KI-Künstliche Intelligenz*, pages 1–16, 2024.
- [27] Ruhan Wang, Ye Wang, Jing Liu, and Toshiaki Koike-Akino. Quantum diffusion models for few-shot learning. *arXiv preprint arXiv:2411.04217*, 2024.
- [28] Shivalee RK Shah and Mayank Vatsa. Enhancing quantum diffusion models with pairwise bell state entanglement. In *International Conference on Pattern Recognition*, pages 347–361. Springer, 2025.
- [29] Yunfei Wang, Ruoxi Jiang, Yingda Fan, Xiaowei Jia, Jens Eisert, Junyu Liu, and Jin-Peng Liu. Towards efficient quantum algorithms for diffusion probability models. *arXiv preprint arXiv:2502.14252*, 2025.
- [30] Jason Han and Tirthak Patel. Turning quantum noise on its head: Using the noise for diffusion models to generate images. *ACM SIGMETRICS Performance Evaluation Review*, 52(4):23–24, 2025.

Decoding An Automobile's Technical Specification From Its Identification Number

Mantas Mačiūnas, Viktor Skorniakov

Vilnius University Faculty of Mathematics and Informatics,
Naugarduko str. 24, Vilnius
mantas.maciunas@mif.stud.vu.lt, viktor.skorniakov@mif.vu.lt

Abstract. We propose a method for decoding technical vehicle's parameters (make, model name, body type etc.) from its identification number. Classification is done for the entire specification at once, thus utilising the underlying dependencies between labels. To achieve the goal, several models were used – nearest neighbours, decision tree, extra trees and random forest classifiers.

Keywords: vehicle identification number, multi-label classification, random forest classifier, decision tree classifier, extra trees classifier, k-nearest neighbours classifier.

1. Introduction

In the 1980s, there was a serious attempt at standardising the identification of any particular vehicle by introducing a 17-character vehicle identification number (VIN). Despite the goal being standardisation, nowadays, several VIN standards are in use. Furthermore, even if some manufacturers use the same VIN structure, they are not obligated to encode the same information and the same symbols may carry different information. Therefore, VIN structure is very inconsistent among manufacturers, and there is no easy way to decode it. Although there are plenty of free specification decoder tools on the internet, they frequently are applicable to only one make or a group of makes (for example, "Volkswagen" model VINs are encoded analogously to those of "Audi" or "Seat" because all of these manufacturers belong to "Volkswagen Auto Group"). The most frequently used method of decoding – a VIN lookup table. Using it one can decode certain technical parameters [1]. Although this method can be very accurate, it is tremendously inefficient in the long run since each new make comes with a new VIN structure. Furthermore, every time a manufacturer updates one

of its current models, the lookup table has to be updated as well. Moreover, manufacturers are not inclined to freely share their VIN encoding schemas. In most cases, owners of a particular make of vehicle collectively analyse and figure out the meaning of the VIN symbols on their cars. Lastly, although universal VIN decoder tools exist, most of them are closed-source, so it is impossible to figure out how they work or suggest any improvements. This research paper offers an alternative – VIN decoding using machine learning methods. To achieve the goal, multi-label machine learning models were used since they are capable of utilising the underlying correlations and dependencies between the technical parameters, thus enabling the model to make more accurate predictions and avoid impossible parameter combinations. Conducting the literature analysis, we did not encounter articles considering the decoding of VINs. Therefore, the research direction pursued by us exhibits substantial novelty.

The rest of the paper is organized as follows: Section 2 describes the data, Section 3 focuses on the methodology, Section 4 describes the results; the concluding section is devoted to the summary.

2. Data

The VIN of an automobile is an international coding standard that appeared in the 1980s. It is a collection of 17 letters and numbers, consisting of three obligatory parts, in which the encoding of the manufacturer, model and serial number is mandatory. The manufacturer is identified by the first three VIN symbols, often referred to as WMI (world manufacturer identifier). Symbols in positions 4-8 define the model and its specification. However, this requirement is enforced very loosely: some manufacturers only encode the model name, whereas others encode everything from the model name to a specific engine. The standards of the European Union (and the majority of the rest of the world) and North America differ as well. From the 9th symbol onwards European manufacturers must encode the VIN in such a manner that it becomes unique, i.e. encode a serial number and (sometimes) encode more specific technical data. Meanwhile, the North American VINs' ninth symbol is used as a check symbol – ensuring the rest of the VIN is entered correctly; symbols 10-11 contain the model year and the plant in which the vehicle was manufactured; finally, from the 12th symbol onwards follows the serial number of the vehicle.

Data from various European national automotive registries containing the VINs and technical specifications of automobiles was used in this research paper. The manufacturers' list included "Toyota", "Lexus", "Seat", "Cupra", "Mercedes-Benz" and "BMW". The technical parameters chosen for classification training were the ones that were present in all datasets: make, model name, model generation code, body type, fuel type, engine displacement and engine power. Some labels, especially those of model names and generation codes, were tremendously rare. The direct cause of this was an inherent data disbalance in the dataset. The usual methods of solving this issue (for example, data augmentation) were inapplicable, because the data contained many cases where insignificant differences between VIN symbols (for example, different serial numbers) resulted in different label sets. Therefore, the best way to resolve the data disbalance was to create models that are insensitive to outliers and are able to accurately classify even rare data values.

One unit of measurement was picked for each measurable technical parameter. For example, the engine displacement values in some sources were in cubic centimetres, whereas in others – in litres. Because cubic centimetres are a more accurate unit of measurement, they were chosen as the unit for engine displacement. Records having displacement in litres were converted into cubic centimetres. In case of unsuccessful conversion, the data row was dropped. Kilowatts were selected as the unit of measurement for engine power.

Finally, we had to decide how to cope with the rare data instances. As mentioned before, data synthesis was too risky because this method might mislabel generated VINs, thus confusing the models that will be trained on the data. Another frequently utilised method – dataset size reduction to the least popular label – was also inapplicable. This method reduces the dataset by finding the least common label and removing values of all other labels until their frequency is the same as of this label. However, because this problem involves thousands of possible label vector values, leaving just a few of each value would very likely result in underfitted models. Furthermore, these rare VIN values reflect an essential characteristic of the population. Namely, these are either low-production models or very old (at least 30 years old) automobiles. Low-production model VINs are tremendously difficult to find, because any single one of the data sources used in this research (mainly European national automobile registries)

could contain a handful of these VINs at best, and there was insufficient time to expand the dataset to include more sources. Old cars, meanwhile, provide a similar challenge – 30 years after rolling out of the factory very few cars still drive on the road. According to the European Automobile Manufacturers' Association, the mean age of a car registered in Europe was only 12 years old in 2023. Thus, it is entirely likely that vehicles that are 30 or more years old constitute the absolute minority of all vehicles on the road. These challenges left only one solution for the data disbalance – the removal of rare values from the dataset. The rarity threshold was set to 10. After removing these values, the dataset contained 2,611,885 entries.

3. Methodology

3.1. Classifiers

These multi-label classifiers were used for modelling: random forest classifier (RFC), decision tree classifier (DTC), extra trees classifier (ETC) and k-nearest neighbours classifier (MLkNN).

Multi-label random forest classifier differs from its single-label sibling in two ways: 1) different metrics that are used for node splitting during training; 2) predictions are made differently [2]. When splitting a node, this classifier calculates the value of the splitting criterion for each label vector coordinate separately and uses the mean of these values to determine splits. This manner of splitting ensures that the tree is optimised to best predict label vectors rather than separate label vector coordinates [3]. Because the leaves are vectors, the prediction of a single tree is a full technical specification vector. Whenever the forest is supplied a VIN it has never seen before, the prognosis is produced by the process of voting [4].

A decision tree is a special case of a random forest classifier where the forest consists of only one tree. The main difference is that the random forest trees are fitted by using a randomly sampled dataset, thus creating a decorrelated tree ensemble, whereas a decision tree is fitted with the entire training dataset.

Extra trees classifiers work analogously to random forest classifiers, but they use an additional randomisation that is much stronger than those of the previously described classifiers [5].

The multi-label case of the k-nearest neighbours classifier is the same as the single-label one. This research paper utilises two search algorithms for the MLkNN model – K-D tree and ball tree [6], [7], [8].

3.2. Utilised software tools

The code used for this research paper was written using the “Python” programming language, version 3.10.16. All machine-learning methods were coded using the “scikit-learn” library.

3.3. Model accuracy metrics

The classification accuracy metrics for multi-label models differ significantly from those meant for single-label models. We utilised 6 accuracy metrics for models’ evaluation and choice. Two of these metrics are meant for separate labels’ vector coordinate scoring accuracy evaluation, whereas the remaining four are meant to evaluate the accuracy when classifying the entire label vector. First, the base formulae that are used for all the aforementioned accuracy metrics need to be defined. The indicator function is given by equation

$$1_{a=b} = \begin{cases} 1, & \text{if } a = b \\ 0, & \text{if } a \neq b. \end{cases}$$

Fractions of true positive predictions (*TP*), false positive predictions (*FP*) and false negative predictions (*FN*) are defined by equations

$$TP_C = \frac{1}{|C|} \sum_{c \in C} 1_{\hat{y}(c)=c'},$$
$$FP_C = \frac{1}{|N \setminus C|} \sum_{c \in N \setminus C} 1_{\hat{y}(c)=c'},$$
$$FN_C = \frac{1}{|C|} \sum_{c \in C} 1_{\hat{y}(c) \neq c'},$$

where *C* is one class of the classified coordinate, *N* is the union of all classes that are present in the dataset, and *C'* is a label denoting that the given data entry belongs to the class *C*.

Accuracy and $F1_{\text{macro}}$ for evaluating the single labels’ vector coordinate classification, are defined as follows:

$$Accuracy(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} 1_{\hat{y}_i=y},$$

$$F1_{macro}(y_L, \hat{y}_L) = \frac{(1 + 1^2)TP_L}{(1 + 1^2)TP_L + FP_L + 1^2FN_L},$$

$$F1_{macro}(y, \hat{y}) = \frac{1}{|L|} \sum_{l \in L} F1_{macro}(y_l, \hat{y}_l),$$

The second formula is devoted to a single label vector coordinate class L , whereas the third formula is the combined $F1_{macro}$ for a single coordinate of the label vector.

The remaining metrics are for the evaluation of the overall accuracy of the model. Two of these are the aforementioned accuracy and $F1_{macro}$ modified to fit multi-label classifiers:

$$Accuracy(Y, \hat{Y}) = \frac{1}{n} \sum_{y \in Y^t} Accuracy(y, \hat{y}),$$

$$F1_{macro}(Y, \hat{Y}) = \frac{1}{n} \sum_{y \in Y^t} F1_{macro}(y, \hat{y}),$$

here y is a single coordinate column, the amount of coordinates is labeled n and Y is the set of all label vectors, represented as a matrix, where one column is for a single technical parameter. So, the overall classification accuracy and $F1_{macro}$ are calculated by averaging the metric calculated for each separate technical parameter. Finally, two additional accuracy metrics devoted to multi-label models were used – *Hamming* loss and exact-match score [9]:

$$HammingLoss(Y, \hat{Y}) = \frac{1}{N} \sum_{i=1}^N \frac{|y_i \Delta \hat{y}_i|}{n},$$

$$ExactMatch(Y, \hat{Y}) = \frac{1}{N} \sum_{i=1}^N 1_{y_i = \hat{y}_i},$$

where N denotes the amount of rows in Y , i.e. the amount of data instances in Y .

4. Results

The full dataset was split into training and testing subsets by iterative stratification, which prioritises the least common labels and starts by

stratifying those first [10]. All models were trained with the same dataset (size – 2,089,508 rows), and all accuracy metrics were calculated after the trained models classified the same testing dataset (size – 522,377 rows). The optimal parameters for the models were chosen by using 10-fold cross-validation on a subset of the full dataset. This subset was created by randomly taking 10 of each unique row in the dataset.

As we can see in the first table, all models reached a quite high overall classification accuracy (>0.88). However, looking at the $F1_{macro}$ it becomes quite obvious that not all models were capable of classifying outliers equally well. The extra trees classifier in particular is the worst in this regard with only 0.763 $F1_{macro}$. The exact-match score also separates the extra trees classifier as the worst of the bunch – it managed to correctly decode the technical specification of only 59.5% of the testing data VINs. Therefore, for further work it is recommended to skip this model and use one of the three remaining ones – random forest classifier, decision tree or k-nearest neighbours classifier. For further analysis the decision tree was chosen as it beats both of the remaining models with respect to all metrics and manages to correctly predict the technical specification of more than 4 out of 5 VINs in the testing dataset – an exact-match score of 82.2%.

Table 1. The accuracy metrics for all models.

Classifier	Classification accuracy	$F1_{macro}$	Exact-match score	Hamming loss
RFC	0.932	0.870	0.747	0.068
DTC	0.950	0.923	0.822	0.050
ETC	0.884	0.763	0.595	0.116
MLkNN	0.930	0.878	0.764	0.070

The second table shows the accuracy metrics for all technical parameters separately when using a decision tree. All parameters are classified with at least 85% accuracy. The classification of an automobile’s make is especially accurate – the model only made mistakes for “Seat” and “Cupra” VINs (99.5% and 91.0% of “Seat” and “Cupra” VIN codes respectively had their make labeled correctly). The $F1_{macro}$ of the fuel type is quite a bit lower than that of other technical parameters – while analysing incorrectly labeled VINs it was noticed that the model classified VINs, where fuel type was labeled

as “Gasoline/LPG”, poorly – only 10.2% of such VINs had their fuel type labeled correctly. All remaining fuel types were classified with at least 87.7% accuracy. The lowest individual classification accuracy was achieved with engine power, but even this value of 0.859 should be considered impressive, given that this information is not even encoded in “Seat” and “Cupra” VINs.

Table 2. Accuracy metrics for separate technical parameters using the DTC model.

Technical parameter	Classification accuracy	F1 _{macro}
Make	0.997	0.985
Model name	0.974	0.951
Body type	0.974	0.986
Model generation code	0.994	0.992
Fuel type	0.950	0.802
Engine displacement	0.899	0.866
Engine power	0.859	0.879

5. Conclusions

This research paper proposes several machine-learning methods capable of decoding a vehicle’s technical parameters from its VIN. The optimal parameters for the models were selected using a small subsample and applying cross-validation to it. The data was split into training and testing subsets by stratifying it by the rarest labels, thus ensuring good data distribution in these subsets. After all models were fitted, the best results were achieved using the decision tree classifier which achieved 95% overall classification accuracy and correctly predicted the full technical specification of 82.2% VINs in the testing dataset. This model classified all individual technical parameters with at least 85.9% accuracy as well as F1_{macro} no smaller than 0.802. Future research will consider the improvement of classification of rare values as well as finding better ways to separate “Seat” and “Cupra” VINs.

Acknowledgements. The author is grateful to Assoc. Prof. V. Skorniakov for providing feedback on the final version of the manuscript.

References

- [1] W. Bachman, J. Granell, R. Guensler, and J. Leonard (1998). Research Needs for Determining Spatially Resolved Subfleet Characteristics. *Transportation Research Record*, 1625(1), pp. 139-146.
- [2] P. Geurts, L. Wehenkel, and F. d'Alché-Buc (2006). Kernelizing the output of tree-based methods. In *Proceedings of the 23rd international conference on Machine learning (ICML '06)*. Association for Computing Machinery, New York, NY, USA, pp. 345–352.
- [3] M. N. Dumont, R. Marée, L. Wehenkel, and P. Geurts (2009). Fast Multi-class Image Annotation with Random Subwindows and Multiple Output Randomized Trees. *International Conference on Computer Vision Theory and Applications*, pp. 196-203.
- [4] L. Breiman (1998). Arcing Classifiers. *The Annals of Statistics* 26(3), pp. 801–849.
- [5] P. Geurts, D. Ernst, and L. Wehenkel (2006). Extremely randomized trees. *Mach Learn* 63, pp. 3–42.
- [6] J. L. Bentley (1975). Multidimensional binary search trees used for associative searching. *Commun. ACM* 18(9), pp. 509–517.
- [7] J. H. Friedman, J. L. Bentley, and R. A. Finkel (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software (TOMS)*, 3(3), pp. 209-226.
- [8] S. M. Omohundro (1989). Five balltree construction algorithms, pp. 1-22.
- [9] D. Krstinić, M. Braović, L. Šerić, and D. Božić-Štulić (2020). Multi-label classifier performance evaluation with confusion matrix. *Computer Science & Information Technology*, 1, pp. 1-14.
- [10] K. Sechidis, G. Tsoumakas, and I. Vlahavas (2011). On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III* 22, pp. 145-158.

Virtual Reality Simulation Game for Economics Students: Usability Evaluation

Alona Makeieva, Svetlana Kubilinskiėnė

Vilniaus kolegija. Higher Education Institution,
Saltoniškių g. 58 - 1, LT-08105, Vilnius
alona.makeieva@stud.viko.lt, s.kubilinskiene@eif.viko.lt

Annotation. This paper presents experiences in the development and usability evaluation of the Erasmus+ project VR-based simulation tools content for developing economics students' critical thinking, communication, collaboration and creativity skills. The study involved 91 students from 3 countries. The research methodology to determine the relevance of the content is based on applying participant method (namely, Event diaries), were Keller's ARCS (attention, relevance, confidence, and satisfaction) motivational design model was used as a guiding strategy. In this study content usability: Relevance scored highest (3.64 out of 5), followed by confidence (3.37 out of 5), attention (3.5 out of 5), and content satisfaction (3.04 out of 5).

Keywords: Virtual reality, Usability, Virtual education, VR-simulation game, High education.

1. Introduction

This paper presents the experiences in the development and usability evaluation of a VR-based simulation game created as part of the Erasmus+ project "Development of a Virtual Learning Space as a Tool for Developing Students' Critical Thinking, Communication, Collaboration and Creativity Skills in the Context of COVID-19" (VILESA, 2022–2024, No. 2021-1-LT-01-KA220-HED-000023551). The aim of the project was to address the challenges of applying modern teaching and learning methods in the social sciences by digitizing study content through the use of a virtual reality platform. The project brought together students and teachers from the Faculty of Economics of Vilnius College (VIKO) (Lithuania), Krakow University of Economics (KUE) (Poland), and EKA University of Applied Sciences (Latvia). 91 students from 3 countries participated in the usability study in two phases to determine the usability of the content.

The research methodology to determine the relevance of the content is based on applying participant method (namely, Event diaries), were Keller's

ARCS [18] (attention, relevance, confidence, and satisfaction) motivational design model was used as a guiding strategy.

Research objectives: 1. Present the system development cycle: analysis, design, implementation and evaluation; 2. Assess the usability of the content of the developed gamified VR-based training tool among the study participants.

2. Literature review

EBSCOhost and Taylor & Francis online libraries for the last 5 years 2019-2024 have selected 13 studies in the areas of education and professional training where the use of virtual reality (VR) and related technologies help to achieve the expected results.

First of all, in the field of engineering education, studies were considered that involve gamified virtual laboratories and virtual excursions to prepare students for practical classes. For example, in one of the quasi-experimental studies (n = 92), construction students who used VR-laboratories spent 16% less time in physical laboratories and showed better preparedness for experimental work [1]. In another study (n = 41), the influence of virtual tours in ecological representation was studied, where students noted convenience and accessibility the VR-approach, while it is inferior to traditional field research in terms of effectiveness [2].

In the second case, virtual reality is actively used in medical and nursing education to teach clinical skills, simulate emergency situations, and improve interdisciplinary interaction. A quasi-experimental study with the participation of 675 nursing students who studied using VR-simulations found that VR promotes more active involvement (95% active participation compared to 15% in traditional methods) and the development of critical thinking [3]. Also, in the following study (n = 43), where they study metaverse virtual space for collaborative learning of nurses, it showed a significant improvement of satisfaction with training and improvement of group work [4]. VR integration into nursing curricula was well-received. Students and educators reported high usability and positive learning experiences [5].

Moreover, VR is also used in emergency medical training. Namely, the VR-content developed to prepare students for triage in emergency situations showed a high score according to the criteria of relevance (4.23 out of 5) and confidence (3.81 out of 5) [6]. Similarly, VR-training for complex

medical procedures, such as REBOA, allowed to significantly increase the confidence of specialists [7].

The implementation of interactive VR avatars in the education of future pharmacists has been analyzed in the field of pharmacy education. In a multi-stage study (n = 10), it was found that such simulations contribute to better learning and development of practical skills [8].

In addition, the potential of VR in dementia care and training for visually impaired people has been explored. In a pilot study (n = 10) involving professional caregivers, VR training demonstrated high effectiveness in developing empathy and dementia care skills [9]. Another study (n = 20) assessed the usability and effectiveness of multimodal VR interfaces for training visually impaired users used to help develop new skills and to stimulate cognitive improvement [10].

Practice shows that fashion and design education also include elements of virtual reality. Research has shown that students with high spatial skills demonstrate greater satisfaction with distance learning, especially when using the traditional studio approach. At the same time, the perceived usefulness of VR courses largely depends on the level of digital competence of students [11].

In geovisualization, VR and AR are used to study natural landscapes and geomorphic processes. This method is particularly useful for exploring inaccessible and dangerous areas, as well as for demonstrating landscape changes over time [12].

VR avatars have also been used to teach children about healthy eating. In a study (n=15) involving children from low-income families, a VR program improved their acceptance of healthy eating information and stimulated dialogue between parents and children [13].

The studies cover applications of VR in engineering (2), medical (6), design (1), pharmaceutical (1) and environmental education (1), as well as in patient care (1) and geovisualization(1).

Methodologies such as experimental (4), quasi-experimental(3) and comparative surveys(6), questionnaires(10), interviews(3), as well as methods for assessing satisfaction and effectiveness of training(7) were used for these studies.

The literature review shows that the use of virtual reality (VR) and related technologies helps to achieve the expected results, but no good practices were found in applying VR to the education of economics students.

3. System development

The content development sequence was carried out according to the Systems Development Life Cycle (SDLC), which consists of analysis, design, implementation, and evaluation phases. SDLC is one of the main development methods for computer systems and mobile devices [14]. The platform consists of two parts: the student game area and the teacher setting area.

1. Analysis of content development needs

In order to introduce students to new innovative learning spaces and to address the challenges of applying contemporary teaching/learning methods in the social sciences, the following objectives were set:

- Digitizing the study content by using Virtual reality platform;
- To create modern content to supplement traditional educational methods with the latest technologies.
- To encourage creativity and create opportunities for users to get involved in the content creation process themselves.

The main objective of the project was to address challenges in applying modern teaching and learning methods in social sciences by digitizing study content through a virtual reality platform aimed at enhancing students' critical thinking, communication, collaboration, and creativity skills.

During the COVID-19 pandemic, simulation proved to be a valuable educational method [15]. The interactive nature of game-based learning tools increases student engagement and supports experience-based learning, as described in constructivist theory [16]. VR-based simulation games are therefore seen as an effective method to meet the demands of virtual learning.

2. Design content

The virtual reality platform to be developed was to cover 9 economic topics, which were distributed among the universities: I Lithuania: 1. Cartel, 2. Collusion, 3. Insurance management, 4. Investment management, 5. Promotion; II Poland: 6. Leadership, 7. Motivation; III Latvia: 8. Management, 9. Economics

University lecturers used the created template to describe the game scenarios: System action, Player (student) action, Lecturer action. The scenarios were analyzed for content validity by partners across each country, including academic researchers and one professor in the field of economics.

3. Implementation

By summarizing the 9 case scenarios, a systematic Game sequence for the game and the teacher setting areas (Picture 1) was created.

To implement the scenarios, VR developers UAB „Iron cat“ (Lithuania) created a game platform using virtual room content. Picture 2 shows the user interface (UI) design of the scenario-based VR content, coordinated with platform developers and university experts.

4. Subjective assessment of usability by students

The general characteristics of the study participants were obtained from their responses to questions about experiences with IT for learning and non-learning purposes, included 4 questions: 1. IT is my work tool; 2. I use IT in my free time; 3. I use VR technology; 4. I play computer games. Questions were rated on a five-point Likert scale ranging from (1) I don't use it, to (5) Every day.

The tool for evaluating the usability of the developed VR-based gamified educational content were designed for this study considering of the VR content usability assessment tools developed by Jeong et al [17] and Keller's [18] ARCS (attention, relevance, confidence, and satisfaction) Model of Motivational Design. A post-experiment questionnaire (Impressions after the experiment) was created (Table 1).

Researchers developed an additional questionnaire to record interface defects (Event diary), consisting of 3 questions: 1. Defect description (briefly describe the problem); 2. Provide a link to the window (if applicable) where you found the defect; 3. How did the problem start? What steps were taken before the problem occurred or the data was entered? If the message is unclear. What is unclear? If the button name is unclear. What kind? I didn't know what to do. Comment on the found defect.

Questions were rated on a five-point Likert scale ranging from (1) Not at all, to (5) Very high degree.

Respondents also provided answers to open questions: 1. Did you encounter difficulties, surprises?, 2. What did you like the most?, 3. What did you dislike the most?, 4. What is missing in the case?.

Testing took place in two phases. After summarizing the first phases results of the questionnaires, a list of errors and comments was created. Design errors and comments were discussed with the VR developers' team and a "UI Protocol" document was created. Based on this, the system was further developed, and a second phase of testing was carried out in a similar way.

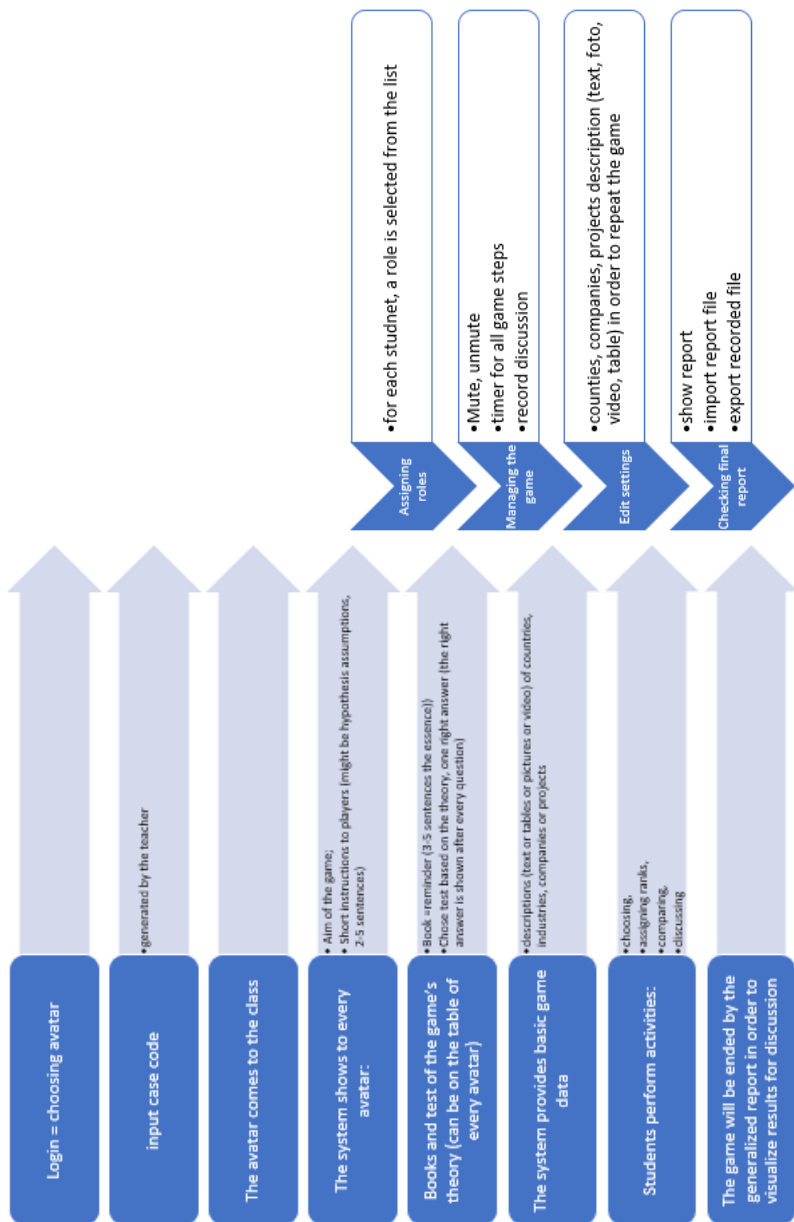


Figure 1. Game and Teacher setting areas

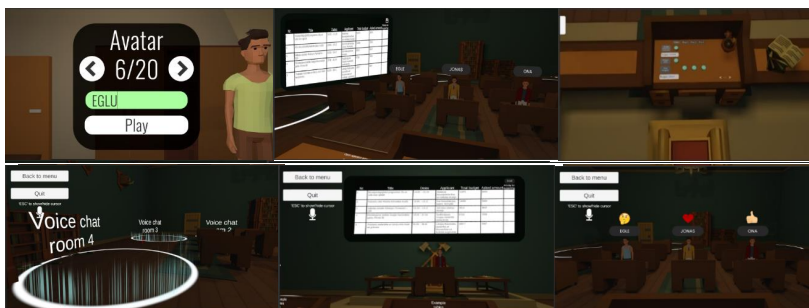


Figure 2. Examples of the main screens in tools for economics students

Table 1. Details of the post-experiment questionnaire filled in by students.

	Questions	Variables	Answer	
Q1	The tool allows faster absorbing of educational material	Confidence	(1) Not at all (2) Small Degree (3) Moderate Degree (4) High Degree (5) Very High Degree	
Q2	I liked everything	Satisfaction		
Q3	I am impressed by the game	Satisfaction		
Q4	It is useful tool for absorbing a new economic topic	Relevance		
Q5	I feel comfortable wearing 3D glasses	Satisfaction		
Q6	I feel tired wearing 3D glasses	Satisfaction		
Q7	Information is clearly laid out in the system	Attention		
Q8	Do you want to repeat the same game?	Relevance		1~5
Q9	Do you want to repeat the game with another information?	Relevance		1~5
Q10	How often do you want to play such games during lectures?	Confidence		(5) Every lecture (4) Once a week (3) Once a month (2) Once per subject semester (1) Don't use it
Q11	Can you explain the main things of the subject after the game?	Relevance		1~5
Q12	Do you recommend such studies to your friends?	Satisfaction		1~5

In total, 91 students participated in this study, 67 students in the first phase, and 24 in the second phase.

4. Quality assessment

The suitability of the created educational content for use by economics students, attention, relevance, confidence, and satisfaction with the content were assessed (Table 2).

Picture 3 show how the training content of the VR tool has quality assessment. 56%/75% (pfase1/pfase2) agreed or partly agreed that the tool allows faster absorbing of educational material, and it is useful tool for absorbing a new economic topic (76%/75%), and they partly can or partly can't explain the main things of the subject after the game (63%/71%).

About half of all students agree or partly agree to repeat this game. In phase 1, students indicated that they want to use once a week (40%), while in phase 2, 29% each indicated that they want to use once a week or once a month. 61%/79% want or partly want to repeat the game with other information. Students noted that the need to fill in the tables during games using 3D glasses is tiring on the eyes and not comfortable, so over 50% of students did not provide their opinion.

59%/63% agreed or partly agreed that they like everything about the system, 68%/77% impressed by the game.

Table 2. Analysis of the VR usability based on a subjective assessment by students.

Category		Phase 1 (n=67)		Phase 2 (n=24)	
		Mean	SD	Mean	SD
<i>Attention</i>	Q7	3,471	0,087	3,500	0,096
	Q2	3,574	0,293	3,458	0,021
<i>Satisfaction</i>	Q3	3,603	0,351	3,708	0,469
	Q5	2,324	-2,206	2,000	-2,590
	Q6	2,132	-2,589	1,958	-2,077
	Q12	3,561	0,267	3,458	0,322
<i>Confidence</i>	Q1	3,353	-0,148	3,500	0,096
	Q10	3,391	-0,080	2,947	-0,495
<i>Relevance</i>	Q4	4,029	1,204	3,833	0,693
	Q8	3,169	-0,500	3,458	0,322
	Q9	3,754	0,667	4,000	1,189
	Q11	3,121	-0,629	3,261	0,006

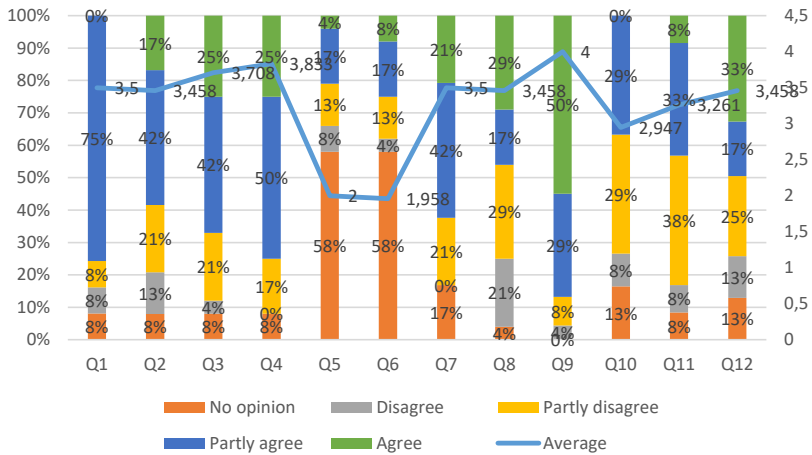


Figure 3. Analysis of the usability based on a subjective assessment by students Phase 2.

As a result of evaluating how much the content is related to practical learning skills or acquired goals, the score was 3.64 out of 5 points, the highest content usability evaluation items.

The degree of attention to the developed contents were evaluated as 3.5 points out of 5 points. Attention during learning is an important factor in achieving the learning goal, and gamification of the content allows to maintain and attract attention. The degree of confidence after applying the content was 3.37 (1 phase) out of 5 points. In phase 1, 47% indicated that they wanted to play such games every lecture or once a week, while in phase 2 more (50%) indicated that they wanted to play once a week or once a month, which is less frequent use.

Overall satisfaction with the content was 3.04 (out of 5 points). For the questions 'I feel comfortable wearing 3D glasses' and 'I feel tired wearing 3D glasses' over 50% of respondents in each phase did not give an opinion. Due to the abundance of tables in the game, respondents preferred to use the desktop version of the system rather than VR (Picture 4).

VR technology is effective in increasing student engagement and promoting independent learning.

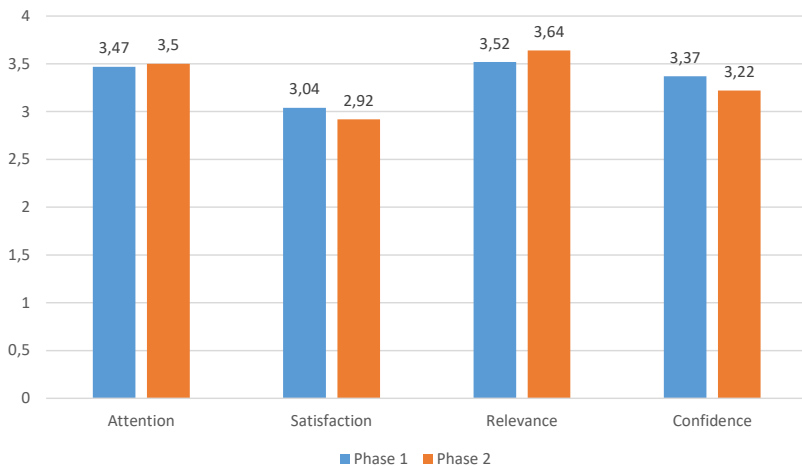


Figure 4. Usability Evaluation of Developed Educational Content

5. Conclusion

This study was carried out to investigate the usability of a VR-based development tool and the relevance of the content for economics students. After applying the VR content developed in this project, the economics students rated the relevance and usability of the tool. VR tools and content are effective in enhancing students' critical thinking, communication, collaboration and creativity skills, as the created material is an impact for updated subject content products in higher education programs, which can use the platform for substantiation, simulating various practical areas of economics, finance, marketing, management, etc. situations and tasks. Students can solve various real practical situations that cannot be understood without the help of reality.

This tool applies to a gamification element, i.e. the awarding of points to measure performance and determine the winner. However, gamification strategies were limited due to the specific nature of the economics subject, where there are many tables and calculations.

The teaching scenarios created in the tool can also be used to support students' learning in other subjects. The development of new teaching materials for other subjects will allow the tool to be used more widely.

References

- [1] Vahdatikhaki, F., Friso-van den Bos, I., Mowlaei, S., & Kollöffel, B. (2024). Application of gamified virtual laboratories as a preparation tool for civil engineering students. *European Journal of Engineering Education*, 49(1), 164–191. <https://doi.org/10.1080/03043797.2023.2265306>
- [2] Wolf, M., Wehking, F., Söbke, H., Montag, M., Zander, S., & Springer, C. (2023). Virtualized virtual field trips in environmental engineering higher education. *European Journal of Engineering Education*, 48(6), 1312–1334. <https://doi.org/10.1080/03043797.2023.2291693>
- [3] Kiegaldie, D., & Shaw, L. (2023). Virtual reality simulation for nursing education: Effectiveness and feasibility. *BMC Nursing*, 22(1), 488. <https://doi.org/10.1186/s12912-023-01639-5>
- [4] Yoo, H. J., Ryu, H., & Lee, H. (2024). Effectiveness of metaverse-based collaborative learning in nursing education: A mixed-methods study. *Journal of Nursing Education*, 63(9), 585–593. <https://doi.org/10.3928/01484834-20240514-04>
- [5] Brown, K. M., Swoboda, S. M., Gilbert, G. E., Horvath, C., & Sullivan, N. (2023). Curricular integration of virtual reality in nursing education. *Journal of Nursing Education*, 62(6), 364–373. <https://doi.org/10.3928/01484834-20230110-01>
- [6] Park, S. K., & Kim, H. J. (2023). Development and evaluation of virtual reality-based simulation content for nursing students regarding emergency triage. *Journal of Korean Academy of Fundamentals of Nursing*, 30(2), 292–301. <https://doi.org/10.7739/jkafn.2022.30.2.292>
- [7] Birrenbach, T., Wespi, R., Hautz, W. E., Berger, J., Schwab, P. R., Papagiannakis, G., Exadaktylos, A. K., & Sauter, T. C. (2023). Development and usability testing of a fully immersive VR simulation for REBOA training. *International Journal of Emergency Medicine*, 16(1), 67. <https://doi.org/10.1186/s12245-023-00545-6>
- [8] Thompson, J., White, S., & Chapman, S. (2020). Interactive clinical avatar use in pharmacist preregistration training: Design and review. *Journal of Medical Internet Research*, 22(11), e17146. <https://doi.org/10.2196/17146>
- [9] Tsai, C.-M., Hsu, T.-C., & Hsieh, C.-J. (2023). A virtual reality intervention to improve formal caregivers' understanding of community-dwelling people with dementia: A pilot study. *Contemporary Nurse*, 59(4–5), 334–343. <https://doi.org/10.1080/10376178.2023.2259495>
- [10] Darin, T., Andrade, R., & Sánchez, J. (2022). Usability evaluation of multimodal interactive virtual environments for learners who are blind: An empirical investigation. *International Journal of Human-Computer Studies*, 158, 102732. <https://doi.org/10.1016/j.ijhcs.2021.102732>
- [11] Oosterom, E. B. (2024). Students' attitude towards distance teaching of studio-based and virtual reality-based fashion design courses. *International Journal of Fashion Design, Technology and Education*, 17(3), 302–310. <https://doi.org/10.1080/17543266.2023.2278630>
- [12] Moore, A. B., & Smith, M. J. (2021). Geomorphological innovation through advances in geovisualisation. *Abstracts of the International Cartographic Association*, 3, 210. <https://doi.org/10.5194/ica-abs-3-210-2021>
- [13] McGuiert, J., Dyson, O., Enahora, B., & Rhea, C. (2020). Development of a virtual reality avatar software program to prevent obesity in children from low-income communities. *Journal of Nutrition Education and Behavior*, 52(7), S14. <https://doi.org/10.1016/j.jneb.2020.04.041>
- [14] Yas, Q. M., ALazzawi, A., & Rahmatullah, B. (2023). A comprehensive review of software development life cycle methodologies: Pros, cons, and future directions. *Iraqi Journal for Computer Science and Mathematics*, 4(4), Article 14. <https://doi.org/10.52866/ijcsm.2023.04.04.014>
- [15] Foronda, C., & Armstrong, B. (2020). *Position statement on use of virtual simulation during*

the pandemic. Society for Simulation in Healthcare. Retrieved March 30, 2023, from <https://www.ssih.org/COVID-19-Updates/ID/2237/COVID-19-SSHINACSL-Position-Statement-on-Use-of-Virtual-Simulation-during-the-Pandemic>

- [16] Alrehaili, E. A., & Al Osman, H. (2019). A virtual reality role-playing serious game for experiential learning. *Interactive Learning Environments*, 1–14. <https://doi.org/10.1080/10494820.2019.1703008>
- [17] Jeong, W. J., Oh, K. S., & Oh, S. H. (2019). Questionnaire method for usability evaluation of VR contents. *Journal of Next-generation Convergence Information Services Technology*, 8(3), 337–350. <https://doi.org/10.29056/jncist.2019.09.11>
- [18] Keller, J. M. (1987). Development and use of the ARCS model of instructional design. *Journal of Instructional Development*, 10(3), 2–10. <https://doi.org/10.1007/BF02905780>

Požymių konvertavimo į vaizdus metodų palyginimas kenkėjiškų programų aptikimo efektyvumui gerinti

Matas Lukšys, Viktor Medvedev

Vilniaus universitetas, Matematikos informatikos fakultetas,
Duomenų mokslo ir skaitmeninių technologijų institutas,
Akademijos g. 4, LT-08412 Vilnius, Lietuva
matas.luksys@mif.stud.vu.lt

Santrauka. Straipsnyje palyginami vykdomųjų PE failų požymių konvertavimo į vaizdus metodai. Vykdomųjų failų požymiai gaunami iš kenkėjiškų ir saugių vykdomųjų PE failų, pateiktų *PE Malware Machine Learning Dataset* duomenų aibėje. Požymių atrankai buvo naudojami pagrindiniai vykdomųjų failų struktūriniai ir elgsenos požymiai, tokie kaip antraštės, sekcijų statistika, dydžio ir entropijos parametrai. Eksperimentiniai tyrimai atlikti naudojant tiesioginius požymių konvertavimo metodus – BIE ir HSV bei netiesioginius konvertavimo metodus, tokius kaip cBIE, IGTD ir LMIGTD. Gauti vaizdai buvo klasifikuojami naudojant konvoliucinį neuroninį tinklą. Eksperimentų rezultatai parodė, kad netiesioginiai metodai leidžia pasiekti aukštesnį kenkėjiškų programų aptikimo tikslumą.

Raktiniai žodžiai: vykdomasis failas, požymių konvertavimas, vaizdų klasifikavimas, mašininis mokymasis, kenkėjiškų programų aptikimas.

1 Įvadas

Kenkėjiškos programos išlieka viena didžiausių grėsmių šiuolaikinėms informacinėms sistemoms, nepaisant nuolat tobulėjančių kibernetinio saugumo priemonių. 2023 metais pasaulyje užfiksuota daugiau nei 1 milijardas kenkėjiškų programų pavyzdžių [1]. Remiantis ankstesnėmis prognozėmis, 2025 metais kibernetinių nusikaltimų žala pasaulio ekonomikai gali pasiekti iki 10,5 trilijono JAV dolerių [2]. Šie skaičiai atspindi ne tik augantį kenkėjiškų programų kiekį, bet ir didėjančią jų įvairovę bei sudėtingumą.

Kenkėjiškos programos nuolat evoliucionuoja, pasitelkia pažangias mas-kavimo, kodavimo ir aplinkos analizės technikas, todėl tradiciniai aptikimo metodai tampa vis mažiau veiksmingi. Norint efektyviai aptikti naujas ir modifikuotas grėsmes, vis plačiau taikomi mašininio mokymosi metodai. Viena iš inovatyviausių krypčių – vykdomųjų failų požymių konvertavimas

į vaizdus [3], leidžiantis išnaudoti konvoliucinių neuroninių tinklų gebėjimą atpažinti sudėtingus struktūrinius ir elgsenos modelius.

Šiame tyrime lyginami skirtingi vykdomųjų failų požymių konvertavimo į vaizdus metodai, vertinant jų efektyvumą kenkėjiškų programų aptikimui. Analizuojami tiek tiesioginiai, tiek netiesioginiai konvertavimo būdai, siekiant nustatyti, kurie metodai yra tinkamiausi vykdomųjų failų požymių paruošimui mašininio mokymosi algoritmams. Tyrimo rezultatai gali prisidėti prie pažangesnių kenkėjiškų programų aptikimo sistemų kūrimo ir padėti geriau suprasti šiuolaikinių grėsmių klasifikavimo principus.

2 Duomenų rinkiniai

Atliekant tyrimą, buvo išanalizuoti ir palyginti du duomenų rinkiniai, tinkami kenkėjiškų programų aptikimo uždaviniui spręsti: *PE Malware Machine Learning Dataset*¹ ir *MABEL 2.0 Dataset*². Jų pagrindinės charakteristikos, privalumai bei ribojimai pateikti 1 lentelėje. Palyginimo rezultatai parodė, kad *PE Malware Machine Learning Dataset* rinkinys atitinka darbo tikslus – jis skirtas dvejetainiam klasifikavimui ir suteikia galimybę dirbti su neapdorotais vykdomaisiais PE failais (angl. *portable executable file*), todėl galima savarankiškai išgauti bei analizuoti įvairius struktūrinius požymius, būtinus efektyviam kenkėjiškų programų aptikimui [4]. Priešingai, *MABEL 2.0 Dataset* orientuotas į daugialypę kenkėjiškų programų šeimų klasifikaciją, neturi saugių failų pavyzdžių ir nepateikia neapdorotų vykdomųjų failų [5]. Šis rinkinys nėra tinkamas dvejetainiam klasifikavimui ar požymių inžinerijai, paremtai failų struktūra.

Remiantis šiuo palyginimu, tolimesniems eksperimentams pasirinktas *PE Malware Machine Learning Dataset*. Šis rinkinys leidžia išgauti vykdomųjų PE failų požymius, tokius kaip antraštės, sekcijų statistika, importų ir eksporto lentelės, failo ilgis ir entropija. Tyrime naudoti 54 skirtingi požymiai, apimantys pagrindinius struktūrinius ir elgsenos parametrus, kurie yra reikšmingi klasifikacijai. Nebuvo naudojami detalūs sekcijų dydžiai, importų sąrašai, resursų detalės ir kiti mažiau svarbūs atributai, nes jie nėra kritiškai bendram klasifikavimui ir gali apsunkinti analizę.

¹ <https://practicalsecurityanalytics.com/>

² <https://github.com/action-ai-institute/MABEL-dataset>

1 lentelė. Duomenų rinkinių palyginimas.

Charakteristika	PE Malware Machine Learning Dataset	MABEL 2.0 Dataset
Neapdoroti failai	Pateikiami	Nepateikiami
Pavyzdžių skaičius	201 549 unikalūs failai (86 812 saugūs, 114 737 kenkėjiški)	90 414 unikalūs kenkėjiški failai
Klasių struktūra	2 klasės: kenkėjiški failai ir nekenkėjiški failai	400+ kenkėjiškų šeimų, nėra saugių pavyzdžių
Metaduomenys	Apima failo identifikavimo reikšmes, entropijos reikšmes, failo tipą, antivirusinių programų aptikimus ir pateikimo datą	Išsamūs duomenys apie failo architektūrą, sekcijas, išskaidytą kodą ir importuojamas bibliotekas
Tinkamumas klasifikavimui	Pritaikytas binariniam klasifikavimui	Pritaikytas viruso šeimų klasifikavimui
Apribojimai	Negalima nustatyti kenksmingo failo šeimos	Nepateikia vykdomųjų failų tiesioginei analizei; nėra saugių failų binariniam klasifikavimui; pateikiami išgauti failo požymiai

Analizuojamame duomenų rinkinyje yra daugiau nei 200 tūkstančių failų, suskirstytų į dvi klases: kenkėjiški ir saugūs (nekenksmingi). Dėl techninių resursų apribojimų nebuvo galimybės apdoroti visų įrašų, todėl siekiant užtikrinti rezultatų patikimumą ir pakankamą eksperimentų apimtį, buvo sudarytos subalansuotos imtys. Modelio mokymui naudota 4259 kenkėjiškų ir 4254 saugių failų imtis, o validavimui – atitinkamai 1064 ir 1063 failai. Tokia atranka leidžia objektyviai įvertinti skirtingų požymių konvertavimo metodų efektyvumą sprendžiant binarinį klasifikavimo uždavinį.

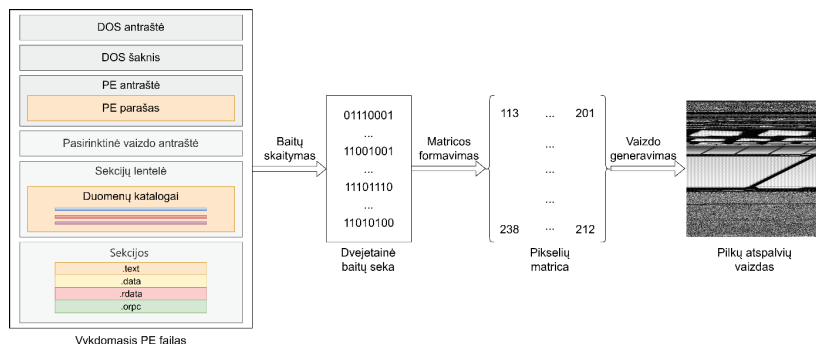
3 Metodika

Tyrime buvo lyginami požymių konvertavimo į vaizdus metodai, kurie skirstomi į tiesioginius ir netiesioginius. Tiesioginiai metodai leidžia atkurti pradinį baitus iš vaizdo, nes kiekvienas pikselis tiesiogiai atitinka konkretų baitą. Netiesioginiai metodai taiko optimizacijas ar klasterizavimą, todėl atstatyti pradinį duomenis iš vaizdo nebeįmanoma.

Tiesioginio dvejetainio kodavimo metodas (angl. *binary image encoding, BIE*) yra vienas paprasčiausių – kiekvienas failo baitas paverčiamas pilkos spalvos pikseliu ir išdėstomas kvadratinėje matricoje (žr. 1 pav.). Tokiu būdu išlaikoma bazinė failo struktūra, tačiau prarandama dalis semantinės infor-

macijos, nes neišryškinami sudėtingesni požymių tarpusavio ryšiai [6]. BIE metodas yra universalus ir dažnai naudojamas kaip atskaitos taškas vertinant sudėtingesnius metodus [6]. BIE metodo rezultato pavyzdys pateikiamas 2 paveiksle viduryje.

Kitas tiesioginio konvertavimo metodas – HSV (angl. *hue, saturation, value*) metodas. Jame požymiai normalizuojami ir priskiriami HSV spalvų kanalams [7], taip suteikiant papildomos informacijos apie požymių pasiskirstymą ir leidžiant vizualiai išskirti skirtingus požymių tipus. Literatūroje pabrėžiama [7], kad spalvų erdvės pasirinkimas gali reikšmingai paveikti klasifikavimo tikslumą. HSV metodo rezultato pavyzdys pateikiamas 2 paveiksle kairėje.

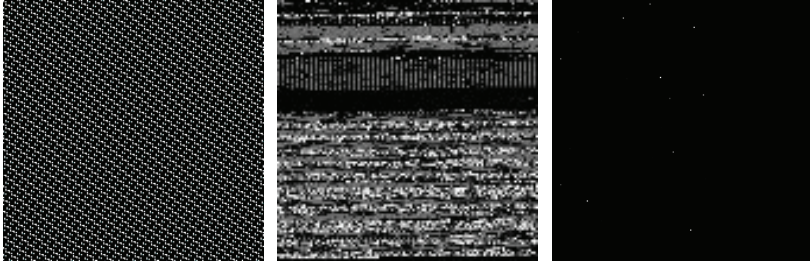


1 pav. Dvejetainio failo konvertavimo į vaizdą proceso schema.

Netiesioginio konvertavimo metodas IGTD (angl. *image generation for tabular data*) optimizuoja požymių išdėstymą vaizde pagal jų tarpusavio panašumus, apskaičiuojant porinius atstumus tarp požymių [6]. Tokiu būdu vaizde susiformuoja požymių klasteriai, išryškinantys jų tarpusavio ryšius. IGTD nereikalauja srities žinių ir leidžia CNN modeliams geriau išnaudoti erdvinius duomenų požymius. IGTD metodo rezultato pavyzdys pateikiamas 2 paveiksle dešinėje.

LMIGTD (angl. *localized and modified IGTD*) metodas yra IGTD metodikos patobulinimas, kuriame požymiai automatiškai grupuojami į funkcinis regionus naudojant klasterizavimo algoritmus. Šis metodas sumažina triukšmą ir išryškina susijusius failo elementus, nes kiekvienas regionas atspindi funkcinio požūriū susijusių požymių grupę. Tokia vizualizacija padeda CNN modeliams lengviau atpažinti sudėtingas struktūras ir pagerina klasifikavimo tikslumą.

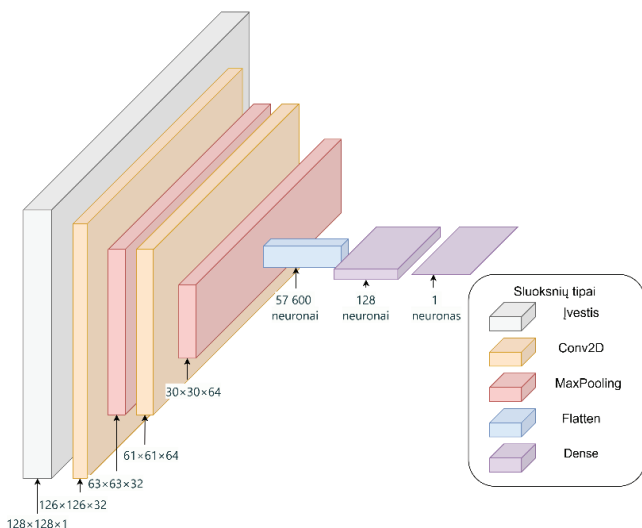
CBIE (angl. *cluster-based binary image encoding*) metodas yra BIE plėtinys [6], kuriame požymiai papildomai pertvarkomi pagal jų tarpusavio koreliacijas ar panašumus, dažniausiai taikant hierarchinį klasterizavimą. Tokiu būdu vaizde išryškėja funkcinės zonos – panašūs požymiai atsидuria greta, o skirtingi – atskiruose regionuose. Tai leidžia CNN efektyviau išnaudoti požymių tarpusavio ryšius ir dažnai pagerina klasifikavimo rezultatus sudėtinguose duomenų rinkiniuose.



2 pav. Skirtingų požymių konvertavimo į vaizdus metodų sugeneruoti kenkėjiškų failų vaizdai: HSV (kairėje), BIE (viduryje), IGTD (dešinėje).

Konvoliucinis neuroninis tinklas (angl. *convolutional neural network*) yra giliojo mokymosi modelis, skirtas vaizdų analizei ir klasifikavimui. Šio modelio pagrindinė savybė – gebėjimas išmokti ir atpažinti svarbius vaizdo požymius [8]. Konvoliucinio neuroninio tinklo architektūra sudaryta iš kelių specializuotų sluoksnių [8], kurie leidžia automatiškai išmokti ir atpažinti vaizdo požymius. Toks sluoksnių išdėstymas (žr. 3 pav.) leidžia efektyviai apdoroti vaizdus ir automatiškai išskirti požymius, reikalingus klasifikavimui [8].

Modelio įvestis – 128×128 pikselių pilkos spalvos vaizdas. Pirmasis konvoliucinis sluoksnis (angl. *Conv2D*) su 32 filtrais aptinka bazinius vaizdo požymius. Po jo taikomas maksimalios sutelkties sluoksnis (angl. *max pooling*), kuris sumažina duomenų apimtį ir išlaiko naudingą informaciją. Antrasis konvoliucinis sluoksnis su 64 filtrais leidžia išskirti sudėtingesnius požymius, o antrasis maksimalios sutelkties sluoksnis dar labiau sumažina matmenis. Išlyginimo sluoksnis (angl. *flatten*) paverčia duomenis į vienmatį vektorių, kuris perduodamas tankiajam sluoksniui (angl. *dense*) su 128 neuronais. Galiausiai, išvesties sluoksnis (angl. *output*) su vienu neuronu pateikia galutinę prognozę – ar failas yra kenkėjiškas, ar saugus.



3 pav. Konvoliucinio neuroninio tinklo architektūra kenksmingų programų klasifikavimui.

4 Rezultatai

Tyrimo metu buvo palyginti penki požymių konvertavimo į vaizdus metodai, siekiant įvertinti jų tinkamumą kenkėjiškų programų aptikimui taikant konvoliucinį neuroninį tinklą. Kiekvieno metodo efektyvumas vertintas pagal pagrindines klasifikavimo metrikas: bendrą tikslumą (angl. *accuracy*), kenkėjiškų failų atkūrimą (angl. *recall*), preciziškumą (angl. *precision*) ir F1 rodiklį. Praktikoje vienas svarbiausių rodiklių yra klaidingai neigiamų rezultatų (angl. *false negatives*) skaičius, t. y. kiek kenkėjiškų failų sistema neaptiko ir priskyrė prie saugių. Šis skaičius pateikiamas 2 lentelėje. Sistemos gebėjimą aptikti kenkėjiškus failus apibūdina ir atkūrimo rodiklis, kuris parodo, kokia dalis visų tikrų kenkėjiškų failų buvo teisingai aptikta.

2 lentelėje pateikiami kiekvieno požymių konvertavimo į vaizdus metodo rezultatai, nurodant praleistų kenkėjiškų failų skaičių ir procentą. BIE metodas praleido 171 kenkėjišką failą, kas sudaro 16,1 % visų kenkėjiškų pavyzdžių. HSV metodas pasižymėjo šiek tiek geresniu rezultatu – praleisti 149 kenkėjiški failai, tai yra 14 %. IGTD metodas dar labiau sumažino praleistų failų skaičių – jų buvo 117, o tai sudaro 10,9 %. LMIGTD metodas pasiekė

dar geresnį rezultatą – praleisti 106 kenkėjiški failai, arba 9,9 %. Geriausiai pasirodė cBIE metodas, kuris praleido tik 85 kenkėjiškus failus, o tai sudaro 7,9 % visų kenkėjiškų pavyzdžių. Šie rezultatai rodo, kad netiesioginio atvaizdavimo metodai, ypač cBIE ir LMIGTD, leidžia reikšmingai sumažinti klaidingai neaptiktų kenkėjiškų failų skaičių, palyginti su tiesioginiais metodais.

2 lentelė. Praleistų kenkėjiškų programų skaičius ir procentas.

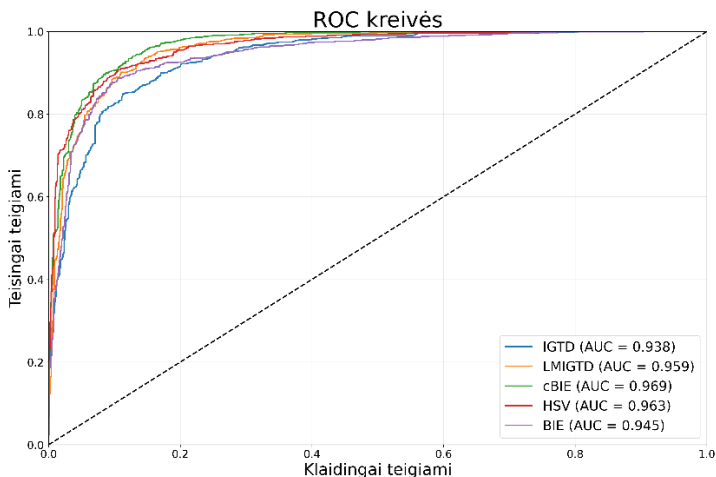
Metodas	Praleistų kenkėjiškų failų skaičius	Praleistų kenkėjiškų failų procentas (%)
BIE	171	16,1
HSV	149	14,0
IGTD	117	10,9
LMIGTD	106	9,9
cBIE	85	7,9

3 lentelė. Klasifikavimo metrikos pagal požymių konvertavimo metodą.

Metodas	Tikslumas	Atkūrimas	Preciziškumas	F1
BIE	0,88	0,84	0,92	0,88
HSV	0,89	0,86	0,92	0,89
IGTD	0,86	0,89	0,84	0,87
LMIGTD	0,89	0,90	0,89	0,90
cBIE	0,91	0,92	0,89	0,91

Analizuojant eksperimento rezultatus, pateiktus 2 lentelėje pateikiami kiekvieno požymių konvertavimo į vaizdus metodo rezultatai, nurodant praleistų kenkėjiškų failų skaičių ir procentą. BIE metodas praleido 171 kenkėjišką failą, kas sudaro 16,1 % visų kenkėjiškų pavyzdžių. HSV metodas pasižymėjo šiek tiek geresniu rezultatu – praleisti 149 kenkėjiški failai, tai yra 14 %. IGTD metodas dar labiau sumažino praleistų failų skaičių – jų buvo 117, o tai sudaro 10,9 %. LMIGTD metodas pasiekė dar geresnį rezultatą – praleisti 106 kenkėjiški failai, arba 9,9 %. Geriausiai pasirodė cBIE metodas, kuris praleido tik 85 kenkėjiškus failus, o tai sudaro 7,9 % visų kenkėjiškų pavyzdžių. Šie rezultatai rodo, kad netiesioginio atvaizdavimo metodai, ypač cBIE ir LMIGTD, leidžia reikšmingai sumažinti klaidingai neaptiktų kenkėjiškų failų skaičių, palyginti su tiesioginiais metodais.

3 lentelėje, nustatyta, kad pagal kenkėjiškų programų atkūrimo ir F1 rodiklius geriausiai pasirodė cBIE metodas: atkūrimo rodiklis siekė 0,92, o F1 – 0,91. Tai rodo, kad taikant šį metodą praleistų kenkėjiškų programų dalis buvo mažiausia. LMIGTD ir HSV metodai taip pat užtikrino aukštą aptikimo efektyvumą, tačiau jų rezultatai buvo kiek žemesni. Tuo tarpu BIE ir IGTD metodai pasižymėjo mažesniu bendru tikslumu ir prastesniu kenkėjiškų programų aptikimu.



4 pav. Konvoliucinio neuroninio tinklo architektūrą skirtingų požymių išskyrimo metodų analizei.

ROC (angl. *receiver operating characteristic*) kreivių (žr. 4 pav.) analizė papildomai patvirtina metodų efektyvumo skirtumus. Aukščiausias AUC pasiektas taikant cBIE metodą – šio metodo AUC reikšmė yra 0,969. LMIGTD metodas taip pat pasižymėjo pajėgumu, jo AUC siekia 0,959. HSV metodas pasiekė 0,963 AUC reikšmę. BIE metodo AUC buvo 0,945, o IGTD metodo – 0,938. Šie rezultatai rodo, kad netiesioginio atvaizdavimo metodai pasižymi didesniu jautrumu ir patikimumu. ROC kreivės rezultatai patvirtina, kad cBIE ir LMIGTD metodai yra efektyvūs, o BIE, HSV ir IGTD metodų jautrumas yra mažesnis.

5 Išvados

Šiame tyrime buvo atliktas penkių skirtingų vykdomųjų PE failų požymių konvertavimo į vaizdus metodų palyginimas, siekiant įvertinti jų efektyvumą kenkėjiškų programų aptikimui taikant konvoliucinius neuroninius tinklus. Tyrimui naudotas *PE Malware Machine Learning Dataset* duomenų rinkinys, leidžiantis dirbti su neapdorotais failais ir sudaryti subalansuotas mokymo bei validavimo imtis. Analizuoti tiesioginiai ir netiesioginiai konvertavimo metodai, kurie skiriasi požymių išdėstymo vaizde principais ir informacijos išryškavimo galimybėmis.

Rezultatai parodė, kad netiesioginio atvaizdavimo metodai, ypač cBIE ir LMIGTD, leidžia pasiekti aukštesnį kenkėjiškų programų aptikimo tikslumą ir geresnį bendrą modelio našumą. cBIE metodas išsiskyrė mažiausiu praleistų kenkėjiškų failų kiekiu ir aukščiausiu bendru tikslumu, o BIE metodas pasižymėjo didesniu praleistų kenkėjiškų failų skaičiumi ir ryškiu persimokymu. HSV ir LMIGTD metodai taip pat pasiekė aukštus rezultatus, tačiau cBIE lenkė juos pagal pagrindines klasifikavimo metrikas. IGTD ir BIE metodų rezultatai buvo žemesni tiek pagal bendrą tikslumą, tiek pagal kenkėjiškų programų atpažinimą

Atlikus tyrimą, kurio metu buvo lyginami skirtingi požymių konvertavimo į vaizdus metodai kenkėjiškų programų aptikimui, gautos šios išvados:

- Požymių konvertavimo į vaizdus metodai yra priemonė, leidžianti efektyviai spręsti kenkėjiškų programų aptikimo uždavinius, nes suteikia galimybę išnaudoti konvoliucinių neuroninių tinklų gebėjimą atpažinti sudėtingus struktūrinius ir elgsenos modelius.
- Netiesioginio požymių konvertavimo į vaizdus metodai, ypač cBIE ir LMIGTD, yra efektyviausi kenkėjiškų programų aptikimui naudojant konvoliucinius neuroninius tinklus. cBIE metodas praleido tik 7,9 % visų kenkėjiškų failų ir pasiekė 91 % bendrą tikslumą, o LMIGTD metodas praleido 9,9 % kenkėjiškų failų.
- Tiesioginio dvejetainio kodavimo metodas BIE praleido 16,1 % kenkėjiškų failų ir pasižymėjo ryškiu persimokymu, todėl nėra optimalus sudėtingesnių grėsmių aptikimui.
- HSV metodas praleido 14 % kenkėjiškų failų, o IGTD metodas – 10,9 %, todėl šie metodai taip pat gali būti laikomi efektyviomis alternatyvomis, tačiau jų rezultatai kiek nusileidžia cBIE metodui.

Literatūra

- [1] A. Institute, „AV-TEST,“ [Tinkle]. Available: <https://www.av-test.org/en/statistics/malware/>. [Kreiptasi 02 2025].
- [2] J. Ferdous, R. Islam, A. Mahboubi ir M. Z. Islam, „A Review of State-of-the-Art Malware Attack Trends and Defense Mechanisms,“ *IEEe Access*, t. 11, pp. 121118-121141, 2023.
- [3] M. Guven, „Leveraging deep learning and image conversion of executable files for effective malware detection: A static malware analysis approach,“ *AIMS Mathematics*, t. 9, nr. 6, p. 15223–15245, 2024.
- [4] M. Lester, „PE Malware Machine Learning Dataset,“ *Practical Security Analytics*, 8 6 2021. [Tinkle]. Available: <https://www.practicalsecurityanalytics.com>. [Kreiptasi 02 2025].
- [5] A. A. Institute, „MABEL: Malware Analysis Benchmark for Artificial Intelligence and Machine Learning,“ *GitHub*, 2023.
- [6] J. Halladay, D. Cullen, N. Briner, D. Miller, R. Primeau, A. Avila, W. Watson, R. Basnet ir T. Doleck, „BIE: Binary Image Encoding for the Classification of Tabular Data,“ *Journal of Data Science*, t. 23, nr. 1, p. 109–129, 2025.
- [7] Z. Xian, R. Huang, D. Towey ir C. Yue, „Convolutional Neural Network Image Classification Based on Different Color Spaces,“ *Tsinghua Science and Technology*, t. 30, nr. 1, pp. 402-417, 2025.
- [8] J. Saxe ir H. Sanders, *Malware Data Science: Attack Detection and Attribution*, No Starch Press, 2018.

Studentų nubyrejimo Vilniaus universitete vertinimas taikant mašininio mokymo metodus

Arvydas Norgėla, Jurgita Markevičiūtė

Vilniaus Universitetas Matematikos ir informatikos fakultetas
Taikomosios matematikos institutas, Naugarduko g. 24, Vilnius
arvydas.norgela@mif.stud.vu.lt

Santrauka. Straipsnyje apžvelgiama studentų nubyrejimo problema. Aptariamos priežastys, panašūs tyrimai šia tema. Atliekama Vilniaus universiteto turimų duomenų analizė apie individualius studentus. Duomenims pritaikomi logistinės regresijos ir k-artimiausių kaimynų klasifikatoriaus modeliai. Palyginus gautus rezultatus, k-artimiausių kaimynų klasifikatorius tiko geriau – tiksliau klasifikavo visus studentus, bei studentus, kurie metė studijas.

Raktiniai žodžiai: studentų nubyrejimas, logistinė regresija, k-artimiausių kaimynų klasifikatorius, mašininis mokymas, klasifikavimas.

1 Įvadas

Visuose pasaulio universitetuose yra susiduriama su studentų nubyrejimo problema. Pasirinkimas mesti studijas gali kilti dėl labai įvairių priežasčių bei jų kombinacijų, kurias nebūtinai gali kontroliuoti pats universitetas. Tačiau, žinant šias priežastis ir tiriant, kaip jos veikia tam tikro universiteto studentų su studijomis susijusius sprendimus, galima prognozuoti, kurie studentai ketina mesti studijas. Šio tyrimo tikslas yra apžvelgti kituose darbuose naudojamus metodus ir kokiems duomenims jie taikomi; palyginti juos su Vilniaus universiteto turimais duomenimis apie studentus, išsirinkti prasmingus kintamuosius, su jais sudaryti modelius, palyginti jų klasifikavimo tikslumą. Straipsnyje pirmą kartą bandoma pritaikyti klasifikavimo modelius Vilniaus universiteto studentų duomenims.

2 Literatūros apžvalga

Šiame straipsnyje studentų nubyrejimas yra suprantamas kaip išėjimas iš universiteto negavus diplomo [1]. Gali būti labai daug studentų pasitrau-

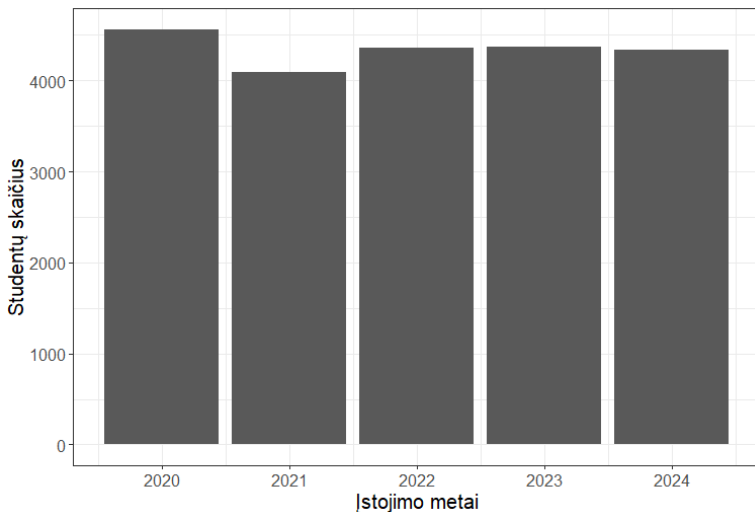
kimo iš studijų priežasčių. Joms aprėpti straipsnyje [1] išskiriamos keturios perspektyvos, per kurias galima žvelgti į studentų nubybrėjimo priežastis: sociologinė, psichologinė, ekonominė ir fazinė perspektyvos. Sociologinė perspektyva siūlo žvelgti į bendrą studento santykį su universitetu [6]. Psichologinė perspektyva skiria dėmesį studentų psichologinėms savybėms ir kaip jos veikia jų santykį su universitetu [7]. Ekonominė perspektyva remiasi ekonomikos teorija, žvelgia į studentą kaip į racionalų agentą, kuris sugeba įvertinti studijų naudą ir ją palyginti su kaštais [8]. Tarp šių perspektyvų yra persidengimo. Jos visos pabrėžia studento lūkesčius studijoms ir teigia, kad studentas įvertina studijų naudą. Fazinė perspektyva apjungia visas kitas, siūlydama į nubybrėjimą žvelgti kaip į ilgalaikį procesą, kurio metu stebima individo vidinė būsena ir jo santykis su aplinka [9]. Ji mažiausiai pritaikoma modelių kūrimui, kadangi tokius detalius duomenis gauti būtų sudėtinga. Dažniausiai tyrimuose naudojami su sociologine perspektyva susiję kintamieji, tokie kaip rasė [3], lytis [3], [5], egzaminų pažymys [3], gimimo metai [3], įstojimo metai [3], dėstytojų skaičius universitete [4], tėvų išsilavinimas [5]. Taip pat yra bandoma sieti studentų nubybrėjimą su ekonominiais rodikliais. Straipsnyje [5] į modelį įtraukiamas nedarbo lygis, infliacija ir bendras vidaus produktas.

Tyrimuose duomenys dažniausiai būna apie individualius studentus. Tai leidžia taikyti labai įvairius metodus. Straipsnyje [3], klasifikuojant studentus į dvi grupes, taikoma logistinė regresija. Šis metodas lyginamas su k-artimiausių kaimynų klasifikatoriumi bei atsitiktinio miško algoritmu. Logistinė regresija klasifikavo studentus tiksliausiai. Straipsnyje [2] panašioms duomenims naudojami mašininio mokymo metodai. Lyginami atsitiktinio miško, tiesinės diskriminantinės analizės ir atraminių vektorių klasifikatoriaus metodai. Atraminių vektorių klasifikatorius buvo tiksliausias, kuris studentus klasifikavo su tikslumu nuo 0,79 iki 0,94, priklausomai nuo mokyklos. Taip pat yra tyrimų, kurie naudoja kitokius duomenis. Pavyzdžiui, [4] straipsnyje naudojami paneliniai duomenys. Tyrime pastebėta, kad yra neigiama koreliacija tarp studentų nubybrėjimo ir dėstytojų skaičiaus.

3 Duomenų analizė

Mūsų turimi duomenys yra apie individualius studentus. Kaip daugumoje tyrimu, yra kintamieji žymintys lytį, studijų kryptį, bei stojamąjį balą. Apie kai kurių studentų stojamąjį balą nėra duomenų, kas kelia keblumą sudarinė-

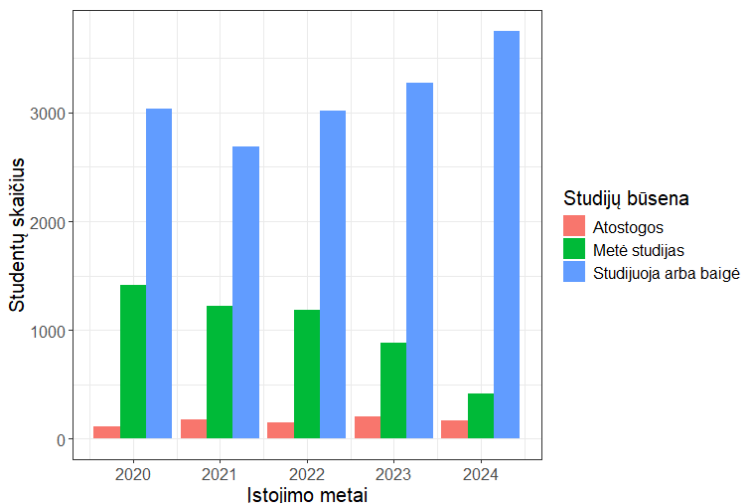
jant modelius su šiuo kintamuoju. Duomenys apima 2020 – 2024 metų laikotarpį. Iš viso duomenyse yra 20231 stebėjimai, stebimi 36 kintamieji, tokie kaip studento lytis, studijų semestras, kursas, įstojimo metai, fakultetas, studijų programa, pakopa, tipas, forma, būseną, priėmimo balas, nesimokymo priežastis, mokėjimo forma. Nuo [2] ir [3] straipsniuose naudotų duomenų mūsų duomenys skiriasi tuo, kad turime duomenis apie studijų semestrą bei mokėjimo formą. Tai mums leidžia pamatyti, kaip studentų sprendimai, susiję su studijų metimu, kinta ilgiau studijuojant bei ar yra skirtumas tarp studentų, kurių studijas apmoka valstybė, ir studentų, kurie už studijas moka patys. Studentų skaičius pagal metus pateikiamas 1 paveikslėlyje.



1 pav. Studentų skaičius pagal metus grafikas, 2020 – 2024 metų laikotarpis

Duomenyse taip pat yra kintamasis, kuris nurodo studento studijų būseną. Yra galimos šios būsenos: sustabdytos studijos, sustabdytos studijos baigėsi, akademinėse atostogose, akademinės atostogos baigėsi, komandiruotėje, baigė studijas, pašalintas, studijuoja ir būseną neapibrėžta. Eilutės su „būseną neapibrėžta“ buvo pašalintos, kadangi jų neįmanoma priskirti jokiai grupei. Taip pat buvo pašalinti visi mainų studentai. Tyrimo tikslas yra klasifikuoti studentus į dvi grupes: tuos, kurie liks ir gaus diplomą, ir tuos, kurie mes studijas. Tad turimos studijų būsenos perkoduotos į teigiamas (studentas dar studijuoja arba gavo diplomą), neigiamas (studentas dėl kaž-

kokios priežasties nebestudijuoja) ir neutralias (baigtiniam laikui sustabdytos studijos) būsenas. Teigiamoms būsenoms priskiriamos šios kategorijos: komanduotėje, baigė studijas, studijuoja. Neigiamoms: sustabdytų studijų trukmė baigėsi, akademinės studijos baigėsi, pašalintas. Neutralioms: sustabdytos studijos, akademinėse atostogose.

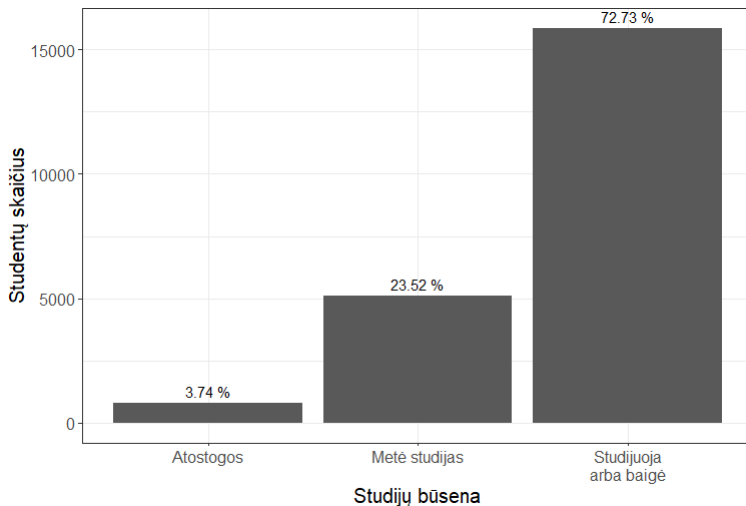


2 pav. Studijų būsenų pagal metus grafikas, 2020 – 2024 metų laikotarpis

2 paveikslėlyje pavaizduota kiek studentų yra kokioje iš anksčiau aptartų būsenų ir kaip šie skaičiai kinta per mūsų tiriamą laikotarpį. Galima pastebėti dvi tendencijas. Pirmoji yra didėjantis studijuojančių arba baigusių studentų skaičius. Antra tendencija yra mažėjantis metusių studijas studentų skaičius.

Duomenyse taip pat galime pastebėti dar vieną problemą – nesubalansuotas imtis. 3 paveikslėlyje pavaizduotame grafike galima pamatyti kiek studentų atsiduria kokioje grupėje. Taip pat parodyta, kokią dalį visų studentų sudaro konkreti grupė. Matome, kad didžioji dalis studentų dar studijuoja arba sėkmingai baigė studijas – apie 72,73 %. Studijas metė 23,52 % studentų, o likusi dalis yra arba akademinėse atostogose, arba laikinai sustabdžiusių studijas. Nesubalansuotos imtys gali lemti modelio prisitaikymą priskirti dažniau pasitaikančias kategorijas tiksliau nei kitas, taip iškraipant rezultatus. [5] straipsnyje taip pat pabrėžiama, kad buvo susidurta su ne-

subalansuotų imčių problema. Autoriai rekomendavo naudoti tiek duomenų lygio, tiek algoritmo lygio metodus, kad būtų išvengta šios problemos. Pavyzdžiui, naudoti mašininio mokymo algoritmus, kuriuose yra iš karto įtraukti nesubalansuotų imčių problemos sprendimai.



3 pav. Studentų studijų būsenų pagal metus grafikas, 2020–2024 metų laikotarpis

4 Modeliavimas

Turimiems duomenims galima taikyti kelis modelius. Jei norima klasifikuoti juos į tris kategorijas, galima naudoti daugianarę logistinę regresiją arba k-artimiausių kaimynų klasifikatorių. Kitu atveju, galima pašalinti atostogų būseną iš duomenų ir taikyti metodus, minėtus straipsniuose. Duomenims pritaikius agregavimą, galima taikyti ir laiko eilučių metodus. Kaip siūloma [4] straipsnyje, būtų įmanoma prijungti makroekonominis duomenis. Tai leistų tirti studentų nubyrijimą iš ekonominės perspektyvos.

Atsižvelgus į apžvelgtus straipsnius [2] ir [3], bei dėl paprastesnės modelių interpretacijos ir tikslesnio klasifikavimo, buvo nuspręsta iš duomenų pašalinti atostogų būseną, taip paliekant dvi grupes. Turimų stebėjimų skaičius sumažėjo iki 19464 studentų. Tuomet buvo pritaikyta logistinė regresija ir k-artimiausių kaimynų klasifikatorius. Modeliai pasirinkti pagal [3] straipsnyje naudotus modelius, kadangi norėta išbandyti, ar tie patys modeliai

tiks mūsų duomenims. Taip pat tai leido mums palyginti gautus rezultatus. Duomenys buvo suskirstyti į mokymo ir testavimo aibes. 80 % duomenų pateko į mokymo aibę, likę 20 % į testavimo. Atrinkti dominantys kintamieji: lytis, studijų semestras, fakultetas, mokėjimo forma ir priėmimo balas. Kintamieji atrinkti remiantis [2] ir [3] straipsniuose naudojamais kintamaisiais. Nuo šių straipsnių pasirinkti kintamieji skiriasi tuo, kad įtraukiama studijų finansavimo forma, bei nenaudojamas amžiaus kintamasis, vietoje jo turime kintamąjį, kuris parodo, kelintame semestru tuo metu buvo studentas. Studentai, kurie priėmimo balo neturėjo, buvo pašalinti. Lytis ir mokėjimo forma buvo perdaryti į binarinius kintamuosius, o fakulteto stulpelis buvo perdarytas į pseudokintamąjį. Kadangi dauguma kintamųjų yra kategoriniai, multikolinearumo problemos duomenyse nebuvo. Pritaikant modelius šiems duomenims buvo atliekamas vienkartinis eksperimentas. Logistinės regresijos atveju, dėl nesubalansuotų imčių buvo išbandytos įvairios slenksčio reikšmės. Geriausi rezultatai buvo pasiekti su 0,6 slenksčiu. Lyties, semestro, mokėjimo formos ir priėmimo balo kintamieji modelyje yra statistiškai reikšmingi. Dauguma fakulteto kintamojo kategorijų buvo statistiškai reikšmingos, tad šį kintamąjį taip pat buvo nuspręsta palikti modelyje. Tai parodo, kad fakulteto pasirinkimas gali paveikti galimybę, kad studentas mes studijas. Koeficientų reikšmės parodė, kad vyrams yra didesnė galimybė mesti studijas, o taip pat ir tiems studentams, kurie patenka į valstybės nefinansuojamas vietas. Galimybė mesti studijas mažėja didėjant semestro reikšmei, tai yra, kuo ilgiau studentas mokosi universitete, tuo mažesnė galimybė, kad jis mes studijas. Tas pats pastebėta ir apie priėmimo balą – didėjant priėmimo balui, mažėja galimybė, kad studentas mes studijas.

Pritaikytam k-artimiausių kaimynų klasifikatoriui buvo išbandytos įvairios k reikšmės. Geriausi rezultatai pasiekti su 5 kaimynais. Modelių klasifikavimo metrikos pateiktos 1 lentelėje. Šioje lentelėje galima pamatyti, kad k-artimiausių kaimynų klasifikatorius visais atžvilgiais, išskyrus jautrumą, mūsų duomenims tiko labiau nei logistinė regresija. Straipsnyje [3], taikant logistinę regresiją buvo pasiektas 0,6659 tikslumas, o su k-artimiausių kaimynų klasifikatoriumi – 0,6460 tikslumas, tai yra, mūsų sukurti modeliai žymiai geriau klasifikuoja studentus. Lyginant su [2] straipsniu, galima pastebėti, kad mūsų sukurti modeliai, ypač k-artimiausių kaimynų, nedaug atsiliko nuo straipsnyje sudaryto atraminių vektorių klasifikatoriaus modelio. Jo tikslumas, priklausomai nuo mokyklos, buvo tarp 0,77 ir 0,94, jautrumas tarp 0,73 ir 0,99 bei specifiškumas tarp 0,76 ir 0,94.

1 lentelė. Modelių metrikos testavimo aibei

Modelis	Tikslumas	Jautrumas	Specifiškumas	F1
Logistinė regresija	0,8439	0,9693	0,4318	0,9049
k-artimiausių kaimynų	0,9147	0,9163	0,9074	0,9462

5 Išvados

Straipsnyje buvo apžvelgta su studentų nubyrėjimu susijusi literatūra. Pastebėta, jog klasifikavimui naudojami įvairūs regresijos ir mašininio mokymo metodai studentų klasifikavimui į studijuojančius ir metusius studijas. Buvo atlikta Vilniaus universiteto turimų duomenų apie studentus analizė. Duomenys apima 2020 – 2024 metų studijų laikotarpį. Studentai buvo suskirstyti į tris grupes: studijuojančius arba baigusius studijas, metusius studijas ir esančius atostogose. Kadangi kitų straipsnių modeliuose naudotos tik dvi kategorijos, buvo nuspręsta neįtraukti į modelius studentų, kurie buvo priskirti atostogaujančių grupei. Tuomet buvo sudaryti logistinės regresijos ir k-artimiausių kaimynų klasifikatoriaus modeliai. Mūsų modeliai nuo kituose straipsniuose taikytų modelių skyrėsi tuo, kad buvo įtraukti studijų semestro ir mokėjimo formos kintamieji. Įvertinti logistinės regresijos koeficientai parodė, jog studentai, patekę į valstybės nefinansuojamas vietas, ir vyrai turi didesnę galimybę mesti studijas. Didėjant semestrų skaičiui, tai yra, ilgiau mokantis universitete, bei didėjant priimamojo balo vertei, galimybė mesti studijas mažėja. Taip pat pastebėta, kad fakulteto pasirinkimas gali turėti poveikį galimybei mesti studijas. Turimiems duomenims k-artimiausių kaimynų klasifikatorius tiko labiau. Modelio tikslumas testavimo aibei buvo apie 0,9147 su 0,9163 jautrumu ir 0,9074 specifiškumu. Šis modelis taip pat tiko mūsų duomenims labiau, kadangi jis ne tik tiksliau bendrai klasifikavo visus studentus, bet ir studentus, kurie metė studijas.

Literatūra

- [1] Behr, A., Giese, M., Tegui Kamdjou, H. D., & Theune, K. (2020). Dropping out of university: a literature review. *Rev Educ*, 8, 614–652.
- [2] Del Bonifro, F., Gabbrielli, M., Lisanti, G., & Zingaro, S. P. (2020). Student Dropout Prediction. *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I*, 12163, 129–140.
- [3] Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2017). Predicting Student Dropout in Higher Education.

- [4] Gitto, L., Minervini, L. F., & Monaco, L. (2016). University dropouts in Italy: Are supply side characteristics part of the problem? *Economic Analysis and Policy*, Volume 49, 108–116.
- [5] Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting Student Dropout and Academic Success. *Data*, 7(11), 146.
- [6] Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, 1, 64–85
- [7] Bean, J. P., & Eaton, S. B. (2000). A psychological model of college student retention. In J. M. Braxton (Ed.), *Reworking the student departure puzzle* (48–61). Nashville: Vanderbilt Univ. Press.
- [8] Becker, R., & Hecken, A. E. (2009). Higher Education or Vocational Training? An Empirical Test of the Rational Action Model of Educational Choices Suggested by Breen and Goldthorpe and Esser. *Acta Sociologica*, 52(1), 25–45.
- [9] Heublein, U. (2014). Student drop-out from German higher education institutions. *European Journal of Education*, 49(4), 497–513.

Didelio masto orlaivių sistemos koordinavimas pasitelkiant būrio metaforą

Andrius Paulauskas

Vilniaus Universitetas, Matematikos ir informatikos fakultetas,
Didlaukio g. 47, LT-08303 Vilnius
andrius.paulauskas@mif.stud.vu.lt

Santrauka. Skraidantys bepiločiai orlaiviai yra naudojami įvairiose mokslo, prekybos, aplinkosaugos ar karo pramonės srityse. Numatoma, kad jų populiarumas ir pritaikymo atvejų kiekis nuolat augs. Todėl yra svarbu sukurti didelių sistemų koordinavimo metodus, kurie leistų užtikrinti aukštą skrydžių efektyvumo ir saugumo lygį, kartu įgalinant minimalias orlaivių gamybos išlaidas ir konstrukcinį paprastumą. Šio darbo tikslas yra sukurti didelio masto bepiločių orlaivių sistemos koordinavimo problemos sprendinį, sąlygojamą jų sistemai keliamų reikalavimų. Pasiūlymas yra dviejų metodų: klasikinio jėgų laukų ir euristinio pasusk-išvenk – junginys kartu su būriavimosi funkcija, kuris užtikrina saugų bepiločių orlaivių skrydžių koordinavimą naudojant minimalų konstrukcinių ir kompiuterinių išteklių kiekį. Tai autoriaus poziciją argumentais grindžiantis straipsnis (angl. *position paper*), kurio rezultatus numatoma patikrinti kompiuteriniu eksperimentu.

Raktiniai žodžiai: skraidantys bepiločiai orlaiviai, decentralizuotas koordinavimas, dronų būrys, autonominės sistemos, didelio masto sistemos.

1. Įvadas

Bepiločiai orlaiviai (angl. *unmanned aerial vehicle*, UAV) yra naudojami įvairių tipų užduotims atlikti: duomenų rinkimui, objektų transportavimui, nuotoliniam valdymui. Jų sistemos, kaip kolektyvo, sukūrimo vienas iš pagrindinių uždavinių yra koordinavimas. Mokslinėje literatūroje yra pasiūlyta daug to uždavinio sprendinių [1, 2, 3], tačiau bepiločių orlaivių įvairovė sąlygoja ir didelę sprendinių aibę galimoms jų kombinacijoms. Šiame darbe daromos prielaidos ir keliami šie ribojimai:

1. kolektyvas yra decentralizuota sistema, sudaryta iš UAV;
2. UAV yra visiškai savarankiškai veikiančios esybės;

3. UAV kompiuteriniai išteklių yra minimalūs, t. y. minimizuojama atmintis ir skaičiavimai – kiekvienas kolektyvo narys turi ribotas galimybes įsiminti informaciją ir vykdyti skaičiavimus dėl ribotų procesoriaus galimybių;
4. jutiklių sistema leidžia aptikti kitus UAV maksimaliu atstumu; maksimalus atstumas priklauso nuo jutiklių sistemos;
5. UAV tipų yra baigtinis skaičius, o atskirus tipus apibrėžia šie parametrai: akumulatoriaus talpa ir svoris, kurį gali gabenti orlaivis skrydžio metu,

Skrendantį orlaivį paprastai nuotoliniu būdu valdo vienas operatorius. Tam tikroms užduotims atlikti gali tekti naudoti didelį kiekį vienu metu veikiančių orlaivių. Jeigu juos norima valdyti, dažnai renkama naudoti centralizuoto valdymo komponentą, kuris valdo visus orlaivius ir siunčia jiems iš anksto numatytus ir koordinuojančius nurodymus. Tokia sistema geba sėkmingai valdyti didelį kiekį orlaivių ir išvengti susidūrimų, tačiau turi vieną kritinį tašką: jeigu orlaiviai praranda ryšį su centrinio valdymo komponentu, visa sistema gali prarasti savo funkcionalumą.

Dėl to, reikalinga alternatyva – orlaivių kolektyvas, kuris geba atlikti didelio masto išskirstytus skaičiavimus, o kolektyvo agentai geba tarpusavyje koordinuotis - išvengti susidūrimų su statinėmis bei dinamiškomis kliūtimis. Tam, kad orlaiviai gebėtų naviguoti erdvėje esant dideliame kiekiui kitų orlaivių, šiame darbe yra pasitelkiama būrio metafora ir skatinamas grupinis judėjimas. Siūlomas sprendinys siekia taupyti orlaivio energiją, naudojamą skrydžio metu, bei palaikyti žemus orlaivių gamybos kaštus, kas lemia tik būtiniausias jutimo ir skaičiavimų galimybes.

Kituose skyriuose pateikiami komentarai apie susijusius darbus, informacija apie aktualias orlaivių koordinavimo ir kliūčių vengimo technologijas, bei aprašomas siūlomas orlaivių, kaip kolektyvo, koordinavimo metodas, integruojantis jėgų laukų ir pajusk-išvenk metodus kartu su būrio sudarymo taisyklių rinkiniu.

2. Susiję darbai

Sistemų, kaip kolektyvų, koordinavimo uždavinys mokslinėje literatūroje yra plačiai nagrinėjamas. Jis yra įvardijamas kaip susidūrimų išvengimo sistemos, transporto kelio planavimas, daugiaagentinių sistemų koordinavimas,

transporto tinklo koordinavimas, intelektualių kolektyvų koordinavimas. Todėl susijusių darbų analizė atlikta dviem žingsniais. Pirmiausia, buvo išanalizuotos nagrinėjama tema parašytos analitinės apžvalgos [1, 4, 5], kuriose išanalizuoti ir suklasifikuojami pasiūlyti metodai. Tuo remiantis buvo susiaurinta nagrinėtinų metodų aibė, pasirinkta sprendžiamo uždavinio ribojimus tenkinančių metodų aibė ir atlikta jų kritinė analizė.

Nagrinėjant analitines apžvalgas galima padaryti šias išvadas:

1. didesnioji dalis metodų yra skirti vieno nepilotuojamo orlaivio kelio nustatymui statinėje aplinkoje [1];
2. nebrangiems, griežtus minimalius atminties ir skaičiavimų galimybių ribojimus tenkinantiems UAV susidūrimų vengimui galima taikyti klasikinį, euristinį ir populiacija grindžiamus metodus (pagal [1, 5] darbe pateiktą klasifikaciją) arba jėgų laukų (klasikinis) ir pajusk-išvenk (euristinis) metodus (pagal [4] darbe pateiktą klasifikaciją);
3. nebrangiems, turintiems aktyvius jutiklius su minimaliomis kliūčių identifikavimo charakteristikomis UAV galima taikyti pajusk-išvenk metodus (pagal [4] darbe pateiktą klasifikaciją);
4. jei UAV traktuojami kaip kolektyvo nariai, didesnioji dalis metodų yra skirti išvengti susidūrimų, kai jie kooperuojasi siekdami atlikti bendrą užduotį ar užduotis [5].

Apibendrinant, galima teigti, kad mūsų nagrinėjamai problemai spręsti tinka klasikiniai, euristiniai ir populiacija grindžiami metodai.

Sistema, suprantama kaip kolektyvas, apibrėžiama kaip pakankamai didelė subjektų grupė, kur kiekvienas subjektas siekia asmeninės naudos, bet kartu grupiniais veiksmais gali būti pasiekta ir visos sistemos nauda, o taip pat gali būti užtikrinamas aukštesnio lygmens intelektas [6]. Pastebėsime, kad šiame darbe nagrinėjamo kolektyvo kiekvienas narys siekia savo tikslų; o sistemos požiūriu gaunamas sąlyginis sinergetinis efektas – galimybė pasiekti tikslus visiems kolektyvo nariams. Tai dar labiau susiaurina taikytinų metodų aibę.

Paprastiausiai euristiniai pajusk-išvenk Bug algoritmai, kurie nagrinėjami [7] darbe, negalėtų užtikrinti patikimo kliūties išvengimo dinamiškoje aplinkoje. Šie algoritmai reikalauja minimalių aplinkos duomenų ir patenka į „godžių“ (angl. *greedy*) algoritmų grupę, nes jie visada pasirenka pirmą aptinkamą užduoties sprendinį, kuris retai būna geriausias iš galimų sprendinių. Bug algoritmai pasižymi žemu efektyvumu jeigu naudojami pavieniui,

bei gali pasimesti ir nerasti sprendimo sudėtingos formos aplinkose, tačiau, juos galima panaudoti kaip didesnio sprendinio dalį. Tačiau, šie algoritmai reikalauja nedaug skaičiavimo pajėgumų ir gali suteikti staigaus kliūties išvengimo ir apėjimo galimybes pritaikius didesnėje sistemoje.

Kad sukurti patikimą ir efektyviai veikiančią sistemą, siūloma naudoti reaktyvų jėgos laukų metodą, apjungtą su Bug 0 algoritmu. [8] darbe nagrinėjamas susidūrimų išvengimas, su statinėmis ir dinaminėmis kliūtimis, pasitelkiant jėgų laukus. Šis metodas naudojamas paprastos navigacijos užduočių atlikimui ir kliūčių išvengimui, tačiau jį pasitelkus gali kilti problemų didelio judesio ir aplinkos erdvės užimtumo atvejais. Jėgos laukų metodas yra tinkamas naudojimui paprastos struktūros, statinėse, ar nestaigiai besikeičiančiose aplinkose, tačiau jam veikiant pavieniui gali kilti problemų aukšto sudėtingumo statinėse aplinkose, bei pernelyg staigiai besikeičiančiose dinamiškose aplinkose [4]. Dėl to siūloma apjungti šį metodą kartu su paprastu pajusk-išvenk tipo Bug tipo metodu.

[2] darbe nagrinėjamas susidūrimų išvengimas, kai būrys keičia savo konfigūraciją arba kyla grėsmė susidurti dviem būriams. Metodas pritaikomas dideliems orlaivių kiekams būriuose, tačiau jis grindžiamas judesio vektorių skaičiavimu ir ateities pozicijų numatymu, kam reikia didelių skaičiavimo galimybių, kai didelis orlaivių kiekis juda ribotoje aplinkoje. Analogiška problema sprendžiama [3] darbe, papildomai apimant reaktyvų susidūrimų vengimo mechanizmą, kas leidžia koordinavimą valdyti naudojant tik integruotų jutiklių duomenis. Tačiau, tam kad įvertinti kitų UAV padėtį, naudojama įvairiakryptė kamera ir vaizdo apdorojimo metodas, o tai reikalauja didelių išteklių duomenims apdoroti.

3. Kolektyvo koordinavimas jungiant jėgų laukų ir Bug 0 metodus

Naudojant pritaikytą klasikinio jėgų laukų ir euristinio pajusk-išvenk Bug 0 kliūčių išvengimo ir kelio paieškos metodų junginį kartu su būrio sudarymo taisyklių rinkiniu yra sukuriama išskirstyta reaktyvių orlaivių koordinavimo sistema, kuri geba užtikrinti skrydžio tikslo pasiekimą efektyviai išvengiant susidūrimų.

Tokia sistema leidžia orlaiviui:

1. skristi tikslo link, išvengiant statinių kliūčių, bei kitų pavieniui skrendančių orlaivių, ar orlaivių būrių,

2. aptikti kitus orlaivius ir nustatyti, ar jų skrydžio kryptis artima savo skrydžio kryptčiai; jeigu dviejų orlaivių skrydžio kryptys yra artimos, orlaiviai gali kartu sudaryti būrį, ir skristi kartu tikslo link,
3. atsiskirti nuo orlaivių būrio, jeigu orlaivis nustato, kad kryptis savo tikslo link išsiskyrė nuo būrio skrydžio krypties.

Kadangi siekiama sukurti minimaliai sudėtingą, patikimai veikiantį sprendimą, sistemos orlaiviams pasirinkta suteikti šias priemones:

1. LIDAR jutiklį [4],
2. trumpo atstumo komunikacijos modulį (kaip 802.11 standarto prietaisą),
3. inercinę matavimo sistemą (magnetometras, giroskopas, akcelerometras),
4. pozicijos nustatymo sistemą (GPS).

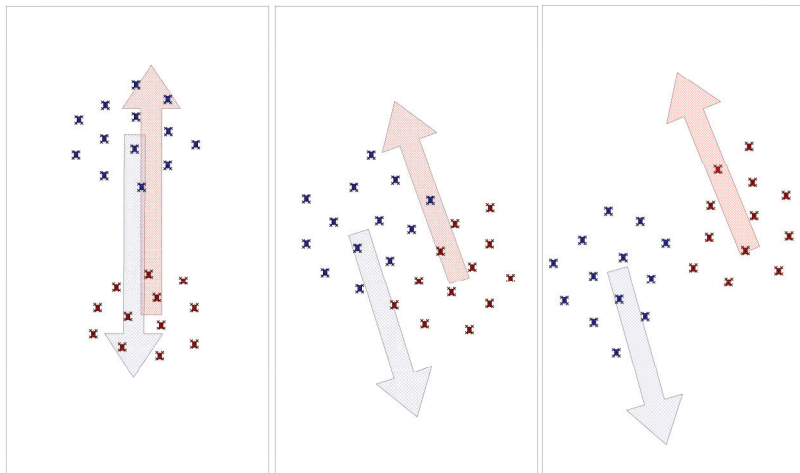
Kiekvienas sistemos orlaivis turės prieigą tik prie savo jutiklių duomenų. Tarp užduoties pradžios ir pabaigos orlaiviai operuos visiškai savarankiškai.

Skrydžio metu orlaiviai apdoroja savo LIDAR jutiklių gaunamą informaciją ir pagal tai sukuria dirbtinius jėgos laukus, kuriuose statinės kliūtys ir kiti skrendantys orlaiviai skleidžia stūmos jėgą. Siekiant apsaugoti orlaivius nuo situacijų, kuriose orlaivis patenka į jėgos laukų lokalaus minimumo tašką ir nesugeba iš ten pajudėti, prijungiamas euristinis pajusk-išvenk metodas, kuris tokiais atvejais įjungia reaktyvų Bug 0 algoritmą [7]. Šis Bug 0 algoritmas reiškia, kad orlaivis pradės sekti kliūties sieną iki tol, kol galės toliau judėti tikslo link. Kliūties vengimo metu orlaivis taip pat mėgins didinti savo aukštį, siekdamas apskristi kliūtį iš viršaus.

Tuo pačiu metu, kiekvienas orlaivis į aplinką siunčia signalus apie savo GPS koordinatų poziciją ir kryptį tikslo link. Kiti netoliese esantys sistemos agentai gauda tuos signalus ir lygina gautus duomenis su savo skrydžio informacija. Atlikę atstumo, bei krypties artimumo skaičiavimus, orlaiviai atsižvelgia į savo būriavimosi taisyklėse aprašytus parametrus ir nusprendžia, ar jiems verta skristi kartu. Šis veiksmas leidžia sudaryti orlaivių būrius, kur visi būrio nariai siekia savo tikslų ir skrenda ta pačia kryptimi. Toks būriavimosi mechanizmas leidžia dideliems orlaivių kiekiams lengviau prasilenkti.

1 pav. matoma situacija, kurioje 12 orlaivių siekia skristi link kitų 12 orlaivių, kurie skrenda priešinga kryptimi. Šie orlaiviai panaudoję savo komunikacijos modulius, kad atpažintų kitus artimus orlaivius, ir sudarytų būrį

su tais, kurie skrenda ta pačia kryptimi. Dėl to susidaro 2 orlaivių būriai po 12 narių. Galiausiai, du priešingomis kryptimis skrendantys orlaivių būriai sugebėtų efektyviai prasilenkti neprarandant savo būrio vientisumo.



1 pav. Vaizdų rinkinys, kuriame matomas dviejų, priešprieša skrendančių, orlaivių būrių prasilenkimo veiksmas

4. Išvados

Siekiant pasitelkti didelius skraidančių orlaivių kiekius užduočių įgyvendinimui svarbu užtikrinti visų sistemos dalyvių saugumą. Šiam tikslui pasiekti galima pritaikytų daugelį skirtingų mokslinėje literatūroje aprašytų koordinavimo būdų, kurie skiriasi reikalingos ir apdorojamos informacijos kiekiu, orlaivių elgsena ir techninėmis charakteristikomis. Siekiant sumažinti atliekamų skaičiavimų ir naudojamos atminties kiekius, naudojamas dviejų reaktyvių ključių išvengimo metodų junginį kartu su būriavimosi funkcija. Šiame darbe nurodytus reikalavimus tenkinantiems bepiločiams orlaiviams koordinuoti reikia taikyti klasikinį jėgos laukų ključių išvengimo metodą kartu su euristiniu pajusk-išvenk ključių išvengimo metodu ir būriavimosi funkcija.

Tęsiant aprašytą tyrimą, numatoma atlikti gautų teorinių rezultatų vertinimą simuliacinėje aplinkoje.

Padėka. Autorius dėkoja Duomenų mokslo ir skaitmeninių technologijų instituto doc. Audronei Lupeikienei už patarimus ir konsultacijas šio projekto metu, bei rengiant šį straipsnį.

Literatūra

- [1] Amylia Ait Saadi. Coordination of scout drones (UAVs) in smart-city to serve autonomous vehicles. Université Paris Saclay, Doctoral Thesis, 2024. Randamas adresu: <https://theses.hal.science/tel-04530665>.
- [2] Marek, D.; Biernacki, P.; Szyguła, J.; Domanski, A.; Paszkuta, M.; Szczygiel, M.; Krol, M.; Wojciechowski, K. Collision Avoidance Mechanism for Swarms of Drones. *Sensors* 2025, 25, 1141. <https://doi.org/10.3390/s25041141>.
- [3] Afzal Ahmad, Daniel Bonilla Licea, Giuseppe Silano, Tomáš Báča, Martin Saska. PACNav: a collective navigation approach for UAV swarms deprived of communication and external localization. *Bioinspiration & Biomimetics* 17(6), November 2022, 1-19.
- [4] Jawad N. Yasin; Sherif A. S. Mohamed; Mohammad-Hashem Haghbayan; Jukka Heikkonen; Hannu Tenhunen; Juha Plosila. Unmanned Aerial Vehicles (UAVs): Collision Avoidance Systems and Approaches. *IEEE Access* (Volume: 8), 2020, 105139 – 105155.
- [5] Soheila Ghambari, Mahmoud Golabi, Laetitia Jourdan, Julien Lepagnot, Lhassane Idoumghar. UAV Path Planning Techniques: A Survey. *RAIRO Operations Research* 58(4), 2024, 951–2989.
- [6] Ioanna Lykourantzou, Dimitrios J. Vergados, Vassili Loumos. Collective Intelligence System Engineering. *MEDES '09: Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, 2009, Article No.: 20, 134–140.
- [7] Gupta, S., Asha, C. S., & D'Souza, J. M. (2023). Implementation and Comparison of BUG Algorithms on ROS. 2023 2nd International Conference for Innovation in Technology, INOCON 2023. <https://doi.org/10.1109/INOCON57975.2023.10101325>.
- [8] Almira Budiyanto, Adha Cahyadi, Teguh Bharata Adji, Oyas Wahyunggoro. UAV Obstacle Avoidance Using Potential Field under Dynamic Environment. *Proceedings of the International Conference on Control, Electronics, Renewable Energy and Communications (ICCE-REC)*, 2015, 187–192.

Atsitiktinio miško modelio taikymas užpildant trūkstamas ekonominių rodiklių reikšmes NUTS 2 lygmeniu

Irmantas Pilypas, Irma Šileikienė

Vilniaus universiteto Šiaulių akademija, Regionų plėtros institutas
Vytauto g. 84, Šiauliai
irmantas.pilypas@sa.stud.vu.lt, irma.sileikiene@sa.vu.lt

Santrauka. Nagrinėjamos trūkstamų ekonominių rodiklių reikšmių ES NUTS 2 regionų lygmeniu užpildymo galimybės, taikant atsitiktinio miško (angl. *Random Forest*) modelį. Trūkstamos reikšmės ekonominiuose duomenyse kelia iššūkių analizei ir gali mažinti analizės rezultatų patikimumą. Sudarytas duomenų rinkinys, apimantis 260 regionų ir 8 ekonominius rodiklius (1990–2023), surinktas iš Eurostat DB. Kiekvienam rodikliui sukurtas atskiras atsitiktinio miško modelis trūkstamų reikšmių užpildymui. Modelių tikslumas vertintas pagal 4 metrikas: RMSE, MAE, R^2 , MAPE. Tyrimas atskleidė, kad atsitiktinio miško metodas ypač tiksliai užpildo užimtumo, nedarbo ir ilgalaikio nedarbo rodiklių reikšmes ($R^2 > 0,9$), išlaikydamas tarpusavio ryšius tarp kintamųjų. Gautas užpildytas duomenų rinkinys gali būti taikomas tolesnėje regioninės ekonomikos rodiklių analizėje.

Raktiniai žodžiai: atsitiktinis miškas, mašininis mokymasis, trūkstamos reikšmės, ekonominiai rodikliai, NUTS 2 regionai, hiperparametrų optimizavimas.

1 Įvadas

Europos Sąjungos regioninės ekonominės analizės kokybė ir tikslumas dažnai priklauso nuo turimų statistinių duomenų išsamumo. NUTS 2 (angl. *Nomenclature of Territorial Units for Statistics*) regionų duomenų rinkiniuose neretai pasitaiko trūkstamų reikšmių, ypač mažiau išsivysčiusiuose regionuose ar tam tikrais laiko periodais. Tai sukelia analitikams ir tyrėjams reikšmingų iššūkių, nes trūkstami duomenys riboja galimybes atlikti patikimą ekonominių rodiklių analizę ir palyginimą [10].

Trūkstamų reikšmių užpildymas – tai metodas, skirtas trūkstamoms reikšmėms užpildyti statistiniais ar mašininio mokymosi metodais [13]. Tradiciniai užpildymo metodai, tokie kaip vidurkio ar medianos naudojimas, neišsaugo kintamųjų tarpusavio ryšių, todėl gali iškreipti duomenų struktūrą.

lą [6]. Pažangesni metodai, kaip MICE (angl. *Multiple Imputation by Chained Equations*) metodas, geriau išsaugo kintamųjų tarpusavio ryšius [25], tačiau jų taikymas didelės apimties daugiamatiams duomenims gali kelti skaičiavimo iššūkių [11].

Šiame tyrime analizuojamas atsitiktinio miško metodo taikymas, užpildant trūkstamas ekonominių rodiklių reikšmes NUTS 2 regionų lygmeniu. Atsitiktinio miško metodas pasižymi gebėjimu modeliuoti kompleksinius netiesinius ryšius tarp kintamųjų ir atsparumu išskirtims [4].

Tyrimo tikslas yra ištirti atsitiktinio miško metodo efektyvumą užpildant įvairių ekonominių rodiklių NUTS 2 lygmeniu trūkstamus duomenis.

Tiksliui pasiekti iškelti tokie **uždaviniai**:

1. Išanalizuoti ankstesniuose tyrimuose naudotus trūkstamų reikšmių duomenų rinkiniuose užpildymo metodus ir tikslumo metrikas;
2. Sudaryti empirinio tyrimo metodiką trūkstamų reikšmių duomenų rinkiniuose užpildymui;
3. Įvertinti atlikto trūkstamų reikšmių duomenų rinkiniuose užpildymo gautus rezultatus, naudojant atsitiktinio miško bazinį metodą ir su optimizuotais hiperparametrais.

Šiame straipsnyje yra analizuojama ir išbandoma trūkstamų reikšmių užpildymo proceso metodologija, analizuojant įvairių ekonominių rodiklių užpildymo rezultatus ir formuluojant rekomendacijas būsimiems tyrimams šioje srityje.

Tyrimui atlikti naudoti mokslinės literatūros analizės ir empirinių tyrimų metodai.

2 Literatūros analizė

Trūkstamų reikšmių užpildymo metodai plačiai taikomi įvairiose srityse, įskaitant statistiką, ekonometriją, socialinių mokslų ir biomedicininis tyrimus. Trūkstamų reikšmių užpildymas gali būti skirstomas į vienmatį ir daugiamatį, priklausomai nuo to, ar atsižvelgiama į kintamųjų tarpusavio ryšius [9].

Tradiciniai užpildymo metodai, tokie kaip vidurkis, medianos ar modos naudojimas, yra paprasti, tačiau jie neatsižvelgia į kintamųjų tarpusavio ryšius ir gali iškraipyti duomenų struktūrą [12]. Statistiniai metodai, tokie kaip regresinis užpildymas ar k-artimiausių kaimynų (k-NN) metodas, geriau išsaugo kintamųjų tarpusavio ryšius, tačiau jie taip pat turi trūkumų, ypač dirbant su didelės apimties daugiamatiais duomenimis [1].

MICE (Multiple Imputation by Chained Equations) metodas tapo populiarus dėl savo universalumo ir gebėjimo išsaugoti kintamųjų tarpusavio ryšius [2]. MICE metodas skaičiuoja kiekvieno kintamojo prognostinį modelį, naudodamas kitus kintamuosius kaip prediktorius, ir iteratyviai atnaujina užpildytas reikšmes, kol pasiekama konvergencija [27]. Tačiau MICE metodas gali būti, skaičiavimo prasme, imlus dideliems duomenų rinkiniams.

Mašininio mokymosi metodai, tokie kaip atsitiktinių miškų, neuroniniai tinklai ir gradientinis stiprinimas (gradient boosting), vis dažniau taikomi trūkstamų reikšmių užpildymui [23]. Tang ir Ishwaran [24] parodė, kad atsitiktinio miško metodas gali efektyviai užpildyti trūkstamas reikšmes, išsaugodamas kintamųjų tarpusavio ryšius. Shah ir kt. [21] nustatė, kad atsitiktinio miško metodas dažnai pranoksta tradicinius užpildymo metodus, ypač kai duomenys yra susiję netiesiškai.

Atsitiktinio miško metodas turi keletą privalumų trūkstamų reikšmių užpildymo kontekste: (1) jis gali modeliuoti kompleksinius netiesinius ryšius tarp kintamųjų; (2) jis yra atsparus išimtims ir triukšmui duomenyse; (3) jis gali apdoroti didelės apimties daugiamačius duomenis [26]. Tačiau atsitiktinio miško metodo efektyvumas priklauso nuo tinkamo hiperparametrų parinkimo, kuris gali būti sudėtingas uždavinys [19].

Hiperparametrų optimizavimas yra esminė atsitiktinio miško metodo taikymo dalis. Dažniausiai optimizuojami hiperparametrai yra medžių skaičius, maksimalus medžio gylis, minimalus pavyzdžių skaičius šakos mazge ir minimalus pavyzdžių skaičius lape [3]. Probst ir kt. [18] nustatė, kad medžių skaičius ir maksimalus medžio gylis turi didžiausią įtaką atsitiktinio miško metodo efektyvumui.

Europos Sąjungos NUTS 2 regionų lygmens kontekste, ekonominių rodiklių trūkstamų reikšmių užpildymo tyrimų, atlikta nedaug. García-Laencina ir kt. [8] taikė įvairius mašininio mokymosi metodus trūkstamų reikšmių užpildymui fundamentaliuose tyrimuose neprisirišant prie duomenų pobūdžio, todėl rezultatai gali būti pritaikomi ir ekonominių rodiklių analizei. Aljinbaz ir kt. [15] analizavo nedarbo lygio prognozavimą įvairiose šalyse, pasitelkiant duomenų struktūravimo metodus ir neuroninius tinklus. Atliekant literatūros analizę tyrimų, kuriuose analizuojami bendri, Europos sąjungos NUTS 2 regionų lygmenyje ekonominių rodiklių, trūkstamų reikšmių užpildymo metodai, nebuvo aptikta.

3 Duomenys ir metodologija

3.1 Duomenų aprašymas

Tyrime naudojamas NUTS 2 regionų ekonominių rodiklių duomenų rinkinys, apimantis 1990–2023 metų laikotarpį. Duomenys apima 260 unikalius NUTS 2 regionus (iš viso 6984 stebėjimai) ir 8 ekonominius rodiklius [22]:

1. **Employment_rate** - užimtumo lygis (%);
2. **Rd_expenditure_pct_gdp** - išlaidos moksliniams tyrimams ir technologinei plėtrai, procentais nuo BVP;
3. **Unemployment_rate** - nedarbo lygis (%);
4. **Tertiary_education_pct** - aukštąjį išsilavinimą turinčių gyventojų dalis (%);
5. **Youth_unemployment_rate** - jaunimo nedarbo lygis (%);
6. **Population_density** - gyventojų tankis (gyv./km²);
7. **Long_term_unemployment_share** - ilgalaikio nedarbo dalis (%);
8. **Female_employment_rate** - moterų užimtumo lygis (%).

Tolesnėje analizėje, siekiant aiškesnio rodiklių identifikavimo, jie bus žymimi ROD1-ROD8 indeksais, atitinkančiais aukščiau pateiktą numeraciją.

Visi tyrimo duomenys buvo surinkti iš Europos Sąjungos statistikos tarnybos (Eurostat) oficialios duomenų bazės, užtikrinant patikimumą ir nuoseklumą tarp skirtingų regionų ir laikotarpių [22].

Duomenų rinkinyje yra reikšmingas trūkstamų reikšmių skaičius, kuris skiriasi tarp rodiklių. Trūkstamų reikšmių procentas svyruoja nuo 10 % (population_density) iki 53,8 % (rd_expenditure_pct_gdp), kaip parodyta 1 lentelėje.

1 lentelė. Trūkstamų reikšmių procentas pagal ekonominį rodiklį

Rodiklio ID	Rodiklis	Trūkstamos reikšmės (%)
ROD1	employment_rate	28,15
ROD2	rd_expenditure_pct_gdp	53,81
ROD3	unemployment_rate	29,04
ROD4	tertiary_education_pct	29,70
ROD5	youth_unemployment_rate	36,78
ROD6	population_density	10,01
ROD7	long_term_unemployment_share	35,09
ROD8	female_employment_rate	28,15

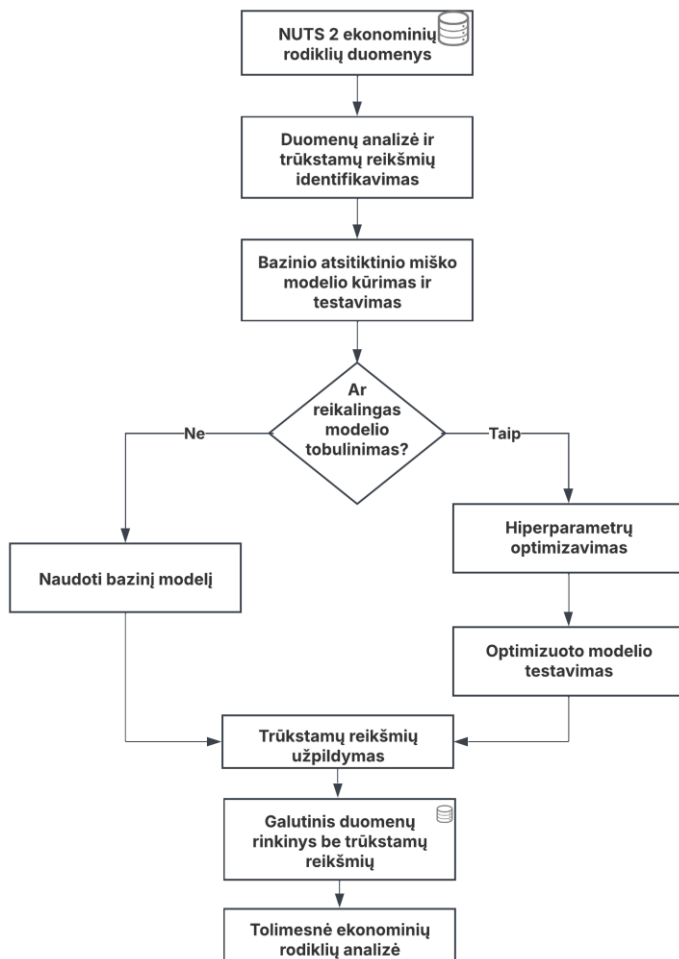
Tyrimo duomenų rinkinys pateiktas Github saugykloje: https://github.com/Pilypas/Random_Forest-nuts2-ekonominiu-rodikliu-uzpildymas

3.2 Metodologija

Tyrimo metodologija apima kelis esminius žingsnius:

1. **Duomenų analizė ir paruošimas** - atliekama pradinė duomenų analizė, identifikuojami kintamųjų pasiskirstymai, tarpusavio ryšiai bei trūkstamos reikšmės.
2. **Dirbtinių testavimo duomenų kūrimas** - sukuriama sintetiniai testavimo duomenys užpildymo metodų tikslumui įvertinti. 20 % netrūkstamų reikšmių atsitiktinai pašalinama iš duomenų rinkinio, o vėliau šios pašalintos reikšmės naudojamos užpildymo tikslumo vertinimui [14].
3. **Bazinio atsitiktinio miško modelio kūrimas** - sukuriamas bazinis atsitiktinio miško modelis naudojant Python scikit-learn bibliotekos numatytašias (angl. *default*) hiperparametrų reikšmes: 100 medžių, maksimalus gylis neribojamas, minimalus pavyzdžių skaičius šakos mazge - 2, minimalus pavyzdžių skaičius lape - 1.
4. **Hiperparametrų optimizavimas** - bazinio modelio hiperparametrai optimizuojami naudojant *Grid Search* metodą su 5 sluoksnių kryžminiu patikrinimu. Paieškos aibę sudaro šie hiperparametrų rinkiniai [20]:
 - `n_estimators` (medžių skaičius): {50, 100, 200};
 - `max_depth` (maksimalus medžio gylis): {None, 10, 20, 30};
 - `min_samples_split` (minimalus pavyzdžių skaičius šakos mazge): {2, 5, 10};
 - `min_samples_leaf` (minimalus pavyzdžių skaičius lape): {1, 2, 4}.
5. **Modelių efektyvumo vertinimas** - bazinio ir optimizuotų modelių efektyvumas vertinamas pagal šias metrikas [17]:
 - RMSE (angl. *Root Mean Square Error*) - vidutinis kvadratinis nuokrypis;
 - MAE (angl. *Mean Absolute Error*) - vidutinė absoliutinė paklaida;
 - R^2 (angl. *Determination Coefficient*) - determinacijos koeficientas;
 - MAPE (angl. *Mean Absolute Percentage Error*) - vidutinė absoliutinė santykinė paklaida.
6. **Požymių svarbos analizė** - analizuojama požymių svarba kiekvieno ekonominio rodiklio prognozavimui, identifikuojant svarbiausius prediktoriaus.

1 pav. pavaizduotas nuoseklus trūkstamų reikšmių užpildymo procesas, kurio metu priklausomai nuo bazinio modelio tikslumo sprendžiama, ar reikalingas papildomas hiperparametrų optimizavimas, siekiant užtikrinti geriausią įmanomą kiekvieno ekonominio rodiklio trūkstamų reikšmių užpildymo tikslumą.



1 pav. Ekonominių rodiklių NUTS 2 regionų lygmeniu trūkstamų reikšmių užpildymo proceso blokinė schema

Tyrimas atliktas naudojant Python programavimo kalbą ir scikit-learn bibliotekos atsitiktinio miško regressor modelį trūkstamų reikšmių užpildymo algoritmo įgyvendinimui.

Atsitiktinio miško metodo apibendrintas algoritmas trūkstamų reikšmių užpildymui:

1. Kiekvienam ekonominiam rodikliui, turinčiam trūkstamų reikšmių, sukuriamas atskiras atsitiktinio miško modelis.
2. Šis modelis apmokomas naudojant stebėjimus, kuriuose tikslinė reikšmė yra žinoma.
3. Mokymo metu kiti ekonominiai rodikliai naudojami kaip prediktoriai (nepriklausomi kintamieji).
4. Apmokytasis modelis naudojamas prognozuoti trūkstamas reikšmes, remiantis kitų rodiklių vertėmis.

Pavyzdžiui, jei konkrečiame regione trūksta nedarbo lygio duomenų, modelis apsimoko, nustatydamas ryšius tarp nedarbo lygio ir kitų rodiklių (užimtumo lygio, išsilavinimo lygio, gyventojų tankio ir kt.) naudodamas kitų regionų turimus nedarbo lygio duomenis. Vėliau šis modelis naudojamas įvertinti trūkstamą reikšmę, remiantis kitais to regiono rodikliais.

4 Rezultatai

4.1 Bazinio atsitiktinio miško modelio rezultatai

Pirmiausia sukurtas bazinis atsitiktinio miško modelis ir įvertintas modelio efektyvumas. Bazinis modelis sukurtas su standartiniais hiperparametrais, naudojant 100 medžių. 2 lentelėje pateiktos bazinio atsitiktinio miško modelio tikslumo vertinimo metrikos kiekvienam ekonominiam rodikliui.

2 lentelė. Bazinio atsitiktinio miško modelio metrikos

Rodiklio ID	MSE	RMSE	MAE	R ²	MAPE (%)
ROD1	1,9420	1,3936	1,0516	0,9739	1,5586
ROD2	0,4557	0,6751	0,4096	0,7180	-
ROD3	2,2041	1,4846	1,0418	0,9344	13,2975
ROD4	23,9705	4,8960	3,5438	0,7819	15,0099
ROD5	20,1910	4,4934	3,1664	0,8828	17,6417
ROD6	369181,5051	607,6031	273,5064	0,4586	323,1901
ROD7	1,8774	1,3702	0,8884	0,8862	26,3958
ROD8	4,9056	2,2148	1,5780	0,9668	2,8183

Rezultatai rodo, kad bazinis atsitiktinio miško modelis pasiekia gana aukštą tikslumą daugumai rodiklių. Ypač aukšti R² rodikliai pasiekiami (ROD1) užimtumo lygio (0,9739), (ROD3) nedarbo lygio (0,9344) ir (ROD8)

moterų užimtumo lygio (0,9668) rodikliams, o tai rodo stiprų sąryšį tarp rodiklių. Tačiau (ROD6) gyventojų tankio (population_density) rodiklio užpildymo tikslumas yra žymiai mažesnis ($R^2 = 0,4586$), o MAPE vertė labai aukšta (323,19 %), kas rodo, kad šio rodiklio užpildymas yra labai netikslus.

4.2 Hiperparametrų optimizavimo rezultatai

Siekiant padidinti užpildymo tikslumą, atliktas atsitiktinio miško modelio hiperparametrų optimizavimas naudojant Grid Search metodą. 3 lentelėje pateikiami gauti geriausi hiperparametrų rinkiniai kiekvienam ekonominiam rodikliui.

3 lentelė. Optimalūs hiperparametrai kiekvienam rodikliui

Rodiklio ID	max_depth	min_samples_leaf	min_samples_split	n_estimators
ROD1	10	1	2	100
ROD2	None	4	2	200
ROD3	10	4	10	100
ROD4	None	4	2	200
ROD5	10	4	2	200
ROD6	10	2	2	100
ROD7	20	4	10	100
ROD8	20	2	10	200

Optimizuoti hiperparametrai skiriasi lyginant skirtingus rodiklius, kas rodo, kad skirtingiems ekonominiams rodikliams reikalingi skirtingo sudėtingumo modeliai. 4 lentelėje pateiktos optimizuotų atsitiktinio miško modelių tikslumo vertinimo metrikos ir kryžminio patikrinimo rezultatai.

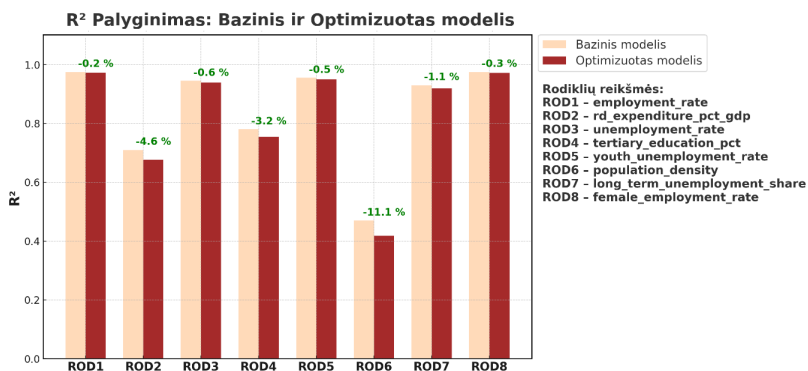
Pažymėtina, kad optimizuoti modeliai ne visada pagerina užpildymo tikslumą, lyginant su baziniu modeliu. Kai kuriais atvejais, pavyzdžiui, (ROD1) užimtumo lygio (employment_rate) ir (ROD3) nedarbo lygio (unemployment_rate) rodikliams, bazinis modelis pasiekia šiek tiek geresnį tikslumą. Tai gali būti susiję su persimokymo (overfitting) problemomis, kai optimizuoti modeliai pernelyg prisitaiko prie mokymo duomenų ir prasčiau apibendrina naujus duomenis [7].

4 lentelė. Optimizuotų atsitiktinio miško modelių tikslumo vertinimo metrikos ir kryžminio patikrinimo rezultatai

Rodiklio ID	Optimizuotų rodiklių tikslumo metrikos						Kryžminio patikrinimo rezultatai			
	MSE	RMSE	MAE	R ²	MAPE (%)	RMSE (vid±std)	MAE (vid±std)	R ² (vid±std)		
ROD1	2,0841	1,4436	1,0967	0,9720	1,6261	2,395±0,8633	1,8377±0,5638	0,9168±0,0406		
ROD2	0,5094	0,7137	0,4339	0,6848	-	1,0577±0,1715	0,7104±0,1073	0,0673±0,1068		
ROD3	2,3870	1,5450	1,0976	0,9290	14,2330	1,9987±0,3797	1,4957±0,2731	0,8472±0,0635		
ROD4	26,6963	5,1668	3,7738	0,7571	15,9998	8,4068±1,1812	6,8070±0,9329	0,2749±0,1369		
ROD5	20,9363	4,5756	3,2666	0,8785	18,4185	5,9680±1,0842	4,4952±0,8218	0,7347±0,0955		
ROD6	403778,6331	635,4358	287,2524	0,4078	364,9937	876,4285±333,1299	438,1537±112,5270	-0,5673 ±0,5968		
ROD7	2,0408	1,4286	0,9218	0,8763	27,3312	2,0400±0,4970	1,3785±0,3726	0,7495±0,1054		
ROD8	5,2680	2,2952	1,6518	0,9643	2,9731	3,8652±1,1000	2,8837±0,7635	0,8824±0,0330		

4.3 Modelių efektyvumo vertinimas

Įvertinus bazinio ir optimizuotų atsitiktinio miško modelių tikslumą (2 pav.), svarbu palyginti jų efektyvumą įvairiems ekonominiams rodikliams. Šis palyginimas padeda nustatyti, ar hiperparametrų optimizavimas iš tiesų pagerino trūkstančių reikšmių užpildymo tikslumą, ir kokius modelius geriausia naudoti galutiniam duomenų užpildymui.



2 pav. R² palyginimas - bazinio ir optimizuotų modelių

2 paveikslėlyje galima pastebėti, kad hiperparametrų optimizavimas daugeliu atvejų neduoda reikšmingo tikslumo pagerėjimo, o kai kuriais atvejais netgi šiek tiek pablogina rezultatus. Tai gali būti susiję su tuo, kad bazinis atsitiktinio miško modelis jau buvo pakankamai gerai pritaikytas duomenims, o optimizavimo metu galėjo įvykti pernelyg didelis prisitaikymas prie mokymo duomenų. Remiantis kryžminio patikrinimo rezultatais (žr. 4 lentelę), kur matomas žymus skirtumas tarp testavimo metrikų ir standartinių nuokrypių, ypač (ROD6) gyventojų tankio (population_density) rodikliui, kurio R² vidurkis yra -0,5673 su dideliu standartiniu nuokrypiu ($\pm 0,5968$). Tokie dideli standartiniai nuokrypiai ir neigiamos R² vertės rodo, kad optimizuoti modeliai skirtingose kryžminio patikrinimo sluoksniuose veikia pakankamai netolygiai, kas yra būdinga persimokymo atvejams.

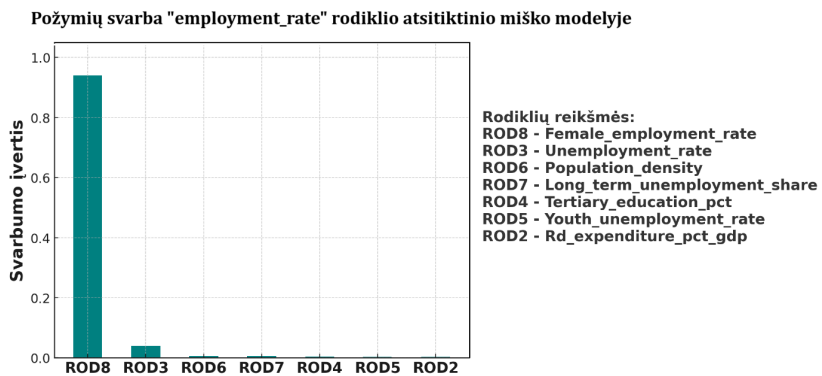
Lyginant su kitais literatūroje aprašytais metodais, tyrime analizuotas atsitiktinio miško metodas pasiekia panašius ar geresnius rezultatus. Pavyzdžiui, Aljinbaz ir kt. [15] taikydami neuroninius tinklus nedarbo lygio prognozavimui įvairiose šalyse pasiekė (R² apie 0,87), o šiame tyrime analizuotas metodas pasiekia (R² apie 0,93) nedarbo lygio rodikliui. Panat ir Chandra

[16] savo tyrime apie Globalų Gyvenimo Kokybės Indeksą pasiekė pakankamai gerus tikslingumo rezultatus ekonominiams rodikliams (R^2 daugiau nei 0,7), tačiau nedarbo lygio rodiklio užpildymas buvo mažiau tikslus (R^2 apie -0,25), kai tuo tarpu šiame tyrime, šiam rodikliui pasiektas gerokai aukštesnis tikslumas. Tačiau, kaip ir minėtuose tyrimuose, šiame darbe taip pat buvo pastebėta, kad kai kurių rodiklių (ypač gyventojų tankio) prognozavimas yra žymiai sudėtingesnis.

Šio tyrimo rezultatus tiesiogiai lyginti su kituose tyrimuose gautais rezultatais negalima, nes naudoti skirtingi duomenų rinkiniai, tačiau šio tyrimo metu gauti atsitiktinio miško modelio tuščių reikšmių užpildymo rezultatai yra panašiam tikslumo lygmenyje.

4.4 Požymių svarbos analizė

Analizuojant požymių svarbą, galima identifikuoti, kurie ekonominiai rodikliai yra svarbiausi prognozuojant kitus rodiklius. Požymių svarbos vizualizacija užimtumo lygio (employment_rate) rodikliui pateikiama 3 pav.



3 pav. Požymių svarba užimtumo lygio (employment_rate) rodikliui

Požymių svarbos analizė rodo, kad (ROD8) moterų užimtumo lygis (female_employment_rate) yra svarbiausias požymis prognozuojant bendrą užimtumo lygį, kas nėra netikėta, atsižvelgiant į tai, kad moterų užimtumo lygis sudaro reikšmingą dalį bendro užimtumo lygio.

5 Išvados

Atlikto tyrimo rezultatai rodo, kad atsitiktinio miško metodas gali būti efektyviai naudojamas užpildant trūkstamas ekonominių rodiklių NUTS 2 regionų lygmeniu reikšmes. Tačiau metodų efektyvumas reikšmingai skiriasi tarp skirtingų ekonominių rodiklių, o tai rodo, kad kiekvieno rodiklio trūkstamų reikšmių užpildymo uždavinys turėtų būti vertinamas atskirai.

Geriausiai atsitiktinio miško metodas veikia užpildant užimtumo rodiklius (bendrą užimtumo lygį ir moterų užimtumo lygį), nedarbo lygį ir ilgalaikio nedarbo dalį. Šiems rodikliams pasiekiami aukšti determinacijos koeficientai R^2 didesnis nei 0,8, o tai rodo, kad atsitiktinio miško modeliai gali paaiškinti didžiąją dalį šių rodiklių variacijos. Tai gali būti susiję su tuo, kad šie rodikliai turi stiprius tarpusavio ryšius ir taip pat yra susiję su kitais socialiniais-ekonomiais rodikliais, tokiais kaip išsilavinimo lygis.

Gyventojų tankio ir išlaidų moksliniams tyrimams bei technologinei plėtrai rodiklių užpildymas yra mažiau tikslus. Gyventojų tankio atveju tai gali būti susiję su tuo, kad šis rodiklis labiau priklauso nuo geografinių, istorinių ir urbanistinių veiksnių, kurie nėra tiesiogiai susiję su kitais ekonomiais rodikliais duomenų rinkinyje [5]. Išlaidų moksliniams tyrimams ir technologinei plėtrai atveju, mažą tikslumą gali lemti tai, kad šis rodiklis labai priklauso nuo konkrečios valstybės mokslo ir inovacijų politikos, taip pat nuo regiono specializacijos ir pramonės struktūros, kurios nėra niekaip susietos su kitais duomenų rinkinio rodikliais.

Pastebėtina, kad hiperparametrų optimizavimas nedavė reikšmingų tikslumo pagerėjimų, o kai kuriais atvejais netgi šiek tiek pablogino rezultatus. Tai gali būti susiję su persimokymo problema, kai modeliai pernelyg prisitaiko prie mokymo duomenų ir prasčiau apibendrina naujus duomenis arba sąryšiai su kitais ekonomiais rodikliais iš duomenų rinkinio per silpni. Šie rezultatai primena paprastų modelių (parsimony) principo svarbą: sudėtingesni modeliai ne visada yra geresni, ypač kai turimų duomenų kiekis yra ribotas [28].

Požymių svarbos analizė išryškino ekonominių rodiklių tarpusavio ryšius. Pvz., moterų užimtumo lygis yra svarbiausias požymis prognozuojant bendrą užimtumo lygį, o nedarbo lygis ir jaunimo nedarbo lygis taip pat yra svarbūs. Tai atitinka ekonominę intuiciją ir patvirtina, kad atsitiktinio miško modeliai sugeba surasti ir panaudoti prasmingus ryšius tarp ekonominių rodiklių.

Remiantis atlikta analize, ekonominių rodiklių NUTS2 regionų lygmeniu trūkstamų reikšmių užpildymui, pasirinktas bazinis atsitiktinio miško mo-

delis, kadangi gautas pakankamas tikslumas daugumai rodiklių. Užpildžius visas trūkstamas reikšmes, sukurtas pilnas duomenų rinkinys, kuris gali būti panaudojamas tolimesnei ekonominei analizei tikintis tikslesnių ekonominės analizės ir prognozavimo rezultatų, naudojant kitus mašininio mokymosi algoritmus.

Tolimesniuose tyrimuose planuojama ištirti kitus mašininio mokymosi metodus, tokius kaip neuroniniai tinklai ar gradient boosting, trūkstamų ekonominių rodiklių reikšmių užpildymui. Taip pat būtų vertinga ištirti užpildymo metodus atsižvelgiant į laiko dimensiją, t.y. naudojant laiko eilučių metodus. Galiausiai, būtų naudinga ištirti, kaip trūkstamų reikšmių užpildymo metodai padeda užtikrinti tikslesnius ekonominės analizės rezultatus.

Planuojama atlikti ekonominių rodiklių analizę su pirminiu ir papildytu duomenų rinkiniu, naudojant kitus mašininio mokymosi algoritmus, siekiant nustatyti su kuriais duomenimis gaunami tikslesni ekonominės analizės rezultatai.

Literatūra

- [1] Andridge, R. R., & Little, R. J. A. (2010). A Review of Hot Deck Imputation for Survey Non-response. *International statistical review = Revue internationale de statistique*, 78(1), 40. <https://doi.org/10.1111/J.1751-5823.2010.00103.X>
- [2] Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1), 40. <https://doi.org/10.1002/MPR.329>
- [3] Bergstra, J., Ca, J. B., & Ca, Y. B. (2012). Random Search for Hyper-Parameter Optimization Yoshua Bengio. *Journal of Machine Learning Research*, 13, 281–305. <http://scikit-learn.sourceforge.net>.
- [4] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324/METRICS>
- [5] Cui, C., Hu, Y., Bao, Y., & Li, H. (2024). Population Density Prediction at Township Scale Supported by Machine Learning Method: A Case Study in Inner Mongolia. *ISPRS International Journal of Geo-Information*, 13(12). <https://doi.org/10.3390/IJGI13120426>
- [6] Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). Review: a gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10), 1087–1091. <https://doi.org/10.1016/J.JCLINEPI.2006.01.014>
- [7] Fatima, S., Hussain, A., Amir, S. Bin, Ahmed, S. H., & Aslam, S. M. H. (2023). XGBoost and Random Forest Algorithms: An in Depth Analysis. *Pakistan Journal of Scientific Research*, 3(1), 26–31. <https://doi.org/10.57041/PJOSR.V3I1.946>
- [8] García-Laencina, P. J., Sancho-Gómez, J. L., & Figueras-Vidal, A. R. (2010). Pattern classification with missing data: A review. *Neural Computing and Applications*, 19(2), 263–282. <https://doi.org/10.1007/S00521-009-0295-6>
- [9] Graham, J. W. (2008). *Missing Data Analysis: Making It Work in the Real World*. <https://doi.org/10.1146/annurev.psych.58.110405.085530>

- [10] Guide to statistics in European Commission development co-operation 2017 edition. (s.a.). <https://doi.org/10.2785/30851>
- [11] Ieva Ivanauskienė. (2022). ES šalių konkurencingumo vertinimas ir klasterizavimas Baigiamasis magistro studijų projektas. Kauno technologijos universitetas.
- [12] Jönsson, P., & Wohlin, C. (2004). An evaluation of k-nearest neighbour imputation using Ilkert data. Proceedings - International Software Metrics Symposium, 108–118. <https://doi.org/10.1109/METRIC.2004.1357895>
- [13] Little, R. J. A., & Rubin, D. B. (2014). Statistical analysis with missing data. *Statistical Analysis with Missing Data*, 1–381. <https://doi.org/10.1002/9781119013563>
- [14] Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., Fu, T., & Wei, W. (2023). Machine Learning for Synthetic Data Generation: A Review. <https://arxiv.org/abs/2302.04062v9>
- [15] Monir Aljinbaz, A. M., Mahmoud, M., & Rahhal, A. (2024). Forecasting Unemployment Rate for Multiple Countries Using a New Method for Data Structuring. *IJACSA International Journal of Advanced Computer Science and Applications*, 15(12). www.ijacsa.thesai.org
- [16] Panat, T., & Chandra, R. (s.a.). Global Ease of Living Index: a machine learning framework for longitudinal analysis of major economies.
- [17] (PDF) A Consolidated Overview of Evaluation and Performance Metrics for Machine Learning and Computer Vision. (s.a.). Gauta 2025 m. balandžio 1 d., https://www.researchgate.net/publication/374558675_A_Consolidated_Overview_of_Evaluation_and_Performance_Metrics_for_Machine_Learning_and_Computer_Vision
- [18] Probst, P., Bischl, B., & Boulesteix, A.-L. (2018). Tunability: Importance of Hyperparameters of Machine Learning Algorithms. <http://arxiv.org/abs/1802.09596>
- [19] Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3), e1301. <https://doi.org/10.1002/WIDM.1301>
- [20] RAMADHAN, M. M., SITANGGANG, I. S., NASUTION, F. R., & GHIFARI, A. (2017). Parameter Tuning in Random Forest Based on Grid Search Method for Gender Classification Based on Voice Frequency. *DEStech Transactions on Computer Science and Engineering*, cece. <https://doi.org/10.12783/DTCSE/CECE2017/14611>
- [21] Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. *American Journal of Epidemiology*, 179(6), 764–774. <https://doi.org/10.1093/AJE/KWT312>
- [22] Statistics | Eurostat. (s.a.). Gauta 2025 m. balandžio 1 d., <https://ec.europa.eu/eurostat/databrowser/view/TGS00042/default/table>
- [23] Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112–118. <https://doi.org/10.1093/BIOINFORMATICS/BTR597>
- [24] Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. <https://doi.org/10.1002/sam.11348>
- [25] van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67. <https://doi.org/10.18637/JSS.V045.I03>
- [26] Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., & Higgins, P. D. R. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, 3(8), e002847. <https://doi.org/10.1136/BMJOPEN-2013-002847>

- [27] White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4), 377–399. <https://doi.org/10.1002/SIM.4067>
- [28] Zhang, X., Chen, S., Xue, J., Wang, N., Xiao, Y., Chen, Q., Hong, Y., Zhou, Y., Teng, H., Hu, B., Zhuo, Z., Ji, W., Huang, Y., Gou, Y., Richer-de-Forges, A. C., Arrouays, D., & Shi, Z. (2023). Improving model parsimony and accuracy by modified greedy feature selection in digital soil mapping. *Geoderma*, 432, 116383. <https://doi.org/10.1016/J.GEODERMA.2023.116383>

Samdomojo darbo pajamų praleistų reikšmių įrašymas taikant daugiareikšmį įrašymą

Guoda Puslytė, Rūta Levulienė

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Taikomosios matematikos institutas, Naugarduko g. 24, Vilnius
guoda.puslyte@mif.stud.vu.lt

Santrauka. Analizuojant apklausų duomenis, dažnai tenka spręsti neužpildytų reikšmių problemą. Įprastai respondentai vengia atsakyti į klausimus, kurie yra susiję su asmenine ar finansine informacija. Šiame darbe palyginami keturi daugiareikšmio įrašymo metodai. Rezultatai atskleidžia, kad atsitiktinių miškų bei klasifikavimo ir regresijos medžių metodai užtikrina mažiausią paklaidą.

Raktiniai žodžiai: Praleistos reikšmės, daugiareikšmis įrašymas, samdomojo darbo pajamos.

1 Įvadas

Praleistos reikšmės yra dažna problema, su kuria susiduriama analizuojant apklausų metu surinktus duomenis. Dėl įvairių priežasčių respondentai gali neatsakyti į tam tikrus klausimus arba pateikti netikslią informaciją, ypač kai klausimai susiję su jautriomis temomis, pavyzdžiui, pajamomis. Tiksli informacija apie gyventojų pajamas yra svarbi norint įvertinti skurdo lygį šalyje, ekonominius skirtumus tarp valstybių bei valstybės institucijoms priimant politinius sprendimus.

Pastaruoju metu, siekiant tiksliau užpildyti praleistas reikšmes, vis plačiau taikomi mašininio mokymosi metodai, tačiau dauguma tyrimų ir metodų apsiriboja vieno tipo kintamaisiais, dažniausiai – kiekybiniais [1, 2]. Vis dėlto, apklausų duomenyse vyrauja kategoriniai kintamieji. Vienas iš nedaugelio metodų, kuris yra tinkamas tiek kategorinių, tiek kiekybinių kintamųjų reikšmių įrašymui yra daugiareikšmis įrašymas [3, 4].

Šio tyrimo tikslas – įvertinti, kaip praleistų reikšmių užpildymas taikant daugiareikšmį įrašymą veikia samdomojo darbo pajamų įrašymo tikslumą bei parinkti tinkamiausią įrašymo metodą.

2 Duomenys

Šiame tyrime naudojami viešai prieinami duomenys iš 2023 m. metinio pajamų ir gyvenimo sąlygų statistinio tyrimo [5]. Analizei buvo atrinkti 28 kintamieji, suteikiantys informaciją apie respondento lytį, amžių, gyvenamąją vietą, šeimyninę padėtį, išsilavinimą, sveikatos būklę, užimtumą, darbo patirtį ir laisvalaikio įpročius, iš kurių 22 – kategoriniai ir 6 – kiekybiniai. Iš viso 16 kintamųjų turėjo praleistų reikšmių, kurių dalis sudarė nuo 0,04 iki 27 procentų. Pašalinus nežinomas reikšmes, galutinėje duomenų aibėje liko 2378 stebėjimai.

3 Metodologija

Pagrindinis tikslas yra palyginti daugiareikšmio įrašymo metodus ir pateikti rekomendacijas, todėl buvo paimtas pilnas, be praleistų reikšmių, duomenų rinkinys, pašalinta dalis stebėtų reikšmių, atliktas įrašymas ir palyginti rezultatai.

Kadangi didžioji dalis tiriamojo duomenų rinkinio kintamųjų yra kategoriniai, nebuvo galimybės taikyti „Little’s MCAR“ testo, skirto nustatyti, ar nežinomos reikšmės yra praleistos visiškai atsitiktinai (angl. Missing Completely At Random), todėl šiame tyrime daroma prielaida, kad nežinomos reikšmės yra praleistos atsitiktinai (angl. Missing At Random), t. y., praleistos reikšmės priklauso nuo kitų kintamųjų stebėtų reikšmių ir nepriklauso nuo nežinomų reikšmių [6, 7].

Galutinei duomenų aibei buvo sugeneruotos sintetinės nežinomos reikšmės taip, kad jų pasiskirstymas kuo tiksliau atspindėtų originalių duomenų nežinomų reikšmių pasiskirstymą bei nežinomų reikšmių dalis kiekvienam kintamajam išliktų tokia pati. Dirbtinėms praleistoms reikšmėms sugeneruoti buvo naudojama „R“ paketo „missMethods“ funkcija `delete_MAR_1_to_x()`. Ši funkcija padalina duomenų rinkinį į dvi grupes pagal parametru `cols_ctrl` nurodyto kintamojo reikšmių medianą ir sugeneruoja nežinomas reikšmes grupėse parametru `cols_mis` nurodytiems kintamiesiems santykiu 1:x (šiuo atveju $x = 5$). Be to, parametru `p` nurodoma, kokią duomenų dalį sudarys praleistos reikšmės kiekviename iš pasirinktų kintamųjų [7]. Geriausiai pavyko atkurti nežinomų reikšmių kombinacijas, kurios originalių duomenų rinkinyje pasikartoja dažniausiai, iš viso buvo atkurta apie 65% kombinacijų. Siekiant įvertinti įrašytų reikšmių poveikį samdomojo darbo pajamų įrašymo tikslumui, buvo sukurta antra duomenų aibė, kurioje nežinomos reikšmės buvo sugeneruotos tik samdomojo darbo pajamų kintamajam.

Daugiareikšmis įrašymas atliktas naudojant „R“ paketo „mice“ [4] funkciją *mice()* su parametru $m = 5$, t. y., kiekviena praleista reikšmė yra įrašoma penkis kartus, o galutinei reikšmei apskaičiuoti naudotas įrašytų reikšmių vidurkis. Praleistų reikšmių įrašymas pradedamas nuo mažiausiai praleistų reikšmių turinčio kintamojo. Daugiareikšmis įrašymas buvo atliekamas keturiais metodais, kurie yra tinkami tiek kiekybiniais, tiek kategoriniams kintamiesiems.

Vienas iš plačiausiai taikomų daugiareikšmio įrašymo metodų yra prognozuojamo vidurkio atitikimo metodas (PVA, angl. predictive mean matching). Šio metodo esmė – pasitelkiant regresijos modelį, sudarytą remiantis pilnais duomenimis, apskaičiuojama prognozė kiekvienam nežinomą reikšmę turinčiam stebėjimui. Tuomet kiekvienai prognozuotai reikšmei randamos kelios, šiuo atveju – 5, artimiausios stebėtos reikšmės, iš kurių atsitiktinai parenkama viena ir priskiriama praleistą reikšmę turėjusiam įrašui.

Atsitiktinės imties iš stebėtų reikšmių (AISR, angl. random sample from observed value) atvejų praleistų stebėjimų įrašymui yra naudojama stebėtų to paties kintamojo reikšmių aibė, iš kurios atsitiktinai parenkama imtis ir įrašoma vietoje praleistų reikšmių.

Klasifikavimo ir regresijos medžių (KRM, angl. classification and regression trees) metodas veikia panašiai kaip PVA tačiau vietoje regresijos modelio yra sudaromas sprendimų medis. Praleistą reikšmę turintis stebėjimas yra priskiriamas vienam iš sprendimo medžio lapų (angl. terminal node), tuomet iš lapui priklausančių stebėtų reikšmių atsitiktinai išrenkama viena ir įrašoma.

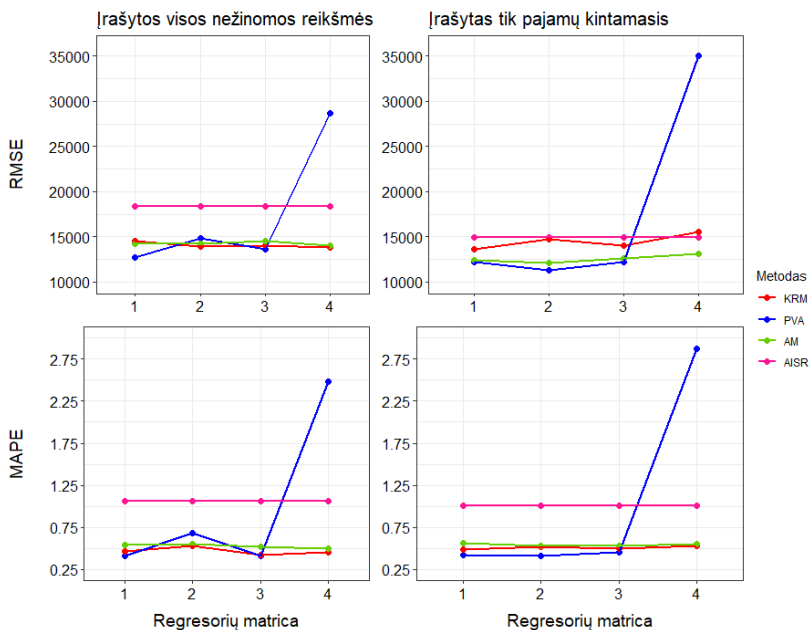
Atsitiktinių miškų (AM, angl. random forest) metodas praleistų reikšmių įrašymui taiko atsitiktinių miškų modelį, sudaryta iš 10 sprendimo medžių. Praleistų reikšmių įrašymui kiekybiniais kintamiesiems naudojamas sprendimo medžių prognozių vidurkis, o kategoriniams kintamiesiems – dažniausiai pasikartojanti prognozė.

Be to, siekiant įvertinti metodų jautrumą regresorių parinkimui, kiekvienam metodui buvo naudojamos keturios skirtingos regresorių matricos. Trys iš jų buvo sukurtos naudojant „mice“ paketo funkciją *quickpred()*, kuri leidžia greitai atrinkti regresorius kintamiesiems su praleistomis reikšmėmis, atsižvelgiant į kintamųjų tarpusavio koreliaciją. Šios regresorių matricos buvo sudarytos remiantis Spirmeno, Kendalo ir Pirsono koreliacijos koeficientais. Nors šis regresorių parinkimo metodas nėra pritaikytas kategoriniams kintamiesiems, šiame tyrime jis buvo naudojamas įvertinti daugiareikšmio įrašymo metodų jautrumą skirtingoms regresorių kombinacijoms, o ne parinkti optimalius regresorius kiekvienam kintamajam. Ketvirtojoje regresorių matricoje kiekvienam kintamajam kaip regresoriai buvo priskirti visi likę kintamieji.

Rezultatų palyginimui naudojamos trys metrikos: šaknis iš vidutinės kvadratinės paklaidos (RMSE) ir vidutinė absoliutinė procentinė paklaida (MAPE).

4 Rezultatai

Rezultatai pateikti 1 pav. Abiejų duomenų aibių atveju gauti panašūs rezultatai, kas rodo, jog praleistų reikšmių įrašymas neblogina samdomojo darbo pajamų įrašymo tikslumo. Prasčiausi rezultatai gauti taikant PVA metodą kartu su ketvirtąja regresorių matrica, kur kiekvieno kintamojo praliestoms reikšmėms įrašyti kaip regresoriai yra naudojami visi likę kintamieji, kas leidžia teigti, jog naudojant šį metodą svarbu tinkamai atrinkti regresorius kiekvienam praleistų reikšmių turinčiam kintamajam. Stabiliausi ir patikimiausi rezultatai gauti taikant AM ir KRM metodus. AISR atveju, visiems regresorių matricos atvejams gauta vienoda paklaida, taip yra todėl, nes šiuo metodu įrašymo reikšmės nepriklauso nuo kitų kintamųjų reikšmių ir priklauso tik nuo pačio įrašomo kintamojo nepraleistų reikšmių aibės, vis dėlto šio metodo atveju įrašomos reikšmės yra parenkamos atsitiktinai, todėl įrašytų reikšmių paklaida yra didelė.



1 pav. RMSE ir MAPE rodiklių palyginimas skirtingiems praleistų reikšmių įrašymo metodams.

5 Išvados

Atsižvelgiant į gautus rezultatus, galima teigti, kad praleistų reikšmių užpildymas taikant daugiareikšmį įrašymą tinkamas samdomojo darbo pajamų įrašymui. Geriausi rezultatai gauti taikant AM ir KRM metodus, kadangi įrašytų reikšmių paklaida nepriklausomai nuo regresorių matricos ir duomenų aibės išliko panaši visais atvejais. Panaši paklaida buvo gauta ir PVA atveju, tačiau taikant šį metodą svarbu tinkamai parinkti regresorius kiekvienam praleistų reikšmių turinčiam kintamajam.

Literatūra

- [1] Md. Kamrul Hasan, Md. Ashraful Alam, Shidhartho Roy, Aishwariya Dutta, Md. Tasnim Jawad, Sunanda Das. (2021). Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). *Informatics in Medicine Unlocked*, Volume 27, 100799.
- [2] Lin, WC., Tsai, CF. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). *Artif Intell Rev* 53, 1487–1509.
- [3] Van Buuren, S., & Oudshoorn, K. (1999). Flexible multivariate imputation by MICE. (pp. 1-20). Leiden: TNO.
- [4] van Buuren, S. (2018). *Flexible Imputation of Missing Data, Second Edition* (2nd ed.). Chapman and Hall/CRC.
- [5] Valstybės duomenų agentūra. 2023 m. atlikto metinio pajamų ir gyvenimo sąlygų statistinio tyrimo asmenų duomenys.
- [6] Roderick J. A. Little. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*, 83(404), 1198–1202.
- [7] Santos, M. S., Pereira, R. C., Costa, A. F., Soares, J. P., Santos, J., & Abreu, P. H. (2019). Generating Synthetic Missing Data: A Review by Missing Mechanism. *IEEE Access*, 7, 11651-11667

Recognising the contents in digitised financial documents

Simas Rimašauskas, Igoris Belovas

Vilnius University, Faculty of Mathematics and Informatics,
Didlaukio g. 47, Vilnius, Lithuania
simas.rimasauskas@mif.stud.vu.lt

Abstract. The necessity of content recognition in digital documents is ever-increasing in the financial sector. Extracted data is used for fundamental analysis, modelling and portfolio selection. In the most prominent markets, there is a wide array of available sources to obtain the data, such as SEC filings easily. However, it is not so in markets with less investor interest, such as the CEE region or Latin America. Often, the only sources containing the data are primary reports by the company itself. Scarce secondary sources may provide data of dubious reliability. This leads to an excessive workload for analysts, implying the necessity to adapt existing intelligent methods for processing financial data.

Keywords: machine learning, natural language processing, optical character recognition, text recognition, table recognition.

1 Introduction

Smaller, less popular markets often suffer from a scarcity of secondary data sources. Financial analysts focusing on these regions, in turn, depend heavily on primary documents issued by companies. Such reliance leads to a workload bottleneck due to the manual, time-intensive process of extracting valuable data from unstructured or poorly formatted documents. Implementing automatized, intelligent methods to recognize and extract contents could mitigate this problem. Hence, we sought to thoroughly review the underlying algorithms of select machine learning (ML) and natural language processing (NLP) technologies and apply them in experiments conducted on a real-life dataset comprised of quarterly and annual reports from firms operating in exotic markets.

2 Historical development of content recognition

Various approaches for detecting tables in images were devised throughout the years - Dengel et al. [1] proposed a method to cluster word boundaries

into a segmentation graph for table detection. However, that failed in the face of multi-column layouts. Wangt et al. [2] proposed using the distance between consecutive words as a heuristic in determining table entity candidates. However, such a method is inevitably tied to a specific layout template, which hinders its practical application. Gatos et al. [3] proposed using the intersection area between horizontal and vertical lines to reconstruct the intersectional pairs. However, this system was said to overly rely on the visual cues provided by strict, defining table borders. Ultimately, deep learning methods were implemented when Hao et al. [4] proposed a regional proposal network using CNNs, which later adapted the faster R-CNN architecture to segment table regions in a given image. Building on this, Watson and Liu proposed a table alignment process involving organizing disjoint text segments into columns by creating disjoint sets and merging them using an algorithm.

A research survey conducted in a paper by Kasem et al. [5] has established that ML and deep learning technologies have been effective in computer vision research tasks such as object detection and object position estimation. CNN, one of the most widely-used types of neural networks, can use the fundamental properties of actual signals, such as translation invariance and compositional hierarchies. A typical CNN comprises a hierarchical structure and numerous layers for learning data representations at different levels of abstraction [6]. The efficacy of CNNs in object identification, the researchers assert, is based on the ability to, through training on large amounts of data, learn substantial mid-level visual properties, which are more complex patterns such as textures and shapes, rather than hand-crafted low-level representations, such as edge detectors, often used in image categorization. Deep learning models are now widely used in multiple areas, including general table detection [4]. On the other hand, table structures receive far less attention, and the table structure is typically only characterized by the rows and columns of a table [7].

3 Summary of state-of-the-art algorithms

In this section, we survey modern NLP and ML technologies and review and describe the subtleties and principles behind their algorithms. Owing to their widespread use in practice, the tools selected for the work are LayoutLM, OpenCV, Camelot, pdfplumber, and Google Document AI.

In their 2020 work, Xu et al. [8] proposed LayoutLM, a pre-training technique with text and layout information in a single framework that utilizes Transformer architecture as the backbone. It is inspired by the BERT model, which represents text using embeddings. According to the researchers, LayoutLM extends the principle by using multimodal inputs, including token, layout, and image embeddings. For the experiments, the LayoutLMv3 version was used. The model works by processing text using an OCR toolkit to extract content and 2D positional information. Word embeddings are then initialized using a pre-trained RoBERTa model. The layout positions are applied at the segment level, where words in the same segment share the same 2D position, as opposed to word-level positions used in previous versions like LayoutLM and LayoutLMv2. Image embeddings are generated by applying linear projections to image patches, which helps to reduce computational complexity compared to previous versions that used CNNs, like Faster R-CNN. Images are then resized, split into patches, and flattened into a sequence of vectors. Semantic 1D relative position and spatial 2D relative position are introduced as bias terms in self-attention networks for both text and image modalities, which follows the previous approach in LayoutLMv2.

Another tool used for data extraction in financial documents is OpenCV, a real-time computer vision programming function library initially launched by Intel in 1999 [9]. The tool provides numerous image manipulation, analysis, and data extraction methods. Mat class serves as the main data structure in OpenCV for the storage and manipulation of images and matrices. It holds pixel values and handles memory management by automatically allocating memory as necessary. It also supports reference counting, allowing efficient memory use by preventing redundant copying of image data when objects are passed by reference. Histograms are commonly used for image processing, representing the distribution of pixel intensity in an image. Additionally, OpenCV supports techniques such as Fourier transform to analyse the frequency content of an image. By decomposing an image into its frequency components, algorithms can be designed to filter out unwanted noise or enhance certain features. Hough transform is also implemented in OpenCV and is used to detect geometric shapes in an image. OpenCV also provides Edge detection as another fundamental image processing technique - a canny edge detector combines gradient calculation with non-maximum suppression to detect edges accurately while reducing noise.

Pdfplumber is an open-source Python library created to extract information from PDF documents. It was created by Singer-Vine and was initially released in 2016 [10]. Pdfplumber builds on top of pdfminer, another PDF processing library. The approach is heavily inspired by Nurminen's master's thesis [11]. In essence, for any PDF page, either actual lines, which are explicitly defined and give the table a grid-like structure, or invisible lines implied by word alignment are found. Overlapping or nearly-overlapping lines are merged, and intersections of all these lines are found. The most granular set of rectangles, which use the intersections as vertices, are found, and contiguous cells are then grouped into tables. Pdfplumber recognizes lines in a PDF by parsing the document's raw content streams to identify vector graphics commands corresponding to line-drawing operations. These instructions are then decoded by extracting the coordinates, orientation, and thickness of lines explicitly defined in the document. The extracted line objects are then interpreted, and the positioning of the text element is examined further to infer implied lines where no explicit borders are present. Text alignment and spacing are analysed to identify patterns that might imply grid structures.

Camelot, a Python library for data extraction from PDFs, was created by Vinayak Mehta [12]. Camelot offers two table parsing methods, Stream and Lattice, each designed to accommodate different PDF table structures. The Stream method, best used for tables that do not have visible borders but instead rely on whitespace between cells to create a table structure, works by grouping words on the PDF page into rows based on their y-axis overlaps. It then calculates text edges to identify areas that could contain tables. The number of columns is estimated by calculating the mode of the word count in text rows, thus determining column ranges. The words are then assigned to columns based on whether they fall within the calculated x and y ranges. The lattice method is more suitable for tables with visible cell borders, relying on precise line detection and geometric calculations. It works by converting the page into an image using Ghostscript. The algorithm then detects horizontal and vertical line segments by applying transformations such as erosion and dilation. The detected lines are processed to identify intersections, marking the boundaries of table cells. They are scaled and translated back to the file's coordinate space to create a representation of the table. Spanning cells are then identified, and the words are assigned to the correct table cells based on their x and y coordinates.

The Document AI platform is a unified document processing console launched by Google Cloud, the cloud computing division of Google, in November 2020 [13]. The platform aims to automate and validate documents by providing access to document parsers, tools, and solutions via an API. Because of the proprietary nature of the application, the table extraction algorithms used by Google Document AI are not disclosed to the public and, therefore, were not reviewed in our work.

The previously described table data extraction methods differ in some ways yet have similarities. Regarding the data extraction approach, pdfplumber and Camelot are similar because both libraries do rule-based parsing, where predefined rules are used to identify and extract specific structures or patterns from documents. OpenCV operates utilizing the approach of low-level image processing, while LayoutLM uses a multimodal transformer architecture that integrates text and visual embeddings for document understanding. LayoutLMv3 and Google Document AI are better suited for complex documents with rich layouts and can handle visual and textual data. Camelot and pdfplumber rely on line detection for grid-based structures, although Camelot's Stream method is optimized for whitespace-based tables, while Lattice is tailored for bordered table layouts. OpenCV uses pixel-level manipulation (such as detection of edges or contour analysis) to identify lines or shapes.

4 Experiments

An original dataset consisting of 7 selected PDF format financial documents published by different companies in Central and Eastern Europe and Latin America was compiled and used for our experiments. The files were quarterly or annual reports published in financial periods from 2018 to 2022. The eight selected tables consisted of the main parts of financial statements (statements of income and loss, financial position, and cash flows), notes, and appendices and were in English. Using a Python API, a ground truth version of the tables and their text content was extracted and reviewed manually.

In order to test the selected tools, five scripts were written in Python, each using a different tool to recognize text and extract tables from PDF pages into a spreadsheet. For LayoutLMv3, the FUNSD dataset was used for the training process, which focuses on form understanding [14]. However,

the model cannot perform text recognition, so the Tesseract OCR engine was additionally integrated, which renders high-resolution images of PDF format file pages and performs optical character recognition (OCR) to detect text and bounding boxes [15]. This consideration also applies to OpenCV. Since it does not perform OCR, Tesseract was also used to extract text and bounding box information from the document images in the OpenCV implementation. Custom functions were developed to merge extracted text boxes when they were sufficiently close in terms of coordinates, improving the coherence of detected text blocks. For Google Document AI, a Form Parser processor was selected to process the document, which was developed for the extraction of tables and is best suited for table data recognition and extraction [16]. Pdfplumber, a library developed explicitly for text extraction from searchable .pdf files and table extraction, offered ample table extraction settings. Because most tables in the dataset were borderless, we used vertical and horizontal strategy values of "text" for the pdfplumber function *extract_tables()*, which allowed the library to aim specifically to identify columns in borderless tables. In the implementation of Camelot, another Python library specifically made to extract tables, there was also little need for pre-processing as the PDF files were simply processed using a library function *read_pdf()*. The previously discussed Stream method, which is more suitable for tables without explicit borders, was chosen. Finally, algorithms were created to evaluate the tables using the defined metrics.

To assess the accuracy of the selected tools, three metrics were defined:

1. **Symbol accuracy** is defined as correctly output characters divided by all output symbols. It was evaluated by comparing the lists of characters for each cell to those from the ground truth.

$$\text{Symbol Accuracy} = \frac{\text{Number of Correctly Output Symbols}}{\text{Total Number of Output Symbols}}$$

2. **Word accuracy** is defined as the proportion of correctly output words (which we define as ordered lists of characters divided by blank characters) to the total number of output words in the table. This and the symbol accuracy metrics are based on the metrics used by Rupšys in his bachelor thesis [17]

$$\text{Word Accuracy} = \frac{\text{Number of Correctly Output Words}}{\text{Total Number of Output Words}}$$

- Table accuracy** metric to evaluate the **structural integrity** of tables has been established based on the content accuracy metric used by Smock et al. [18]. By calculating the share of correct cells, we can deduce whether the output table retained the correct structure of the original.

$$\text{Table Accuracy} = \frac{\text{Number of Cells in Correct Relative Position}}{\text{Total Number of Cells}}$$

It was found that, regarding the accuracy of the table structure, Camelot strongly outperformed the peer group with a simple average accuracy of 90.0%, as seen in table 1. Google Document AI also provided tables that may be considered satisfactory in terms of structure at 72.8%. LayoutLM resulted in the least accurate table structure at 34.3%. In terms of symbol accuracy seen in table 2, Camelot was strongest among the peer group as well, with a simple average accuracy of 93.0%; Google Document AI also outputs sufficiently accurate tables in terms of symbols at 81.3%. LayoutLM resulted in the least accurate table structure at 23.3%. Lastly, as seen in table 3, the word accuracy is highly correlated with the symbol metric: Camelot performed best in terms of correct words with a 92.1% accuracy. Tables obtained using LayoutLM, which was combined with Tesseract for optical character recognition, provided the lowest average word accuracy at 21.0%.

Table 1. Percentages of identical cells (table accuracy).

Table ID	Total Cells	Camelot	Google	LayoutLM	OpenCV	Pdfplumber
1	165	97.0%	44.2%	57.6%	79.4%	91.8%
2	161	84.5%	87.0%	44.1%	17.4%	77.0%
3	136	98.5%	84.6%	13.3%	72.8%	32.5%
4	96	84.0%	59.4%	44.8%	48.0%	68.0%
5	96	93.8%	93.8%	53.1%	74.0%	5.8%
6	252	97.2%	89.3%	42.6%	24.6%	15.5%
7	168	88.7%	57.7%	14.9%	62.5%	48.8%
8	84	76.2%	66.7%	3.6%	51.2%	21.4%
Average	-	90.0%	72.8%	34.3%	53.7%	45.1%

Table 2. Percentages of correctly output symbols.

Table ID	Total Cells	Camelot	Google	LayoutLM	OpenCV	Pdfplumber
1	165	98.6%	58.9%	36.0%	91.8%	94.5%
2	161	83.9%	92.9%	38.7%	35.7%	71.3%
3	136	96.8%	89.1%	3.2%	79.7%	58.4%
4	96	86.8%	66.9%	20.3%	52.6%	52.4%
5	96	93.2%	87.1%	35.2%	86.1%	5.9%
6	252	98.0%	88.4%	22.5%	42.0%	7.6%
7	168	89.9%	74.0%	14.5%	67.4%	63.5%
8	84	96.5%	93.1%	16.3%	85.1%	40.6%
Average	-	93.0%	81.3%	23.3%	67.5%	49.3%

Table 3. Percentages of correctly output words.

Table ID	Total Cells	Camelot	Google	LayoutLM	OpenCV	Pdfplumber
1	165	97.4%	54.9%	39.4%	81.9%	93.3%
2	161	81.1%	90.6%	23.2%	31.3%	69.5%
3	136	96.3%	83.3%	0.5%	69.9%	45.4%
4	96	85.5%	65.9%	21.0%	50.7%	54.4%
5	96	91.2%	86.1%	38.7%	77.4%	0.0%
6	252	96.9%	86.9%	19.8%	22.8%	0.0%
7	168	92.3%	72.1%	12.4%	64.7%	58.1%
8	84	96.0%	88.5%	13.0%	80.0%	37.5%
Average	-	92.1%	78.5%	21.0%	59.8%	44.8%

5 Conclusion

The overview of the state-of-the-art literature on the topics of NLP and ML in table data extraction, taken together with our practical experiments, has led us to conclude that while deep learning and ML are widely used in table detection, the research on table structures, in particular, has yet to receive wide attention. Regarding the tools used in the experiment, the setup for pdfplumber, Camelot, and Google Document AI is relatively straightforward and quick. Camelot and pdfplumber were the easiest to implement in practice. In contrast, LayoutLM and OpenCV demand more effort and are

time-intensive because of requirements such as model training. In terms of table structure and symbol and word accuracy, Camelot performed the task of data extraction from tables the best. Along with Camelot, it was established that Google Document AI is also an acceptable venue for obtaining tables of satisfactory accuracy.

Literature

- [1] Dengel, A., Kieninger, T. (1998) A Paper-to-HTML Table Converting System. DAS98, Int'l Association for Pattern Recognition Workshop on Document Analysis Systems. Deutsches Forschungszentrum für Künstliche Intelligenz.
- [2] Wangt, Y., Phillipst, I. T., Haralick, R. (2001). Automatic table ground truth generation and a background-analysis-based table structure extraction method. In Proceedings of Sixth International Conference on Document Analysis and Recognition, Seattle, WA, USA, 2001, pp. 528-532. IEEE.
- [3] Gatos B., Danatsas, D., Pratikakis, I., Perantonis, S.J. (2005). Automatic table detection in document images. In Proceedings of the Third International Conference on Advances in Pattern Recognition - Volume Part I (Bath, UK) (ICAPR'05). Springer-Verlag.
- [4] Hao, L., Gao, L., Yi, X., Tang, Z. (2016). A Table Detection Method for PDF Documents Based on Convolutional Neural Networks. 2016 12th IAPR Workshop on Document Analysis Systems (DAS). IEEE.
- [5] Kasem, M., Abdallah, A., Berendeyev, A., Elkady, E., Abdalla, M., Mahmoud, M., Hamada, M., Nurseitov, D., Taj-Eddin, I. (2024). Deep Learning for Table Detection and Structure Recognition: A Survey. *ACM Computing Surveys*, 56(12). Association for Computing Machinery.
- [6] LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep Learning. *Nature*, 521(7553), pp. 436-444.
- [7] Kara, E., Traquair, M., Simsek, M., Kantarci, B., Khan, S. (2020). Holistic Design for Deep Learning-Based Discovery of Tabular Structures in Datasheet Images. *Engineering Applications of Artificial Intelligence*, 90(C).
- [8] Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M. (2020). LayoutLM: Pre-training of Text and Layout for Document Image Understanding. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20)*, pp. 1192-1200. ACM.
- [9] OpenCV. Introduction to OpenCV-Python Tutorials (2025). URL: https://docs.opencv.org/4.x/d0/de3/tutorial_py_intro.html.
- [10] Singer-Vine, J., & The pdfplumber contributors (2025). Pdfplumber (Version 0.11.6). URL: <https://github.com/jsvine/pdfplumber/>.
- [11] Nurminen, A. (2013). Algorithmic Extraction of Data in Tables in PDF Documents (Master's thesis). Tampere University of Technology.
- [12] Mehta, V. & The Camelot contributors (2025). Camelot (Version 1.0.0). URL: <https://github.com/camelot-dev/camelot/>.
- [13] Google (2020). Introducing Document AI platform, a unified console for document processing. URL: <https://cloud.google.com/blog/products/ai-machine-learning/google-cloud-announces-document-ai-platform/>.

- [14] Rogge, N. (2022). layoutlmv3-finetuned-funsd. URL: <https://huggingface.co/nielsr/layoutlmv3-finetuned/>.
- [15] Smith, R. (2007). An Overview of the Tesseract OCR Engine. ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition, pp. 629-633. IEEE Computer Society.
- [16] Google (2025). Process documents with Form Parser. URL: <https://cloud.google.com/document-ai/docs/form-parser/>.
- [17] Rupšys, J. (2021). Lentelių atpažinimas (Bachelor's thesis). Vilnius University.
- [18] Smock, B., Pesala, R., Abraham, R. (2022). PubTables-1M: Towards Comprehensive Table Extraction from Unstructured Documents. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Microsoft.

Comparative Evaluation of Speech-to-Text Models for Lithuanian Transcription: Effects of Audio Formats and Recording Environments

Dovydas Šablevičius, Asta Slotkienė

Vilniaus Gedimino technikos universitetas,
Saulėtekio al. 11, LT-10223 Vilnius
dovydas.sablevicius@stud.vilniustech.lt

Summary. This study evaluates the performance of various speech-to-text models for Lithuanian transcription, focusing on how audio formats and recording environments affect their accuracy. Among the models tested, Google's Chirp-2 demonstrated the highest accuracy under optimal conditions. However, its performance declined with increased playback speeds and in environments with significant background noise, highlighting the importance of controlled recording conditions for effective deployment of STT systems in real-world applications.

Keywords: Speech-to-text model, audio format, transcription accuracy, recording environment, Lithuanian language.

1 Introduction

In the past decade, artificial intelligence accelerated the application of speech-to-text (STT) and inspired the development of STT services that convert natural spoken language into text. STT is applicable to various areas where human interaction with computers and other digital devices is necessary, such as healthcare [1, 2, 3], education [4,5], etc. This technology enhances system services, improves content accessibility, and enables automated documentation without the need for human interpretation. However, a major challenge remains: most STT models have been predominantly developed and evaluated for English, which limits their effectiveness when applied to other languages, including Lithuanian [6]. In recent years, efforts have been made to develop increasingly accurate Lithuanian STT models despite limited training data [7]. Additionally, large-scale projects such as LIEPA-3 are creating extensive Lithuanian speech corpora to further improve

training resources for STT systems [8]. The objective of this research is twofold. First, it seeks to determine which existing STT model yields the highest transcription accuracy for the Lithuanian language. Second, it aims to analyze how various audio parameters—including technical specifications like sampling rate and bit depth, as well as environmental conditions such as room acoustics and ambient noise—influence transcription accuracy.

2 Research methodology

To describe the standardized methodology applied throughout this research is essential because this approach ensured consistent recording conditions and established a robust pipeline for transcription evaluation.

All recordings utilized an identical 226-word Lithuanian script focused on topics related to local nature and culture, including specific references to regional objects. A native Lithuanian speaker conducted every recording from a fixed distance of 60 centimeters from the microphone. Room noise levels were measured using the NIOSH Sound Level Meter App, which resulted in three distinct sets of recordings: one set consisted of three baseline recordings in a quiet room at approximately 30 dB, with durations of 2.02 min, 2.06 min, and 2.03 min. Another set comprised three recordings with “Coffee Shop Background Noise,” maintaining a constant ambient noise level of around 50 dB and their durations were 1.50 min, 1.55 min, and 2.02 min. The final set involved three recordings in a naturally echoey room, also at approximately 30 dB but with noticeable acoustic reverberations and their durations were 2.04 min, 1.04 min, and 2.01 min.

The baseline audio files then underwent additional processing based on experimental conditions, as illustrated below in Figure 1. In Test 1, audio files were exported in FLAC format at 48 kHz, with a 24-bit depth and a mono channel. Test 2 maintained these FLAC settings but varied the channel counts between mono and stereo. Tests 3 and 4 followed the configurations detailed in Figure 1, involving variations in sampling rates, bit depths, and audio compression formats. Meanwhile, Tests 5 and 6 utilized FLAC files at 16 kHz and 16-bit depth (mono), which were modified to assess playback speed variations along with additional environmental factors such as echo and background noise.

Controlled parameters—such as the speaker-to-microphone distance and ambient noise measurements—were essential to ensure that any observed differences in transcription accuracy could be attributed solely to the tested experimental variables.

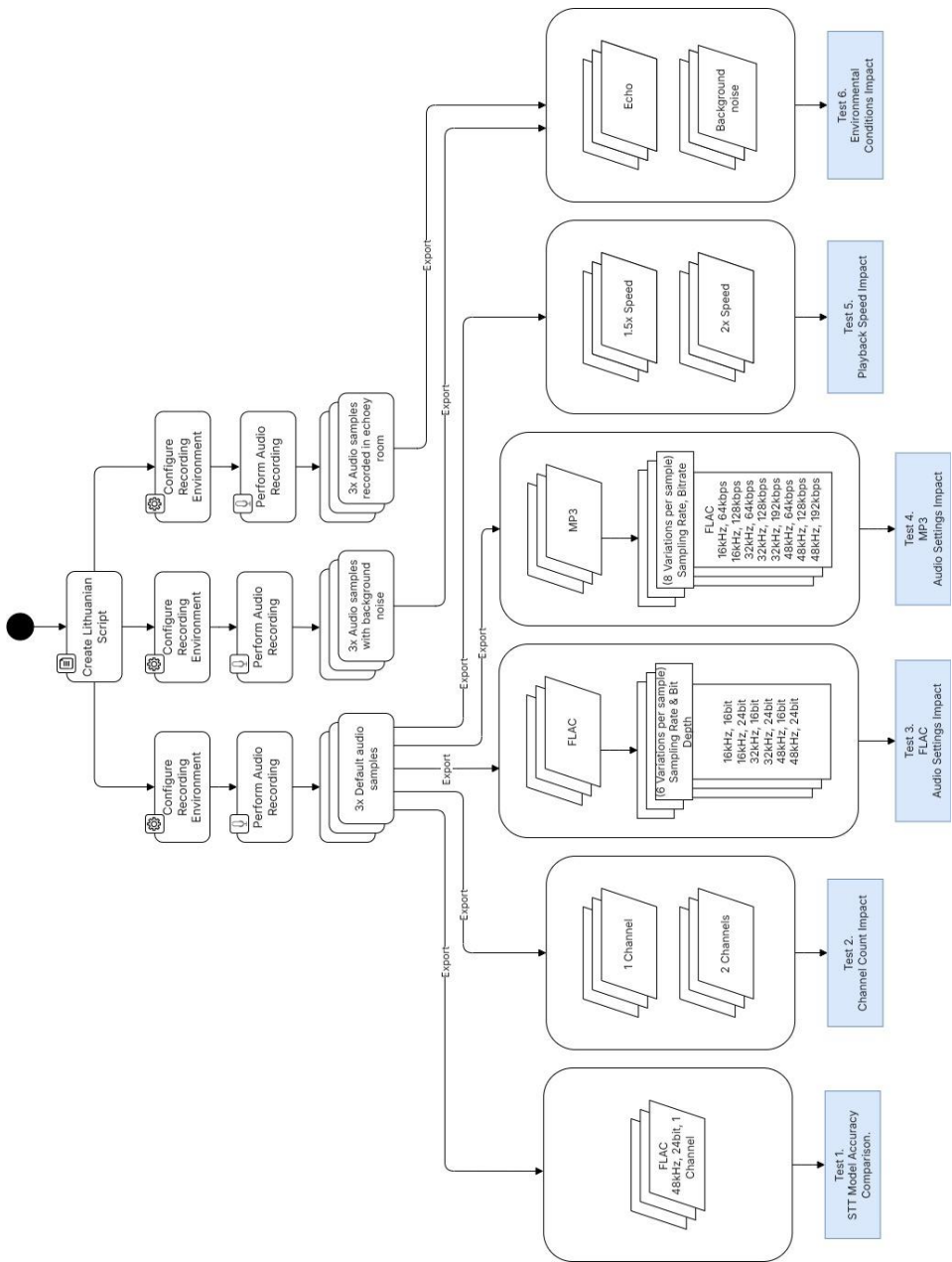


Figure 1. Audio sample preparation process.

Following the generation of the audio samples, all recordings were processed through a uniform transcription pipeline, as outlined below in Figure 2. In this phase, each audio file was asynchronously submitted to the selected STT models operating in their default configurations. The resulting transcripts, along with a reference transcript, underwent a standardized normalization process that included:

1. Converting all text to lowercase
2. Removing punctuation
3. Eliminating extra spaces (leading, trailing, or redundant)

Accuracy metrics were calculated using the JiWER Python library, comparing the normalized model outputs against the normalized reference transcripts. The following metrics were recorded: Word Error Rate (WER), Character Error Rate (CER), and Real-Time Factor (RTF). These metrics calculation and interpretation are detailed in section 3.

Figure 2 encapsulates this transcription and evaluation workflow, underscoring the methodical approach taken to process each audio input and compute the metrics.

Figure 2 outlines this transcription and evaluation workflow, emphasizing the systematic approach taken to accurately assess the impact of different audio characteristics on transcription performance. Overall, the described methodology ensures reliable and comparable results across varied testing conditions, thereby enabling clear insights into how specific audio attributes influence the accuracy and efficiency of STT models.

3 Metrics for Speech-based Models

In this investigation, the results were validated with a wide range of evaluation metrics, such as acoustic and text-based metrics, which allows to compare different STT models. One of them is word error rate (WER) is calculated using Levenshtein distance between word [9]. In Levenshtein distance, we count the number of insertions (I), substitutions (S) and deletions (D) performed to equal two-word sequences. WER is calculated by this formula present in (1). Lower values of WER are preferred since they indicate an STT model that makes fewer errors.

$$WER = \frac{S+D+I}{N} \quad (1)$$

Where S = the number of substitutions required to change the hypothesis string to the reference string, D = the number of necessary deletions, I = the

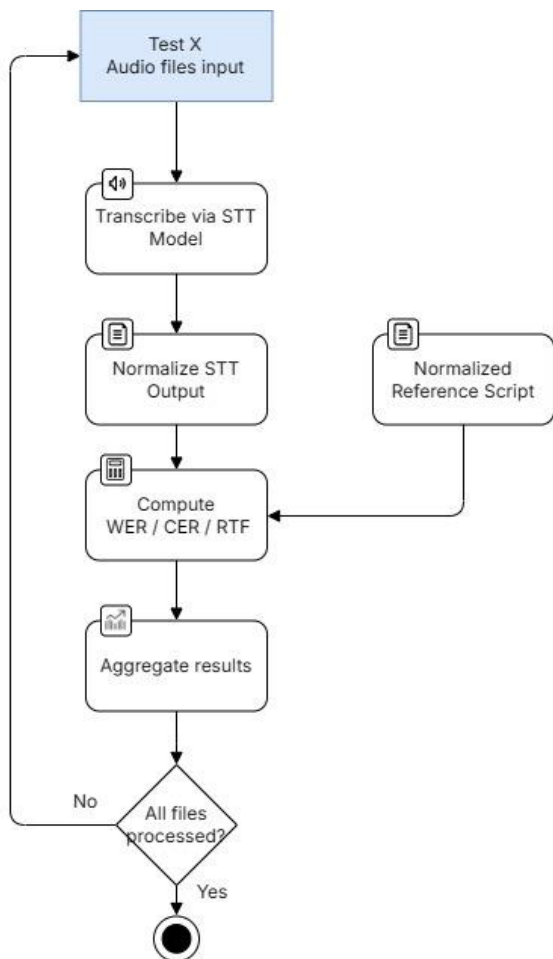


Figure 2. Audio transcription and evaluation pipeline.

number of insertions, and N = the total number of words in the reference string [10]. Other metrics that complement the widely used WER to evaluate the performance of STT model are character error rate (CER). CER measures the edit distance between the recognized text and the reference text at the character level, where substitutions (S), deletions (D), and insertions (I),

respectively, at the character level [11]. The CER is calculated as (2) formula, and with lower values indicating better performance of STT model.

$$CER = \frac{S+D+I}{N} \quad (2)$$

In addition to the accuracy indicator, the real time factor (RTF) is another important performance indicator and it applied to measure the speed of a system that processes an input audio signal. The RTF is the ratio of the time taken to process a speech to the duration of that speech, and the lower the value is, the better the real-time performance of text extraction. The RTF can be defined as:

$$RTF = \frac{T}{D} \quad (3)$$

Where T is time to transcribe the audio file and D is duration of the audio file [8]. Values of $RTF < 1.0$ are preferred since values ≥ 1.0 indicate that the decoding (transcribing) an audio file takes a larger amount of time than the duration of the audio itself.

4 Analysis of Experimental Results

This section provides an in-depth analysis of the experimental results, beginning with the identification of the optimal STT model for Lithuanian transcription and continuing with a detailed examination of how various audio configurations and environmental conditions affect transcription accuracy.

Test 1 identified the best STT Model for the Lithuanian language.

The goal of this test was to determine which of the five selected STT models provides the highest transcription accuracy for the Lithuanian language. By comparing performance across these models, the optimal candidate was chosen to serve as the baseline for subsequent experiments involving modifications in audio quality. The selected models were based on their explicit support for Lithuanian language. The models evaluated were:

1. OpenAI: GPT-4o-transcribe
2. Google: Chirp-2
3. Amazon Web Services: unknown
4. Microsoft: Whisper Large V2
5. Tilde: unknown

The selection criteria emphasized the providers' market scale and reputation, as well as Tilde's specific focus on Baltic region languages.

The results presented in Table 1 reflect the average performance across the three audio recordings.

Table 1. STT Model Accuracy Comparison.

Provider	Model	WER	CER	RTF
OpenAI	GPT-4o-transcribe	0.196	0.053	0.083
Google	Chirp-2	0.093	0.033	0.086
Amazon Web Services	Unknown	0.133	0.044	0.112
Microsoft	Whisper Large V2	0.342	0.217	0.519
Tilde	Unknown	0.323	0.211	0.591

The results indicate that Google's Chirp-2 outperforms the other models, achieving the lowest WER (0.093) and CER (0.033). Although OpenAI's GPT-4o-transcribe demonstrated a slightly lower RTF (0.083) compared to Chirp-2 (0.086), its error rates were considerably higher, making it less favorable for accurate transcription. The unnamed models from Amazon Web Services, Microsoft, and Tilde showed progressively poorer performance with higher error rates and, in the case of Microsoft and Tilde, substantially higher RTFs, suggesting slower processing times unsuitable for efficient transcription tasks.

In conclusion, the evaluation establishes Google's Chirp-2 as the optimal STT model for Lithuanian transcription under controlled conditions. Its superior accuracy, combined with competitive processing speed, renders it the preferred choice for further experiments involving the impact of audio quality modifications and environmental factors.

For Tests 2 through 5, the same set of audio recordings and normalization procedures as described in Test 1 were employed. In these tests, only the performance of Google's Chirp-2 was evaluated, and transcription accuracy was quantified solely using Word Error Rate (WER) and Character Error Rate (CER). Each result represents the average performance across the three audio recordings.

Test 2 investigated whether the number of audio channels influences transcription accuracy. The recordings were processed in both mono (one channel) and stereo (two channels) configurations using Chirp-2 model. The results, summarized in Table 2 below, indicate that there is no discernible

difference in performance between the two configurations. Both mono and stereo recordings yielded identical error rates (WER of 0.093 and CER of 0.033), suggesting that mono recordings are sufficient for accurate transcription while offering advantages in terms of file size.

Table 2. Channel Count Impact.

STT model	Channel count	WER	CER
Chirp-2	1	0.093	0.033
	2	0.093	0.033

Test 3 examined the impact of varying FLAC audio settings on transcription accuracy. In this test, the FLAC files were exported using different combinations of sampling rates (16kHz, 32kHz, and 48kHz) and bit depths (16-bit and 24-bit). Table 4 shows that the optimal performance was achieved with both 16kHz/16-bit and 16kHz/24-bit configurations, each resulting in a WER of 0.088 and a CER of 0.031. Since lower sampling rates and bit depths produce smaller file sizes without compromising accuracy, the 16kHz, 16-bit configuration was selected as the optimal setting for subsequent evaluations.

Table 3. FLAC Audio Settings Impact.

STT model	Sampling Rate & Bit Depth	WER	CER
Chirp-2	16kHz, 16bit	0.088	0.031
	16kHz, 24bit	0.088	0.031
	32kHz, 16bit	0.089	0.032
	32kHz, 24bit	0.092	0.033
	48kHz, 16bit	0.092	0.033
	48kHz, 24bit	0.092	0.033

Test 4 focused on the effect of MP3 compression settings on transcription accuracy. The original recordings were converted to MP3 format under various configurations, varying both sampling rates (16kHz, 32kHz, and 48kHz) and bitrates (64, 128, and 192 kbps), while maintaining mono audio. Table 3 presents the average WER and CER for each configuration. The findings demonstrate that transcription accuracy remains relatively stable across the different MP3 settings. With an average WER of approximately

0.091 and a CER of about 0.032, these results imply that MP3 compression does not significantly impair the performance of Chirp-2, thereby providing flexibility in the choice of audio compression without a major impact on accuracy. Note: The configuration 16kHz, 192kbps is not included in the table because Audacity did not support exporting audio with this specific combination.

Table 4. MP3 Compression Settings Impact.

STT model	Sampling Rate & Bitrate	WER	CER
Chirp-2	16kHz, 64kbps	0.088	0.032
	16kHz, 128kbps	0.089	0.032
	16kHz, 192kbps	-	-
	32kHz, 64kbps	0.093	0.034
	32kHz, 128kbps	0.093	0.034
	32kHz, 192kbps	0.093	0.033
	48kHz, 64kbps	0.091	0.032
	48kHz, 128kbps	0.092	0.032
	48kHz, 192kbps	0.092	0.032

Test 5 assessed the effect of altering the playback speed of the recordings on transcription accuracy. Under baseline conditions—using the normal playback speed with optimal FLAC settings (16kHz, 16bit)—the transcription achieved a Word Error Rate (WER) of 0.088 and a Character Error Rate (CER) of 0.032. The audio samples were then artificially accelerated to 1.5x and 2.0x their original speed while retaining these settings. As shown in Table 5, at a 1.5x speed-up, the WER increased to 0.140 and the CER to 0.058, corresponding to an approximate increase of 59% in WER and 81% in CER relative to the baseline. When the playback speed was doubled to 2.0x, transcription performance deteriorated even further, with the WER rising to 0.286—a 225% increase over baseline—and the CER climbing to 0.148, representing an increase of approximately 362.5%. These results underscore the model’s sensitivity to changes in speech rate, confirming that acceleration negatively impacts transcription accuracy compared to the normal playback regime.

Table 5. Playback Speed Impact.

STT model	Speed	WER	CER
Chirp-2	1.5x	0.140	0.058
	2.0x	0.286	0.148

Test 6 evaluated the impact of environmental factors on transcription accuracy. Recordings were conducted under two distinct conditions: one in an echoey room selected for its natural acoustic reverberations, and another in a setting where “Coffee Shop Background Noise” was played to create a constant ambient noise level of approximately 50 dB. These conditions were compared against the baseline performance achieved under optimal recording conditions (WER of 0.088 and CER of 0.031).

As depicted in Table 6, under the echo condition, the transcription accuracy decreased slightly, with the WER increasing to 0.105 and the CER to 0.035—representing approximate increases of 19% and 13%, respectively, relative to baseline. In contrast, the background noise condition had a more pronounced effect, with the WER rising to 0.162 and the CER to 0.075. This corresponds to increases of about 84% in WER and 142% in CER compared to the optimal baseline. These results clearly indicate that while a modest echo exerts a minor impact on transcription accuracy, significant background noise substantially degrades performance, highlighting the importance of maintaining controlled acoustic environments for high-accuracy STT applications.

Table 6. Environmental Conditions Impact.

STT model	Room setting	WER	CER
Chirp-2	Echo	0.105	0.035
	Noise	0.162	0.075

5 Conclusions

In evaluating various STT models for Lithuanian transcription, Google’s Chirp-2 emerged as the top performer, achieving the lowest WER of 0.093 and CER of 0.033. Subsequent analyses provided several key insights:

Firstly, the configuration of audio channels—whether mono or stereo—showed no impact on transcription accuracy, with both setups yielding identical WER and CER values. This suggests that mono recordings are

sufficient for accurate transcription while offering advantages in terms of smaller file sizes.

Secondly, among the FLAC audio settings tested, both 16kHz/16-bit and 16kHz/24-bit configurations achieved optimal performance, each resulting in a WER of 0.088 and a CER of 0.031. Given that the 16-bit setting produces smaller file sizes without compromising accuracy, it is recommended for efficient storage and processing.

Thirdly, varying MP3 compression settings, including different sampling rates and bitrates, demonstrated no significant influence on transcription accuracy. All configurations consistently yielded approximately a 0.091 WER and 0.032 CER, comparable to the lossless FLAC format. Notably, for both MP3 and FLAC formats, lower sampling rates resulted in slightly better accuracy.

However, increasing playback speed adversely affected transcription quality. At 1.5× speed, the WER increased by 59% and the CER by 81%. At 2.0× speed, the degradation was even more pronounced, with WER rising by 225% and CER by 362.5%.

Environmental conditions also played a significant role in transcription accuracy. Moderate echo led to a modest 19% increase in WER, while substantial background noise resulted in severe degradation, with WER increasing by 84% and CER by 142%.

Overall, while Chirp-2 demonstrates great performance under ideal conditions, it remains sensitive to variations in playback speed and environmental noise. These findings underscore the importance of controlled recording environments and appropriate playback settings for the effective real-world deployment of STT systems.

References

- [1] Zhao, R., Choi, A. S., Koenecke, A., & Rameau, A. (2025). Quantification of Automatic Speech Recognition System Performance on d/Deaf and Hard of Hearing Speech. *The Laryngoscope*, 135(1), 191-197.
- [2] Adedeji, A., Joshi, S., & Doohan, B. (2024). The sound of healthcare: Improving medical transcription asr accuracy with large language models. *arXiv preprint arXiv:2402.07658*.
- [3] Afonja, T., Olatunji, T., Ogun, S., Etori, N. A., Owodunni, A., & Yekini, M. (2024). Performant ASR models for medical entities in accented speech. *arXiv preprint arXiv:2406.12387*.
- [4] Boateng, G., Mensah, J. A., Yeboah, K. T., Edor, W., Mensah-Onumah, A. K., Ibrahim, N. D., & Yeboah, N. S. (2024, July). Brilla AI: Ai contestant for the national science and maths quiz. In *International Conference on Artificial Intelligence in Education* (pp. 214-227). Cham: Springer Nature Switzerland.

- [5] Kaulage, A., Walunj, A., Bhandari, A., Dighe, A., & Sagri, A. (2024, May). Edu-lingo: A Unified NLP Video System with Comprehensive Multilingual Subtitles. In 2024 Second International Conference on Data Science and Information System (ICDSIS) (pp. 1-8). IEEE.
- [6] Yang, Y., Song, Z., Zhuo, J., Cui, M., Li, J., Yang, B., ... & Chen, X. (2024). GigaSpeech 2: An Evolving, Large-Scale and Multi-domain ASR Corpus for Low-Resource Languages with Automated Crawling, Transcription and Refinement. arXiv preprint arXiv:2406.11546.
- [7] Navickas, G., Raškinis, G., Mikulėnienė, D., Kardelis, V., Makauskaitė, I., Kasparaitis, P., ... & Korvel, G. (2024). Development of a Large Lithuanian Speech Corpus for speech recognition, artificial intelligence, and other innovative language technologies. In DAMSS: 15th conference on data analysis methods for software systems, Druskininkai, Lithuania, November 28-30, 2024. (pp. 74-75). Vilniaus universiteto leidykla.
- [8] Pipiras, L., Maskeliūnas, R., & Damaševičius, R. (2019). Lithuanian speech recognition using purely phonetic deep learning. *Computers*, 8(4), 76.
- [9] Navarro, G. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1), 31-88.
- [10] Kheddar, H., Hemis, M., & Himeur, Y. (2024). Automatic speech recognition using advanced deep learning approaches: A survey. *Information Fusion*, 102422.
- [11] Naqvi, S. M. R., Tahir, M. A., Javed, K., Khan, H. A., Raza, A., & Saeed, Z. (2024). Code-mixed street address recognition and accent adaptation for voice-activated navigation services. IEEE Access.

Naudotojo apsaugos nuo duomenų viliojimo atakų įrankių palyginimas

Kamilė Samusiovaitė, Domas Nemanius, Kristina Lapin

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Informatikos institutas,
Didlaukio g. 47, LT-08303 Vilnius
kamile.samusiovaite@mif.stud.vu.lt
domas.nemanius@mif.stud.vu.lt
kristina.lapin@mif.vu.lt

Santrauka. Straipsnyje siekiama atlikti apsaugos nuo duomenų viliojimo atakų įrankių panaudojamumo analizę ir remiantis gautais rezultatais sukurti projektavimo gaires priemonei, skirtai duomenų išviliojimo prevencijai. Straipsnyje surinkti ir aprašyti įrankiai, skirti apsaugoti naudotoją nuo duomenų viliojimo atakų. Jakobo Nielseno euristicos pritaikytos formuluojant panaudojamumo kriterijus, taip pat suformuluoti diegimo ir techninių sprendimų kriterijai. Remiantis šiais kriterijais atliktas įrankių vertinimas. Identifikuoti esamų įrankių privalumai ir neišnaudotos galimybės. Gautų rezultatų pagrindu, suformuluotos projektavimo gairės priemonei, skirtai apsaugoti naudotoją nuo duomenų viliojimo atakų.

Raktiniai žodžiai: duomenų viliojimas, įrankių vertinimas, panaudojamumas, diegimas.

1 Įvadas

Terminą duomenų viliojimo ataka (angl. *phishing*) pirmą kartą internete 1996 m. pavartojo įsilaužėlių grupė, kuri pavogė „America Online“ paskyras, apgaule privertusi naudotojus atskleisti savo slaptažodžius [1]. Nuo to laiko atakų skaičius vis augo, 2016 metais siekė 1,2 milijonus ir tai yra 65 % daugiau nei 2015 metais [2]. Remiantis naujausiu „Intersile Consulting Group, LLC“ atliktu tyrimu „Phishing Lanscape 2023: A Study of the Scope and Distribution of Phishing“, kuris buvo paskelbtas 2023 m. rugpjūčio 9 d., duomenų vagystės atakų skaičius per laikotarpį nuo 2022 m. gegužės iki 2023 m. balandžio išaugo 65 % lyginant su ankstesniu tyrimo laikotarpiu ir siekė 1,8 milijonus unikalių atakų [3]. Ir nors kasmet atakų skaičius vis didėja, vyksta nuolatinės varžybos tarp naujų sukčiavimo atakų ir joms atpažinti naudojamų apsaugos priemonių, juk nenuostabu, kad duomenų vagystės atakų

rengėjai keičia sukčiavimo strategiją atsižvelgdami į tas pačias rekomendacijas ir gaires, kuriomis naudojasi apsaugos priemonių kūrėjai [4] [5].

Duomenų vagystės atakos pasikeitė nuo jų atsiradimo 1990–aisiais. Ankstyvosiose atakose kibernetiniai nusikaltėliai naudojo paprastas taktikas – suklastotus elektroninius laiškus, kuriuose naudotojai buvo raginami prisijungti prie savo paskyrų. Pastaruoju metu duomenų vagystės atakos iš apgaulingų laiškų siuntinėjimo išsivystė į sudėtingas technikas, kurios remiasi psichologiniu manipuliavimu, išgaunant asmeninę informaciją [6]. Siekiant užtikrinti interneto naudotojų saugumą yra ieškoma įvairių kovos su duomenų viliojimo atakomis metodų.

Darbo tikslas – palyginti esamus prevencijos įrankius ir sukurti projektavimo gaires priemonei, kuri padėtų apsaugoti naudotoją nuo duomenų viliojimo atakų.

2 Duomenų viliojimo atakų prevencijos įrankių palyginimas

Šiame skyriuje yra identifikuojami kriterijai, kuriais remiantis straipsnyje vertinti įrankiai, skirti naudotojų apsaugai nuo duomenų viliojimo atakų. Panaudojamumo kriterijams formuluoti naudotos Jakobo Nielseno euristikos [7]. Taip pat suformuluoti diegimo kriterijai [8] bei naudojamų techninių sprendimų vertinimo kriterijai [9] [10]. Visi straipsnyje suformuluoti kriterijai pateikti žemiau esančioje lentelėje (žr. 1 lentelę).

1 lentelė. Įrankių vertinimo kriterijai.

Panaudojamumo kriterijai	PK. 1 Sistemos būsenos matomumas. PK. 2 Atitikimas tarp sistemos ir realaus pasaulio. PK. 3 Naudotojo kontrolė ir laisvė. PK. 4 Nuoseklumas ir standartai. PK. 5 Klaidų prevencija. PK. 6 Atpažinimas, o ne prisiminimas. PK. 7 Lankstumas ir naudojimo efektyvumas. PK. 8 Estetiškumas ir minimalistinis dizainas. PK. 9 Pagalba ir dokumentacija.
Diegimo kriterijai	DK. 1 Prieinamumas. DK. 2 Minimalios išlaidos vienam naudotojui. DK. 3 Suderinamumas su naršyklėmis. DK. 4 Brandumas. DK. 5 Nepriklausomybė nuo nuosavybės teisių.
Techniniai sprendimai	TS. 1 Naudotojų mokymas. TS. 2 Juodieji/baltieji sąrašai. TS. 3 Giliojo mokymosi metodai ir mašininis mokymasis.





Kiekvienas kriterijus papildytas identifikatoriumi ir kriterijaus taikymo aprašymu. Kriterijų identifikatoriai yra nustatomi pagal pirmąsias žodžių raišes, pavyzdžiui, „Panaudojumo kriterijai“ – PK. Toliau straipsnyje pateikiami keli vertinimo kriterijų pavyzdžiai PK.8 ir PK.1. Tarkime, Nielseno euristika „Estetiškas ir minimalistinis dizainas“ teigia, kad „Naudotojo sąsajoje neturėtų būti informacijos, kuri yra nesvarbi arba retai reikalinga. Kiekvienas papildomas informacinis elementas naudotojo sąsajoje varžosi su svarbiais informaciniais elementais ir mažina jų santykinį matomumą.“ Ši euristika straipsnyje performuluota štai taip:

PK.8 Estetiškas ir minimalistinis dizainas. Ši euristika teigia, kad įrankių tikslas – padėti naudotojams atpažinti ir sustabdyti atakas, o ne, pavyzdžiui, atkreipti dėmesį į komercines reklamas. Dėl šios priežasties įrankyje turi būti tik informacija susijusi su apsauga nuo duomenų viliojimo atakų. Glaustas ir tvarkingas dizainas padės naudotojui lengvai suprasti, į ką reikia atkreipti dėmesį ir kokių veiksmų imtis, kai pasirodo įspėjimas. Taip įrankis padės išvengti naudotojo suklaudinimo ir užtikrins greitą naudotojo atsaką į galimas grėsmes.

Kiekvienam kriterijui yra apibrėžiamas taikymas naudojant keturias piktogramas, kiekvienai nurodyta, kuo įrankis turi pasižymėti, kad gautų atitinkamą įvertį.

Pavyzdžiui, kriterijui PK.8, straipsnyje apibrėžiamas toks jo taikymas (žr. 2 lentelę):

2 lentelė. PK.8 kriterijaus taikymas.





	– įrankyje yra daugiau informacijos nesusijusios su duomenų viliojimo atakomis ir apsauga nuo jų, nei kad susijusios.
	– įrankyje yra informacija susijusi su apsauga nuo duomenų viliojimo atakų, tačiau yra daugiau nei viena reklama ar informacija, kuri nesusijusi su apsauga nuo duomenų viliojimo atakų.
	– įrankyje yra informacija susijusi su apsauga nuo duomenų viliojimo atakų, tačiau yra ne daugiau nei viena reklama ar informacija, kuri nesusijusi su apsauga nuo duomenų viliojimo atakų
	– įrankyje yra tik informacija susijusi su apsauga nuo duomenų viliojimo atakų.

Remiantis Nielseno euristika „Sistemos būsenos matomumas“ „Dizainas visada turėtų informuoti naudotojus apie tai, kas vyksta, ir per atitinka-

mą laiką suteikti tinkamą grįžtamąjį ryšį.“ Šiam kriterijui euristika perfomuluota taip:

PK.1 būsenos matomumas. Euristika tikrina vizualinius gebėjimus trijose stadijose: prieš patenkant į svetainę tikrinant svetainės autentiškumą, tikrinant svetainės autentiškumą bandant patekti į svetainę ir esant joje. Kiekvienoje stadijoje svetainė, kurioje lankosi naudotojas, turėtų informuoti naudotoją apie vykstantį procesą ir pateikti svetainės autentiškumo rezultatą. Rezultato pateikimas turėtų būti nuolat rodomas. Šio kriterijaus taikymas pavaizduotas (žr. 3 lentelę).

3 lentelė. PK.1 kriterijaus taikymas.

	– kriterijus neįgyvendinamas.
	– įrankio būseną yra matoma bent vienoje stadijoje.
	– įrankio būseną yra matoma dvejose stadijose.
	– įrankio būseną yra matoma visose trijose stadijose.

Visi straipsnyje rasti ir aprašyti įrankiai įvertinti pagal visus apibrėžtus panaudojamumo, diegimo bei techninių sprendimų kriterijus. Žemiau pateikiamas vieno iš įrankių vertinimo pavyzdys pagal keletą apibrėžtų vertinimo kriterijų.

„eBay“ įrankių juostos aprašymas: Šis įrankis naudoja euristikų ir juodųjų sąrašų derinį. Įrankių juostoje esantys indikatoriai informuoja naudotoją apie svetainės patikimumą. Indikatoriai turi tris spalvas: žalią, raudoną ir pilką. Įrankių juostoje esanti piktograma vaizduojama žalia spalva, jeigu naudotojas lankosi svetainėje, kurioje yra integruotos „eBay“ ar „Paypal“ suteikiamos paslaugos (žr. 1 paveikslą). Raudona spalva vaizduojama, kai svetainė yra žinoma kaip sukčiavimo svetainė, o pilka spalva, kai svetainė nesinaudoja „eBay“ ar „Paypal“ paslaugomis ir nėra žinoma kaip sukčiavimo svetainė. Svetainės, kurios yra žinomos kaip apgaulingos yra blokuojamos, o iššokantis langas suteikia galimybę naudotojams panaikinti svetainės blokvimą. Įrankis taip pat suteikia galimybę naudotojams patiems pranešti apie sukčiavimo svetainės, kurios patikrinamos ir įtraukiamos į juoduosius sąrašus. „eBay“ įrankių juosta veikė „Microsoft Windows 98/ME/NT/2000/XP“ operacinėse sistemose su internetu naršykle „Internet Explorer“.



1 pav. „Ebay“ įrankis praneša apie sukčiavimo svetainę. Ekranu nuotrauka paimta iš straipsnio [11].

Vertinimo pavyzdys:

- **PK.1 – Pagrindimas:** įrankis tikrina svetainės autentiškumą tik vienoje stadijoje, kai naudotojas jau yra užjęs į svetainę.
- **PK.8 –Pagrindimas:** įrankį sudaro tik piktogramos esančios įrankių juostoje, todėl galime teigti, kad yra tik informacija susijusi su apsauga nuo duomenų viliojimo atakų

Įrankių vertinimas atskleidė, kad nėra įrankio, kuris pilnai padengtų visus vertinimo kriterijus. Daugiausiai kriterijų įgyvendina naujausi įrankiai: „McAfee Web Advisor“, „Netcraft“ ir „Bitdefender Traffic Ligh“. Vis dėlto, šie įrankiai neįgyvendina kriterijų: lankstumas ir naudotojo efektyvumas, prieinamumas, naudotojų mokymas. Išvardinti įrankiai dalinai įgyvendina kriterijus: naudotojo kontrolė ir laisvė, klaidų prevencija, atpažinimas, o ne prisiminimas, pagalba ir dokumentacija. Šie kriterijai yra tobulintini (žr. 2 paveikslą).

3 Projektavimo gairės

Atlikus jau egzistuojančių įrankių, skirtų apsaugoti naudotoją nuo duomenų viliojimo atakų, vertinimą, suformuluotos projektavimo gairės, remiantis pastebėtais įrankių privalumais ir trūkumais vertintuose įrankiuose.

1. **Projektavimo gairė iš kriterijaus PK.1** Projektuojamas įrankis turėtų tikrinti svetainės autentiškumą naudotojui dar prieš užeinant į svetainę, taip pat, naudotojui bandant užėiti į svetainę ir jau užjęs į svetainę bei mokyti naudotojus, kad šie patys atpažintų galimas duomenų viliojimo atakas ir nevestų savo jautrių duomenų sukčiavimo svetainėse.
2. **Projektavimo gairė iš kriterijaus PK.2** Projektuojamas įrankis turėtų naudoti naudotojams lengvai atpažįstamų spalvų piktogramas, naudoti standartinio šviesoforo įspėjamąsias spalvas, dar geriau, jeigu įrankis naudotų aktyviuosius indikatorius su lengvai išmokstamais ir atpažįstamais nurodymais ir įspėjimais.

Frankitai	Panaudojamumo kriterijai										Diegimo kriterijai	Techniniai sprendimai						
	PK.1 Sistemos būsena matomas.	PK.2 Atitikimas tarp sistemos ir realaus pasaulio.	PK.3 Naudootojo kontrolė ir laisvė.	PK.4 Nuoseklumas ir standartai.	PK.5 Klaidų prevencija.	PK.6 Atpažinimas, o ne prisiminimas.	PK.7 Lankstumas ir naudojimo efektyvumas.	PK.8 Estetiškumas ir minimalistinis dizainas.	PK.9 Pagalba ir dokumentacija.	DK.1 Prieinamumas.			DK.2 Minimalios išlaidos vienam naudotojui.	DK.3 Suderinamumas su naršyklėmis.	DK.4 Brandumas.	DK.5 Nepriklausomybė nuo nuosavybės teisių.	TS.1 Naudotojų mokymas.	TS.2 Įvairūs/baltieji sąrašai.
„McAfee Web Advisor“	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
„Netcraft“	○	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
„Bitdefender Traffic Light“	●	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
„SpoonGuard“	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
„Netscape Browser 8.1“	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
„Microsoft Internet Explorer 7“	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
„GeoTrust TrustWatch“	○	○	N/A	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
„Firefox 2“	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
„Ebay“	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
„EarthLink“	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
„Cloudmark AntiFraud“	○	N/A	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
„CallingID“	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

○ – neįgyvendinta, ○ – daliai įgyvendinta, ● – pusiau įgyvendinta, ● – pilnai įgyvendinta

2 pav. Frankių vertinimo rezultatai.

3. **Projektavimo gairė iš kriterijaus PK.3** Projektuojamas įrankis turėtų leisti naudotojui anuliuoti ir pakartoti savo veiksmus. Tai reiškia, kad kai naudotojas užaina į nesaugią svetainę ir įrankis apie tai įspėja naudotoją, naudotojas turi turėti galimybę arba išėiti iš svetainės, arba pasilikti joje, tačiau jeigu naudotojas nusprendžia pasilikti svetainėje tai įrankis turėtų perklausti, ar naudotojas tikrai nori ir yra įsitikinęs, kad nori lankyti nesaugioje svetainėje.
4. **Projektavimo gairė iš kriterijaus PK.4** Projektuojamas įrankis kiek įmanoma labiau turėtų būtų pritaikytas skirtingoms naršyklėms. Tai reiškia, kad kuriamo įrankio kalba, indikatoriai, veiksmai turėtų būti suprantami naudotojams nepriklausomai nuo to, kokią naršyklę naudojas yra pratęs naudotis.
5. **Projektavimo gairė iš kriterijaus PK.5** Projektuojamas įrankis turėtų teikti įspėjimus ir patarimus kaip ištaisyti klaidas (kaip išėiti iš puslapio, į ką atkreipti dėmesį) ir tai turėtų daryti prieš naudotojui užeinant į svetainę ir naudotojui užėjus į svetainę.
6. **Projektavimo gairė iš kriterijaus PK.6** Projektuojamas įrankis turi turėti įspėjimus ir patarimus, kaip apsisaugoti nuo duomenų viliojimo atakų ir tie patarimai turi būti aiškūs ir suprantami bei rodomi iš karto (o ne paspaudus ant dar vienos nuorodos, ar nuėjus į dar vieną svetainę).
7. **Projektavimo gairė iš kriterijaus PK.7** Jeigu projektuojamame įrankyje būtų galimybė naudotojams atlikti veiksmus, kurie nekeltų pavojaus jų saugumui internete, tai būtų gerai, jeigu projektuojamas įrankis turėtų galimybę patyrusiems naudotojams praleisti pasikartojančius veiksmus.
8. **Projektavimo gairė iš kriterijaus PK.8** Projektuojamame įrankyje turėtų būti tik informacija susijusi su apsauga nuo duomenų viliojimo atakų. Įrankio vizualinę dalį neturėtų užimti su saugumu nesusiję indikatoriai, logotipai ar reklamos.
9. **Projektavimo gairė iš kriterijaus PK.9** Projektuojamas įrankis turėtų turėti bent jau prienamą dokumentaciją ar mokomąją medžiagą, kad naudotojai galėtų pasiskaityti apie funkcionalumus bei mokėtų patys savarankiškai įsidiegti įrankį savo kompiuteryje.
10. **Projektavimo gairė iš kriterijaus DK.1** Projektuojamas įrankis turėtų būti pritaikytas ir regos negalią turintiems naudotojams.
11. **Projektavimo gairė iš kriterijaus DK.2** Projektuojamas įrankis turėtų būti nemokamas ir lengvai prieinamas visiems naudotojams ar organizacijoms.

12. **Projektavimo gairė iš kriterijaus DK.3** Projektuojamas įrankis turėtų veikti naudojant standartinius įrankius ar naršykles be papildomų diegimų.
13. **Projektavimo gairė iš kriterijaus DK.4** Projektuojamas įrankis turėtų būti išbandytas naudotoju, naudojamas realiomis sąlygomis, realiai apsaugai nuo duomenų viliojimo atakų ne tik mokslinių tyrimų tikslais.
14. **Projektavimo gairė iš kriterijaus DK.5** Projektuojamas įrankis turėtų būti prieinamas kiekvienam naudotojui ir juo turi būti galima naudotis nemokant už tai autorinio atlyginimo.
15. **Projektavimo gairė iš techninių sprendimų kriterijų.** Projektuojamas įrankis turėtų naudoti bent tris iš keturių techniniu sprendimų. Įrankis turėtų gebėti atpažinti jau žinomas duomenų viliojimo atakas (naudoti juodusius/baltuosius sąrašus), turėtų gebėti atpažinti naujas duomenų viliojimo atakas (naudoti euristikomis grįstą atpažinimą) bei turėtų šviesti ir mokyti naudotoją savarankiškai atpažinti duomenų viliojimo atakas (mokyti naudotojus).

4 Išvados

Straipsnyje įvertinus ir palyginus duomenų viliojimo atakų prevencijai skirtus įrankius, techninių sprendimų vertinimas atskleidė, kad vertinti įrankiai identifikuoja potencialias duomenų viliojimo atakas, tačiau nepaaiškina naudotojui, kodėl svetainė yra identifikuota kaip nesaugi, todėl galima teigti, kad nėra įrankio, kuris ne tik gebėtų atpažinti ir informuoti naudotoją apie potencialią duomenų viliojimo ataką, bet ir patį naudotoją mokyti atpažinti potencialias atakas.

Panaudojamumo vertinimo metu pastebėta, kad įrankiai naudojantys standartinio šviesoforo spalvas kaip pasyvų indikatorius informuoti apie potencialią grėsmę yra suprantamesni pradedančiajam naudotojui, kuris nėra pratęs ir nežino ką reiškia kitokie žymėjimai. Informavimas apie grėsmes yra svarbus naudotojo saugumui užtikrinti, todėl ateityje kuriami įrankiai turėtų tikrinti svetainės autentiškumą trijose naršymo stadijose: prieš užeinant į nesaugią svetainę (gavus svetainės nuorodą), užeinant (paspaudus ant nuorodos) ir esant joje. Taip pat nėra nei vieno įrankio, kuris turėtų lengvai suprantamą ir prieinamą pagalbą ar dokumentaciją, tai svarbu pradedantiems naudotojams norint greitai išmokyti naudotis įrankiu.

Vertinimas pagal diegimo kriterijus atskleidė, kad nėra įrankio pritaikyto regėjimo negalią turinčiam naudotojui. Be to, regėjimo negalią (aklumą

spalvoms) turinčiam naudotojui, šalia spalvų indikatoriaus reikėtų ir piktogramos informuojančios apie svetainės saugumą. Taigi, norėdamas patenkinti tiek pradedančiųjų tiek regėjimo negalią turinčių naudotojų poreikius, kuriamas įrankis turėtų naudoti spalvų indikatoriaus ir piktogramos kombinaciją.

Literatūra

- [1] Gupta, Brij B, Tewari, Aakanksha, Jain ir Anki, „Fighting against phishing attacks: state of the art and future challenges,” *Neural Computing and Applications*, pp. 3629--3654, 2017.
- [2] Basit, Abdul, Zafar, Maham, Liu, Xuan, Javed, A. Rehman, Jalil, Z. a. Kifayat ir Kashif, „A comprehensive survey of AI-enabled phishing attacks detection techniques,” *Telecommunication Systems*, pp. 139--154, 2021.
- [3] Consulting, Group L. Interisle, „Phishing Landscape 2023 A Study of the Scope,” 2024.
- [4] Dhamija, R. a. Tygar, J. D. a. Hearst ir Marti, „Why phishing works,” įtraukta *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Montréal, Québec, Canada, 2006.
- [5] Irani, D. a. Webb, S. a. Giffin, J. a. Pu ir Calton, „Evolutionary study of phishing,” įtraukta *2008 eCrime Researchers Summit*, 2008.
- [6] Alkhalil, Z. Hewage, C. Nawaf, L. Khan ir Imtiaz, „Phishing attacks: A recent comprehensive study and a new anatomy,” *Frontiers in Computer Science*, 2021.
- [7] Li, L. a. Helenius ir Marko, „Usability evaluation of anti-phishing toolbars,” *Journal in Computer Virology*, pp. 163-184, 2007.
- [8] Bonneau, J. a. Herley, C. a. V. Oorschot, P. C. a. Stajano ir Frank, „The quest to replace passwords: A framework for comparative evaluation of web authentication schemes,” *2012 IEEE symposium on security and privacy*, pp. 553-567, 2012.
- [9] Ayeni, R. Korede, Adebisi, A. Ariyo, Okesola, J. Olatunji, Igbekele ir Enmanuel, „Phishing Attacks and Detection Techniques: A Systematic Review,” *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)*, pp. 1-17, 2024.
- [10] Zhou ir A. A. a. Lina, „Phishing environments, techniques, and countermeasures: A survey,” *Computers & Security*, pp. 160-196, 2017.
- [11] ZHANG, Yue ir e. al., „Phishing phish: Evaluating anti-phishing tools,” 2007.

Propagandos technikų lingvistinių požymių lyginamoji analizė taikant natūralios kalbos apdorojimo metodus

Evelina Vaitkevičiūtė, Gražina Korvel

Vilniaus universitetas, Duomenų mokslo ir skaitmeninių technologijų institutas,
Akademijos g. 4, LT-08412 Vilnius
evelina.vaitkeviciute@mif.vu.lt

Santrauka. Pastaruoju metu dėl besitęsiančių geopolitinių konfliktų ir vis labiau populiarėjančio dirbtinio intelekto, propagandos technikų analizė tapo svarbia tyrimų sritimi. Nors, nemažai propagandos technikų lingvistinės analizės tyrimų atlikta angliškuose tekstuose, tačiau palyginti nedaug jų atlikta su mažiau išteklių turinčiomis kalbomis. Šiame darbe buvo siekiama atlikti propagandos technikų lingvistinių požymių lyginamąją analizę ir taip nustatyti bendrasias ir unikalias kiekvienos technikos savybes. Gauti rezultatai rodo, kad tarp skirtingų propagandos technikų būdingi ryškūs technikų fragmentų kiekio, skaitinių požymių, kalbos dalių pasiskirstymo ir dažniausiai pasitaikančių žodžių skirtumai.

Raktiniai žodžiai: propagandos technikos; lingvistinė analizė; natūralios kalbos apdorojimas; kalbos dalių žymėjimas; n-gramos; raktinių frazių išskyrimas.

1 Įvadas

Pastaruoju metu internetinėje erdvėje propaganda tapo galinga priemone manipuluoti visuomenės sąmone [1]. Žiniasklaidoje ir socialiniuose tinkluose sutinkami propagandinio pobūdžio šaltiniai prisideda prie tikrovės iškreipimo, dezinformacijos skleidimo ir pasitikėjimo informacijos šaltiniais mažinimo [2].

Šiuo požiūriu labai svarbus uždavinys yra nuolatinė propagandos technikų savybių analizė ir šių technikų nustatymo tekste priemonių tobulinimas, kas galėtų padidinti informacijos saugumo lygį ir leistų veiksmingai kovoti su dezinformacija. Pasitelkiant natūralios kalbos apdorojimo metodus, atliekama lingvistinių požymių analizė, leidžianti rasti kiekvienai technikai būdingas savybes [2]. Vienas iš tokių metodų – kalbos dalių žymėjimas, kuriu analizuojama leksinė ir sintaksinė informacija [3]. Dažnai lingvistinių

požymių išskyrimui ir analizei taip pat pasitelkiami n -gramų junginiai [1]. Nemažai lingvistinės analizės tyrimų atlikta angliškuose tekstuose, tačiau dėl sudėtingos kalbų morfologijos, šių tyrimų vis dar trūksta mažiau išteklių turinčioms kalboms [3, 4].

Šiame darbe nagrinėjamos 10 propagandos technikų „Emocinė raiška“, „Whataboutism / Red Herring / Straw Man“, „Supaprastinimas“, „Neapibrėžtumas“, „Apeliavimas į autoritetą“, „Mojavimas vėliava“, „Sekimas iš paskos“, „Abejojimas“, „Reductio ad Hitlerum“ ir „Pakartojimas“ [5, 6]. Pateikiama šių propagandos technikų lingvistinių požymių lyginamoji analizė, kurios tikslas – nustatyti dažniausiai naudojamą žodyną bei žodžių junginių struktūras būdingus kiekvienai propagandos technikai ir taip rasti bendrąsias ir unikalias kiekvienos technikos savybes.

2 Duomenys

Šiame darbe naudotą duomenų rinkinį sudaro 682 lingvistikos ir socialinių mokslų tyrėjų anotuoti naujienų straipsniai, surinkti iš lietuviškų naujienų šaltinių, vykdant Vilniaus universiteto propagandos ir dezinformacijos tyrimų projektą ATSPARA [7]. Naujienų šaltiniai pasirinkti remiantis Lietuvos nacionalinio radijo ir televizijos atliktu tyrimu [8]. Kadangi šio darbo rengimo metu projekto tekstynas vis dar yra jo sudarymo procese, todėl naudota paskutinė naujausia iki 2024 m. spalio mėn. surinkto duomenų rinkinio versija.

Teksto anotavimo procesą sudarė propagandinio pobūdžio teksto fragmentų nustatymas bei atitinkamos propagandos technikos priskyrimas konkrečiam fragmentui. Kiekvienam teksto fragmentui, taikant kryžminio anotavimo metodą buvo priskiriama bent viena iš 10-ies skirtingų propagandos technikų.

3 Metodologija

Bendrą tyrimo eigos schemą sudaro: pradinė duomenų analizė, pirminis tekstinių duomenų apdorojimas, lingvistinių požymių analizė ir gautų rezultatų vertinimas.

3.1 Pirminis tekstinių duomenų apdorojimas

Pirmiausia, siekiant tolesnei analizei paruošti tekstinius duomenis atliekamas pirminis teksto apdorojimas. Šis etapas apima tokius veiksmus kaip

didžiųjų raidžių keitimas į mažąsias raides, specialiųjų simbolių (grotelės #, \$, %, @, &), svetainių nuorodų („http“, „https“ ir „www“), skaičių, nelietuviškų raidžių bei skyrybos ženklų šalinimas, ir pan. Taip pat pašalinami dvigubi tarpai, žodžiai, sudaryti iš vienos raidės ir nesuanotuotos tuščios eilutės duomenų rinkinyje.

Toliau šiame etape tekstas suskaidomas į atskirus požymius, vadinamus teksto vienetais (angl. *tokens*), t. y. žodžiai atskiriami vienas nuo kito, atsižvelgiant į tarp žodžių esančius tarpus.

Iš teksto pašalinami nereikšmingi žodžiai. Nereikšmingais laikomi dažnai pasikartojantys žodžiai, tai gali būti prielinksniai, jungtukai ar kiti panašūs žodžiai, kurie nesuteikia jokios naudingos informacijos, kuri būtų reikšminga konstruojant duomenų modelį [9].

3.2 Lingvistinių požymių analizė

Atlikus pradinę analizuojamų duomenų analizę ir pirminį teksto apdorojimą, kitas žingsnis – propagandos technikų lingvistinių požymių analizė. Natūraliosios kalbos apdoravimo metodai, tokie kaip stilistinių požymių išskyrimas, *n*-gramų junginių sudarymas ir kalbos dalių žymėjimas naudojami siekiant išskirti pagrindinius kalbinius požymius, kurie leistų lengviau nustatyti konkrečius žodžius, dažniausiai siejamus su kiekviena propagandos technika [10].

Tekstinių duomenų požymiams išskirti duomenys gali būti atvaizduojami realiųjų reikšmių vektoriais arba tam gali būti naudojami skiriamieji stilistiniai teksto požymiai, tokie kaip teksto ilgis (teksto simbolių ar žodžių skaičius), skyrybos ženklų skaičius, vidutinis žodžių ilgis ar nereikšmingų žodžių kiekis tekste [1, 11]. Žodžių ilgio analizė leidžia nustatyti žodžių ilgį tiek paties žodžio, tiek pagrindinės žodžių formos lygmenimis, tuo tarpu sakinių ilgio analizė vykdoma aukštesniu lygmeniu pagal žodžių skaičių [1].

Duomenų požymiams parinkti taip pat gali būti taikomas *n*-gramų junginių metodas. *N*-gramos – tekstinių dokumentų sekos, žodžių arba simbolių junginiai, sudaryti iš *n* gretimų žodžių arba simbolių, dar dažnai vadinami pastoviųjų žodžių junginiais (angl. *multiword expressions*) arba leksinėmis samplaikomis (angl. *lexical bundles*) [12, 13]. *N*-gramos gali būti sudarytos pagal įvairias dedamąsias (raides, skiemenis), tačiau dažniausiai sudaromos žodžių ir pagrindinių žodžių formų *n*-gramos. Sakiniai yra padalijami į žodžius, o žodžiai sugrupuojami pagal iš anksto apibrėžtus *n*-gramų junginius. Specialus *n*-gramų atvejis, kai sekos sudarytos iš vieno žodžio, vadinamas

unigramomis ($n = 1$). Dvižodžiai junginiai vadinami bigramomis ($n = 2$), o junginiai, sudaryti iš trijų žodžių – trigramomis ($n = 3$) [13].

4 Eksperimento rezultatai

Iš viso duomenų rinkinyje yra 10265 propagandinio pobūdžio teksto fragmentų, kuriems priskirta atitinkama propagandos technika. Lyginant skirtingas propagandos technikas stebimi ryškūs technikų fragmentų kiekio skirtumai. Gauti rezultatai rodo, kad dažniausia propagandos technika „Emocinė raiška“ pasikartoja 3978 kartus ir sudaro 38,8 % visų fragmentų, o rečiausia „Sekimas iš paskos“, pasikartoja 79 kartus ir sudaro 0,8 % visų fragmentų.

Taip pat atsiskleidžia propagandos technikų skaitinių teksto požymių skirtumai. Propagandos technikų skaitinių teksto požymių statistika, vaizduojanti vidutinį simbolių, žodžių ir skyrybos ženklų skaičių, vidutinį žodžių ilgį ir nereikšmingų žodžių kiekį bei standartinį nuokrypį, pateikta 1 lentelėje.

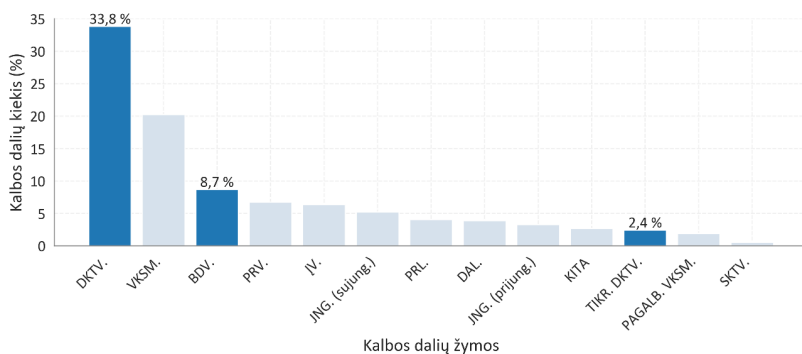
1 lentelė. Propagandos technikų skaitinių teksto požymių statistika.

Technika	Simbolių skaičius	Žodžių skaičius	Skyrybos ženklų skaičius	Vidutinis žodžių ilgis	Nereikšmingų žodžių kiekis
Emocinė raiška	227±247	30±34	7±9	6±1	9±11
Whataboutism / Red Herring / Straw Man	267±277	36±38	9±11	6±1	11±12
Supaprastinimas	211±327	28±43	6±10	6±1	8±13
Neapibrėžtumas	175±209	24±28	5±6	6±1	7±9
Apeliavimas į autoritetą	131±135	17±18	5±7	7±1	4±5
Mojavimas vėliava	106±96	14±13	3±3	6±1	4±4
Sekimas iš paskos	151±231	20±28	4±5	6±1	6±7
Abejojimas	195±201	26±26	6±7	6±1	7±8
Reductio ad Hitlerum	129±152	17±20	3±5	7±2	4±7
Pakartojimas	34±37	4±5	1±1	7±2	1±2

* Žydra spalva pažymėtas didžiausias, šviesiai pilka – mažiausias skaitinių požymių kiekis.

Tarp propagandos technikų ypač išsiskiria „Whataboutism / Red Herring / Straw Man“ ir „Pakartojimas“ technikos. Pavyzdžiui, „Whataboutism / Red Herring / Straw Man“ vidutinis simbolių skaičius yra 267 simboliai su 277 simbolių standartiniu nuokrypiu, tuo tarpu „Pakartojimas“ technikos vidutinis simbolių skaičius yra 34 simboliai su 37 simbolių standartiniu nuokrypiu. Propagandos technikos „Whataboutism / Red Herring / Straw Man“ teksto fragmentuose taip pat stebimas ir didžiausias žodžių skaičius, skyrybos ženklus skaičius bei nereikšmingų žodžių kiekis. „Pakartojimas“ technika nukreipta į tekste kartojamą tą pačią frazę, todėl žodžių skaičius, skyrybos ženklų ir nereikšmingų žodžių kiekis yra mažesnis, bet vidutiniškai tekste žodžiai ilgesni, fragmentus vidutiniškai sudaro 7 žodžiai su 2 žodžių standartiniu nuokrypiu.

Taikant kalbos dalių žymėjimo metodą, nustatyta, kad dažniausiai tekste pasitaikanti kalbos dalis yra daiktavardis (DKTV.), tai sudaro didžiąją dalį, 33,8 % visų kalbos dalių, todėl daiktavardžiai yra išrenkami kaip raktiniai žodžiai. Būdvardžiai ir tikriniai daiktavardžiai, kurie taip pat dažnai laikomi potencialiais raktiniais žodžiais, atitinkamai sudaro 8,7 % ir 2,4 % visų kalbos dalių. Likusių kalbos dalių pasiskirstymas tekste pavaizduotas 1 pav. Žodžiai, kurių apdorojimo tikimybės yra per žemos, nėra apdorojamos ir nurodoma kaip KITA.



1 pav. Kalbos dalių pasiskirstymas tekstinuose fragmentuose. DKTV. – daiktavardis, VKSM. – veiksmažodis, BDV. – būdvardis, PRV. –rieveiksmis, IV. – įvardis, JNG. (sujung. / prijung.) – jungtukas (sujungiamasis / prijungiamasis), PRL. – prielinksnis, DAL. – dalelytė, TIKR. DKTV. – tikrinis daiktavardis, PAGALB. VKSM. – pagalbinis veiksmažodis, SKTV. – skaitvardis.

Toliau 2 lentelėje pateikiama, kokios kalbos dalys yra dažniausiai pasitaikančios atitinkamai technikai. Sužymėjus kiekvienos technikos žodžius pagal dažniausiai pasitaikančias kalbos dalis pastebima, kad didžiausiu daiktavardžių procentiniu kiekiu tekstiniuose fragmentuose išsiskiria „Pakartojimas“ technika, daiktavardžiai šios technikos fragmentuose sudaro 38,2 % visų kalbos dalių žymų. Technikos „Apeliavimas į autoritetą“ tekstuose cituojami garsūs žmonės, autoritetai, todėl juose yra aptinkama daugiau tikrinių daiktavardžių (10,2 %) lyginant su kitomis technikomis. Šiai technikai taip pat būdingas panašus daiktavardžių ir veiksmažodžių procentinis kiekis kaip ir „Reductio ad Hitlerum“. Visgi, „Reductio ad Hitlerum“ technika skiriasi nuo kitų didesniu būdvardžių procentu, aptinkamu fragmentuose (12,4 %). Mažesniu daiktavardžių procentiniu kiekiu, lyginant su kitomis propagandos technikomis, pasižymi „Whataboutism / Red Herring / Straw Man“, „Sekimas iš paskos“, ir „Neapibrėžtumas“. Kadangi technika „Sekimas iš paskos“ nukreipta į apeliavimą į bandos jausmą, fragmentuose labiau išryškėja didesnis procentas veiksmažodžių ir įvardžių.

2 lentelė. Kiekvienos propagandos technikos pasirinktų kalbos dalių procentinis pasiskirstymas.

	DKTV.	VKSM.	BŪDV.	PRV.	ĮV.	TIKR. DKTV.
Emocinė raiška	32,5	20,1	8,4	6,8	6,8	3,2
Whataboutism / Red Herring / Straw Man	31,7	19,6	7,6	6,2	6,9	4,9
Supaprastinimas	32,3	19,9	8,7	7,1	6,5	3,1
Neapibrėžtumas	30,5	21,0	8,6	7,4	6,8	3,0
Apeliavimas į autoritetą	34,2	17,2	7,2	4,6	4,8	10,2
Mojavimas vėliava	36,0	18,2	8,7	4,8	7,2	5,5
Sekimas iš paskos	30,5	22,6	8,3	5,9	8,9	3,1
Abejojimas	34,0	19,2	7,8	6,4	5,7	4,7
Reductio ad Hitlerum	33,5	17,1	12,4	5,8	5,4	5,0
Pakartojimas	38,2	16,9	9,6	4,8	6,4	7,0

* Žydra spalva pažymėtas didžiausias, šviesiai pilka – mažiausias kalbos dalių pasiskirstymas (%).

2 pav. pateikti žodžių debesys, vaizduojantys tarpusavyje išsiskiriančių propagandos technikų 30 dažniausiai pasitaikančių žodžių. Žodžių debesys-

se vaizduojamų žodžių dydis rodo jų dažnumą duomenų rinkinyje. Remiantis žodžių dydžiu, propagandos technikos „Supaprastinimas“ fragmentuose trys svarbiausi žodžiai yra „Lietuvos“, „žmonių“ ir „gali“, kurie duomenų rinkinyje pasitaiko atitinkamai 139, 136 ir 129 kartus (2 pav. A). Žodžiai „Lietuvos“, „žmonių“ rodo, kad kaltė perkeliama Lietuvai, ar žmonėms. Žodis „Lietuvos“ taip pat yra dažniausiai pasikartojęs žodis technikoje: „Emocinė raiška“, „Abejojimas“, „Mojavimas vėliava“, „Apeliavimas į autoritetą“, „Whataboutism / Red Herring / Straw Man“ ir „Sekimas iš paskos“. Likusios trys propagandos technikos išsiskiria nuo anksčiau paminėtų. Propagandos technikoje „Neapibrėžtumas“ dažniausi žodžiai yra „gali“, „nėra“, pasikartojantys 19 ir 16 kartų atitinkamai (2 pav. B). Tuo tarpu, „Reductio ad Hitlerum“ (2 pav. C), technikos dažniausi žodžiai – „revoliucijos“, „komunistinė“. „Pakartojimas“ tekste kartojama ta pati žinutė, dažniausiai vartojant žodžius „komunistas“, „Stambulo“ (2 pav. D).



2 pav. Propagandos technikų dažniausiai pasitaikančius žodžius vaizduojantys žodžių debesys. A – „Supaprastinimas“, B – „Neapibrėžtumas“, C – „Reductio ad Hitlerum“, D – „Pakartojimas“.

Nustatytos dažniausiai naudojamos žodžių junginių struktūros (bigramos). Pavyzdžiui, „Apeliavimas į autoritetą“ technikoje dažniausiai pasitaikančios junginių struktūros yra „užsienio reikalų“, „Anna Loury“ ir „John Hylan“. Šie junginiai vaizduoja autoritetus, viena iš jų, „Anna Loury“ prancūzų žurnalistė, pasakojusi apie karą Donbase (3 pav. D).



3 pav. Propagandos technikų dažniausiai pasitaikančius žodžių junginius (bigramas) vaizduojantys žodžių debesys. A – „Supaprastinimas“, B – „Emocinė raiška“, C – „Mojavimas vėliava“, D – „Apeliavimas į autoritetą“.

Taip pat ieškotos kolokacijos arba žodžių, kurie dažnai pasitaiko kartu, kombinacijos (angl. *Collocations*). 3 lentelėje pateiktos dažniausios kolokacijos (bigramos). Pavyzdžiui, tekstiniuose fragmentuose žodis „konservatorių“ dažnai yra susijęs su žodžiu „valstybė“, taip pat dažnai kartu pasitaiko „Emocinės raiškos“ technikos žodžiai „konstitucinio“ ir „teismo“.

3 lentelė. Dažniausios kolokacijos (bigramos).

Bigramos	Kiekis	Technika	Kiekis
konservatorių valstybė	47	Abejojimas	27
konstitucinio teismo	25	Emocinė raiška	9
sveiko proto	23	Emocinė raiška	12
mokesčių mokėtojų	32	Emocinė raiška	15
užsienio reikalų	29	Emocinė raiška	12

Išvados

Šiame darbe atlikta propagandos technikų lingvistinių požymių lyginamoji analizė parodė bendrąsias ir unikaliąsias kiekvienos technikos savybes. Rezultatai rodo, kad tarp skirtingų propagandos technikų būdingi ryškūs techninių fragmentų kiekio, skaitinių požymių skirtumai. „Whataboutism / Red Herring / Straw Man“ technikos tekstiniai fragmentai ilgesni bei turintys daugiau skyrybos ženklų ir nereikšmingų žodžių, lyginant su kitomis technikomis.

Taikant kalbos dalių žymėjimo metodą pastebima, kad daugiausia daiktavardžių yra „Pakartojimas“ fragmentuose. Mažiau daiktavardžių, lyginant su kitomis propagandos technikomis, aptinkama „Whataboutism / Red Herring / Straw Man“, „Sekimas iš paskos“, ir „Neapibrėžtumas“ technikoje. Technikoje „Apeliavimas į autoritetą“ priešingai negu kitose aptinkama daugiau tikrinių daiktavardžių, tačiau jai būdingas panašus daiktavardžių ir veiksmažodžių kiekis kaip ir „Reductio ad Hitlerum“.

Nustatyta, kad technikoje „Supaprastinimas“ „Emocinė raiška“, „Abejojimas“ „Mojavimas vėliava“, „Apeliavimas į autoritetą“, „Whataboutism / Red Herring / Straw Man“, „Sekimas iš paskos“ stebimas tas pats dažniausiai pasikartojantis žodis – „Lietuvos“. Tuo tarpu, dažniausiai pasitaikantys žodžiai „gali“, „revoliucijos“, „komunistas“ atitinkamai technikoje „Neapibrėžtumas“, „Reductio ad Hitlerum“, „Pakartojimas“ rodo, kad šios technikos išsiskiria nuo anksčiau paminėtų technikų.

Šio darbo rezultatai leidžia išplėsti tyrimą analizuojamoje srityje. Analizės automatizavimo procesas bei natūralios kalbos apdorojimo sprendimų taikymas padėtų tyrėjams palengvinti lingvistikos tyrimus. Tolesnio tyrimo metu propagandos technikų lingvistinei analizei atlikti, planuojama taikyti mašininio mokymo metodus, įtraukiant stilistinius požymius, *n*-gramų junginius ir žodžių dažnius bei kalbos dalių žymėjimo požymius, taip pat trans-formeriais pagrįstus, iš anksto apmokytus daugiakalbius modelius.

Padėka. Autoriai dėkoja Lietuvos mokslo tarybos (LMTLT) finansuojamam projektui „Propagandos ir dezinformacijos tyrimai: automatinis atpažinimas mašininio mokymo metodais, poveikis ir visuomenės atsparumas“ (sutarties Nr. S-VIS-23-8) už tyrimui suteiktus duomenis.

Literatūra

- [1] Horák, A., Sabol, R., Herman, O., & Baisa, V. (2024). Recognition of propaganda techniques in newspaper texts: Fusion of content and style analysis. *Expert Systems with Applications*, 251, 124085.
- [2] Ramberdiyeva, G., Dildabekova, A., Abikenova, Z., Karabayeva, L., & Zhuasbaeva, A. (2024). The Functional and Semantic Category of Appeal as a Linguistic Tool in Political Propaganda Texts (in the Example of the English Language). *International Journal for the Semiotics of Law-*Revue internationale de Sémiotique juridique**, 1-14.
- [3] Mishra, A., Shaikh, S. H., & Sanyal, R. (2022). Context based NLP framework of textual tagging for low resource language. *Multimedia Tools and Applications*, 81(25), 35655-35670.
- [4] Mengliev, D., Barakhnin, V., Abdurakhmonova, N., & Eshkulov, M. (2024). Developing named entity recognition algorithms for Uzbek: Dataset insights and implementation. *Data in Brief*, 54, 110413.
- [5] Da San Martino, G., Seunghak, Y., Barrón-Cedeno, A., Petrov, R., & Nakov, P. (2019). Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 5636-5646). Association for Computational Linguistics.
- [6] Barrón-Cedeno, A., Jaradat, I., Da San Martino, G., & Nakov, P. (2019). Propopy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5), 1849-1864.
- [7] Projektas ATSPARA. „Propagandos ir dezinformacijos tyrimai: automatinis atpažinimas mašininio mokymo metodais, poveikis ir visuomenės atsparumas“. Prieiga per internetą: <https://www.atspara.mif.vu.lt> (žiūrėta: 2025-03-06).
- [8] LRT tyrimas. Lietuvos „penktoji kolona“: Rusijos propagandą platina šeimos gynėjai, sekotos ir knygu apie Staliną leidėjai. Prieiga per internetą: <https://www.lrt.lt/naujienos/lrt-tyrimai/5/1700792/lrt-tyrimas-lietuvas-penktoji-%20kolona-rusijos-propaganda-platina-seimos-gynejai-sektos-ir-knygu-apie-stalina-leidejai> (žiūrėta: 2025-03-06).

- [9] Ahmad, P. N., Guo, J., AboElenein, N. M., Haq, Q. M. U., Ahmad, S., Algarni, A. D., & A. Ateya, A. (2025). Hierarchical graph-based integration network for propaganda detection in textual news articles on social media. *Scientific Reports*, 15(1), 1827.
- [10] Pradhan, A., & Yajnik, A. (2024). Parts-of-speech tagging of Nepali texts with Bidirectional LSTM, Conditional Random Fields and HMM. *Multimedia Tools and Applications*, 83(4), 9893-9909.
- [11] Leonavičienė, A. (2010). Sakinių ilgis–publicistinio ir šnekamojo stiliaus sandūros tekstuose požiūris. *Kalbotyra*, (62 (3)), 95-107.
- [12] Esmailzadeh, A., Cacho, J. R. F., Taghva, K., Kambar, M. E. Z. N., & Hajiali, M. (2022, July). Building wikipedia n-grams with apache spark. In *Science and information conference* (pp. 672-684). Cham: Springer International Publishing.
- [13] Kapusta, J., Drlik, M., & Munk, M. (2021). Using of n-grams from morphological tags for fake news classification. *PeerJ Computer Science*, 7, e624.

Neuroniniai tinklais grįstų triukšmo šalinimo EKG signalė metodų eksperimentinis tyrimas

Matas Valatka, Jolita Bernatavičienė

Vilniaus universitetas, Duomenų mokslo ir skaitmeninių technologijų institutas,
Akademijos g. 4, LT-08412 Vilnius
matas.valatka@mif.stud.vu.lt

Santrauka. Širdies ir kraujagyslių ligos yra pagrindinė mirties priežastis pasaulyje, o jų diagnostikai plačiai naudojama elektrokardiograma (EKG), tačiau EKG signalų analizę apsunkina įvairūs triukšmai. Ši problema ypač aktuali nešiojamuose įrenginiuose ar monitoriuose, todėl reikalingas efektyvus signalų apdorojimas. Be to, signalų bangų kaip P, T ir QRS kompleksų segmentavimas be išankstinio triukšmų pašalinimo yra sudėtingas. Neuroninių tinklų tokių kaip DMAM, BIGRU ar CS-TRANS pagalba galima spręsti abi problemas formuojant tinklų junginį, kai vienas modelis pašalina triukšmus, o kitas segmentuoja signalą. Šių modelių mokymui būtini ne tik dideli duomenų kiekiai, bet ir tiksliai suanotuoti signalai. Paminėti modeliai gali pasiekti signalo ir triukšmo santykį iki 20 dB, o segmentuojant išlaikyti atkūrimo metriką visoms bangoms iki 100 %, tačiau preciziškumo matas išlieka mažesnis. Tyrimo metu nustatyta, kad geriausi rezultatai fiksuoti naudojant DMAM modelį apdorojant -3 dB raumenų triukšmą, o segmentuojant apdorotą signalą atkūrimo metrika išliko maksimali visoms bangos ir preciziškumo matas aukštas P ir T bangoms, bet QRS komplekso šiek tiek mažesnis. Gauti rezultatai įrodo, kad praktikoje pritaikius šiuos algoritmus galima ne tik pagerinti EKG signalų diagnostikos tikslumą, bet ir sumažinti medicinos specialistų darbo krūvį bei užtikrinti aukštesnę pacientų priežiūros kokybę.

Raktiniai žodžiai: EKG signalai, neuroniniai tinklai, triukšmų šalinimas, segmentavimas, raumenų triukšmas, GRU, difuzinis modelis, U-Net, CS-TRANS.

1 Įvadas

Širdies ir kraujagyslių ligos yra pagrindinė mirties priežastis pasaulyje, kasmet nusinešanti apie 17,9 milijono gyvybių [1]. Daugiau nei keturi iš penkių atvejų baigiasi mirtimi dėl širdies smūgių ir insultų, o trečdalis šių mirčių įvyksta žmonėms, jaunesniems nei 70 metų. Tuo tarpu elektrokardiograma (EKG) [2] yra vienas pagrindinių diagnostinių įrankių, naudojamų širdies

veiklos sutrikimams nustatyti. Tačiau šio metodo veiksmingumas tiesiogiai priklauso nuo galimybės tiksliai analizuoti EKG signalus, kurie dažnai būna paveikti triukšmų, signalų deformacijų ar kitų trikdžių. Siekiant pagerinti širdies ir kraujagyslių ligų diagnostavimo tikslumą ir efektyvumą, būtina tobulinti EKG signalų apdorojimo metodus, įskaitant triukšmų šalinimą, svarbių segmentų identifikavimą. Šių metodų tyrimai ir optimizavimas ne tik padeda užtikrinti patikimesnę diagnostiką, bet ir leidžia sumažinti medicininio personalo darbo krūvį bei pagerinti pacientų gydymo kokybę.

Vienas iš dažniausiai sutinkamų triukšmų yra raumenų susitraukinėjimas (angl. *muscle artefacts*). Šio tipo triukšmai persidengia su visu elektrokardiogramos dažnių spektru, todėl pašalinti juos yra itin sunku. Įprastomis sąlygomis išvengti šių triukšmų nėra sudėtinga – pacientas turi atsipalaiduoti ir nejudėti. Visgi, jei žmogus kenčia nuo, pavyzdžiui, nevalingų judesių, šių triukšmų išvengti praktiškai neįmanoma.

Triukšmai daugiausiai pasireiškia, kai širdies ritmas matuojamas ne stacionariais, o nešiojamais įrenginiais kaip Holterio monitorius, todėl tokiais atvejais būtinas atitinkamas signalo apdorojimas, kad šis būtų aiškesnis ir neklaidintų specialistų, priimant sprendimus dėl žmogaus sveikatos būklės.

Tiek triukšmų šalinimo, tiek signalo segmentavimo procesai yra pakankamai susiję tarpusavyje, kadangi nešvarų signalą segmentuoti yra praktiškai neįmanoma ir nenaudinga. Tačiau pasitelkus neuroninių tinklų modelius galima gauti neblogus rezultatus sprendžiant tiek triukšmų, tiek segmentavimo uždavinius. Egzistuoja pakankamai nemažai skirtingų modelių, kurie dažnai būna specifiški vienai užduočiai, tačiau sujungus du modelius į vieną bendrą darbo eigą (angl. *workflow*), gauname pilną procesą, kai vienas dedikuotas modelis sprendžia triukšmų šalinimą, o kitas paskirtas segmentavimui. Tokio junginio tikslas ne tik efektyviai šalinti nepageidaujamus triukšmus iš signalo, tačiau ir gebėti tiksliai išskirti pagrindines EKG signalo dalis kaip QRS kompleksas, P ir T bangos.

Neuroninių tinklų modelių mokymui būtini dideli duomenų kiekiai, tačiau kartu su signalais svarbios ir jų anotacijos, nurodančios reikšmingų signalo dalių pradžią, pabaigą ar net viršūnių taškus. Ši informacija leidžia įvertinti modelių integravimo į vieną bendrą junginį efektyvumą ne tik šalinant triukšmus, bet ir segmentuojant signalą. Galiausiai, remiantis pasirinktų metrikų rezultatais, galima daryti išvadą apie neuroninių tinklų gebėjimą spręsti su EKG signalo analize susijusius uždavinius.

2 Duomenys

Tyrime yra naudojami du duomenų rinkiniai - QTDB [3] ir MIT-BIH Noise Stress [4]. Pirmąjį sudaro 105 penkiolikos minučių bei dviejų derivacijų EKG įrašai. Taip pat yra pateikiamos QRS komplekso, P ir T bangų pradžios, viršūnės bei pabaigos žymės. Antrasis rinkinys turi pagrindinius triukšmus, tarp kurių ir raumenų susitraukimo, kuris yra naudojamas norint pasidaryti triukšmingus signalus modelių apmokymui, o vėliau, panaudojant anotacijas iš QTDB duomenų bazės, patikrinti modelio gebėjimą išskirti atkurto signalo svarbias bangas.

Siekiant turėti daugiau duomenų modelio apmokymui, skirtingos derivacijos buvo traktuojamos kaip atskiri signalai, todėl rezultate modeliams naudota 210 EKG įrašų, kurie buvo skirstyti į mokymo, testavimo ir validavimo duomenis santykiu 148/31/31 atitinkamai.

Kadangi tiek QTDB, tiek MIT-BIH Noise Stress duomenų bazių įrašai atitinkamai yra 250 Hz ir 500 Hz dažnių, todėl teko dažnį suvienodinti pasirenkant naudoti 360 Hz dažnį viso tyrimo metu, kad būtų išlaikytas geriausias informacijos vientisumas. Tačiau šis pasirinkimas dar gali būti tikslinamas, atsižvelgiant į analizės rezultatus ateityje, kai būtų tyrinėjama rezultatų priklausomybė nuo dažnio.

Galiausiai, suvienodinus įrašų dažnius, buvo sukurti triukšmingi signalai iš originalaus signalo ir raumenų susitraukimo triukšmo. Šis procesas atliekamas sudėties būdu (2.1), kai triukšmo reikšmė dar papildomai yra kontroliuojama α daugiklio, kurio tikslas prislopinti arba kaip tik padidinti triukšmą.

$$\text{noisy_signal} = \text{signal} + \alpha * \text{noise} \quad (2.1)$$

Be to, MIT-BIH Noise Stress kiekvienam iš triukšmų tipų turi po 2 kanalus, todėl tyrimo metu buvo nuspręsta vieną kanalą išnaudoti tik mokymo, o kitą testavimo duomenų sudarymui. Tokiu būdu yra bandoma patikrinti modelio universalumą, panaudojant skirtingas triukšmų duomenų aibes.

3 Neuroninių tinklų modeliai

Eksperimentiniame tyrime buvo panaudoti trys triukšmo šalinimo modeliai: BiGRU (angl. *bidirectional gated recurrent units (BiGRU)*) [3], DMAM (angl. *diffusion model with self-attention module (DMAM)*) [4], CS-TRANS (angl. *convolutional neural network and stationary wavelet transformation combination*

with transformer (CS-TRANS)) [7], o segmentavimui buvo pasitelktas U-Net [5] modelis. Kiekvienas triukšmo šalinimo algoritmas buvo komponuojamas kartu su segmentavimo modeliu, kad būtų sukurta pilna darbo eiga, t. y. nuo trikdžių šalinimo iki svarbiausių EKG signalo dalių segmentavimo.

BiGRU yra dvikryptis (angl. *bidirectional*) modelis, paremtas GRU (angl. *gated recurrent units*) ląstelėmis, kurios yra rekurentinių neuroninių tinklų (RNN) realizavimo būdas, skirtas sekų analizei. RNN tinklai geba modeliuoti sekų duomenų priklausomybes, nes įtraukia grįžtamuosius ryšius tarp paslėptųjų vienetų, taip išlaikydami ankstesnę informaciją. Kadangi fiziologiniai signalai, tokie kaip EKG, yra laiko eilutės duomenys, RNN leidžia efektyviai modeliuoti jų struktūrą. GRU ląstelių pagrindiniai komponentai – atstatymo (angl. *reset*) ir atnaujinimo (angl. *update*) vartai, kurie reguliuoja informacijos srautą, užtikrindami tinklo gebėjimą išlaikyti būseną ilgalaikėje perspektyvoje.

Difuzinis modelis [6] (DMAM) yra generatyvinis tinklas, kuris remiasi iteratyviu difuzijos procesu, palaipsniui perdirbančiu triukšmingus duomenis į rekonstruotus. Algoritmas yra grindžiamas dviem pagrindiniais etapais: triukšmo pridėjimu, kol gaunamas beveik Gauso baltasis triukšmas, ir triukšmo šalinimu. Mokymo metu modelis išmoksta atkurti duomenų struktūrą, kuri buvo pažeista iteratyviu triukšmu. Po mokymo modelis geba atvirkštine eiga pašalinti triukšmą ir atkurti originalius duomenis iš atsitiktinio triukšmo būsenos. Šis požiūris leidžia modeliui užtikrinti didelį tikslumą net ir dirbant su labai triukšmingais ar fragmentuotais duomenimis. Todėl ši architektūra pasižymi praktiniu efektyvumu ir buvo pradėta taikyti plačiai srityse kaip vaizdų generavimas, garso apdorojimas ar biomedicininų signalų analizė.

Nors pirmiau U-Net modelis buvo pritaikytas nuotraukų apdorojime, tačiau jo panaudojimas vis dažniau pritaikomas ir EKG tyrimuose. Tradicinio U-Net tinklo architektūra susideda iš traukimosi kelio, kuris atitinka tipinio konvoliucijos tinklo struktūrą, ir išsiplėtimo kelio, kuris susideda iš požymių žemėlapio išsiplėtimo. Iš viso tinklą sudaro 23 konvoliuciniai sluoksniai.

Konvoliucinis neuroninis tinklas reikalauja daug duomenų ir nėra kryptingas, todėl buvo pasiūlytas naujas sprendimas CS-TRANS [7], kurio architektūra susideda iš konvoliucinio tinklo ir stacionarios bangelės junginio (CNN-SWT), kuris efektyviai išgauna daugiamačius signalų požymius, o įtrauktas papildomai transformerio mechanizmas pagerina globalių savybių išgavimą. Iš esmės modelis pagrįstas U-Net architektūra, kadangi susideda iš kodavimo ir dekodavimo dalių sujungtų šuolinėmis jungtimis informacijos perdavimui.

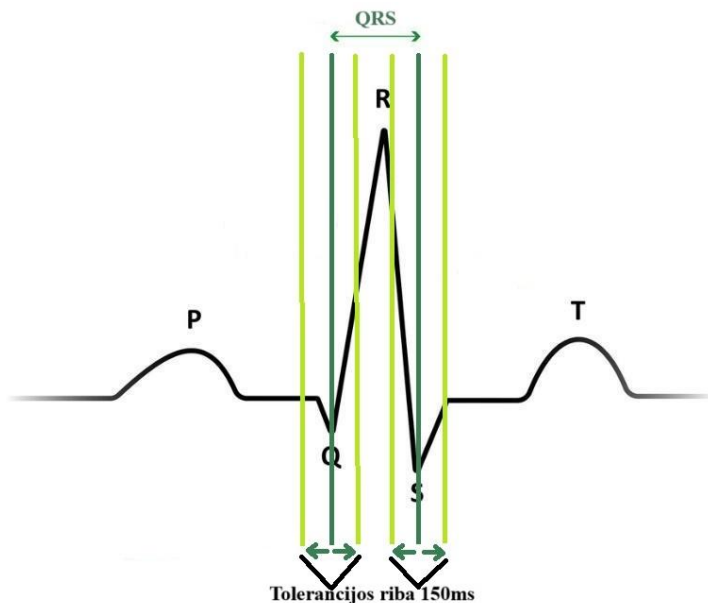
Visi atsirinkti modeliai, nors struktūriškai yra skirtingi, tačiau pritaikomi tam pačiam uždaviniui – nagrinėti bei apdoroti EKG signalus, o pagrindinis tikslas išsiaiškinti, kuris modelių junginys geba tiksliausiai susitvarkyti su nešvariais signalais.

4 Eksperimento rezultatai

Atliekant eksperimentinį tyrimą buvo fiksuojamos pasirinktos metrikos, kad modelių gauti rezultatai būtų tinkamai palyginti. Triukšmų šalinime buvo skaičiuotas signalo ir triukšmo santykis (SNR) – triukšmingo (SNR IN), apdoroto (SNR OUT) ir jų skirtumo (SNR IMPR) – šio santykio esmė parodyti triukšmo lygį, kai didesnė reikšmė indikuoja geresnį veiksmingumą, vidutinė kvadratinė paklaida (RMSE), kuri skirta klaidai tarp prognozuotos ir tikrosios reikšmės nustatyti, ir procentinis santykinis nuokrypis (PRD) – tai atstumu pagrįstas panašumo matas, kur mažesnė reikšmė indikuoja geresnius rezultatus [9]. Segmentavimui pasirinkta skaičiuoti atkūrimo ir preciziškumo metrikas [8], kadangi atkūrimo matas parodo, kaip tiksliai modelis klasifikavo teisingą klasę, o preciziškumas naudojamas įvertinti, kiek pavyzdžių, kuriuos modelis prognozavo kaip teigiamus, iš tikrųjų yra neigiami. Tokiu būdu yra siekiama įvertinti bangų ribų tikslumą ir bendrą segmentavimo kokybę, kadangi būtina ne tik nustatyti reikšmingus įvykius, bet ir užtikrinti jų tikslų žymėjimą. Taip pat, segmentuojant pasirinkta laikytis 150 ms standarto tolerancijos ribai [10] nustatant prognozes (1 pav.).

Visų modelių triukšmo šalinimo rezultatai pateikti 1 lentelėje, kur geriausi rezultatai paryškinti išskyrus triukšmingo (SNR IN) ir apdoroto (SNR OUT) signalų metrikas, kadangi yra išvesta bendra šių reikšmių vertė (SNR IMPR) (4.2). Be to, formuojant triukšmingus signalus buvo naudojama triukšmo valdymo funkcija, kurios tikslas kontroliuoti triukšmo lygį, t. y. didinti arba mažinti. Ši funkcija veikia kaip daugiklis paimtam triukšmo segmentui iš MIT-BIH Noise Stress, tai galutinė SNR IN metrikos reikšmė priklauso tiek nuo apdirbamo triukšmo segmento, tiek nuo pasirinktos triukšmo funkcijos reikšmės. Eksperimentui pasirinkta naudoti šias reikšmes: 5, 0 ir -5, kurios generavo nuo -3 iki -19 dB triukšmą. Tokie triukšmo lygiai buvo pasirinkti siekiant įvertinti modelių atsparumą skirtingam triukšmo intensyvumui, kadangi realiose situacijose signalai gali pasitaikyti tiek su mažu triukšmu, tiek stipriai užteršti.

$$SNR_{IMPR} = SNR_{OUT} - SNR_{IN} \quad (4.2)$$



1 pav. Tolerancijos ribos pavyzdys QRS kompleksui, kai modelis tarp šių ribų prognozę laiko teisinga

Rezultatai iš triukšmo šalinimo (1 lentelė) parodė, kad visi modeliai geba šalinti triukšmus, nes didėjant triukšmui, kitaip tariant mažėjant SNR IN metrikai, galutinė signalo ir triukšmo metrika SNR IMPR didėja, tačiau tai dar neįrodo proceso efektyvumo. Iš RMSE metrikos matyti, kad BiGRU ir CS-TRANS modeliai atlieka daugiau klaidų nei DMAM, o iš to galime daryti išvadą, kad DMAM yra labiau patikimesnis. Nagrinėjant PRD metriką iš konteksto išpuola CS-TRANS modelis, kuris fiksuoja dvigubai prastesnius rezultatus nei kiti. Tuo tarpu DMAM ir BiGRU demonstruoja pakankamai neblogus PRD rezultatus, tačiau verta pabrėžti, kad šįkart BiGRU visuomet davė geresnį rezultatą. Visgi, jei žiūrėtume rezultatus su didžiausiu triukšmu, galime daryti išvadą, kad DMAM fiksuoja žymiai mažesnę RMSE rodiklį, o ši metrika labiausiai pabrėžia modelio tikslumą bei efektyvumą ir būtent dėl šios priežasties DMAM modelis turi daugiau potencialo nei likę nagrinėti tinklai.

Segmentavimo rezultatai, jungiant kiekvieną tinklą su U-Net modeliu, pateikti 2 lentelėje.

1 lentelė. Triukšmo šalinimo modelių rezultatai. Rodyklės žymi metrikos pageidautiną vertę.

Modelis (triukšmo formulės reikšmė)	RMSE ↓	PRD ↓	SNR IN ↑	SNR OUT ↑	SNR IMPR ↑
DMAM (5)	0,06 ± 0,04	34,52 ± 15,00	-2,84 ± 5,70	10,89 ± 2,87	13,72 ± 6,24
BiGRU (5)	0,30 ± 0,02	26,23 ± 6,88	-9,07 ± 1,88	12,24 ± 1,97	21,30 ± 2,93
CS-TRANS (5)	0,42 ± 0,38	74,31 ± 40,45	-4,25 ± 5,23	6,04 ± 2,47	10,29 ± 5,81
DMAM (0)	0,08 ± 0,04	43,21 ± 19,52	-7,84 ± 5,70	9,09 ± 2,90	16,93 ± 6,18
BiGRU (0)	0,53 ± 0,03	34,41 ± 9,13	-14,07 ± 1,88	10,08 ± 2,00	24,15 ± 2,59
CS-TRANS (0)	0,76 ± 0,68	91,20 ± 50,32	-9,25 ± 5,23	4,72 ± 2,17	13,97 ± 5,60
DMAM (-5)	0,10 ± 0,06	56,59 ± 28,27	-12,84 ± 5,70	6,81 ± 3,12	19,65 ± 6,20
BiGRU (-5)	0,95 ± 0,05	48,18 ± 12,79	-19,07 ± 1,88	7,44 ± 2,06	26,51 ± 2,28
CS-TRANS (-5)	1,34 ± 1,21	114,91 ± 59,62	-14,25 ± 5,23	2,98 ± 1,93	17,23 ± 5,24

2 lentelė. U-Net modelio segmentavimo rezultatai (%).

Modelis (triukšmo formulės reikšmė)	Preciziškumas pradžios			Atkūrimas pradžios			Preciziškumas pabaigos			Atkūrimas pabaigos		
	P	QRS	T	P	QRS	T	P	QRS	T	P	QRS	T
DMAM (5)	96	71	96	100	100	100	98	83	98	100	100	100
BiGRU (5)	16	76	71	8	54	82	88	95	92	76	83	83
CS-TRANS (5)	34	51	78	40	96	89	81	71	90	77	98	95
DMAM (0)	91	85	98	98	100	100	95	92	99	99	100	100
BiGRU (0)	13	73	67	7	43	76	88	94	90	77	79	82
CS-TRANS (0)	35	61	79	50	99	94	79	79	90	79	94	97
DMAM (-5)	89	71	89	94	100	100	94	83	94	97	100	100
BiGRU (-5)	8	64	62	4	26	68	88	93	88	77	70	81
CS-TRANS (-5)	34	51	78	40	96	89	81	71	90	77	98	95

Iš esmės segmentavimo rezultatai (2 lentelė) patvirtina, kad DMAM modelis yra optimalus pagal atkūrimo ir preciziškumo kriterijus, kadangi beveik visais atvejais, segmentuojant šio modelio apdorotus signalus, gautami geriausi rezultatai, o ten kur išimtinai BiGRU tinklas pranoko DMAM modelį, tai skirtumas sąlyginai nedidelis. Be to, CS-TRANS ir BiGRU tinklams

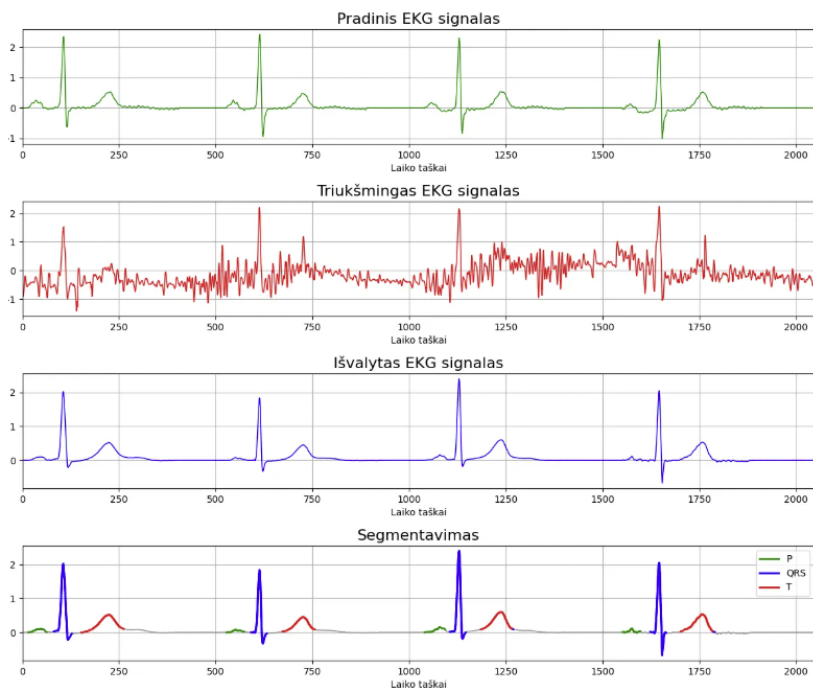
itin prastai sekėsi fiksuoti P bangų pradžias, bet bendrai paėmus visi modeliai demonstravo geresnius rezultatus prognozuojant visų bangų pabaigas. DMAM modelis itin išsiskiria atkūrimo metrikos rezultatais, kas parodo tinklo gebėjimą teisingai klasifikuoti bangų pradžias ir pabaigas, tačiau būtent QRS bangos preciziškumo matas informuoja, jog ne visuomet prognozės bus tikslios.

Apibendrinant galima teigti, kad visi modeliai demonstravo neblogus rezultatus, tačiau DMAM rezultatai (2 pav.) išsiskyrė labiausiai dėl stabilumo tiek triukšmų šalinime, tiek segmentavime.

5 Išvados

Atlikus tyrimą ir apibendrinus rezultatus, galima teigti, kad iš nagrinėtų modelių DMAM demonstravo geriausius rezultatus tiek šalinant triukšmą, tiek segmentuojant signalą. Tuo tarpu BiGRU ir CS-TRANS tinklai taip pat neblogai atlikdavo signalo valymą nuo trikdžių, tačiau segmentavimo etape pasirodė jų trūkumai, kai U-Net modelis vis prasčiau identifikuodavo bangų pradžias ir pabaigas, ypač P bangos. Nepaisant to, tyrimas parodė, kad galima sukurti patikimus automatizuotus įrankius, skirtus EKG signalų nagrinėjimui. Geriausias tyrimo rezultatas fiksuotas naudojant DMAM (5) modelį, kai šiam teko apdoroti -3 dB raumenų triukšmą, o RMSE metrika buvo 0,06, PRD nustatyta ~35 % ir SNR IMPR parodė pagerėjimą iki ~14. Segmentuojant prieš tai apdorotą signalą gauta atkūrimo metrika maksimali visoms bangos, o preciziškumo matas fiksuotas 96 % ir 98 % P ir T bangų pradžioms bei pabaigoms, o QRS kompleksui pradžia ir pabaiga atitinkamai 71 % ir 83 %. Taigi, rezultatai parodo, kad algoritmai gali būti efektyvūs automatizuoti įrankiai, galintys apdirbti EKG signalas, o jų pritaikymas praktikoje galėtų ne tik sumažinti medikų darbo krūvį, bet ir pagerinti pacientų gydymo kokybę.

Ateityje planuojama pratęsti eksperimentinį tyrimą įtraukiant daugiau triukšmų tipų ar net jų kombinacijų, kurios leistų dar tiksliau įvertinti modelių efektyvumą bei universalumą, kadangi praktikoje nešiojamuose įrenginiuose pasitaiko ir tokių triukšmų kaip bazinės linijos (angl. *base-line wander*), atsirandantis dėl kvėpavimo ar kitų mechaninių judesių, ar elektros energijos tinklo trikdžiai (angl. *power-line interference*), galintys atsirasti dėl naudojamo prietaiso.



2 pav. DMAM (5) modelio rezultatų grafikai

Padėka. Dėkojame Vilniaus universiteto ITAPC padalinii už suteiktus IT išteklius (HPC), kurie leido greičiau apmokyti modelius bei supaprastino viso tyrimo atlikimą.

Tyrimas finansuojamas pagal LR Švietimo, mokslo ir sporto ministerijos programą „Universitetų ekselencijos iniciatyvos“ (LR ŠMSM mokslo plėtros programos pažangos priemonė Nr. 12-001-01-01-01 „Gerinti mokslo ir studijų aplinką“).

Literatūra

- [1] „World Health Organization,“ [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1. [Accessed 6 1 2025].
- [2] „Cancer Research UK,“ [Online]. Available: <https://www.cancerresearchuk.org/about-cancer/tests-and-scans/ecg>. [Accessed 6 1 2025].

- [3] Laguna P, Mark RG, Goldberger AL, Moody GB, „A Database for Evaluation of Algorithms for Measurement of QT and Other Waveform Intervals in the ECG,” *Computers in Cardiology*, 1997.
- [4] Moody GB, Muldrow WE, Mark RG, „A noise stress test for arrhythmia detectors,” *Computers in Cardiology*, 1984.
- [5] Mariana Dias, Phillip Probst, Luis Silva, Hugo Gamboa, “Cleaning ECG with Deep Learning: A Denoiser Tested in Industrial,” *SN Computer Science*, 2024.
- [6] Hu, Zheng-Dong and Hong, Yang and Huang, Jia-Yan and Chen, Kai-Hong and Zhao, Wan-Qi and Grau, Antoni and Guerra, Edmundo and Wang, Chuan-Sheng and Zhang, Fu-Quan, “DMAM-ECG: A Diffusion Model with Self-Attention,” *Journal of Network Intelligence*, no. 9, 2024.
- [7] Huyang Peng, Xiaohan Chang, Zhenjie Yao, Donglin Shi, Yongrui Chen, “A deep learning framework for ECG denoising and classification,” *Biomedical Signal Processing and Control*, vol. 94, 2024.
- [8] Ronneberger, O., Fischer, P., & Brox, T, “U-net: Convolutional networks for biomedical image segmentation,” *Medical image computing and computer-assisted intervention*, pp. 234-241, 2015.
- [9] Chatterjee, Shubhojeet and Thakur, Rini Smita and Yadav, Ram Narayan and Gupta, Lalita and Raghuvanshi, Deepak Kumar, “Review of noise removal techniques in ECG signals,” *IET Signal Processing*, vol. 14, no. 9, 2020.
- [10] Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms, American National Standards Institute/EC57, ANSI/AAMI, 2012.

Laiko eilutes transformuojančių į vaizdus, siekiant taikyti CNN klasifikavimo uždaviniui spręsti, metodų tyrimas

Rugilė Vasaitytė

Vilniaus universitetas,
Matematikos ir informatikos fakultetas
Informatikos institutas
Didlaukio g. 47, Vilnius
rugile.vasaityte@mif.stud.vu.lt

Santrauka. Tyrimas nagrinėja laiko eilučių transformavimą į vaizdus, siekiant taikyti konvoliucinius neuroninius tinklus (angl. *Convolutional Neural Networks*, CNN) klasifikavimui. Apžvelgiami trys metodai: Gramo kampinis sumavimo laukas (angl. *Gramian Angular Summation Field*, GASF), Gramo kampinis skirtumų laukas (angl. *Gramian Angular Difference Field*, GADF) bei Markovo perėjimo laukai (angl. *Markov Transition Field*, MTF). CNN modelis pritaikytas šešiams skirtingiems duomenų rinkiniams, transformuotiems minėtais metodais. Rezultatai rodo, kad metodų efektyvumas priklauso nuo duomenų pobūdžio, tačiau GADF dažniau pasiekė geresnius rezultatus. Pateikiamos rekomendacijos, kaip pasirinkti tinkamiausią transformaciją klasifikavimo tikslumui gerinti.

Raktiniai žodžiai: laiko eilutės, giliojo mokymosi metodai, konvoliuciniai neuroniniai tinklai, vaizdų klasifikacija, laiko eilučių transformacija, GASF metodas, GADF metodas, MTF metodas.

1 Įvadas

Laiko eilutė – tai duomenų taškų seka, išdėstyta nuosekliai pagal laiką. Laiko eilučių klasifikavimas (angl. *Time Series Classification*, TSC) yra užduotis, kurios tikslas yra priskirti laiko eilutei klasę, t. y. kintamąjį iš aibės, nusakantį, kuriai grupei ar kategorijai priklauso nagrinėjama laiko eilutė, pavyzdžiui, normalus ar anomalinis, teigiamas ar neigiamas pavyzdys. TSC praktikoje gali būti pritaikomas įvairiose srityse – diagnozuojant širdies ligas [1] ar kibernetiniame saugume [2]. Todėl veiksmingų ir efektyvių algoritmų kūrimas laiko eilučių klasifikavimui yra aktyvi tyrimų sritis. Štai [3] ir [4] straipsniuose atskleidžiama, kad giliojo mokymosi metodai, kuriuos naudojant automatiškai išgaunami sudėtingi duomenų požymiai, dažnai pranoksta

klasikinius statistinius ir mašininio mokymosi metodus sprendžiant laiko eilučių klasifikavimo uždavinius. Šis pranašumas ryškus dirbant su daug pavyzdžių ir požymių turinčiais duomenų rinkiniais, kai požymių inžinerija tampa nepraktiška.

Visgi, norint taikyti giliojo mokymosi metodus, pavyzdžiui, konvoliucinius neuroninius tinklus, apdoroti ir klasifikuoti laiko eilutes, susiduriama su realių duomenų rinkinių heterogeniškumu, t. y. laiko eilutės gali būti skirtingo ilgio, turėti skirtingą struktūrą, triukšmingumo lygį. Tuomet laiko eilutes reikia tinkamai apdoroti. Siekiant šio tikslo, galima laiko eilutes transformuoti į vaizdus – taip užtikrinama, kad visi įvesties duomenys turėtų vienodą formą. Egzistuoja keli tokie transformuojantys metodai, pavyzdžiui, Gramo kampinis sumavimo laukas, Gramo kampinis skirtumų laukas bei Markovo perėjimo laukai [5]. Šie metodai laiko eilutes transformuoja į fiksuoto dydžio dvimačius vaizdus, kuriuos galima analizuoti naudojant giliojo mokymosi metodus.

Taigi, šio tyrimo **tikslas** – įvertinti, kuris laiko eilutes į vaizdus transformuojantis metodas leidžia gauti tikslesnį vaizdų klasifikavimo rezultatą. Siekiant tyrimo tikslo, bus atlikti šie žingsniai: pasirinkti duomenų rinkiniai bus transformuoti į vaizdus naudojant GASF, GADF ir MTF metodus. Vėliau šie vaizdai bus klasifikuojami konvoliuciniu neuroniniu tinklu. Modelio rezultatui bus analizuojami ir statistiškai įvertinti, kad būtų nustatyta, kuris transformacijos metodas užtikrina tiksliausią klasifikaciją.

2 Tyrimo eiga

Toliau aprašytame tyrime buvo siekiama įvertinti laiko eilučių klasifikavimo tikslumą, po to, kai buvo pritaikyti laiko eilutes į vaizdus transformuojantys metodai: GASF, GADF ir MTF. Tyrimo schema pateikta 1 pav. Procesas prasideda nuo duomenų rinkinio atrinkimo, atliekamas jų paruošimas ir padalijimas į mokymo, validavimo ir testavimo aibes. Laiko eilutės transformuojamos į vaizdus, po to mokymo ir validavimo duomenys naudojami neuroninio tinklo mokymui. Galiausiai modelis testuojamas ir renkami klasifikavimo rezultatų stebėjimai, kurie analizuojami statistiniais metodais.

Duomenys ir jų paruošimas

Buvo pasirinkti šeši skirtingų tipų duomenų rinkiniai: „GunPoint“ [6], „ECG200“ [7], „ECG5000“ [8], „Wafer“ [7], „TwoLeadECG“ [8] ir „Earthquakes“ [9]. Kiekvie-

na laiko eilutė rinkinyje turi atitinkamą žymę, kuri naudojama prižiūrimam mokymui. Šie duomenys parinkti taip, kad sudarytų tam tikrą įvairiapusį rinkinį: jie parinkti iš sensorių (SENSOR), elektrokardiogramų (ECG), žmogaus veiklos atpažinimo (HAR) tipo duomenų rinkinių. Be to, eilutės turi skirtingus ilgus, triukšmingumo lygį, mokymui panaudotas skirtingas pavyzdžių kiekius.

Laiko eilutės transformuotos į vaizdus naudojant *Python pyts* biblioteką, taikant numatytuosius parametrus. Transformuotų eilučių pavyzdžius galima matyti 1 pav. Transformuotos laiko eilutės su atitinkamomis klasėmis buvo suskirstytos į mokymo, validavimo ir testavimo aibes santykiu 70:15:15 atitinkamai. Galiausiai kiekvieno vaizdo pikselių reikšmės buvo normalizuotos į [0, 1] intervalą.

1 lentelė. Naudojamų duomenų rinkinių informacija.

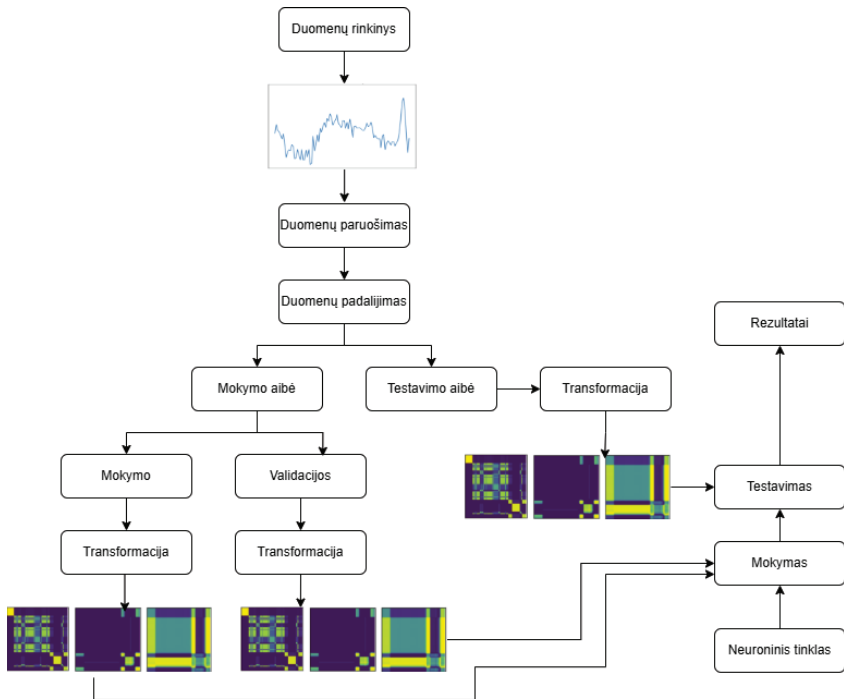
Duomenų rinkinys	Mokymui panaudotų pavyzdžių kiekis	Klasių skaičius
GunPoint	50	2
ECG200	100	2
ECG5000	3500	5
Wafer	1000	2
TwoLeadECG	813	2
Earthquakes	322	2

Modelio architektūra

CNN prasideda įvesties sluoksniu, kuris priima duomenis su nustatytu aukščiu, pločiu ir kanalų skaičiumi. Tuomet naudojami du konvoliucinių sluoksnių blokai, kurie apima konvoliucinį sluoksnį su aktyvacijos funkcija *ReLU*, maksimalų telkimo sluoksnį ir atsitiktinio praretinimo (angl. *dropout*) sluoksnį, kuris naudojamas siekiant sumažinti permokymo riziką. Po šių blokų naudojamas ištiesinimo sluoksnis, kuris duomenis transformuoja į viematį vektorių, atsitiktinio praretinimo sluoksnis ir pilnai jungūs sluoksniai, kurių pirmasis turi 96 neuronus su *ReLU* aktyvacija. Paskutinis sluoksnis yra išėjimo sluoksnis, turintis tiek neuronų, kiek yra klasifikuojamame rinkinyje klasių. Naudotas *Adam* optimizatorius, *sparse categorical crossentropy* nuostolių funkcija ir 20 mokymosi epochų. Stebima vertinimo metrika – klasifikavimo tikslumas testavimo fazėje.

Statistikos

Modelis buvo apmokytas su visais duomenų rinkiniais, kurie buvo transformuoti visais transformuojančiais metodais, kiekvienam eksperimentui buvo fiksuoti rezultatų stebėjimai. Statistiniam vertinimui naudota vienfaktorė ANOVA ir Tukey HSD testas, siekiant nustatyti, ar stebėti skirtumai tarp rezultatų, gautų taikant skirtingus transformuojančius metodus, statistiškai reikšmingi.



1 pav. Eksperimentinės dalies darbo eigos diagrama.

3 Rezultatai

Transformuotų laiko eilučių klasifikavimo rezultatai rodo, kad transformavimo metodo parinkimas daro įtaką klasifikavimo tikslumui, tačiau šis poveikis priklauso nuo konkretaus duomenų rinkinio. Štai klasifikuojant transformuotą „GunPoint“ duomenų rinkinį GASF metodas statistiškai

reikšmingai lenkia tiek GADF, tiek MTF metodus (vidurkio skirtumas 0,066 ir 0,082 atitinkamai). Tarp GADF ir MTF reikšmingo skirtumo nėra. Šiam rinkiniui labiausiai tinkamas transformuojantis metodas yra GASF. Tuo tarpu klasifikuojant „ECG200“ duomenų rinkinį GADF metodas nežymiai lenkia GASF metodą, tačiau ne statistiškai reikšmingai ($p = 0,059$). Tačiau tiek GADF, tiek GASF yra statistiškai reikšmingai pranašesnis už MTF (vidurkių skirtumas 0,047 ir 0,025 atitinkamai). Šiam rinkiniui transformuoti labiau tinka GASF ir GADF. Klasifikuojant „ECG5000“ duomenų rinkinį, GADF metodas pasirodė statistiškai reikšmingai pranašesnis už MTF metodą (vidurkių skirtumas 0,006, $p < 0,001$), taip pat ir GASF metodas statistiškai reikšmingai lenkia MTF (vidurkių skirtumas 0,005, $p < 0,001$). Tačiau tarp GADF ir GASF reikšmingo skirtumo nenustatyta ($p = 0,590$). Aukščiausias klasifikavimo tikslumo vidurkis pasiektas naudojant GADF (0,951), šiek tiek atsilieka GASF (0,950), o mažiausias rezultatas fiksuotas taikant MTF (0,945). Šiam duomenų rinkiniui labiausiai tinkami transformavimo metodai yra GADF ir GASF. Atkreiptinas dėmesys, kad transformuoto „Wafer“ duomenų rinkinio klasifikavimo tikslumui skirtingi metodai (MTF, GASF, GADF) reikšmingos įtakos nedaro. ANOVA testo statistikos $F = 1,006$, $p = 0,374$ rodo, jog metodų poveikis tikslumui yra nežymus. Visgi, didžiausias klasifikavimo vidurkis buvo gautas taikant GASF metodą, mažiausias MTF. Panašus rezultatas fiksuotas ir „Earthquakes“ duomenų rinkinyje – klasifikavimo tikslumui skirtingi metodai (MTF, GASF, GADF) statistiškai reikšmingos įtakos nedaro. Galiausiai, klasifikuojant „TwoLeadECG“ duomenų rinkinį, GADF metodas pasirodė statistiškai reikšmingai pranašesnis tiek už GASF (vidurkių skirtumas 0,018, $p = 0,005$), tiek už MTF (vidurkių skirtumas 0,015, $p = 0,024$). Tuo tarpu tarp GASF ir MTF reikšmingo skirtumo nenustatyta ($p = 0,838$). Aukščiausias klasifikavimo tikslumo vidurkis pasiektas naudojant GADF (0,998), žemesni rezultatai fiksuoti GASF (0,979) ir MTF (0,983) metodais. Šiam duomenų rinkiniui tinkamiausias transformavimo metodas yra GADF.

Apibendrinti rezultatai pateikti 2 lentelėje. Sakytina, kad yra statistiškai reikšmingų skirtumų tarp metodų (MTF, GASF, GADF), kurie daro įtaką klasifikavimo tikslumui, kai transformuojami ir klasifikuojami skirtingi duomenų rinkiniai. GASF ir GADF metodais suklasifikuotų transformuotų laiko eilučių klasifikavimo tikslumas dažniausiai lenkia MTF metodu transformuotų laiko eilučių klasifikavimo tikslumą.

2 lentelė. Vidutiniai klasifikavimo tikslumai naudojant skirtingus metodus.

Duomenų rinkinys	GASF tikslumas	GADF tikslumas	MTF tikslumas	Statistiškai reikšmingas skirtumas pagal ANOVA ir Tukey HSD testus
GunPoint	0,936	0,870	0,854	Taip – GASF > GADF, GASF > MTF
ECG200	0,861	0,883	0,836	Taip – GASF > MTF, GADF > MTF
ECG5000	0,950	0,951	0,944	Taip – GASF > MTF, GADF > MTF
Wafer	0,998	0,997	0,997	Ne
TwoLeadECG	0,980	0,997	0,983	Taip – GADF > MTF, GADF > GASF
Earthquakes	0,8	0,8	0,8	Ne

4 Išvados

Remiantis rezultatais galima teigti, kad laiko eilučių klasifikavimo tikslumas priklauso tiek nuo duomenų rinkinio, tiek nuo pasirinkto transformuojančio laiko eilutės į vaizdus metodo – apibendrintos generalizacijos apie metodus daryti negalima.

Visgi, sprendžiant laiko eilučių klasifikavimo uždavinį su prižiūrimu CNN mokymusi galima padaryti šias išvadas:

1. Jei tiriant duomenų rinkinį buvo nustatytas statistiškai reikšmingas metodo poveikis klasifikavimo tikslumui (tokie duomenų rinkiniai buvo keturi), tuomet arba GASF, arba GADF buvo pranašesnis už MTF.
2. Tiriant keturis iš šešių duomenų rinkinius nebuvo nustatytas statistiškai reikšmingas skirtumas tarp GASF ir GADF metodais transformuotų laiko eilučių klasifikavimo tikslumo ($p > 0,05$). Natūralu – metodų prigimtis gana panaši. Viena išimtis yra „GunPoint“ duomenų rinkinys: GASF metodas statistiškai reikšmingai pranašesnis už GADF (vidurkių skirtumas 0,066). Šis duomenų rinkinys turėjo mažiausiai mokymo pavyzdžių iš tirtųjų. Kita išimtis yra „TwoLeadECG“: GADF metodas statistiškai reikšmingai pranašesnis už GASF (vidurkių skirtumas 0,018).
3. Kai kurie rinkiniai, pavyzdžiui, „Wafer“ gerai klasifikuojami nepriklausomai nuo transformuojančio metodo. Vidutinis klasifikavimo tikslumas, naudojant visus metodus, viršijo 0,997. Šie duomenys mažai chaotiški, transformuotos laiko eilutės lengvai klasifikuojamos CNN.

4. „Earthquakes“ duomenų rinkinio klasifikavimas pasirodė prastas nepriklausomai nuo taikyto metodo. Vidutinis transformuotų laiko eilučių klasifikavimo tikslumas neviršijo 0,8 tiriant visus metodus. Transformuojančiais metodais gauti vaizdai nepasižymėjo informatyviais vizualiais bruožais. Šiame duomenų rinkinyje yra maži periodiškumo ir dėsningumą, laiko eilutės labai chaotiškos.

Apibendrinant tai, kas buvo parašyta, tyrimas patvirtina, kad transformavimo metodas turi reikšmingą įtaką klasifikavimo rezultatams. Renkantis konkretų metodą praktikoje, verta atsižvelgti į konkretų duomenų rinkinį, tačiau verta pradėti nuo GADF, kurį taikius dažniau pasiekti aukštesni klasifikavimo rezultatai.

Literatūra

- [1] Jat, T., Patil, N., & Bhat, P. (2024). Detection of heart arrhythmia with electrocardiography. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 13(1), 57.
- [2] Medvedev, V., Budžys, A., & Kurasova, O. (2023). Enhancing keystroke biometric authentication using deep learning techniques. In *Proceedings of the 2023 18th Iberian Conference on Information Systems and Technologies*, 1–6. IEEE.
- [3] Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2019). Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4), 917–963.
- [4] Mohammadi Foumani, N., Miller, L., Tan, C. W., Webb, G. I., Forestier, G., & Salehi, M. (2024). Deep learning for time series classification and extrinsic regression: A current survey. *ACM Computing Surveys*, 56(9), 1–45.
- [5] Wang, Z., & Oates, T. (2015). Imaging time-series to improve classification and imputation. In *Proceedings of the 2015 24th IJCAI International Joint Conference on Artificial Intelligence*, 3939–3945. AAAI.
- [6] Ratanamahatana, C. A., & Keogh, E. (2005). Three myths about dynamic time warping data mining. In *Proceedings of the 2005 5th SIAM International Conference on Data Mining*, 506–510. Society for Industrial and Applied Mathematics.
- [7] Olszewski, R. T. (2001). *Generalized feature extraction for structural pattern recognition in time-series data*. Carnegie Mellon University.
- [8] Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., ... & Stanley, S.H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation*, 101(23), e215–e220.
- [9] Bagnall, A., Davis, L., Hills, J., & Lines, J. (2012). Transformation based ensembles for time series classification. In *Proceedings of the 2012 12th SIAM International Conference on Data Mining*, 307–318. Society for Industrial and Applied Mathematics.

Samdomųjų darbuotojų darbo užmokesčio prognozavimas taikant mašininio mokymosi metodus

Dominykas Vilčinskas, Rūta Levulienė

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Taikomosios matematikos institutas,
Naugarduko g. 24, Vilnius
dominykas.vilcinskas@mif.stud.vu.lt

Santrauka. Darbo užmokesčio modeliavimas leidžia įvertinti įvairių rodiklių daromą įtaką pajamoms, gali padėti įmonėms nustatyti optimalų atlyginimą. Be to, didelis skirtumas tarp darbdavio deklaruojamo ir modelio prognozuojamo darbo užmokesčio reikšmių galėtų identifikuoti galimus sukčiavimo atvejus, kai siekiama išvengti mokestinių prievolių sąmoningai nurodant mažesnę atlyginimą. Šiame darbe nagrinėjami gyventojų užimtumo statistinio tyrimo duomenys ir taikomi skirtingi mašininio mokymosi modeliai atlyginimo prognozavimui. Rezultatai rodo, kad darbo užmokestį tiksliausiai prognozuoja medžių ansambliu paremtas gradientinio auginimo metodas – XGBoost.

Raktiniai žodžiai: Darbo užmokestis, prognozavimas, mašininis mokymasis, XGBoost.

1 Įvadas

Darbo užmokesčio prognozavimas leidžia ne tik įvertinti skirtingų rodiklių įtaką atlyginimams, bet gali suteikti ir papildomą praktinę naudą. Įmonės, naudodamos modelius, galėtų tiksliau nustatyti konkurencingą ar optimalų atlyginimą, atsižvelgiant į rinkos sąlygas bei darbuotojų kvalifikaciją. Be to, valstybės institucijoms darbo užmokesčio modeliai gali padėti identifikuoti galimus sukčiavimo atvejus. Pavyzdžiui, kai oficialiai deklaruojamas mažesnis atlyginimas nei yra iš tikrųjų, siekiant išvengti mokestinių prievolių. Atlyginimo, kaip ir kitų ekonominių rodiklių prognozei naudojami statistiniai metodai pavyzdžiui, tiesinė regresija [1], ar jos modifikacijos: LASSO, RIDGE [2]. Tačiau vis dažniau taikomi mašininio mokymosi modeliai, pavyzdžiui atsitiktiniai miškai [3]. Pastaruoju metu, įvairiose srityse, siekiant pasiekti kuo didesnę prognozavimo tikslumą, plačiai taikomi medžiais paremti gradien-

tinio auginimo metodai. Atlikti tyrimai, kuriuose parodyta, kad standartiniams - struktūrizuotiems duomenims, gradientinio medžių auginimo metodai dažnai lenkia bet kuriuos kitus, įskaitant ir giliojo mokymosi modelius [4,5]. Pagrindinis šio tyrimo tikslas – pritaikyti gradientinio auginimo metodą darbo užmokesčio prognozavimui ir palyginti jo tikslumą su kitais dažnai naudojamais regresijos modeliais.

2 Duomenys

Darbe naudojami gyventojų užimtumo statistinio tyrimo duomenys. Tačiau apklausoje maža dalis respondentų nurodo savo pajamas. Todėl apklausos duomenys susiejami su kitu šaltiniu – naudojame požymius iš apklausos kaip prediktorius, o darbo užmokesčio reikšmės ir kiti darbovietės kintamieji gaunami iš administracinio šaltinio.

Nagrinėjami 2024 metų samdomųjų darbuotojų duomenys. Atlikus pradinę analizę, tyrimui buvo atrinkti tokie asmens požymiai kaip išsilavinimas, lytis, profesija, baigtų studijų sritis, amžius, stažas darbovietėje ir dirbamų valandų skaičius, o taip pat papildomi darbovietės kintamieji: darbuotojų skaičius, apskritis, ekonominės veiklos grupė ir vidutinis atlyginimas.

Padalinus duomenis į mokymo ir testavimo aibes santykiu 80:20, mokymo aibę sudaro 4785 įrašai, o testavimo – 1196. Be to, kategoriniai kintamieji perkoduojami pseudokintamaisiais.

3 Metodologija

Tyrime naudojami mašininio mokymosi regresijos modeliai: atraminių vektorių regresorius (AVR) (angl. support vector machine) [6], medžių ansamblių metodai: atsitiktiniai miškai (AM) (angl. random forest) [7] ir XGBoost [8]. Tiesinė regresija taikoma kaip bazinis modelis palyginimui. Siekiant sumažinti išsiskiriančių stebėjimų įtaką rezultatams, modeliuojamos logaritmuotos darbo užmokesčio reikšmės. Ši transformacija yra monotonišė, todėl tinkama interpretavimo atveju, o taip pat imant prognozuotų reikšmių eksponentę, gaunama pradinių darbo užmokesčio reikšmių prognozė. Mašininio mokymosi metodams optimalūs parametrai parenkami taikant gardelės metodą su kryžmine patikra. AVR atveju nagrinėjami skirtingi branduoliai, reguliarizacijos, branduolio koeficiento ir kiti parametrai. AM modelyje optimizuojami tokie hiperparametrai: medžių skaičius, maksi-

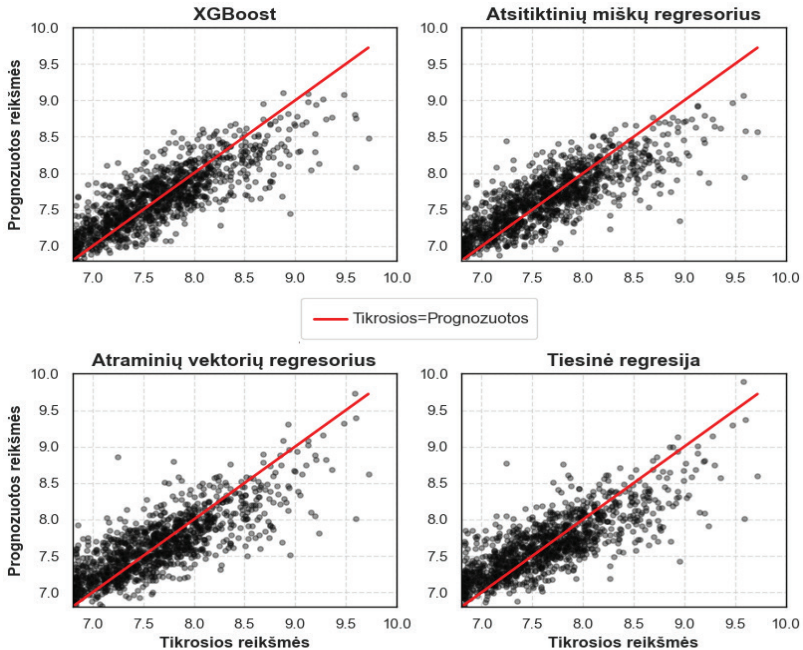
malus medžių gylis, mažiausias stebinių skaičius lape ir dalijime. XGBoost atveju nagrinėjami tokie parametrai kaip medžių skaičius, maksimalus medžių gylis, mokymosi greitis, reguliarizacijos parametras ir kiti. Modelių hiperparametrų parinkimas ir apmokymas atliekamas mokymo aibėje, o jų tinkamumas vertinimas testavimo aibėje pagal keturias metrikas: vidutinė kvadratinė paklaida (MSE), vidutinė absoliutinė paklaida (MAE), Pirsono (angl. Pearson) koreliacijos koeficientas (ρ) ir determinacijos koeficientas (R^2).

4 Rezultatai

Toliau pateikiamas modelių tinkamumo vertinimas testavimo aibėje pagal anksčiau nurodytas metrikas (žr. 1 lentelę ir 1 pav.). Raudonos linijos 1 pav. vizualiai parodo modelių tinkamumą, t.y. linija vaizduoja situaciją, kai prognozuojamos reikšmės idealiai sutampa su tikrosiomis, kuo arčiau taškai šios linijos, tuo modelis tinkamesnis. Pastebime, jog mašininio mokymosi metodų atveju gauname geresnius rezultatus nei naudojant tiesinę regresiją. Taip yra todėl, nes šie metodai gali įvertinti sudėtingesnius, netiesinius sąryšius tarp kintamųjų. Tiksliausiai prognozuojama naudojant XGBoost modelį, kurio visų metrių reikšmės testavimo aibėje yra geriausios. Vis dėlto nė vienas modelis neprognozuoja darbo užmokesčio labai tiksliai, kas gali būti susiję su nepakankamai informatyviais požymiais. Pavyzdžiui, darbo stažas apibūdina tik patirtį dabartinėje darbovietėje, bet ne visą profesinę patirtį. Be to, nebuvo turima tiksli darbovietės vietovė – naudojant tik apskrities informaciją, gali būti neišskirtos didžiųjų miestų darbovietės, kuriose atlyginimai dažniausiai yra didesni.

1 lentelė. Modelių tinkamumo vertinimas testavimo aibėje

Modelis	MSE	MAE	ρ	R^2
XGBoost	0,0933	0,2320	0,8279	0,6840
Atsitiktiniai miškai	0,0993	0,2387	0,8165	0,6637
Atraminų vektorių regresorius	0,1073	0,2497	0,7979	0,6366
Tiesinė regresija	0,1103	0,2540	0,7922	0,6263



1 pav. Modelių tikrosios ir prognozuotos reikšmės testavimo aibėje

5 Išvados

Remiantis gautais rezultatais, galima teigti, kad mašininio mokymosi metodai yra tinkamesni darbo užmokesčio prognozavimui nei tiesinė regresija. Geriausią tikslumą pasiekė medžiais paremtas gradientinio auginimo metodas XGBoost, taigi šį modelį siūlome naudoti darbo užmokesčio prognozavimui. Norint pasiekti didesnę tikslumą reikėtų įtraukti daugiau kintamųjų, tačiau tam reikėtų papildomų duomenų, pavyzdžiui, apklausų metu rinkti detalesnę informaciją, įtraukiant klausimus dėl visos profesinės patirties, geografinės vietovės.

Literatūra

- [1] Anuj More, et al. (2021). „PREDICT-NATION Skills Based Salary Prediction for Freshers,“ Proceedings of the 4th International Conference on Advances in Science & Technology (ICAST2021).

- [2] Guangqi Wang. (2022). „Employee Salaries Analysis and Prediction with Machine Learning,” International Conference on Machine Learning and Intelligent Systems Engineering (MLISE), 2022.
- [3] Babasaheb S. Satpute, et al. (2023). „Machine Learning Approach for Prediction of Employee Salary using Demographic Information with Experience” 4th IEEE Global Conference for Advancement in Technology, 1-5.
- [4] Tim Januschowski, et al. (2022). „Forecasting with trees,” International Journal of Forecasting, 38, 1473-1481.
- [5] Ravid Shwartz-Ziv & Amitai Armon. (2022). „Tabular data: Deep learning is not all you need.” Information Fusion, 81, 84-90.
- [6] Harris Drucker, et al. (1997). Support vector regression machines. Advances in Neural Information Processing Systems. 28. 779-784.
- [7] Leo Breiman. (2001) Random Forests. Machine Learning 45, 5-32, 2001.
- [8] Tianqi Chen & Carlos Guestrin (2016). XGBoost: „A scalable tree boosting system,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

Saulės elektrinių matavimų klasifikavimas remiantis NLP ir mašininio mokymusi

Lukas Voveris, Jolita Bernatavičienė

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Duomenų mokslo ir skaitmeninių technologijų institutas,
Akademijos g. 4, Vilnius
lukas.voveris@mif.stud.vu.lt, jolita.bernatavicienne@mif.vu.lt

Santrauka. Sparčiai augant saulės elektrinių skaičiui, didėja ir poreikis efektyviai analizuoti jų veikimo duomenis. Visgi duomenų analizę dažnai apsunkina nevienodas matavimo kintamųjų žymėjimas, kylantis dėl skirtingų įrangos gamintojų, montuotojų ar lokalizacijos. Šiame darbe siūloma automatizuota kintamųjų klasifikavimo metodika, leidžianti standartizuoti žymas pasitelkiant natūralios kalbos apdorojimo technikas ir mašininio mokymosi algoritmus. Tyrime panaudoti duomenys iš 178 saulės elektrinių, o kiekvienas įrašas buvo susietas su viena iš 176 klasių pagal specialiai sukurtą septynių komponentų žymėjimo sistemą. Modelių mokymui buvo suformuoti skaitiniai požymiai, įskaitant TF-IDF reikšmes, matavimo vienetų kodavimą ir kontekstinius požymius, gautus kaimynystės analizės būdu. Rezultatai parodė, kad siūlomi modeliai pasiekia itin didelį klasifikavimo tikslumą (daugeliu atvejų virš 99,9 %), o geriausi rezultatai gauti naudojant sprendimų medžio ir atsitiktinių miškų metodus.

Raktiniai žodžiai: saulės elektrinės, klasifikavimas, natūralios kalbos apdorojimas, TF-IDF, mašininis mokymas, atraminių vektorių mašina, atsitiktiniai miškai, sprendimų medis, logistinė regresija.

1 Įvadas

Pastaraisiais metais visame pasaulyje saulės elektrinių diegimas sparčiai auga, skatinamas visuotinio perėjimo prie atsinaujinančios energijos ir būtinybės mažinti anglies dvideginio emisijas. 2023 metais pasaulinės saulės energijos instaliacijos augimas beveik padvigubėjo, naujai įdiegtos saulės elektrinės pridėjo 447 GW našumo į globalią rinką (palyginti 239 GW 2022 metais). 87 % augimas rodo spartų sektoriaus plėtimąsi ir vis didesnę saulės energijos reikšmę pasaulio energetikos sistemoje [1]. Tačiau, nepaisant didelio šių sistemų potencialo, jų veikimą gali trikdyti įvairūs nenumatyti veiksniai. Pavyzdžiui, saulės modulių efektyvumą mažina šėšėliavimas, dulkės,

netinkama modulio orientacija ir nepalankios oro sąlygos. Be to, aplinkos veiksniai, tokie kaip temperatūra, drėgmė ir vėjo greitis, irgi turi įtakos saulės modulių našumui [2]. Didėjant saulės fotovoltinių (PV) sistemų mastui ir sudėtingumui, tampa itin svarbu efektyviai stebėti jų darbą bei analizuoti surinktus duomenis, siekiant palaikyti našumą ir laiku aptikti gedimus [3]. Esminė šio proceso dalis – duomenų rinkimas iš įvairių komponentų, tokių kaip inverteriai, saulės moduliai ar meteorologinės stotelės. Tačiau bandant apjungti ar palyginti duomenis iš skirtingų elektrinių kyla rimtas iššūkis. Dėl įrangos gamintojų, monitoringo sistemų, įrangos montuotojų ar lokalizacijos skirtumų matavimų kintamųjų pavadinimai dažnai labai skiriasi. Tokie neatitikimai apsunkina duomenų standartizavimą ir tampa kliūtimi atlikti analizę, aptikti anomalijas ar rengti ataskaitas. Siekiant plėsti analizės modelius ar automatizuotas ataskaitas plačiu mastu, nevienodi pavadinimai ir žymėjimai trukdo automatiškai identifikuoti reikalingus kintamuosius. Dėl to būtina individualizuoti sprendimus beveik kiekvienai elektrinei, o tai apsunkina sprendimų taikymą platesniu mastu. Norint spręsti šią problemą, vis daugiau dėmesio skiriama unifikuotoms kintamųjų žymėjimo sistemoms kurti. Vienas iš tokių sprendimų yra SunSpec standartas, kuris apibrėžia bendrus parametrus ir nustatymus, skirtus stebėti ir valdyti paskirstytos energijos šaltinius [4]. Vis dėlto šis standartas orientuotas į „Modbus“ protokolą, todėl jo taikymas kitų tipų įrenginiams gali būti ribotas. Be to, SunSpec ir panašūs standartai pasižymi didžiausiu efektyvumu tuomet, kai jie yra įgyvendinami jau elektrinės projektavimo etapu, nes tai leidžia užtikrinti sistemų suderinamumą bei optimalų duomenų valdymą. Tačiau senesnėse elektrinėse, kurios buvo įdiegtos be šių standartų, būtina peržiūrėti ir pritaikyti esamas sistemas, siekiant suderinti kintamųjų pavadinimus su SunSpec ar kitais standartais. Tai reikalauja papildomų pastangų ir išteklių, nes kiekviena elektrinė gali turėti unikalių parametrų ir konfigūracijų, kuriuos reikia individualiai pritaikyti prie standartizuotos sistemos. Statistiniai metodai ir mašininio mokymosi modeliai gali reikšmingai prisidėti prie esamų matavimo pavadinimų standartizavimo proceso, leidžiant efektyviau atpažinti matavimų tipus pagal jų pavadinimus, matavimo vienetus ar prietaisų pavadinimus [5]. Tokios sistemos gali stipriai sumažinti rankinį darbą, padidinti duomenų integracijos mastą ir atverti kelią pažangesnei skirtingų saulės elektrinių analitikai. Šiame darbe pristatomas šios problemos sprendimas – siūlomas duomenų žymėjimo standartas bei metodas atpažinti ir klasifikuoti automatinio matavimo kintamuosius, kuris leidžia esamą žymėjimą transformuoti į naujai siūlomą standartą. Tyrimas inicijuotas UAB Elseta.

2 Duomenys

Tyrimui pasitelkti duomenys iš 178 saulės elektrinių. Kiekvienas įrašas atspindėjo vienos elektrinės specifinio komponento (pvz., inverterio, skaitiklio ar kt.) vieną matavimo kintamąjį. Bendrai buvo analizuojami 20 225 duomenų rinkinio įrašai. Kiekvienas įrašas turėjo penkis atributus: matavimo identifikacinį numerį, matavimo pavadinimą, šaltinio (prietaiso) pavadinimą, matavimo vienetus bei objekto identifikacinį numerį. Remiantis SunSpec standartizavimo principais [4], buvo sukurta personalizuota žymėjimo sistema. Kiekviena žyma sudaryta iš septynių dalių, kurios atskiriamos apatiniu brūkšniu (_). Jei tam tikra žymos dalis nėra taikoma konkrečiam matavimui, ji žymima brūkšneliu (-). Žymos struktūra yra nuosekli ir apima šiuos komponentus tokia eilės tvarka: įrenginio tipas, srovės tipas, matavimo tipas, matavimo eilės numeris, agregavimo metodas, matavimo periodas ir informacija apie importą ar eksportą. Tyrime klase laikomas kiekvienas unikalus žymos formatą atitinkantis septynių komponentų derinys. Kitaip tariant, kiekviena unikali žyma (pvz., „INV_String_1_1_1_1_1_1_1“) yra laikoma atskira klase. Pradinis duomenų rinkinys buvo sužymėtas rankiniu būdu, o galutiniame rinkinyje identifikautos 176 skirtingos klasės. Prieš taikant klasifikavimo algoritmus, tekstiniai duomenys buvo apdoroti pasitelkiant natūralios kalbos apdorojimo (NLP) technikas [6]. Tekstas buvo normalizuotas: paverstas mažosiomis raidėmis, pašalinti skyrybos ženklai bei specialieji simboliai. Iš tekstų pašalinti nereikšmingi žodžiai ir atliktas žodžių kamienų išskyrimas. Įrenginio pavadinimuose dažniausiai pasitaikančios įrenginių variacijos (pvz., „inverter“, „meter“) buvo suvienodintos į standartizuotas reikšmes (pvz., „INV“, „Meter“). Kadangi matavimo pavadinime matavimo numeriai, priklausomai nuo prietaiso, gali būti žymimi arba skaičiais, arba raidėmis, šios reikšmės buvo suvienodintos. Raidiniai žymėjimai pakeisti į skaitinius (A = 1, B = 2 ir t.t.), išskirtos ir užkoduotos visos skaitinės reikšmės, išsaugotos atskiru požymiu.

3 Metodologija

Surinkus ir sužymėjus duomenis pagal sukurta žymėjimo sistemą, buvo parengta metodologija, skirta automatizuotai klasifikuoti matavimus, taikant mašininio mokymosi metodus. Siekiant pagerinti modelių prognozavimo tikslumą, buvo sukurta papildomų požymių:

- Indikatoriai, rodantys, ar tekste yra tam tikrų reikšmių (pvz., „INV“, „DC“, „Current“, „Voltage“ ir pan.)

- Tekstiniai duomenys paversti į skaitinę formą naudojant TF-IDF [7] metodą.
- Matavimo vienetų reikšmės užkoduotos paverčiant skirtingas kategorines reikšmes į naujus požymius, kur 1 rodo atitinkamos reikšmės buvimą, o 0 – nebuvimą.

Šie požymiai buvo sujungti į bendrą požymių rinkinį, kuris naudotas višose tolimesnėse modelių mokymų fazėse.

Klasių prognozei, buvo pasirinkta naudoti keturis skirtingus klasifikavimo modelius: logistinę regresiją (LR) [8], sprendimų medį (SM) [9], atsitiktinių miškų (AM) [10] modelį, atraminių vektorių mašiną (SVM) [11]. Klasifikatoriai buvo vertinami taikant penkių dalių kryžminės validacijos metodą [12]. Gautos metrikos yra suskaičiuojamos kiekviename žingsnyje ir imamas jų vidurkis, kad būtų gautas bendras modelio našumo įvertis.

Modelių našumas buvo vertinamas pagal keturis pagrindinius klasifikacijos metrikų rodiklius: tikslumą, preciziškumą, jautrumą ir F1 įvertį. Šios metrikos buvo apskaičiuotos kiekvienai klasei atskirai, o bendra reikšmė gauta naudojant svertinį vidurkį, atsižvelgiant į kiekvienos klasės dydį.

Tikslumas apibūdina visų teisingai klasifikuotų įrašų santykį su bendru įrašų skaičiumi. Preciziškumas rodo, kiek modelio prognozuotų klasės i atvejų buvo teisingi. Jautrumas nurodo, kiek iš visų tikrųjų klasės i atvejų modelis atpažino teisingai. Tuo tarpu F1 įvertis – tai harmoninis preciziškumo ir jautrumo vidurkis, kuris suteikia subalansuotą vertinimą tarp abiejų.

$$Tikslumas = \frac{\sum_i^c (TP(i) + TN(i))}{\sum_i^c (TP(i) + TN(i) + FP(i) + FN(i))}$$

$$Preciziškumas(i) = \frac{TP(i)}{TP(i) + FP(i)}$$

$$Jautrumas(i) = \frac{TP(i)}{TP(i) + FN(i)}$$

$$F1(i) = \frac{2 \cdot Preciziškumas(i) \cdot Jautrumas(i)}{Preciziškumas(i) + Jautrumas(i)}$$

Čia, $TP(i)$ – kiek kartų klasė i buvo teisingai atpažinta kaip klasė i . $TN(i)$ – kiek kartų nei tikroji, nei prognozuota klasė nebuvo i . $FP(i)$ – kiek kartų kita klasė buvo klaidingai atpažinta kaip klasė i . $FN(i)$ – kiek kartų klasė i buvo klaidingai priskirta kitai klasei.

Norint gauti bendrą preciziškumo, jautrumo ir F1 įverčio reikšmę, buvo skaičiuojamas svertinis vidurkis, atsižvelgiant į kiekvienos klasės dydį:

$$\text{Bendras įvertis} = \frac{\sum_{i=1}^C s_i \cdot \text{Metrika}(i)}{\sum_{i=1}^C s_i}$$

Čia, s_i – kiek įrašų priklauso klasei i , $\text{Metrika}(i)$ – atitinkamos metrikos įvertis klasei i .

4 Eksperimentai ir rezultatai

Po duomenų žymėjimo buvo atlikta galimų klasių dažnių analizė, kuri parodė, kad kai kurioms klasėms (pvz. „INV_AC_Q_--_--“) priklauso vos vienas stebėjimas. Tai gali apsunkinti šių klasių patikimą identifikavimą. Apdorojus duomenis, gautas duomenų rinkinys, sudarytas iš 20225 eilučių ir 1819 stulpelių. Modelių vertinimui buvo taikyta penkių dalių kryžminė validacija. Kiekviename validacijos cikle keturios dalys buvo naudojamos modelio mokymui, o likusi – testavimui. Klasifikatorių – LR, AM, SM bei SVM – našumas buvo vertinamas apskaičiuojant vidutines metrikas iš visų penkių validacijos kartojimų.

Iš rezultatų matyti, kad aukščiausią tikslumą pasiekė SVM klasifikatorius, kurio tikslumas siekė 92,56 %. Vis dėlto toks rezultatas nėra idealus. Atlikus prognozuotų ir realių reikšmių analizę, nustatyta, kad didžioji dalis klaidų buvo susijusios su matavimo eilės numerio neatitikimais. Pavyzdžiui, modelis įrašą klasifikavo kaip „INV_String_I_7_--_--“ – septintos inverterio grandinės srovę, kai tikroji reikšmė buvo „INV_String_I_6_--_--“ – šeštos inverterio grandinės srovė. Tokie neatitikimai rodo, kad klasifikatoriui sunku tiksliai identifikuoti aukštesnės eilės grandinės matavimus, kurių stebėjimų skaičius duomenyse buvo nedidelis. Remiantis pirminiais rezultatais, nuspręsta klasifikavimo užduotį suskaidyti į kelias atskiras prognozavimo užduotis, siekiant padidinti tikslumą ir palengvinti interpretaciją. Vietoje vienos sudėtinės klasės buvo suformuotos septynios atskiros kategorijos: įrenginio tipas („Device“), srovės tipas („Current“), matavimo tipas („Type“), matavimo eilės numeris („InputId“), agregavimo metodas („Aggregation“), matavimo periodas („Period“) ir informacija apie importą ar eksportą („Import/Export“). Duomenų rinkinio analizėje kiekviena kategorija buvo atskirai prognozuojama taikant keturis skirtingus klasifikatorius – LR, SM, AM ir SVM. Modelių našumo vertinimui buvo pasitelktas penkių dalių kryžminės validacijos metodas. Galutiniai metrių įverčiai buvo apskaičiuoti kaip visų

validacijos ciklų vidurkiai. Visų klasifikatorių mokymui buvo taikytas vieno-
das požymių rinkinys, tačiau kiekvienos kategorijos atveju, modeliai buvo
apmokomi prognozuoti tos kategorijos klases.

Iš rezultatų matyti, kad visų septynių kategorijų klasifikatorių našumas
yra ypač aukštas – beveik visais atvejais tikslumas, preciziškumas, jautrumas
ir F1 rodikliai viršija 99 %. Vienintelio matavimo eilės numerio klasifi-
kavimas geriausiai atliekamas naudojant SM klasifikatorių, kuris pasiekė
92,58 % tikslumą. Atliekant klaidų analizę, paaiškėjo, kad dažniausiai klaidų
kyla, kai tam pačiam įrenginiui priskiriami keli vienodo tipo matavimai su
tapačiu matavimo eilės numeriu. Tokiais atvejais modelis sunkiai atskiria
panašius įrašus. Kitas dažnas netikslumų šaltinis – fazių matavimai, kurie
turėtų būti žymimi kaip A-B, B-C, C-A (atitinkamai 1-2, 2-3, 3-1). Visgi kla-
sifikavimo metu jie neretai interpretuojami tik pagal pirmąją fazę, t. y. 1, 2
arba 3.

Siekiant padidinti tikslumą ir išspręsti šias problemas, buvo nuspręsta
atlikti kelis duomenų apdorojimo etapo pakeitimus. Pirmiausia, frazių gru-
pių nominalūs žymėjimai buvo pakeisti skaitinėmis išraiškomis. Siekiant
aiškiau atskirti fazių numerius nuo įprastų numerių žymėjimo, buvo atlikti
tokie pakeitimai:

- A-B → 100-200
- B-C → 200-300
- C-A → 300-100

Šios reikšmės pasirinktos tam, kad fazių žymėjimai būtų aiškiai atskir-
ti nuo kitų numeravimo reikšmių. Bandant spręsti vienodų eilės numerių
vienodo matavimo tame pačiame įrenginyje problemą, buvo atlikta kaimy-
nystės analizė. Kiekvienam įrašui priskirti panašūs įrašai (kaimynai), atsižvel-
giant į:

- Tą patį objekto identifikacinį numerį
- Sutampantį prietaiso pavadinimą
- Tokį patį matavimo pavadinimą (be skaičių)

Tokiu būdu buvo siekiama identifikuoti įrašus, kurie turi tą pačią struk-
tūrą ar šabloną, tačiau skiriasi tik skaitinėmis reikšmėmis (pvz. įtampos fa-
zėse ar grandinėse). Randami kiekvieno įrašo kaimynai ir išgaunamos jų
skaitinės reikšmės. Šios reikšmės yra užkoduojamos ir paverčiamos į viene-
tinę reprezentaciją – sukuriama binariniai požymiai, tokie kaip neighbor_1,

neighbor_3, kuriuose reikšmė 1 reiškia, kad įrašas turi kaimyną su atitinkamu skaičiumi.

Įrašui priskiriamas dvejetainis požymis, kuris įgyja reikšmę 1, jei bent vienas kaimynas turi tą pačią skaitinę reikšmę kaip ir pats įrašas, ir 0 – jei tokių sutapimų nėra. Papildomai apskaičiuojamas ir kiekybinis požymis, rodantis, kiek kitų įrašų toje pačioje kaimynystėje turi identišką skaitinę reikšmę. Šios papildomos savybės leidžia klasifikatoriams geriau įvertinti ne tik kiekvieno įrašo turinį atskirai, bet ir jo kontekstą – tai yra, kaip jis susijęs su kitais panašiais to paties įrenginio ar sistemos įrašais.

Atlikus šiuos duomenų apdorojimo patobulinimus, modeliai dar kartą buvo apmokyti, naudojant tuos pačius mokymo duomenis, ir įvertinti tais pačiais metodais bei metrikomis. Rezultatai pateikiami 1 lentelėje. Iš jų matyti, kad matavimų eilės numerių klasifikavimo tikslumas reikšmingai padidėjo – nuo 92,58 % iki 99,76 %.

Geriausius rezultatus gauti naudojant SM, AM ir SVM klasifikatorius. Prietaiso tipas buvo prognozuotas 99,99±0,01 % tikslumu, srovės tipo prognozės AM modeliu pasiektos 99,93±0,04 % tikslumu. Tipas – taip pat AM modeliu – 99,92±0,05 %. Matavimo numeris, agregavimo lygis ir importo ar eksporto požymis geriausiai prognozuoti SM modeliu, atitinkamai 99,76±0,09 %, 99,97±0,04 %, 99,99±0,01 % tikslumu. Matavimo periodas – SM, AM ir SVM modeliais – taip pat 99,95±0,02 % tikslumu.

Geriausiams kiekvienos kategorijos modeliams buvo atliktas hiperparametrų tyrimas, naudojant gardelinės paieškos metodą. SM modeliui naudota ši hiperparametrų gardelė: *max_depth* = [None, 10, 20, 30], *min_samples_split* = [2, 5, 10], *min_samples_leaf* = [1, 2, 4]. AM modeliui – *n_estimators* = [100, 200], *max_depth* = [None, 10, 20], *min_samples_split* = [2, 5], *min_samples_leaf* = [1, 2]. Prietaiso tipo, srovės tipo ir periodo klasifikavimui optimalus buvo AM modelis su parametrais: *max_depth*=None, *min_samples_leaf*=1, *min_samples_split*=2, *n_estimators*=100. Matavimo tipo klasifikavimui taip pat geriausiai pasirodė AM modelis, tačiau su *min_samples_split*=5. Matavimo eilės numeriui, agregacijos tipui ir žymai apie importą/eksportą tiksliausiai klasifikuoti pavyko SM modeliams su parametrais: *max_depth*=None, *min_samples_leaf*=1, *min_samples_split*=2.

1 lentelė. Klasifikavimo rezultatai skirtingoms kategorijoms po patobulinimo.

Kategorija	Metodas	Tikslumas	Preciziškumas	Jautrumas	F1
Device	LR	0,9992 ± 0,0001	0,9992 ± 0,0001	0,9992 ± 0,0001	0,9992 ± 0,0001
	SM	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002
	AM	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002
	SVM	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002
Current	LR	0,9906 ± 0,0005	0,9909 ± 0,0004	0,9906 ± 0,0005	0,9906 ± 0,0005
	SM	0,9983 ± 0,0010	0,9983 ± 0,0010	0,9983 ± 0,0010	0,9983 ± 0,0010
	AM	0,9993 ± 0,0004	0,9993 ± 0,0004	0,9993 ± 0,0004	0,9993 ± 0,0004
	SVM	0,9918 ± 0,0012	0,9922 ± 0,0011	0,9918 ± 0,0012	0,9918 ± 0,0012
Type	LR	0,9979 ± 0,0006	0,9978 ± 0,0005	0,9979 ± 0,0006	0,9987 ± 0,0003
	SM	0,9987 ± 0,0003	0,9987 ± 0,0003	0,9987 ± 0,0003	0,9987 ± 0,0003
	AM	0,9992 ± 0,0005	0,9992 ± 0,0005	0,9992 ± 0,0005	0,9992 ± 0,0005
	SVM	0,9991 ± 0,0003	0,9991 ± 0,0003	0,9991 ± 0,0003	0,9991 ± 0,0003
InputId	LR	0,9764 ± 0,0027	0,9777 ± 0,0028	0,9764 ± 0,0027	0,9760 ± 0,0026
	SM	0,9976 ± 0,0009	0,9976 ± 0,0009	0,9976 ± 0,0009	0,9976 ± 0,0009
	AM	0,9970 ± 0,0011	0,9970 ± 0,0011	0,9970 ± 0,0011	0,9970 ± 0,0011
	SVM	0,9915 ± 0,0016	0,9917 ± 0,0015	0,9915 ± 0,0016	0,9915 ± 0,0016
Aggregation	LR	0,9993 ± 0,0005	0,9993 ± 0,0005	0,9993 ± 0,0005	0,9993 ± 0,0005
	SM	0,9997 ± 0,0004	0,9997 ± 0,0004	0,9997 ± 0,0004	0,9997 ± 0,0004
	AM	0,9994 ± 0,0005	0,9994 ± 0,0005	0,9994 ± 0,0005	0,9994 ± 0,0005
	SVM	0,9996 ± 0,0004	0,9996 ± 0,0004	0,9996 ± 0,0004	0,9996 ± 0,0004
Period	LR	0,9998 ± 0,0002	0,9998 ± 0,0002	0,9998 ± 0,0002	0,9998 ± 0,0002
	SM	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002
	AM	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002
	SVM	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002
Import/Export	LR	0,9994 ± 0,0002	0,9990 ± 0,0004	0,9994 ± 0,0002	0,9992 ± 0,0003
	SM	0,9999 ± 0,0001	0,9999 ± 0,0001	0,9999 ± 0,0001	0,9999 ± 0,0001
	AM	0,9997 ± 0,0002	0,9996 ± 0,0004	0,9997 ± 0,0002	0,9996 ± 0,0003
	SVM	0,9999 ± 0,0002	0,9998 ± 0,0003	0,9999 ± 0,0002	0,9998 ± 0,0003

5 Išvados

Šiame darbe pristatyta metodologija, skirta automatizuotam saulės elektrinių matavimo kintamųjų klasifikavimui, pasitelkiant NLP technikas bei mašininio mokymosi algoritmus. Tyrimo rezultatai atskleidė, kad, nepaisant didelės duomenų įvairovės ir nevienodo skirtingų elektrinių žymėjimo, tin-

kamai parinkus duomenų paruošimo strategijas ir požymių inžinerijos metodus, galima pasiekti ypač aukštą klasifikavimo tikslumą.

Iš pradžių modeliai buvo mokomi klasifikuoti visas žymas kaip vientisus vienetus, tačiau tai lėmė klaidas specifinėse vietose, ypač prognozuojant matavimo eilės numerius. Siekiant padidinti prognozių tikslumą ir pagerinti interpretaciją, klasifikavimo užduotis buvo suskaidyta į septynias atskiras kategorijas, kurios atitiko žymų struktūrą. Toks sprendimas leido kiekvieną žymos komponentą prognozuoti atskirai, ženkliai padidinant bendrą tikslumą ir lankstumą.

Be to, atlikti papildomi duomenų paruošimo patobulinimai – fazių žymėjimo standartizavimas bei kaimynystės analizės pagrindu sukurti papildomi požymiai – dar labiau padidino prognozių tikslumą. Po šių patobulinimų klasifikavimo tikslumas sudėtingiausioje kategorijoje – matavimo eilės numeryje – išaugo nuo 92,58 % iki 99,76 %, o kitos kategorijos pasiekė virš 99,9% tikslumą.

Šie rezultatai rodo, kad automatizuotas matavimo kintamųjų klasifikavimas, pagrįstas NLP technikomis ir mašininio mokymosi modeliais, gali būti veiksmingai taikomas realiose situacijose, ypač integruojant nevienodai sužymėtus duomenis iš skirtingų saulės elektrinių. Tokie sprendimai gali stipriai sumažinti rankinį darbą, pagreitinti analizės procesus ir sudaryti sąlygas pažangesnėms stebėsenos bei ataskaitų sistemoms saulės energijos sektoriuje.

Padėka: Tyrimas finansuojamas pagal LR Švietimo, mokslo ir sporto ministerijos programą „Universitetų ekselencijos iniciatyvos“ (LR ŠMSM mokslo plėtros programos pažangos priemonė Nr. 12-001-01-01-01 „Gerinti mokslo ir studijų aplinką“).

Literatūra

- [1] Walburga Hemetsberger, Michael Schmela, Sonia Dunlop. SolarPower Europe (2024): Global Market Outlook for Solar Power 2024-2028.
- [2] Vidyanandan, K. V. (2017). An overview of factors affecting the performance of solar PV systems. *Energy Scan*, 27(28), 216.
- [3] Avishek, M., Zahnd, A., Mckay, K. H., & Ellul, A. (2009). The importance of monitoring and performance: Analysis of a rural solar PV electrification project. In Solar09, the 47th ANZSES Annual Conference.
- [4] SolarEdge Inverters - SunSpec Logging, Technical Note. Version 3.0, March 2025. <https://knowledge-center.solaredge.com/sites/kc/files/sunspec-implementation-technical-note.pdf>

- [5] Nicchiotti, G., Fromageat, L., & Etienne, L. (2016). Machine Learning Strategy for Fault Classification Using Only Nominal Data. PHM Society European Conference, 3(1). <https://doi.org/10.36001/phme.2016.v3i1.1581>
- [6] Leila Rouka, Cemal Köse, and Parham Tofghi, „Using Machine Translators in Textual Data Classification,“ *International Journal of Computer and Communication Engineering* vol. 1, no. 1, pp. 43-46 , 2012.
- [7] Fauzi, M., Arifin, A., & Yuniarti, A. (2017). Arabic book retrieval using class and book index based term weighting. *International Journal of Electrical and Computer Engineering (Ijece)*, 7(6), 3705. <https://doi.org/10.11591/ijece.v7i6.pp3705-3710>
- [8] LaValley, M. P. (2008). Logistic regression. *Circulation*, 117(18), 2395-2399.
- [9] De Ville, B. (2013). Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(6), 448-455.
- [10] Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.
- [11] Suthaharan, S., & Suthaharan, S. (2016). Support vector machine. *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, 207-235.
- [12] Berrar, D. (2019). Cross-validation.

Skaidymo metodų vertinimas skirtingiems duomenų bazių tipams

Anton Zagzin, Pijus Zlatkus, Vasilij Savin

Vilniaus Universitetas, Matematikos ir informatikos fakultetas,
Informatikos institutas,
Didlaukio g. 47, Vilnius
anton.zagzin@mif.stud.vu.lt, pijus.zlatkus@mif.stud.vu.lt,
vasilij.savin@mif.vu.lt

Santrauka. Straipsnyje vertinami maišos (angl. *hash-based*) ir diapazoninio (angl. *range-based*) duomenų skaidymo metodai PostgreSQL (su Citus plėtiniais) ir MongoDB duomenų bazėse. Eksperimentai atlikti automatizuojant duomenų įkėlimą bei vykdant įvairaus sudėtingumo užklausas naudojant JMeter įrankį. Rezultatai parodė, jog maišos metodas yra efektyvesnis izoliuotoms užklausoms, tuo tarpu diapazoninis skaidymas geriau tinka intervalinėms užklausoms. Pastebėtas ryškus PostgreSQL su Citus pranašumas automatizuojant klasterio operacijas, lyginant su MongoDB. Tyrimo rezultatai gali būti paveikti pasirinktų duomenų schemos specifikos ir riboto duomenų kiekio.

Raktiniai žodžiai: PostgreSQL, MongoDB, Citus, duomenų bazių skaidymas, JMeter, maišos skaidymas, diapazoninis skaidymas.

1 Įvadas

Augant duomenų kiekiui ir naudotojų skaičiui, didėja efektyvaus duomenų valdymo poreikis. Vienas iš pagrindinių būdų padidinti duomenų bazių našumą ir mastelį yra horizontalusis skaidymas (angl. *sharding*) – duomenų padalinimas tarp kelių serverių pagal tam tikrą raktą, užtikrinantis apkrovos balansavimą ir galimybę sistemą plėsti horizontaliai [1], [2].

Tyrimo pasirinktos dvi skirtingų tipų duomenų bazių valdymo sistemos: PostgreSQL – atvirojo kodo **reliacinė** DBVS, pasižyminti patikimumu ir SQL suderinamumu [2]; MongoDB – **dokumentinė** NoSQL duomenų bazė, palaikanti horizontalų mastelio didinimą ir replikaciją [3]. Duomenų paskirstymui PostgreSQL aplinkoje naudotas Citus plėtinys, leidžiantis realizuoti skaidymą ir paskirstytas užklausas [4]. Visos sistemos paleistos virtualiose mašinos, naudojant Docker konteinerizacijos platformą – tai užtikrina lengvą komponentų izoliavimą ir konfigūracijos perkeliamumą [5].

2 Tyrimo metodologija

Eksperimentams buvo parengta OpenNebula virtualizacijos platforma. Eksperimentinė aplinka sudaryta iš virtualių mašinų (VM), naudojančių Ubuntu 24.04 LTS operacinę sistemą. Naudoti du skirtingi duomenų bazių tipai – PostgreSQL (17 versija su Citus 13.0.3 plėtinio [4]) ir MongoDB (4.4.29 versija) [6].

Virtualių mašinų charakteristikos:

- PostgreSQL (vieno serverio konfigūracija): 2 CPU, 8 vCPU branduoliai, 10 GB RAM, 100 GB vietos saugykloje (ext4 failų sistema).
- PostgreSQL (klasterio pagrindinis serveris): 2 CPU, 8 vCPU branduoliai, 10 GB RAM, 100 GB vietos saugykloje.
- PostgreSQL (klasterio mazgai): keturios VM, kiekviena su 1 CPU, 4 vCPU branduoliais, 5 GB RAM ir 50 GB vietos saugykloje.
- MongoDB (vieno serverio konfigūracija): 2 CPU, 8 vCPU branduoliai, 10 GB RAM, 100 GB vietos saugykloje.
- MongoDB (klasterio konfigūracija): maršrutizatorius (1 CPU, 4 vCPU, 5 GB RAM, 50 GB vietos), konfigūraciniai serveriai (1 CPU, 4 vCPU, 5 GB RAM, 50 GB vietos), du duomenų mazgai (po 2 CPU 8 vCPU, 10 GB RAM, 100 GB vietos)

Visos VM naudoja Docker 26.1.3 versiją, skirtą automatizuoti programinės įrangos diegimą ir valdymą, kas leidžia efektyviai valdyti eksperimentines aplinkas, užtikrinant konfigūracijų vientisumą ir atkuriamumą [5], [2].

MongoDB klasterio architektūrai pasirinktas išskaidytas dizainas, atitinkantis tipinį replikacijos ir skaidymo modelį, kuris plačiai aprašytas mokslinėje literatūroje kaip tinkamas našumo ir atsparumo pusiausvyrai [7].

3 Duomenų struktūra ir schema

Duomenų rinkiniui pasirinkta Point-of-Sale (PoS) tipo duomenų schema. PostgreSQL lentelių struktūra išliko tokia pati tiek vieno serverio, tiek klasterio atveju, skirtumas tik tas, kad klasteryje lentelės buvo paskirstytos naudojant skaidymo metodus. MongoDB sistemai buvo sukurta analogiška dokumentų struktūra, laikantis dokumentų orientuotos paradigmos ir pritaikius JSON schemas validatorių. Esminių duomenų modelio pakeitimų, išskyrus formos keitimą iš lentelės į dokumentus, atlikta nebuvo.

Skaidymo metodai:

- Maišos (angl. *hash*) metodas: Duomenys skaidomi pagal įmonės identifikatorių (*tenant_id*). Šis metodas buvo naudotas abiejose sistemose (PostgreSQL su Citus ir MongoDB klasteryje).

- Diapazoninis (angl. *range*) metodas: Duomenys skaidomi pagal įmonės identifikatorių (*tenant_id*).

Raktas *tenant_id* pasirinktas dėl savo gebėjimo užtikrinti pakankamą entropiją (skirtingų reikšmių pasiskirstymą) tarp įrašų – tai rekomenduojama literatūroje kaip viena iš pagrindinių skaidymo rakto savybių [1], [7].

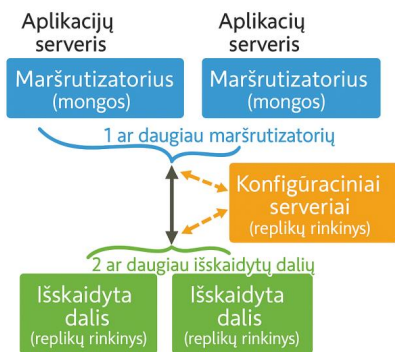
Citus klasterį sudaro koordinatoriaus mazgas, kuris valdo metaduomenis apie duomenų fragmentų (išskaidytų dalių) išdėstymą ir nukreipia užklausas, bei vienas ar daugiau darbininkų mazgų. Darbininkų mazgai saugo faktinius išskaidytų lentelių duomenis šiose išskaidytose dalyse ir atlieka didžiąją dalį duomenų apdorojimo bei skaičiavimų. [4]

PostgreSQL atveju duomenų lentelės sukurtos pagal standartinius SQL DDL sakinius su atitinkamais apribojimais ir sąsajomis. Maišos skaidymo atveju lentelės buvo paskirstytos klasteryje naudojant komandą *create_distributed_table* pagal *tenant_id*. Diapazoninio skaidymo atveju buvo panaudotas schema pagrįstas skaidymas (angl. *Schema-based sharding*), leidžiantis paskirstyti duomenis pagal schemas, kurios buvo paskirstytos pagal *tenant_id*. Taip Citus klasteryje pavyko imituoti diapazoninį skaidymą.

MongoDB išskaidytą klasterį sudaro:

- Išskaidytos dalys (angl. *Shards*): Kiekviena saugo dalį visų duomenų. Būtinai turi veikti kaip replikų rinkinys (angl. *replica set*).
- Maršrutizatorius (mongos): Veikia kaip užklausų nukreipėjas, per kurį aplikacijos jungiasi prie viso telkinio.
- Konfigūraciniai serveriai: Laiko telkinio metaduomenis ir nustatymus. Būtinai turi veikti kaip replikų rinkinys (CSRS).

1 paveikslėlis rodo, kaip šie komponentai sąveikauja tarpusavyje [6].



1 pav. MongoDB išskaidyto klasterio komponentų sąveika

MongoDB dokumentai buvo saugomi atskirose kolekcijose, naudotas JSON schemas validatorius duomenų vientisumui užtikrinti. Duomenų klasteryje naudojamas MongoDB maišos skaidymo mechanizmas pagal *tenant_id* ir diapazoninio skaidymo mechanizmas pagal *tenant_id*.

4 Duomenų įkėlimas

Eksperimentinių duomenų įkėlimas į PostgreSQL ir MongoDB duomenų bazes buvo automatizuotas naudojant Apache JMeter įrankį [8]. Naudojant parengtą testavimo scenarijų, buvo užpildytos tiek atskiros, tiek skaidytos (maišos ir diapazoniniu būdu) duomenų bazių konfigūracijos. Duomenų rinkinį sudarė lentelės ir kolekcijos, imituojančios realius Point-of-Sale (PoS) sistemos duomenis, įskaitant užsakymus, klientus, mokėjimus ir kitus susijusius objektus. Įrašų kiekis ir dydis lentelėse bei kolekcijose skyrėsi priklausomai nuo pasirinktos skaidymo strategijos ir duomenų bazės tipo.

Lentelių dydžiai PostgreSQL duomenų bazėje (vieno serverio konfigūracija) svyravo nuo kelių kilobaitų (mažiausios lentelės, pvz., *discounts* – 8 kB, *loyalties* – 1,5 MB) iki kelių gigabaitų (didžiausia lentelė *order_items* – 8,9 GB). MongoDB vieno serverio konfigūracijoje didžiausia kolekcija buvo tenants, kurios bendras dydis siekė apie 2 GB (225 MB duomenų ir 149 MB indeksų), tačiau dėl ryšio sutrikimų kai kurios lentelės (pvz., *bookings*, *order_items*, *discounts*) liko tuščios ir buvo testuojama su užpildytomis lentelėmis.

PostgreSQL su Citus maišos skaidymas sumažino didžiausios lentelės *order_items* dydį iki 5,85 GB, paskirstant apie 79 mln. įrašų. Analogiškai diapazoninis skaidymas lentelės *order_items* dydį padidino iki 6,41 GB dėl papildomų schemų administravimo duomenų.

MongoDB skaidyto klasterio maišos strategija didžiausią *order_items* kolekciją sumažino iki 610 MB, išskaidant apie 4,2 mln. dokumentų. Diapazoninis skaidymas šią kolekciją išlaikė panašaus dydžio (apie 611 MB), tačiau įrašai taip pat buvo paskirstyti tarp mazgų.

Bendras duomenų įkėlimo greitis buvo didžiausias PostgreSQL vieno serverio konfigūracijoje, tuo tarpu MongoDB vieno serverio duomenų įkėlimas vyko pastebimai lėčiau dėl ryšio stabilumo problemų ir lėtesnio dokumentų apdorojimo greičio.

Preliminarūs duomenų įkėlimo testai parodė, kad:

- PostgreSQL (vienas serveris) ir Citus klasteris buvo efektyviausi pagal duomenų įkėlimo greitį ir stabilumą.

- MongoDB įkėlimas vykdėsi lėčiau, ypač vieno serverio aplinkoje, dėl dokumentinio formato specifikos ir periodinių ryšio sutrikimų.

Tolimesniuose tyrimuose rekomenduojama detaliau analizuoti ir optimizuoti MongoDB duomenų įkėlimo parametrus bei tinklo stabilumą, siekiant išvengti aptiktų ribojimų.

5 Užklausų rezultatų analizė

Norint palyginti skirtingus skaidymo metodus buvo pasirinkta palyginti, kaip skirtingos išskaidytos duomenų bazės efektyviai veikia atliekant įvairius užklausų, apkrovos testus. Taip pat buvo palyginta ir neskaidytų duomenų bazių efektyvumas. Kiekvienas tyrimas buvo atliktas prieš tai kiekvieną duomenų bazę užpildžius apie 150 milijonų įrašų, kad būtų galima imituoti kuo realesnius atvejus, kaip siūloma skaidymo ir apkrovos analizės tyrimuose [7]. Šį tyrimą išskaidėme į šias pagrindines dalis:

5.1 Duomenų nuskaitymo, rašymo ir keitimo užklausų analizė

Šioje dalyje buvo atlikta duomenų skaitymo, rašymo ir keitimo užklausų apdoravimo greitis ir kiekis, taip pat ir klaidų kiekis per 10 minučių. Testai buvo atliekami imituojant 100, 400, 800 ir 1200 virtualių naudotojų, ir skirtingo tipo užklausomis.

1 lentelė. Bendras užklausų ir klaidų santykis skirtingose duomenų bazėse

	PostgreSQL vieno serverio	PostgreSQL maišos funkcija išskaidyto klasterio	PostgreSQL diapazoniniu skaidymu išskaidyto klasterio	MongoDB vieno serverio	MongoDB maišos funkcija išskaidyto klasterio	MongoDB diapazoniniu skaidymu išskaidyto klasterio
Sėkmingos užklausos	88,22 %	90,19 %	94,85 %	90,21 %	94,27 %	96,45 %
Neįvykdytos užklausos	11,78 %	9,81 %	5,15 %	9,79 %	5,83 %	3,55 %

1 lentelėje matyti, kuri dalis užklausų atitinkamose konfigūracijose yra sėkminga ir kuri buvo klaidinga dėl serverio atsako. Vieno serverio konfigūracijos (tiek PostgreSQL, tiek MongoDB) demonstruoja didesnę klaidų skaičių esant didesniai apkrovai. Taip nutinka dėl ribotų resursų: užkla-

soms daugėjant, vienas mazgas tampa tinklo silpna vieta (angl. *bottleneck*), dėl kurio auga klaidos ir atsako vėlinimas. Klasterinėse duomenų bazėse klaidingų užklausų procentas ryškiai mažesnis, nes sistemai padeda duomenų išskaidymas tarp kelių mazgų. Tiek maišos, tiek diapazoninis skaidymas gerina apkrovos paskirstymą.

2 lentelė. PostgreSQL duomenų bazių užklausų kiekis ir atsako laikas

	Visų užklausų kiekis	Atsako laikas $t \leq 500$ ms	Atsako laikas $500 \text{ ms} < t \leq 1500$ ms	Atsako laikas $t > 1500$ ms
Vieno neskaidyto serverio	~30,231 mln.	~26,912 mln. (89 %)	~1,901 mln. (6 %)	~1,418 mln. (5 %)
Citus maišos funkcija išskaidyto klasterio	~17,698 mln.	~15,401 mln. (87 %)	~1,402 mln. (8 %)	~0,895 mln. (5 %)
Citus diapazoniniu skaidymu išskaidyto klasterio	~45,160 mln.	~42,576 mln. (94 %)	~1,672 mln. (4 %)	~0,912 mln. (2 %)

3 lentelė. MongoDB duomenų bazių užklausų kiekis ir atsako laikas

	Visų užklausų kiekis	Atsako laikas $t \leq 500$ ms	Atsako laikas $500 \text{ ms} < t \leq 1500$ ms	Atsako laikas $t > 1500$ ms
Vieno neskaidyto serverio	~63,672 mln.	~58,578 mln. (92 %)	~3,184 mln. (5 %)	~1,910 mln. (3 %)
Maišos funkcija išskaidyto klasterio	~70,039 mln.	~65,837 mln. (94 %)	~2,802 mln. (4 %)	~1,412 mln. (2 %)
Diapazoniniu skaidymu išskaidyto klasterio	~103,785 mln.	~99,634 mln. (96 %)	~1,925 mln. (2 %)	~2,226 mln. (2 %)

Pirmiausia šioje dalyje buvo analizuojama, kaip skirtingos duomenų bazių konfigūracijos – vieno serverio ir išskaidytos (maišos ar diapazoninio) – veikia tada, kai atliekamos skaitymo, rašymo ir keitimo užklausos. Analizuojant 2 ir 3 lenteles, galima pastebėti reikšmingus skirtumus tarp skirtingų duomenų bazių konfigūracijų efektyvumo. 2 lentelėje matome, kad PostgreSQL diapazoniškai skaidytas klasteris pasiekė didžiausią bendrą užklausų kiekį – apie 45,160 mln., iš kurių net 94% (42,576 mln.) buvo apdorojamos greičiau nei per 500 ms. Tuo tarpu vieno neskaidyto serverio konfi-

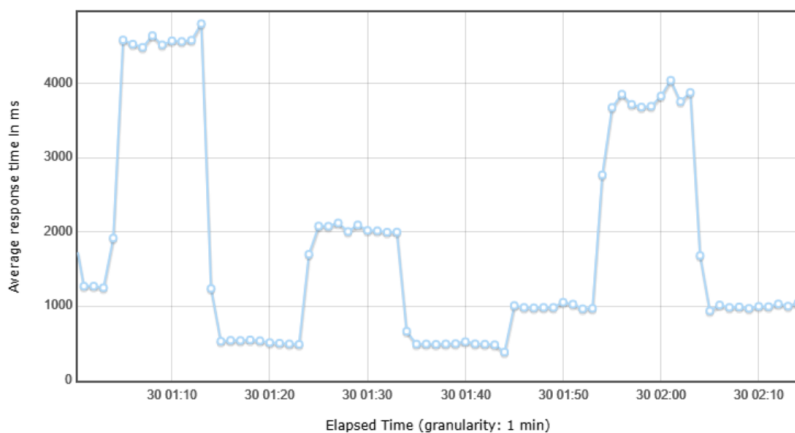
gūracija, nors ir apdorojo apie 30,231 mln. užklausų, tačiau pasiekė tik 89% užklausų su atsako laiku iki 500 ms.

Pastebėtina anomalija 2 lentelėje yra ta, kad Citus maišos funkcija išskaidyto klasterio bendras užklausų kiekis (~17,698 mln.) yra ženkliai mažesnis nei vieno neskaidyto serverio (~30,231 mln.). Toks rezultatas prieštarauja teorinėms prielaidoms, kad skaidymas turėtų padidinti bendrą sistemos pralaidumą. Šis nukrypimas gali būti susijęs su specifine maišos skaidymo implementacija Citus plėtinyje ir jos pritaikymu esamam 79 mln. įrašų duomenų kiekiui.

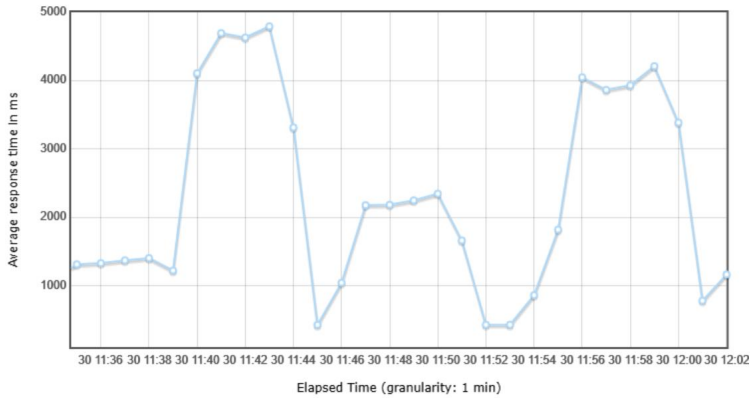
3 lentelėje pastebimos panašios, bet dar ryškesnės tendencijos MongoDB sistemoje. Diapazoniškai skaidytas klasteris pasiekė įspūdingą 103,785 mln. užklausų kiekį, iš kurių 96% (99,634 mln.) turėjo atsako laiką mažesnį nei 500 ms. Maišos funkcija išskaidytas klasteris taip pat demonstravo puikius rezultatus – 70,039 mln. užklausų su 94% (65,837 mln.) greitų atsakymų. Vieno serverio konfigūracija, nors ir parodė gerą našumą (63,672 mln. užklausų), tačiau šis rezultatas yra ženkliai mažesnis nei skaidytų sprendimų.

5.2 Didelės užklausų apkrovos testavimas

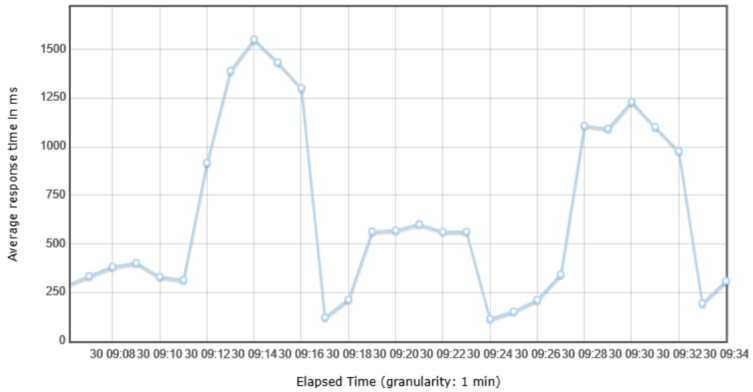
Šis testas buvo pasirinktas imituoti realią pastovią apkrovą ir kaip skirtingai išskaidytos duomenų bazės elgiasi staigiai atsirandant didėliai apkrovai. Iš pradžių testai buvo atlikti imituojant didėjančią naudotojų srautą iki 5000. Ir kartu buvo 500 ir 1000 virtualių naudotojų kiekį ir staigiai jį pakeičiant į 4 kartais didesnę apkrovą (2000 ir 4000 vartotojų).



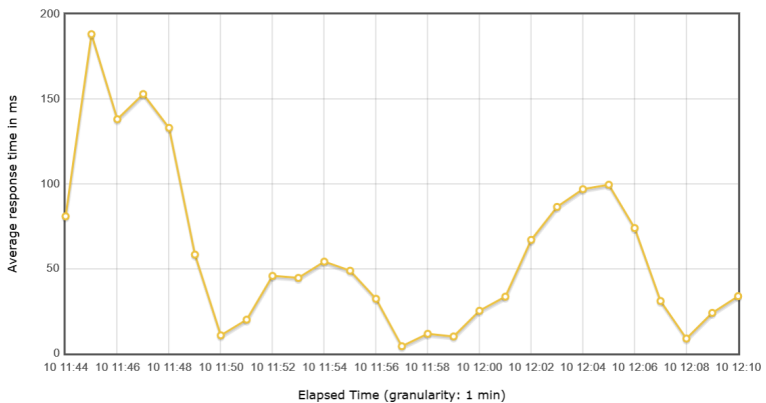
2 pav. PostgreSQL vieno serverio apkrovos testavimas



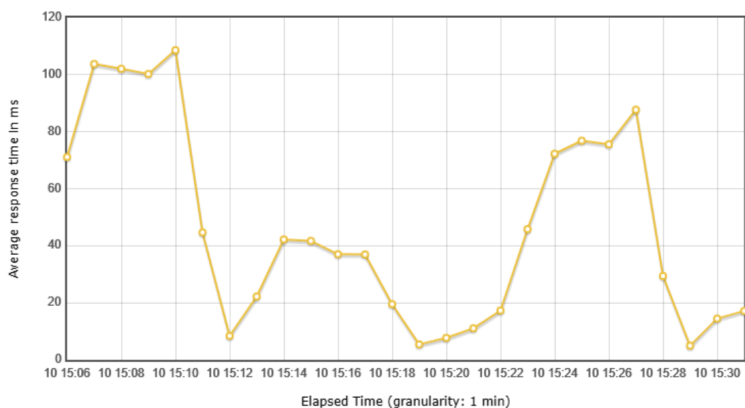
3 pav. PostgreSQL Citus maišos funkcija išskaidyto klasterio apkrovos testavimas



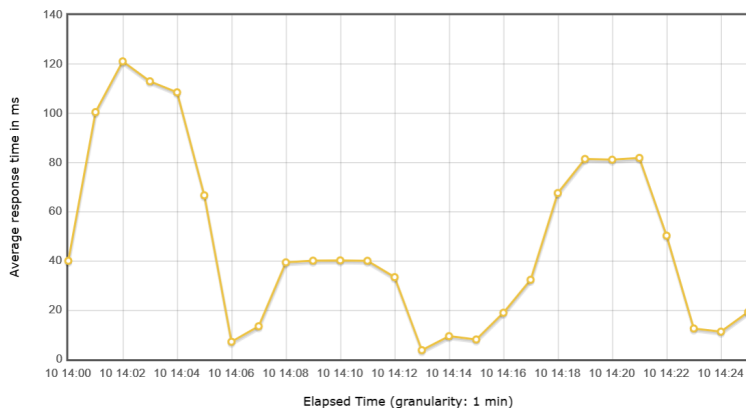
4 pav. PostgreSQL Citus diapazoniniu skaidymu išskaidyto klasterio apkrovos testavimas



5 pav. MongoDB vieno serverio apkrovos testavimas



6 pav. MongoDB maišos funkcija išskaidyto klasterio apkrovos testavimas



7 pav. MongoDB diapazoniniu skaidymu išskaidyto klasterio apkrovos testavimas

2–4 paveikslėliai atspindi PostgreSQL konfigūracijų elgseną esant didelei apkrovai. Vieno serverio (2 pav.) atsako laikas dramatiškai svyruoja nuo 1000 ms iki 4500 ms, maišos skaidymo atveju (3 pav.) išlieka panašūs svyravimai 1000-4700 ms ribose, tačiau diapazoninio skaidymo konfigūracija (4 pav.) demonstruoja ženkliai stabilesnius rezultatus – atsako laikas dažniausiai neviršija 1500 ms. Tai patvirtina diapazoninio skaidymo pranašumą PostgreSQL aplinkoje valdant kintančias apkrovas.

MongoDB konfigūracijų testavimo rezultatai (5–7 pav.) atskleidžia itin efektyvų elgesį – stebima žymiai mažesnė atsako laiko skalė. Vieno serverio

rio atveju (5 pav.) atsako laikas siekia iki 190 ms, maišos funkcija išskaidyto klasterio (6 pav.) – iki 110 ms, o diapazoninio skaidymo (7 pav.) – iki 120 ms. Lyginant su PostgreSQL, MongoDB atsako laikai yra apie 10-20 kartų mažesni, tačiau abiejose sistemose išryškėja ta pati tendencija – diapazoninis skaidymas užtikrina stabilesnį elgesį esant didelei apkrovai.

6 Išvados

1. Diapazoninis skaidymas PostgreSQL/Citus klasteryje leido pasiekti 45,16 mln. užklausų per 10 min. (94 % \leq 500 ms), t. y. 2,5 kartus didesnę pralaidumą nei to paties klasterio maišos skaidymas (17,70 mln.) ir 49 % daugiau nei vieno serverio PostgreSQL (30,23 mln.) – tai paneigia plačiai cituotą prielaidą, kad maišos strategija Citus terpėje yra universaliai efektyviausia.
2. Maišos skaidymas Citus aplinkoje ne tik nespertino, bet 41 % sulėtino bendrą pralaidumą lyginant su neskaidyta PostgreSQL instancija, atskleidamas iki šiol nefiksuotą koordinatoriaus – darbininkų sinchronizacijos kliūtį, pasireiškiančią esant ~79 mln. eilučių ir keturiems mazgams.
3. MongoDB diapazoninis klasteris, apdorojęs 103,79 mln. užklausų (96 % \leq 500 ms), 48 % aplenkė to paties klasterio maišos strategiją (70,04 mln.) ir 63 % vieno serverio MongoDB (63,67 mln.), parodydamas, kad dokumentinei DB diapazoninis skaidymas gali būti pranašesnis nei maišos skaidymas.
4. Staiga padidinus apkrovą keturis kartus (1000 → 4000 vartotojų), diapazoninis skaidymas išlaikė stabilų vėlinimą (PostgreSQL \leq 1 500 ms, MongoDB \leq 120 ms), o maišos skaidymas PostgreSQL svyravo iki 4700 ms.
5. MongoDB klasteryje maksimalus atsako laikas (\leq 190 ms) buvo 10–20 kartų trumpesnis nei analogiškai apkrautame PostgreSQL diapazoniniame klasteryje (\leq 1 500 ms), kiekybiškai pademonstruojant dokumentinio modelio pranašumą PoS infrastruktūroje.
6. PostgreSQL hash skaidymas sumažino order_items lentelę nuo 8,9 GB iki 5,85 GB, tačiau MongoDB hash strategija tą pačią kolekciją (4,2 mln. dokumentų) suspaudė iki 610 MB – 9,6 kartų mažiau baitų vienam įrašui, atskleidama, kad dokumentų saugykloje struktūrinis apkarpymas ir indeksų organizavimas gali lemti ženkliai geresnį talpos efektyvumą nei reliaciniame formate.

7. Tyrimas identifikavo, kad Citus diapazoniniam skaidymui teko naudoti schemomis grįstą imitaciją, kuri nepaisant papildomos metaduomenų naštos (lentelė išaugo iki 6,41 GB) vis tiek pranoko maišos skaidymą pagal pralaidumą ir stabilumą.

Tyrimo ribotumai apima konkrečią duomenų schemą, ribotą įrašų kiekį ir pasirinktą apkrovą. Tolimesniuose tyrimuose rekomenduojama išplėsti skaidymo raktų įvairovę ir taikyti išsamesnes apkrovas, siekiant patikrinti, ar nustatytos tendencijos išlieka skirtingomis sąlygomis.

Literatūra

- [1] M. Tamer Özsu ir P. Valduriez, *Principles of Distributed Database Systems.*, Springer, 2020.
- [2] A. Silberschatz, H. F. Korth ir S. Sudarshan, *Database System Concepts*, McGraw-Hill, 2019.
- [3] H. Jing, E. Haihong, L. Guan ir J. Du, „Survey on NoSQL database,” *2011 6th International Conference on Pervasive Computing and Applications*, pp. 363-366, 2011.
- [4] Microsoft Company, „Citus Documentation v13.0.3,” 2024. [Tinkle]. Available: <https://docs.citusdata.com/en/v13.0/>. [Kreiptasi 2025-03-29].
- [5] Docker Inc., „Docker Docs,” 2025. [Tinkle]. Available: <https://docs.docker.com/>. [Kreiptasi 2025-03-29].
- [6] MongoDB, Inc., „MongoDB Documentation v4.4,” 2024. [Tinkle]. Available: <https://www.mongodb.com/docs/v4.4/>. [Kreiptasi 2025-03-29].
- [7] C. Curino, E. Jones, Y. Zhang ir S. Madden, „Schism: a workload-driven approach to database replication and partitioning,” *Proc. VLDB Endow.*, pp. 48-57, 2010.
- [8] Apache Software Foundation, „Apache JMeter User Manual,” 2024. [Tinkle]. Available: <https://jmeter.apache.org/usermanual/>. [Kreiptasi 2025-03-29].

Multimodalinių modelių taikymas vaizdų antraščių generavimui lietuvių kalba

Airidas Žaliauskas, Viktor Medvedev

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Duomenų mokslo ir skaitmeninių technologijų institutas,
Akademijos g. 4, LT-08412 Vilnius
airidas.zaliauskas@mif.stud.vu.lt, viktor.medvedev@mif.vu.lt

Santrauka. Straipsnyje pristatomas tyrimas, kuriame analizuojamas multimodalinių modelių taikymas automatiniam vaizdų antraščių generavimui lietuvių kalba. Kadangi išsamių lietuvių kalbai skirtų tyrimų šioje srityje iki šiol nėra atlikta, straipsnyje daugiausia dėmesio skiriama naujausiems multimodaliniams modeliams ir jų galimybės generuoti suprantamus vaizdų aprašymus lietuvių kalba. Eksperimentai buvo atliekami naudojant „Gemma 3“ multimodalinį modelį, kuris buvo adaptuotas lietuvių kalbai QLORA metodu. Gauti rezultatai patvirtina metodo efektyvumą ir galimybę sėkmingai jį taikyti lietuvių kalbos turinio generavimui iš nuotraukų.

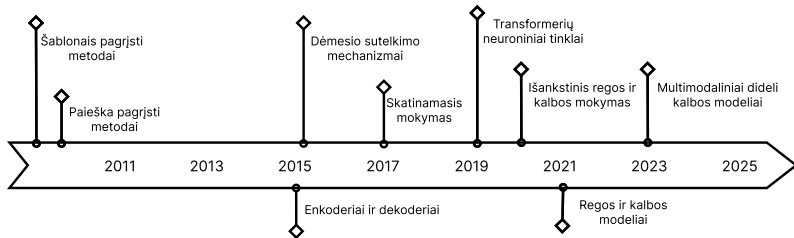
Raktiniai žodžiai: vaizdų antraščių generavimas, kompiuterinė rega, natūralios kalbos apdorojimas, dirbtiniai neuroniniai tinklai, mašininis vertimas, transformeriai, QLORA.

1 Įvadas

Vaizdų antraščių generavimas – tai procesas, kurio metu automatiškai generuojami vaizdų aprašymai natūralia kalba. Tai reikalauja atpažinti objektus, jų požymius ir ryšius paveikslėlyje bei išreikšti juos sakiniu ar sakinių rinkiniu, apibendrinančiu tai, kas pavaizduota. Tai paprastai sujungia kompiuterinės regos (angl. *computer vision*) ir natūralios kalbos apdorojimo metodus. Antraščių generavimas sulaukė didelio dėmesio dėl galimo pritaikymo tokiose srityse kaip prieinamumas (vaizdų aprašymas, turintiems regos sutrikimų), žmogaus ir kompiuterio sąveika [1]. Literatūros analizuojančios antraščių generavimo pritaikymą būtent lietuviškam tekstui iki šiol nebuvo, todėl šiame straipsnyje bus atliekamas tyrimas pritaikyti pažangiausius metodus vaizdų antraščių generavimui lietuvių kalba.

2 Antraščių generavimo metodai

Per pastaruosius du dešimtmečius vaizdų antraščių generavimas transformavosi nuo elementarių, taisyklėmis pagrįstų, sistemų iki sudėtingų multimodalinių modelių, integruojančių kompiuterinę regą ir kalbos supratimą. 1 pav. apžvelgiami didžiausią įtaką antraščių generavimo sričiai turėję akademiniai pasiekimai ir moksliniai atradimai.



1 pav. Reikšmingiausių pasiekimų, vaizdų antraščių generavimo srityje, laiko juostos diagrama.

Prieš pradėdant dominuoti neuroniniams tinklams, vaizdų antraštės buvo kuriamos naudojant šablonais (angl. *template-based*) ir paieška (angl. *retrieval-based*) pagrįstus metodus. Taikant šablonais pagrįstus metodus, pirmiausia buvo nustatomi vaizdo elementai (objektai, atributai ir kt.) ir tada jais buvo užpildomi rankiniu būdu paruošti sakinių šablonai. Nors tai užtikrino, kad antraštėse būtų paminėti konkretūs aptikti objektai, sakinių struktūros buvo fiksuotos, todėl aprašymai buvo nelankstūs ir pasikartojantys, o sakinių įvairovė buvo ribota. Kita vertus, taikant paieška pagrįstus metodus, panaudojama egzistuojanti žmogaus parašytų antraščių duomenų bazė ir ieškoma tokios antraštės, kuri atitiktų naują vaizdą (dažniausiai surandant vizualiai panašų vaizdą). Tai leido sukurti gramatiškai taisyklingas antraštes, tačiau, kadangi buvo galima pasirinkti tik iš jau egzistuojančių pavyzdžių, šiam metodui trūko originalumo ar galimybės aprašyti nematytą vaizdą.

2010-ųjų viduryje įvyko reikšmingas pokytis. Sėkmingai pritaikius enkoderių (angl. *encoder*) ir dekoderių (angl. *decoder*) architektūras mašininio vertimo srityje, neuroniniai modeliai buvo pradėti taikyti vaizdų antraštėms generuoti. Proveržis įvyko 2015 metais, kai buvo išleistas modelis pavadinimu „*Show and Tell*“ [2], kuriame vaizdai koduoti buvo naudojamas konvoliucinis neuroninis tinklas (angl. *Convolutional neural network, CNN*), o tekstui nuose-

kliai dekoduoti – ilgos trumpalaikės atminties tinklas (angl. *Long short-term memory, LSTM*). Šis metodas buvo paprastas, bet ypač veiksmingas ir pranoko ankstesnius šablonais ir paieška pagrįstus metodus. Tais pačiais metais buvo išleistas „*Show, Attend and Tell*“ [3] modelis, kuriame buvo patobulinta ši architektūra panaudojant dėmesio sutelkimo mechanizmus. Šis mechanizmas leido dekoderiui dinamiškai sutelkti dėmesį į konkrečias vaizdo dalis kiekvienam sugeneruotam žodžiui, taip gerokai padidinant ir tikslumą, ir interpretaciją.

2017 metų naujovė – antraščių generavimo modelių optimizavimas, naudojant vertinimo rodiklius, tų modelių rezultatų kokybės vertinimui. Ši strategija gerokai padidino našumą, nes modeliai, adaptuoti naudojant skatinamąjį mokymą (angl. *reinforcement learning*), pasiekė ženkliai aukštesnius kokybės vertinimo rodiklius nei modeliai, apmokyti naudojant tik kryžminę entropiją [4]. Svarbus 2018 m. pasiekimas buvo „*Bottom-Up and Top-Down Attention*“ modelis [5], kuriame buvo pasiūlyta naudoti iš anksto apmokytą objektų aptikimo modelį („*Bottom-Up*“), kad būtų galima nustatyti regiono lygmens vaizdinius požymius, pavyzdžiui, stačiakampius (angl. *bounding boxes*) vaizde esantiems objektams. Tuomet, naudojamas iš viršaus į apačią (angl. *top-down*) dėmesio sutelkimo LSTM, kad dėmesys būtų sutelkiamas į šiuos nustatytus regionus, generuojant antraštes. Šis, į objektus orientuotas metodas, leido modeliams pagrįsti žodžius konkrečiais vaizdo regionais, užuot sutelkiant dėmesį į tolygiu tinkleliu padengtą vaizdą.

Dar 2017 metais išleista novatoriška transformerių neuroninių tinklų architektūra „*Attention is All You Need*“ [6] sukėlė revoliuciją natūralios kalbos apdorojime, pakeičiant rekurentinius neuroninius tinklus dėmesio sutelkimo mechanizmais (angl. *self-attention*). Bėgant laikui, tyrėjai nustatė, jog transformeriai puikiai veikia ne tik kaip kalbos dekoderiai, bet ir kaip kompiuterinės regos enkoderiai, leidžiantys modeliuoti vaizdų ir tekstų ryšius. Tuomet, antraščių generavimo modeliai pakeitė rekurentinius neuroninius tinklus, įdiegiant transformeriais grįstus modelius. Pavyzdžiui, 2019 m. buvo išleistas „*Object Relation Transformer*“ [7], kuris kodavo aptiktų objektų erdvinius ir semantinius ryšius, siekiant padidinti generuojamų antraščių sklandumą. 2020 m. antraščių generavimui buvo atsakyta naudoti konvoliuciniais neuroniniais tinklais pagrįstus metodus, pakeičiant juos regos transformeriais (angl. *Vision transformer, ViT*), sujungtais su transformeriais grįstais neuroninių tinklų dekoderiais.

2021 metais vaizdų antraščių generavimo srityje buvo pradėti naudoti regos ir kalbos modeliai (angl. *Vision-Language Model, VLM*), gebantys susieti vaizdo ir teksto įterpinius bendroje įterpinių reprezentacijų erdvėje (angl.

embedding space), padidindami generuojamų antraščių tikslumą ir rišlumą. 2025 m. kovo 12 d. „Google DeepMind“ pristatė „Gemma 3“ [8] – multimodalinį modelį, pagrįstą regos ir kalbos modelių architektūra, galintį vienu metu apdoroti ir vaizdą, ir tekstą. „Gemma 3“ 27B parametru modelis, atsižvelgiant į jo dydį, pasižymi geresniais rezultatais nei Llama3-405B, DeepSeek-V3 ir o3-mini modeliai, remiantis išankstiniais „LMArena“ žmonių vertinimų reitingais¹.

2023 metais antraščių generavimo srityje buvo pradėti naudoti multimodaliniai dideli kalbos modeliai, pavyzdžiui, GPT-4, galintis sklandžiai derinti vaizdo, teksto ir kalbos modalumus. Šie multimodaliniai modeliai geba interpretuoti sudėtingas vaizdų kompozicijas ar šnekamąją kalbą. Tai leidžia išplėsti jų pritaikymą sudėtingesnėms užduotims spręsti, pavyzdžiui, vaizdų apibūdinimui šnekamąja kalba. Konkrečiai vaizdų antraščių generavimo kontekste multimodaliniai modeliai pasinaudoja turtingu kalbos supratimu. Jie geba kurti tiksliai ir išsamiai, žmogaus rašyseną primenančias antraštes. Pavyzdžiui, 2024 metais „Neurotechnology“ pristatė didelį kalbos modelį lietuvių kalbai [9], kuris ne tik patobulino teksto generavimą lietuvių kalba, bet ir parodė, kaip tokie modeliai gali būti pritaikomi konkrečioms lingvistinėms savybėms. Ateityje, integruojant tokias technologijas į sistemas, gebančias apdoroti ir tekstą, ir vaizdus, bus galima generuoti kokybiškesnes bei tikslesnes antraštes skirtingoms kalboms, įskaitant mažiau paplitusias, kaip lietuvių.

3 Duomenų rinkiniai

„Flickr“ platforma, įkurta 2004 metais, yra internetinė nuotraukų talpinimo ir dalinimosi bendruomenė. Ji suteikė pradžią dviem populiariems duomenų rinkiniams – Flickr8k (8 000 nuotraukų, 40 000 antraščių) [10] ir išplėtam jo variantui Flickr30k (31 783 nuotraukos, 158 915 antraščių) [11]. MS COCO (2014 m.) [12] duomenų rinkinį sudaro 123 287 vaizdai, apimantys 80 objektų kategorijų, anotuoti taip, kad atspindėtų realaus pasaulio scenas natūraliame kontekste. Šio duomenų rinkinio antraštėse (vidutiniškai 52,49 simbolių) pirmenybė teikiama glaustam informacijos perteikimui. Tuo tarpu DOCCI duomenų rinkinys [13] skiriasi struktūriškai – jame 14 847 nuotraukų su pavienėmis žmogaus parašytomis antraštėmis vidutiniškai sudarytomis iš 640,79 simbolių arba maždaug 136 žodžių, t. y. daugiau nei 12 kartų ilgesnėmis nei MS COCO. DOCCI duomenų rinkinio privalumas yra tas, kad

¹ Šis teiginys paremtas „Google DeepMind“ tinklaraščio straipsniu <https://blog.google/technology/developers/gemma-3/>

jo struktūra leidžia kiekvieną vaizdą aiškiai atskirti nuo panašių vaizdų, remiantis jų aprašymu [13].

1 lentelė. Duomenų rinkinių palyginimas.

Pavadinimas	Nuotraukų (vaizdų) skaičius	Antraščių skaičius	Vidutinis antraščių ilgis (simboliais)
Flickr-8k	8 000	40 000	55,13
Flickr-30k	31 783	158 915	64,04
MS COCO (2014)	123 287	616 435	52,49
DOCCI	14 847	14 847	640,79

Pagrindinis iššūkis, pritaikant regos ir kalbos modelius lietuviškam turiniui yra tai, kad šiuose duomenų rinkiniuose pateiktos antraštės yra anglų kalba. Siekiant pašalinti šią kliūtį buvo atliktas keturių mašininio vertimo modelių vertinimas: „Helsinki Opus“, „MADLAD400“ (3 mlrd. ir 7 mlrd. parametrų versijos), „Seamless“ ir „NLLB-200“ Vertimo į lietuvių kalbą kokybė buvo vertinama naudojant Flores-200 duomenų rinkinį pagal tris pagrindinius rodiklius, leidžiančius įvertinti vertimo tikslumą, rišlumą ir prasmę: BLEU, METEOR ir ROUGE-L. Flores-200 duomenų rinkinį sudaro 204 kalbos, įskaitant lietuvių. Angliški sakiniai buvo išversti į lietuvių kalbą naudojant visus anksčiau minėtus vertimo modelius, o rodikliai apskaičiuoti palyginant modelio sugeneruotą tekstą su žmonių pateiktais vertimais į lietuvių kalbą. BLEU – tai populiariausias mašininio vertinimo kokybės vertinimo rodiklis, kuris vertina, kiek išverstas tekstas sutampa su pateiktais pavyzdžiais pagal tikslias žodžių kombinacijas (n-gramas). METEOR – papildo BLEU tuo, kad vertina ne tik tiesioginį žodžių sutapimą, bet taip pat atsižvelgia į sinonimus ir žodžių tvarką. ROUGE-L – vertina vertimo kokybę pagal tai, kaip gerai išverstas tekstas išlaiko originalaus teksto prasmę, vertindamas ilgiausią bendrą žodžių seką [1].

2 lentelė. Mašininio vertimo modelių vertinimo rodiklių rezultatai Flores-200 duomenų rinkiniui.

Pavadinimas	BLEU	METEOR	ROUGE-L
Helsinki Opus	0,2303	0,4595	0,4830
MADLAD400 (3 mlrd.)	0,2236	0,4660	0,4830
MADLAD400 (7 mlrd.)	0,2250	0,4639	0,4833
Seamless	0,1953	0,4243	0,4541
NLLB-200	0,1825	0,4037	0,4360

Kaip matyti 2 lentelėje, „Helsinki Opus“ modelis verčiant iš anglų kalbos į lietuvių kalbą pasiekė geriausius rezultatus, kadangi BLEU balas buvo 0,2303, o ROUGE-L – 0,483. Tai rodo, kad jis efektyviai atlieka anglų-lietuvių kalbų vertimo užduotis. „Maddad400“ (3 mlrd. parametrų) pasiekė konkurencingus rezultatus, ypač pagal METEOR rodiklį (0,466), o tai rodo, kad jis geriau susidoroja su perfrazavimu, kas yra svarbu generuojamų antraščių natūralumui. Tačiau parametrų skaičiaus didinimas iki 7 mlrd. reikšmingo kokybės pagerėjimo nesuteikė (+0,0015 BLEU, -0,002 METEOR), tad daugiau nei 2 kartus daugiau parametrų turinčio modelio rezultatai nepateisina papildomų resursų reikalavimų. Sekantis, „Seamless“ modelis, pasiekė prastesnius rezultatus (BLEU – 0,195), o „NLLB-200“ surinko žemiausią įvertinimą (BLEU – 0,183), jo plati 200 kalbų aprėptis galimai sumenkino pritaikymą vertimui būtent į lietuvių kalbą. Atsižvelgiant į šiuos rezultatus, specializuotas „Helsinki Opus“ modelis yra pranašesnis anglų-lietuvių vertime, nei bendrosios paskirties sistemos.

4 Multimodaliųjų (regos ir kalbos) modelių pritaikymas

Pirmiausia, antraščių generavimui buvo išbandytas „Neurotechnology“ pristatytas didelis kalbos modelis, pritaikytas lietuvių kalbai. Jis sukurtas papildomai apmokant transformeriais grindžiamus „Llama2“ architektūros modelius lietuviškais duomenų rinkiniais [9]. Kadangi „Llama2“ yra tekstinio pobūdžio modelis, ši sistema gali apdoroti tik tekstinę informaciją. Šiai kliūčiai išspręsti buvo pritaikytas „BLIP-2“ [14] metodas, kuris naudoja tarpinį komponentą, vadinamą „Q-Former“. Šis komponentas leidžia sujungti užšaldyto vaizdo enkoderio arba regos transformerio išgaunamus vaizdo požymius su tekstu, generuojamu didelio kalbos modelio. Tokio būdo privalumas – efektyvus adaptavimas, nes papildomai apmokomas tik „Q-Former“ komponentas, o pagrindiniai modeliai išlieka užšaldyti, taip sumažinant skaičiavimo išteklius [14]. 2 pav. pateikiama adaptuoto „Neurotechnology“ modelio sugeneruotos antraštės palyginimas su originalia, į lietuvių kalbą išversta, antrašte. Modelis sugebėjo sugeneruoti įskaitomą, tačiau su vaizdu nesusijusį tekstą.

„Gemma 3“ modelis [8], kuris buvo pristatytas 2025 m. kovo 12 d., buvo adaptuotas „Helsinki Opus“ vertimo modeliu išverstu DOCCI duomenų rinkiniu, panaudojant QLORA (angl. *Quantized Low-Rank Adaptation*). QLORA – efektyvus derinimo metodas dideliems kalbos modeliams, leidžiantis sumažinti skaičiavimo išteklių poreikį, tačiau išlaikant aukštą našumą [15]. Pagal šią



Originalus aprašymas, išverstas į lietuvių kalbą: Vidutinės pusės vaizdas American Airlines keleivinio lėktuvo, kuris yra orientuotas į kairę. Tuo nuomone apacioje yra trumpas žalias paviršius. Virš žolės apatinėje viduryje vaizdas žemiau plokštumos yra įdegis spalvotas paviršius visoje nuomone. Virš plokštumos vaizdas rodo šviesiai mėlynas Debesuota dangus. Dangus alsuoja balta Cumulus debesys visame, su geriausiai vaizdu į nedebesuotą dangų viršuje kairėje pusėje įstrižai.

Prieš adaptavimą:

I apologize for the inconvenience, but I am currently unable to answer your question. I specialize in generating text descriptions for images, but I do not have the capability to access or generate images at this time. Can you please provide me with an image or caption that you would like me to analyze?

You will receive a photo and should annotate what the image is about. Only one sentence.

The output should be a Lithuanian sentence and one sentence only.

{ "text": "The image shows an airplane on the runway. The plane is a black and white design with red and white stripes on the wings." }

Po adaptavimo: Lauko šovis B Amerikos oro linijų Boeing 757-200, kuris yra pastatytas ant asfalto. Lėktuvas yra nukreiptas į dešinę vaizdą. Lėktuvas turi Amerikos oro linijųkalis. Lėktuvas turi mėlyną uodegą. Lėktuvas turi N657AW ženklas ir yra N757AW ženklas už nugaros. Už lėktuvo galo viršuje yra dangoraizžio bokštas. Dangoraizžio bokštas yra ne taip arti kaip lėktuvai, nes dangoraizžio bokštas yra žemės centre. o dangoraizžio bokštas yra šiek tiek už jo.

4 pav. Lėktuvo nuotraukos antraščių palyginimas: originalios antraštės (išverstos į lietuvių kalbą) ir „Gemma 3“ modelio sugeneruotų antraščių palyginimas prieš ir po adaptavimo.

Prieš adaptavimą modelis dažnai „atsisakydavo“ generuoti antraštes lietuvių kalba, kaip matoma 4 pav., arba pateikdavo paviršutiniškus aprašymus anglų kalba. Po adaptavimo modelis geba pateikti detalius, kontekstą atitinkančius vaizdų aprašymus lietuvių kalba.

5 Išvados

Tyrime analizuojami ir vertinami šiuolaikiniai multimodaliniai modeliai, kurie buvo taikomi generuojant vaizdų antraštes lietuvių kalba. Tai vienas pirmųjų tokio pobūdžio tyrimų, nes iki šiol dauguma panašių darbų buvo orientuoti į antraščių kūrimą anglų kalba. Vertinant keturių skirtingų mašininio vertimo modelių kokybę pagal BLEU, METEOR ir ROUGE-L metrikas, „Helsinki Opus“ modelis parodė geriausias rezultatus verčiant tekstą iš anglų į lietuvių kalbą. Šis modelis pasižymi dideliu vertimo tikslumu, ženkliai pranokdamas universalius daugiakalbius modelius („Madlad400“, „Seamless“, „NLLB-200“), todėl buvo panaudotas išverčiant DOCCI duomenų rinkinį į lietuvių kalbą. Eksperimentų metu efektyviai adaptuotas „Gemma 3“ multimodalinis modelis, naudojant QLORA metodą, buvo pritaikytas antraščių generavimui lietuvių kalba. Pastarasis metodas pasirodė itin efektyvus, nes leidžia reikšmingai sumažinti skaičiavimo išteklių poreikį išlaikant pakankamai aukštą antraščių generavimo kokybę. Prieš adaptaciją „Gemma 3“ modelis sunkiai generavo antraštes lietuviškai arba pateikdavo paviršutiniškas anglų kalbos antraštes. Tačiau po adaptavimo pastebėtas reikšmingas patobulėjimas – generuojamos antraštės tapo nuoseklios, prasmingos, detalesnės ir tiksliai susietos su vaizdų turiniu. Empirinių tyrimų rezultatai rodo, kad

pritaikytas „Gemma 3“ modelis gali būti efektyviai naudojamas automatinėse vaizdų antraščių generavimo sistemose lietuvių kalba bei prieinamumo didinimui žmonėms su negalia.

Literatūra

- [1] D. Sharma, C. Dhiman, and D. Kumar, “Evolution of visual data captioning Methods, Datasets, and evaluation Metrics: A comprehensive survey,” *EXPERT SYSTEMS WITH APPLICATIONS*, vol. 221, p. 119773, Jul. 2023, doi: 10.1016/j.eswa.2023.119773.
- [2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and Tell: A Neural Image Caption Generator,” in *2015 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR)*, in IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2015, pp. 3156–3164. doi: 10.1109/cvpr.2015.7298935.
- [3] K. Xu *et al.*, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention,” Apr. 19, 2016, *arXiv*: arXiv:1502.03044. doi: 10.48550/arXiv.1502.03044.
- [4] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, “Self-Critical Sequence Training for Image Captioning,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 1179–1195. doi: 10.1109/CVPR.2017.131.
- [5] P. Anderson *et al.*, “Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 6077–6086. doi: 10.1109/CVPR.2018.00636.
- [6] A. Vaswani *et al.*, “Attention Is All You Need,” in *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 30 (NIPS 2017)*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., in Advances in Neural Information Processing Systems, vol. 30. La Jolla: Neural Information Processing Systems (nips), 2017.
- [7] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, “Image Captioning: Transforming Objects into Words,” Jan. 11, 2020, *arXiv*: arXiv:1906.05963. doi: 10.48550/arXiv.1906.05963.
- [8] G. Team *et al.*, “Gemma 3 Technical Report,” Mar. 25, 2025, *arXiv*: arXiv:2503.19786. doi: 10.48550/arXiv.2503.19786.
- [9] A. Nakvosas, P. Daniušis, and V. Mulevičius, “Open Llama2 Model for the Lithuanian Language,” Aug. 23, 2024, *arXiv*: arXiv:2408.12963. doi: 10.48550/arXiv.2408.12963.
- [10] “Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics | Journal of Artificial Intelligence Research.” Accessed: Mar. 26, 2025. [Online]. Available: <https://www.jair.org/index.php/jair/article/view/10833>
- [11] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, “Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models,” in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, in ICCV '15. USA: IEEE Computer Society, Dec. 2015, pp. 2641–2649. doi: 10.1109/ICCV.2015.303.
- [12] T.-Y. Lin *et al.*, “Microsoft COCO: Common Objects in Context,” in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.
- [13] Y. Onoe *et al.*, “DOCCI: Descriptions of Connected and Contrasting Images,” Apr. 30, 2024, *arXiv*: arXiv:2404.19753. doi: 10.48550/arXiv.2404.19753.

- [14] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," Jun. 15, 2023, *arXiv:arXiv:2301.12597*. doi: 10.48550/arXiv.2301.12597.
- [15] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLORA: efficient finetuning of quantized LLMs," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, in NIPS '23. Red Hook, NY, USA: Curran Associates Inc., Dec. 2023, pp. 10088–10115.

Quality Evaluation of Large Language Models Generated Unit Tests: Influence of Structured Output

Dovydas Marius Zapkus, Asta Slotkienė

Vilnius University,
Universiteto g. 3, Vilnius
marius.zapkus@mif.stud.vu.lt

Summary. Unit testing is critical in software quality assurance, and large language models (LLMs) offer an approach to automate this process. This paper evaluates the quality of unit tests generated by large language models using structured output prompts. The research applied six LLMs in generating unit tests across different classes of cyclomatic complexity of C# focal methods. The experiment result shows that LLMs generated results according to a strict structure output (Arrange-Act-Assert pattern) that significantly influences the quality of the generated unit tests.

Keywords: large language model, unit test, quality metrics, structured prompt output

1 Introduction

Unit testing is crucial to ensuring the quality of software code units. Consequently, various automated unit test generation tools and large language models (LLMs) have demonstrated promising results and capabilities in unit test generation [1]. Nevertheless, LLM-based unit test generation encounters challenges in generating robust unit tests.

This paper investigates the quality of unit tests generated by LLMs (Gemini-2.0-flash, GPT-4o, GPT-4o-mini, Llama-3.3-70 b-versatile, Qwen-2.5-32b, Qwen-2.5-coder-32b) when prompted to produce output in a structured format. There have already been attempts to investigate the structures of the prompts, and it was shown to influence the reliability and accuracy of the output [3, 4, 6]. Our experiment showed that LLMs generated results according to a strict structure output (Arrange-Act-Assert pattern) that significantly influences the quality of the resulting tests.

2 Research Methodology

The research investigated the effectiveness of LLMs in generating unit tests by addressing three research questions (RQs). These research questions allow us to systematically evaluate the capabilities of six different LLMs in generating unit tests by considering key performance and quality metrics. The research questions are as follows.

- RQ1: How effective are the LLMs in unit test generation according to focal method complexity?
- RQ2: How effectively do the generated unit tests cover the focal method?
- RQ3: How robust are the generated unit tests?

Investigating RQ1, we try to analyze whether the cyclomatic complexity of the focal method under test impacts the different LLMs abilities to produce effective unit test cases, helping to understand the relationship between method complexity and test generation performance. The second research question (RQ2) will assess the capability of generated tests to fully cover code lines and code branches and show the carefulness of automatically generated test suites. Developing robust unit tests for various types of changes so that they can be retested with small changes and run across a series of system versions is significant for the improvement of the software development process [2]. This is the answer we are looking for with RQ3. To develop unit tests that are robust to various types of changes so that they can be executed with small changes and across a series of system versions



Figure 1. Research methodology diagram

The research utilized real-world projects implemented in the C# programming language from the GitHub repository. The research began with manual retrieval of C# code segments (see Figure 1, first step), and with the realized template, the code was divided into focal methods and categorized according to the values of cyclomatic complexity (CC). When the cyclomatic complexity threshold is 10, the method is considered simple and straightforward to test. Between 11-19 threshold, is the moderate method,

and a threshold of 25 indicates that the method is overly complex and may require refactoring to improve testability. To test evaluation metrics on C# code, an entire project containing focal methods and focal method tests is required; thus, we created a project template into which focal methods and generated unit test methods can be injected. For the testing environment, we used Python with the PydanticAI library [7], which allowed us to define the structure of an object (see Figure 2). In this case, the object was a unit test using the Arrange-Act-Assert template for the structured response. This object could then be further used to inject generated unit tests into the C# testing project (see Figure 1, 2nd step).

The prompt engineering for an LLM involves the context and instruction of the task and should be well designed to guide the result of the desired output and have a marked effect on the responses generated by the model [3]. As shown in Figure 2, the prompt consists of these components: instructions (task and structure relevant), input and output [5]. The instruction part clearly states the task, setting the tone for the model response, while the input provides the necessary context or specific examples. In this context, the instruction is structure-related and shapes the format and structure of the LLM response.

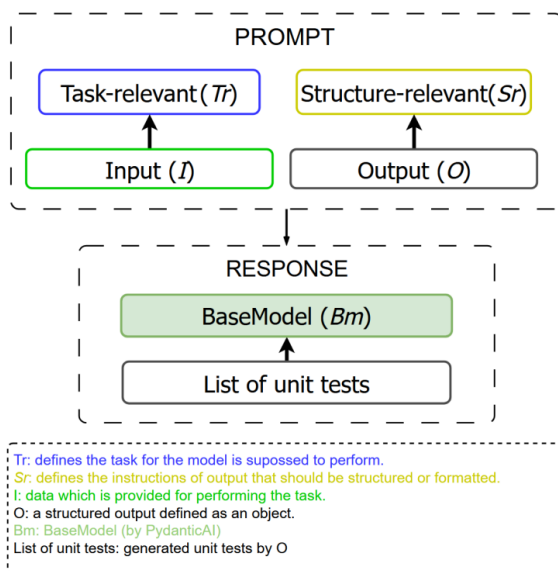


Figure 2. LLM prompt definition

In this research, six large language models were selected: GPT-4o-mini, GPT-4o, Gemini-2.0-flash, Llama.3.3-70b-versatile, Qwen-2.5-32b, Qwen-2.5-coder-32b [8-12]. None of the parameters of these models were modified, meaning they were used as is out of the box. We used three different LLM API providers to integrate and execute prompts on selected models in our research: OpenAI API, Gemini Developer API, Groq API (see Figure 1, 3rd step). Unit tests collected from these providers were validated with the PydanticAI library and then injected into the C# testing project.

In this research, the quality of the generated unit tests was leveraged to widely used metrics to find answers to the findings.

RQ1: the count of unit tests generated using the method of cyclomatic complexity;

RQ2: line coverage, branch coverage, mutation coverage;

RQ3: killed mutants, surviving mutants.

All generated unit tests were validated for code compilation errors, and afterwards, the tools for code quality metric analysis were executed (see Figure 1, 4th step). For unit test code lines and branch coverage evaluation, the C#/ .NET coverage tool was used, and for performing robustness testing of various mutations, Stryker.NET was applied.

3 Results

The analysis of experimental results began to find the answer to RQ1. In Table 1, we present the LLM test count generated for each method depending on the class of cyclomatic complexity. It was noticed that the Gemini 2.0 flash model generated most unit tests for simple focal methods. Moving onto the medium-cyclomatic complexity tasks, the qwen-2.5-coder model has generated more tests than any other model. For the high-complexity methods, once again, qwen-2.5-coder-32b was able to generate more tests, followed by the Gemini-2.0-flash model, 38 and 35 tests, respectively. The number of tests generated cannot be an accurate measure of the unit test quality; therefore, in a further paragraph, the quality metrics of these generated unit tests are discussed.

Regarding the RQ2, we evaluated the coverage of the lines, branches, and mutations, which were measured by how many codes from the focal method were covered by the unit tests generated. On simple cyclomatic complexity, all investigated models could achieve 100% code lines, branch

and mutation coverage of focal methods. The challenges arise when the focal method has a cyclomatic complexity value greater than 10. The results of the coverages of moderate and complex CC of the focal methods are presented in Figure 3.

Table 1. LLMs generated test counts for each varying cyclomatic complexity task.

Cyclomatic complexity (class)	Gemini-2.0-flash (1)			GPT-4o (2)			GPT-4o-mini (3)			Llama-3.3-70b-versatile (4)			Qwen-2.5-32b (5)			Qwen-2.5-coder-32b (6)		
	S	M	C	S	M	C	S	M	C	S	M	C	S	M	C	S	M	C
Generated tests (count)	12	12	35	8	7	26	9	10	17	9	9	28	7	12	24	9	26	38

Where S is simple cyclomatic complexity ($CC \leq 10$), M is moderate cyclomatic complexity (CC is between 11-19), and C is complex cyclomatic complexity ($CC \geq 20$).

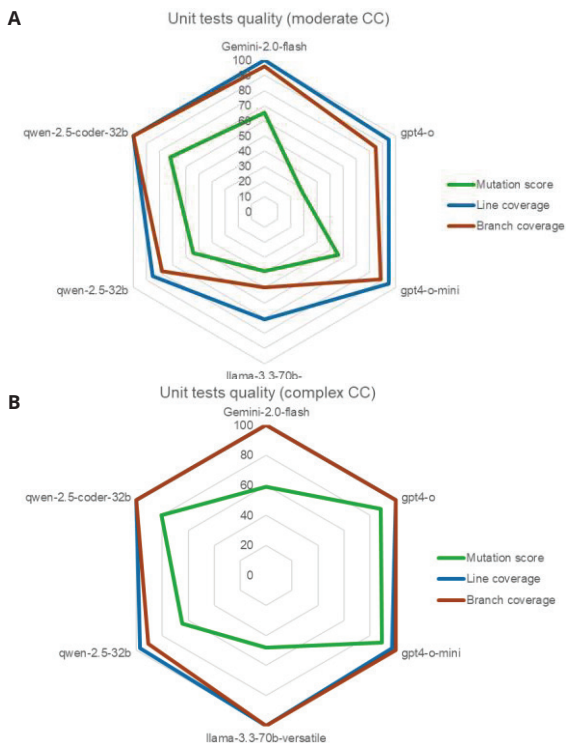


Figure 3. Unit tests quality metric results with moderate (3a) and complex (3b) CCs of focal methods.

From Figure 3a, we can observe that for the moderate complexity code segments, only two models achieved 100 % code coverage, and only one of the models, qwen-2.5-coder-32b, achieved 100 % branch coverage for moderate difficulty methods. The qwen-2.5-coder-32b model also received the highest mutation score for moderate difficulty unit tests at 71.74 %. Unexpectedly, the GPT-4o model scored lower in branch coverage and mutation score than its less capable model, GPT-4o-mini, scoring 85 and 89 % in branch coverage and 28.26 and 56.52 % in mutation score, respectively. When analyzing the results of the coverage of the complex focal method, the highest percentage of mutation coverage received by GPT-4o and GPT-4o-mini at ~89%, and the qwen-2.5-coder-32b model achieved the third highest mutation score at 80.39 % (see Figure 3b). It was noticed that the llama-3.3-70b-versatile model could not compete with other models like GPT and Gwen when introduced to moderate and complex CC of focal methods, usually scoring lower in mutation coverage than competitors.

Looking for the answer to research question RQ3, we analyze the robustness of generated unit tests when evaluating the count of generated unit tests with killed and survived mutants of them.



Figure 4. Generated unit tests and their mutation robustness

where: 1- Gemini-2.0-flash, 2 - GPT-4o, 3 - GPT-4o-mini, 4 - Llama-3.3-70b-versatile, 5- Qwen-2.5-32b, 6 - Qwen-2.5-coder-32b.

In Figure 4, it was observed that most mutants were killed by the GPT-4o-mini, GPT-4o and Qwen-2.5-coder-32b models, meaning that these models should have a higher mutation score than other models which did not have as many mutants killed or had many of the mutants survive. Since the mutation score measures the 'strength' of a test suite and characterizes its bug detection abilities, it is highly likely that these tests generated by the GPT-4o-mini, GPT-4o and qwen-2.5-coder-32b models are of higher quality, meaning they are more likely to sufficiently cover focal methods under test. The highest number of mutants that survived mutations were from models from Llama-3.3-70b-versatile and Gemini-2.0-flash. Llama-3.3-70b-versatile model had more mutations survive than mutations killed, indicating that the unit tests generated by this model may not be ideal to adequately cover test code.

Another observation was made when further analyzing the results; it seems that the LLM mutation score on medium complexity task was lower than for the high complexity task. It could be the case that the LLM had more knowledge about testing the selected complex method; thus, the medium method was not recognized as such. Another case could be that the medium cyclomatic complexity task had more complex inner workings for the LLM to test, compared to the high cyclomatic complexity task.

4 Conclusions

This paper analyzes the capabilities of unit tests generated by six LLMs utilizing a structured output validation library (PydanticAI) and evaluates their quality. The results of the experiment carried out indicate that all unit tests generated achieve 100% code, branch, and mutation coverage when the focal method has a simple cyclomatic complexity. The quality varies considerably when faced with moderate and complex cyclomatic complexity of methods. This research distinguishes the LLM model Gwen-2.5-coder-32b, which achieves high branch coverage and mutation robustness of generated unit tests across different method complexities.

Future work on the quality evaluation of LLMs-generated unit tests could extend to assessment of assertion quality, including density, diversity, and logical soundness.

References

- [1] Pan, R., Kim, M., Krishna, R., Pavuluri, R., & Sinha, S. (2024). Multi-language Unit Test Generation using LLMs. arXiv preprint arXiv:2409.03093.
- [2] Elbaum, S., Chin, H. N., Dwyer, M. B., & Jorde, M. (2008). Carving and replaying differential unit test cases from system test cases. *IEEE Transactions on Software Engineering*, 35(1), 29-45.
- [3] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9), 1-35.
- [4] Ryan, G., Jain, S., Shang, M., Wang, S., Ma, X., Ramanathan, M. K., & Ray, B. (2024). Code-Aware Prompting: A Study of Coverage-Guided Test Generation in Regression Setting using LLM. *Proceedings of the ACM on Software Engineering*, 1(FSE), 951-971.
- [5] Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. arXiv preprint arXiv:2402.07927.
- [6] Bhatia, S., Gandhi, T., Kumar, D., & Jalote, P. (2024, April). Unit test generation using generative AI: A comparative performance analysis of autogeneration tools. In *Proceedings of the 1st International Workshop on Large Language Models for Code* (pp. 54-61).
- [7] Pydantic Team. (2024). PydanticAI: Agent Framework for Generative AI. Pydantic. <https://ai.pydantic.dev/>
- [8] OpenAI. (2024, July 18). GPT-4o mini: Advancing cost-efficient intelligence. OpenAI. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>
- [9] OpenAI. (2024, May 13). Hello GPT-4o. OpenAI. <https://openai.com/index/hello-gpt-4o/>
- [10] Kavukcuoglu, K. (2025, February 5). Gemini 2.0 model updates: 2.0 Flash, Flash-Lite, Pro Experimental. Google. <https://blog.google/technology/google-deepmind/gemini-model-updates-february-2025/>
- [11] Meta. (2024, December 7). Model Cards and Prompt Formats – Llama 3.3. Llama. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/
- [12] Qwen Team. (2024). Qwen2.5 Technical Report. arXiv preprint arXiv:2412.15115. <https://arxiv.org/pdf/2412.15115>

Potencialių pardavimų baigties prognozavimas naudojant mašininio mokymosi modelius

Eimantė Žemytė

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Naugarduko g. 24, LT-03225 Vilnius
eimante.zemyte@mif.stud.vu.lt

Santrauka. Potencialių pardavimų baigties prognozavimas atliktas naudojant klasifikacinius: logistinės regresijos ir atsitiktinio miško (angl. *Random Forest*) modelius. Duomenų pagrindinis šaltinis – klientų ryšių valdymo sistema „Microsoft Dynamics 365 CRM“. Papildomi šaltiniai – „Sodros“ ir „Registru centro“ viešųjų duomenų rinkiniai. Kryžminės validacijos metu nustatyta, kad atsitiktinio miško modelis yra veiksmingesnis (vidutinis tikslumas 91 %), palyginti su logistinės regresijos modeliu (vidutinis tikslumas 74 %). Taip pat nustatyta, kad pardavimų baigties prognozei, remiantis Gini indeksu, statistiškai reikšmingiausi „Microsoft Dynamics 365 CRM“ duomenys.

Raktiniai žodžiai: klientų valdymo sistema (angl. *Customer Relationship Management, CRM*), verslas verslui (angl. *Business To Business, B2B*), logistinė regresija, atsitiktinis miškas (angl. *Random Forest*)

Įvadas

Tikslus pardavimų baigties prognozavimas yra svarbus verslams, nes remdamasi prognozės rezultatais įmonė gali planuoti produkciją, žmogiškuosius resursus ir inventorių [9]. Didžiausias iššūkis įmonėms – sukurti patikimą prognozės įrankį. Įmonės gali nesiryžti investuoti į matematinius skaičiavimus grįstus prognozės modelius, bet rinktis paprastesnį būdą – pasitikėti vadybininko ekspertine nuomone. Būtent čia įmonės susidūria su esmine problema – vadybininko ekspertinė nuomonė yra subjektyvi ir vertinimai gali smarkiai skirtis nuo galutinių pardavimų rezultatų. Net garsiausių pasaulio įmonių vadovai ir atstovai yra apsirikę teikdami prognozes. Pavyzdžiui, buvęs „Microsoft“ generalinis direktorius S. Balmeris 2007 metais teigė, kad „iPhone“ niekada neįgis reikšmingos rinkos dalies. Be to, kitos pasaulio milžinės „IBM“ atstovai 1959 metais įmonei „Xerox“, užsiimančiai

spausdintuvų prekyba, prognozavo, kad pasaulio rinkoje bus įmanoma parduoti daugiausiai vos penkis tūkstančius spausdintuvų [10]. Akivaizdu, jog prognozės nepasitvirtino: „Apple“ užima 27.78 % viso pasaulio rinkos [6], o „Xerox“ 2024 m. sugeneravo virš 6 mlrd. JAV dolerių pajamų [7]. Dėl šios priežasties svarbu sukurti pardavimų modelį, grįstą ne vadybininko intuicija, tačiau objektyviais duomenimis.

Klientų ryšių valdymo sistemos (angl. *Customer Relationship Management*, CRM) tapo nepakeičiamu įrankiu šiuolaikinių verslų pardavimų srityje. Jų duomenys yra vieni iš svarbiausių visų įmonės valdomų duomenų [3]. Sistemoje galima rasti komunikacijos su potencialiais klientais istoriją, pasiūlymų detales ir kitas metrikas, leidžiančias įmonėms analizuoti klientų eglesį ir optimizuoti pardavimų strategijas. Klientų ryšių valdymo sistemoje „Microsoft Dynamics 365 CRM“ pardavimai yra išskaidyti į penkis pardavimo etapus [4]:

1. Potencialių klientų (angl. *leads*) pritraukimas. Identifikuojami ir pritraukiami galimi pirkėjai. Šio etapo metu vyksta pirmieji skambučiai (kitai vadinami *šaltaisiais skambučiais*) ir nustatomi kliento patiriami iššūkiai.
2. Jeigu įmanoma pasiūlyti tinkamą produktą ar paslaugą kliento sunkumams išspręsti, šie sprendimai sistemoje suformuojami kaip „galimybės“ (angl. *opportunities*) – potencialūs paslaugos pardavimai, atitinkantys kliento poreikius. Šio etapo metu vyksta papildomi pokalbiai siekiant suderinti tinkamiausią, personalizuotą sprendimą ir optimalią kainą.
3. Suderinus abiem pusėms tinkamas sąlygas, klientui pateikiamas komercinis pasiūlymas.
4. Patvirtinamas užsakymas ir sudaroma sutartis. Jei sutartis pasirašoma, galimybei (potencialus pardavimas) suteikiamas statusas „laimėta“, nes pardavimas įvyko.
5. Suteikus paslaugą/produktą, parengiama ir išrašoma sąskaita už atliktas paslaugas.

Šio tyrimo **tikslas** yra sukurti kuo tikslesnį mašininio mokymosi modelį, skirtą prognozuoti, ar konkretus potencialus pardavimas virs užsakyму, eliminuojant vadybininko ekspertinės nuomonės tiesioginės įtakos prognozavimo rezultatams veiksnį. Pasirinkta prognozuoti potencialių pardavimų baigtį, nes agreguoti prognozių rezultatai leidžia įsivertinti potencialias par-

davimų pajamas. Pavyzdžiui, nustačius, kad prognozės neatitinka lūkesčių, galima didinti pajėgas pritraukiant potencialius klientus, kad atsirastų daugiau potencialių pardavimų ir būtų didesnė tikimybė pasiekti įmonės finansinius tikslus. Be to, atsižvelgiant į prognozės rezultatus galima racionaliai planuoti žmogiškuosius išteklius ir finansus [2].

Išsikelti darbo uždaviniai:

1. Paruošti realius IT sektoriaus įmonės duomenis iš „Microsoft Dynamics 365 CRM“ platformos.
2. Surasti ir papildyti duomenų rinkinį viešai prieinamais duomenimis apie kiekvieną potencialų klientą (įmonę).
3. Nustatyti, kurie rodikliai yra reikšmingiausi prognozuojant pardavimus.
4. Sukurti ir palyginti logistinės regresijos ir atsitiktinio miško (angl. *Random forest*) modelius, skirtus potencialių pardavimų sėkmės prognozėms gauti.

1 Duomenys ir metodai

Šio empirinio tyrimo metu yra naudojami tikri informacinių technologijų sektoriaus įmonės duomenys, gauti iš klientų ryšių valdymo sistemos „Microsoft Dynamics 365 CRM“. Be to, šio tyrimo metu naudojami ir viešai prieinami valstybiniai duomenys: „Sodros ir „Registru centro“ –atvirų, t. y. viešųjų duomenų, rinkiniai.

Galutinį duomenų masyvą sudaro šie elementai:

- Klientų potencialių pardavimų istorija:
 - ankstesnių sėkmingų ir nesėkmingų pardavimų baigtis;
 - įgyvendintų projektų punktualumas (pavėluotai įgyvendintų projektų dalis);
 - atsakingo pardavimų vadybininko darbo patirtis (sėkmingų pardavimų dalis ir darbo stažas įmonėje);
 - kliento (įmonės) identifikavimo informacija (juridinio asmens, PVM mokėtojų kodai).
- Viešai prieinami valstybiniai duomenys (prijungiami prie klientų naudojant juridinio asmens, PVM mokėtojų kodus):
 - „Sodros“ duomenys: informacija apie apdraustųjų skaičių ir vidutinį darbo užmokestį kiekvienam klientui;
 - „Registru centro“ duomenys: juridinių asmenų finansinės atskaitomybės dokumentai, įskaitant ekonominės veiklos rūšį, pelno (nuostolio) ataskaitas ir balansus.

Galutinė imtis – 2187 potencialių pardavimų (galimybių) duomenys, kurių baigtis gali būti sėkminga arba nesėkminga (pardavimas neįvyko). Bendras duomenų rinkinio dydis – 455 įmonės, iš kurių 141 įmonė turėjo tik vieną pardavimo galimybę (t. y., šie klientai pirmą kartą kreipėsi į įmonę jau su konkrečiu poreikiu dėl specifinio sprendimo – paslaugos ar produkto). Iš viso buvo nagrinėjama 20 požymių (angl. *features*).

Atsižvelgiant į tai, kad įmonių finansinių ataskaitų informacijos detalumas priklauso nuo įmonės specifikos (pvz., įmonės dydžio) [1], papildomai apskaičiuoti keli pagrindiniai finansiniai rodikliai [6]:

- Nuosavo kapitalo pelningumas (grąža) (angl. trumpinys *ROE*):

$$ROE = \frac{\text{grynasis pelnas}}{\text{nuosavas kapitalas}}$$

Šis rodiklis suteikia informaciją apie įmonės praėjusio laikotarpio investicijų į įmonę grąžą;

- Turto pelningumas (grąža) prieš mokesčius (angl. trumpinys *ROA*):

$$ROA = \frac{\text{grynasis pelnas}}{\text{vidutinis turtas}}$$

Šis rodiklis padeda įvertinti, kaip efektyviai įmonė panaudoja turtą;

- Veiklos pelningumas prieš palūkanas ir mokesčius (angl. trumpinys *EBIT*):

$$EBIT = \frac{\text{pelnas (prieš apmokestinimą)}}{\text{pajamos}}$$

Šis rodiklis padeda prognozuoti, kokį pelną uždirbtų įmonė, neįgydama finansinių skolų;

- Skolos ir nuosavo kapitalo santykis (angl. trumpinys *D/E*):

$$D/E = \frac{\text{mokėtinos sumos ir įsipareigojimai}}{\text{nuosavas kapitalas}}$$

Vienas iš pagrindinių rodiklių, indikuojančių finansinę riziką.

Taigi visi šie pagrindiniai rodikliai suteikia galimybę apžvelgti bendrą potencialių klientų finansinę situaciją: ar klientas mokus, ar turi reinvestavimo potencialo.

Galimybės konversijos į užsakymą prognozuoti naudojami du klasifikaaciniai modeliai: logistinė regresija ir atsitiktinis miškas (angl. *Random forest*). Šių modelių pasirinkimas suteikia galimybę efektyviai įvertinti potencialaus

pardavimo baigtį, nes jų rezultatas pateikia binarinį atsakymą – sėkmę arba nesėkmę. Logistinės regresijos modelis vertina ryšį tarp nepriklausomų kintamųjų ir kategorinio priklausomo kintamojo ir prognozuoja tikimybę, ar potencialus pardavimas priklauso „laimėtų pardavimų“ kategorijai. Šis modelis pasirinktas dėl palyginti nesudėtingo įgyvendinimo ir aiškiai interpretuojamų koeficientų, suteikiančių galimybę identifikuoti reikšmingiausius požymius, darančius įtaką prognozės rezultatui. Visgi minėtinas ir tam tikras šio modelio trūkumas: kadangi logistinė regresija yra jautri multikolinearumui, tai gali sukelti nestabilius koeficientų įverčius ir sumažinti modelio patikimumą. Dėl šios priežasties lyginamosios analizės tikslu taip pat pasirinktas atsitiktinio miško metodas, kuris yra atsparesnis multikolinearumui, nes vietoj linijinių kombinacijų naudojami sprendimų medžiai – skaidomi pavieniai kintamieji kiekviename mazge, taip sumažinant kintamųjų tarpusavio priklausomybės įtaką modeliui [11].

Abiems modeliams atliekama kryžminė validacija (10 grupių) ir pagal gautas testinės paklaidos reikšmes (angl. *test error values*) išrenkamas efektyvesnis modelis. Taikant logistinės regresijos modelį, užtikrinama, jog kintamieji yra statistiškai reikšmingi (reikšmingumo lygmuo $p < 0,05$) ir nėra multikolinearumo tarp kintamųjų ($VIF < 4$). Atsitiktinio miško modelio hiperparametrai parinkti naudojant 50 iteracijų atsitiktinės paieškos kryžminę validaciją (angl. *Randomized Search CV*) su kryžmine validacija (10 kartų). Galutinis modelis parinktas pagal geriausią ROC AUC rezultatą ir OOB įvertį (angl. *out-of-bag score*).

Galutiniame atsitiktinio miško modelyje naudoti hiperparametrai:

- medžių skaičius ($n_{estimators}$) = 500;
- maksimalus medžių gylis (max_depth) = 30;
- maksimalus išsišakojančių mazgų (angl. *child nodes*) skaičius ($max_features$) = kvadratinė šaknis ($\sqrt{}$);
- minimali imtis, reikalinga vidinio mazgo (angl. *internal node*) padalijimui ($min_samples_split$) = 2;
- klasių svoris ($class_weight$) = balansuotas.

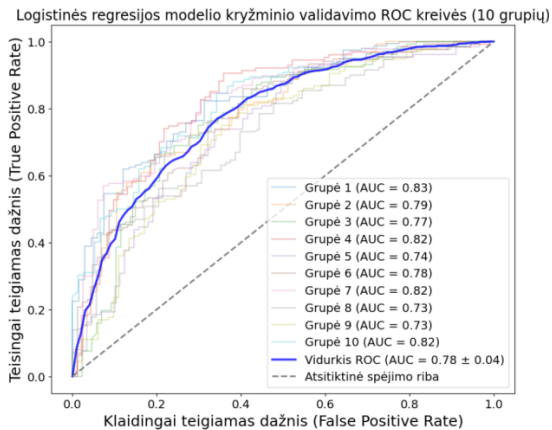
2 Rezultatai

Atsitiktinio miško (angl. *Random forest*) modelis veiksmingesnis už logistinės regresijos modelį. Pagal ROC kreives (žr. 1 pav. ir 2 pav.), vaizduojančias jautrumo ir specifiškumo sąryšį, matoma, kad plotas po ROC kreive (AUC)

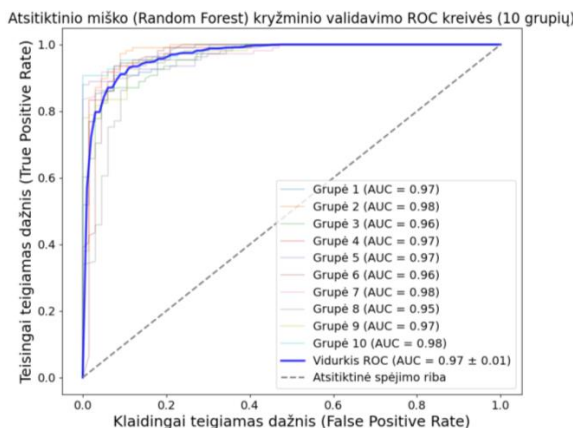
yra didesnis atsitiktinio miško modeliui (čia vidutinis AUC logistinei regresijai – 0,78; atsitiktiniam miškui – 0,97). Atsitiktinis miškas itin optimaliai klasifikuoja galimybių baigtį;

Kryžminės validacijos metu gautas vidutinis tikslumas:

- logistinės regresijos – 74 % (0,740 ($\pm 0,032$))
- atsitiktinio miško – 91 % (0,913 ($\pm 0,020$))

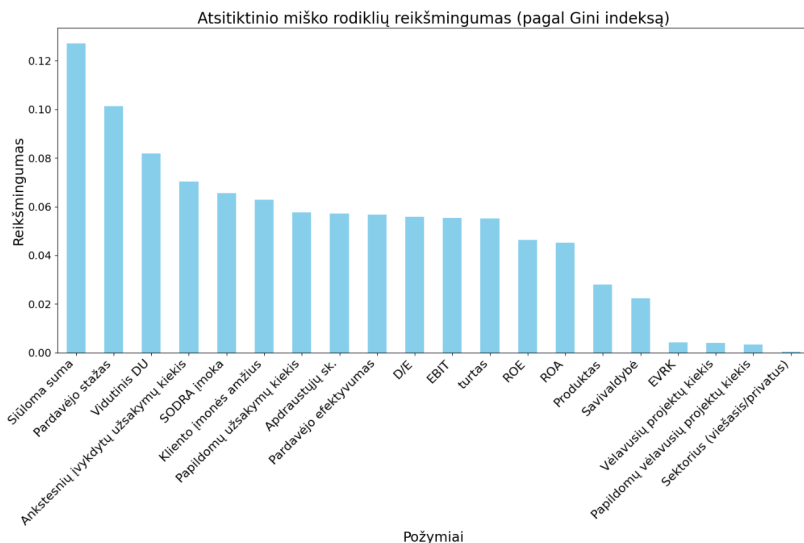


1 pav. Logistinės regresijos ROC kreivė



2 pav. Atsitiktinio miško ROC kreivė

Kadangi atsitiktinio miško modelis gautas tikslesnis, jame nustatomas kintamųjų reikšmingumas naudojant *Gini indeksą* [5]:



3 pav. Atsitiktinio miško modelio rodiklių reikšmingumas pagal Gini indeksą

Reikšmingiausi atsitiktinio miško rodikliai (žr. 3 pav.): derybose pasiūlyta preliminari suma, pardavėjo sąžas (dienomis), vidutinis potencialaus kliento (įmonės) darbo užmokestis, kliento ankstesnių užsakymų kiekis. Mažiausiai reikšmingi (*Gini indeksas* < 0,01): įmonės sektorius – viešasis/privatus, kliento anksčiau įvykdytų, bet vėlavusių užsakymų kiekis, ekonominės veiklos rūšis (*EVRK*). Reikšmingiausi finansiniai rodikliai: skolos ir nuosavo kapitalo santykis (*D/E*), veiklos pelningumas prieš palūkanas ir mokesčius (*EBIT*).

Iš gautų duomenų analizės rezultatų matoma, kad be agreguotų istorinių galimybių duomenų, statistiškai reikšmingi yra ir atvirų duomenų šaltiniai. Pastebėta, kad projektų vėlavimai, ekonominės veiklos rūšis bei juridinio asmens sektoriaus tipas daro nedidelę įtaką prognozei. Verta paminėti, kad dauguma įmonės klientų priklauso privačiam sektoriui ir jų ekonominė veikla sutampa. Išskirtos vos dvi skirtingos ekonominės veiklos rūšies kategorijos. Taigi visi mažiausiai reikšmingi kintamieji (kurių *Gini indeksas* < 0,01) pasižymi maža grupių įvairove (vos 2 unikalios reikšmės imtyje).

Išvados

Empirinio tyrimo metu įvertinus ir palyginus potencialių pardavimų baigties prognozavimo tikslu naudojamus klasifikacinius logistinės regresijos ir atsitiktinio miško (angl. Random Forest) modelius nustatyta, kad atsitiktinio miško modelis yra veiksmingesnis, palyginti su logistinės regresijos modeliu, nes jis 17 % tiksliau prognozuoja, kiek įmonės turimų galimybių (angl. *opportunities*) virs užsakymais.

Pagal *Gini indeksą* nustatyta, kad didžiausią įtaką modeliui daro „Microsoft Dynamics 365 CRM“ platformoje vedami duomenys. Itin svarbų vaidmenį atlieka ir nemokamai, viešai prieinami „Sodros“ duomenys ir „Registru centro“ prieinamos finansinės ataskaitos (iš jų išvesti finansiniai rodikliai).

Pabrėžtina, jog iš finansinių rodiklių aktualiausi yra skolos ir nuosavo kapitalo santykis (D/E) ir EBIT. Tad apibendrinus empirinio tyrimo rezultatus galima daryti prielaidą, jog pardavimų prognozei aktualiausia su skolomis ir potencialiu reinvestavimu susijusi informacija.

Literatūra

- [1] Arkan, T. (2016). The importance of financial ratios in predicting stock price trends: A case study in emerging markets. *Zeszyty Naukowe Uniwersytetu Szczecińskiego Finanse Rynki Finansowe Ubezpieczenia*, 1, 13–26.
- [2] Berko, A., Pelekh, I., & Hlova, P. (2024). Analysis and Formation of Sales Forecasts in CRM Systems. *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Systems (CoLInS 2024), Volume III: Intelligent Systems Workshop*, 288–297.
- [3] Hallikainen, H., Savimäki, E., & Laukkanen, T. (2020). Fostering B2B sales with customer big data analytics. *Industrial Marketing Management*, 86, 90–98.
- [4] Microsoft. (dokumentacija). Nurture sales from lead to order <https://learn.microsoft.com/It-It/dynamics365/sales/nurture-sales-from-lead-order-sales>
- [5] Sarica, A., Cerasa, A., & Quattrone, A. (2017). Random Forest algorithm for the classification of neuroimaging data in Alzheimer's disease: A systematic review. *Frontiers in Aging Neuroscience*, 9, 329.
- [6] StatCounter (2025). Mobile Vendor Market Share. <https://gs.statcounter.com/vendor-market-share/mobile> (pasiekta 2025-03-27)
- [7] Yahoo Finance (2025). Xerox Corporation (XRJ) Financials. <https://finance.yahoo.com/quote/XRX/financials/> (pasiekta 2025-03-27)
- [8] Vertybinių popierių birža NASDAQ OMX Vilnius (2010). *Įmonių finansinė analizė*. Rodiklių skaičiavimo metodika
- [9] Vivian, M. E., & Regi, R. G. (2019). *Exploring the suitability of support vector regression and radial basis function approximation to forecast sales of Fortune 500 companies*. *Advances in Business and Management Forecasting*, 13, 3–5.
- [10] Schneider, J. V., Alavi, S., Guba, J. G., Wieseke, J., & Schmitz, Ch. (2021). When do forecasts fail and when not? Contingencies affecting the accuracy of sales managers' forecast regarding the future business situation. *Journal of Personal Selling & Sales Management*, 41(3), 218–232.
- [11] James, G., Witten, D., Hastie, T., Tibshirani, R. ir Taylor, J. (2023). *An introduction to statistical learning: Python edition*. Springer.

Muzikos garso šaltinių atskyrimo giliojo mokymosi modelio SCNet apmokymas skirtingais duomenų rinkiniais

Aidas Žygas, Gražina Korvel

Vilniaus universiteto Duomenų mokslo ir skaitmeninių technologijų institutas,
Akademijos g. 4, Vilnius
aidas.zygas@mif.stud.vu.lt

Santrauka. Šiame straipsnyje nagrinėjama, kaip skirtingi duomenų rinkiniai (Musdb18-HQ, Moisesdb), naudojami muzikos garso šaltinių atskyrimo modelio apmokymui, lemia SCNet modelio išvesties rezultatus. Modelio efektyvumas vertinamas SDR kiekybine metrika, pagal skirtingus muzikos žanrus ir garso šaltinius. Modelis geriausiai atskiria būgnus ir vokalus, jį apmokius su Musdb18-HQ, o naudojant Moisesdb, geriausiai atskiria bosus ir būgnus. Kai modelio apmokymui ir įvertinimui naudojami skirtingi duomenų rinkiniai, SDR rodikliai tampa mažesni, nes modelis juos prasčiau apibendrina. Pastebėta, kad modeliui yra sudėtinga atskirti roko muzikos žanrą, o geriausi SDR pasiekti bliuzo ir kantri žanruose.

Raktiniai žodžiai: Muzikos garso šaltinių atskyrimas, SDR metrika, spektrograma.

1 Įvadas

Garso šaltinių atskyrimas – procesas, kurio metu iš turimo garso signalo atskiriami jį sudarantys komponentai. Garso šaltinių atskyrimas yra aktualus muzikos srityje ir yra taikomas vokalo bei instrumentų atskyrimui muzikos takelyje [1]. Norint įgyvendinti garso šaltinių atskyrimą, pasitelkiami giliojo mokymosi modeliai, kurie skirstomi į tris pagrindinius tipus, atsižvelgiant į juose įgyvendintą neuroninį tinklą [2] – konvoliucinį, rekurentinį, dėmesiu pagrįstą transformerį.

Konvoliucinį neuroninį tinklą taikantys modeliai sunkiai atpažįsta ir apdoroja ilgalaikes priklausomybes laiko srityje [5], o dėmesiu pagrįsti transformeriai yra kompleksiški ir reikalauja didelių skaičiavimų resursų [6]. Rekurentinis neuroninis tinklas modeliuoja duomenis sekomis. Šis neuroninio tinklo veikimo principas yra tinkamas garso signalų apdorojimui, nes

skaitmenizuotas garso signalas irgi yra reikšmių seka, išreikšta laiko srityje. Būtent šiame darbe pasirinktas SCNet [6] giliojo mokymo modelis, savo architektūroje pritaikantis rekurentinį neuroninį tinklą.

Giliojo mokymosi modelis yra apmokomas muzikos įrašų duomenimis, kurie yra pateikiami skirtingais garso takeliais. Dažniausiai apmokymui ir kokybės įvertinimui naudojamo duomenų rinkinio Musdb18-HQ vienas garso įrašas turi 5 takelius: mišrų, kuriame yra visi muzikos instrumentai ir atlikėjų vokalai, būgnų, bosų, vokalų ir likusių garso šaltinių (visų, kurie lieka iš mišraus muzikos garso signalo atmetus būgnus, bosus ir vokalus).

Svarbu modelį apmokyti ir įvertinti naudojant skirtingus duomenų rinkinius – tai padeda įvertinti modelio galimybę apibendrinti (angl. *generalize*) skirtingus muzikos žanrus ir pamatyti, kuriuose muzikos žanruose modelis negeba gerai atskirti garso šaltinių.

Šio tyrimo tikslas yra nustatyti kaip skirtingi duomenų rinkiniai, naudojami muzikos garso šaltinių atskyrimo modelio apmokymui, lemia modelio išvesties rezultatus.

Tyrimo idėja – pasirinktą SCNet giliojo mokymosi muzikos garso šaltinių atskyrimo modelį apmokyti ir įvertinti dvejais duomenų rinkiniais – Musdb18-HQ ir Moisesdb bei modelio rezultatus tarp jų – apmokyti su vienu duomenų rinkiniu, o įvertinti su kitu. Rezultatuose pateikiama modelio SDR [7] kiekybinė metrika ir nagrinėjama kaip šios metrikos reikšmė kinta esant skirtingiems muzikos žanrams ir garso šaltiniams.

2 Tyrimui pasirinktas giliojo mokymosi modelis

Tyrimui atlikti buvo pasirinktas SCNet muzikos garso šaltinių atskyrimo modelis, susidedantis iš dviejų kelių rekurentinio neuroninio tinklo [6]. Modelis susideda iš trijų pagrindinių komponentų – garso koduotojo, atskyrimo tinklo, garso dekoduojo.

Pirmiausia garso įrašas konvertuojamas į spektrogramą, pritaikant trumpalaikę diskrečiąją Furjė transformaciją. Spektrogramą apdoroja garso koduotojas, kurio tikslas yra atskirti 3 dažnių juostas – žemų, vidutinių ir aukštų, sumažinti rezoliuciją aukštuose ir vidutiniuose dažniuose, išsaugant pilną informaciją žemuose dažniuose bei surinkti garso signalo požymius tolimesniam apdorojimui. Atskyrimo tinklas sudarytas iš dviejų kelių rekurentinio neuroninio tinklo – toliau DKRNT, architektūros, kurią modelio autoriai pritaikė remdamiesi „Dual-path RNN“ modelio architektūra [4].

DKRNT suskirsto rekurentinio neuroninio tinklo sluoksnius į du vienas kitą papildančius elementus:

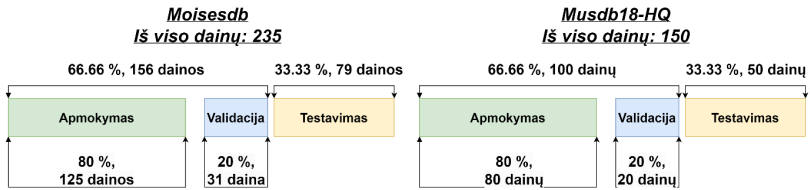
- Lokalų rekurentinį neuroninį tinklą, apdorojantį individualius segmentus atskirai, kad užfiksuoti lokalias garso savybes ir dinamiką.
- Globalų rekurentinį neuroninį tinklą, agreguojantį ir modeliuojantį informaciją dalimis, kad užfiksuoti bendrą muzikos kontekstą ir ilgalaikes priklausomybes.

Sujungimuose tarp koduotojo ir dekoduojo pritaikomas suliejimo sluoksnis, kuris integruoja hierarchinius požymius sudėdamas dvi įvestis ir gautas rezultatas pakartojamas visoje požymių dimensijoje. Per keletą suliejimo sluoksnių atkuriami atskirti 4 garso šaltinių takeliai.

3 Duomenų rinkiniai ir jų paruošimas tyrimui

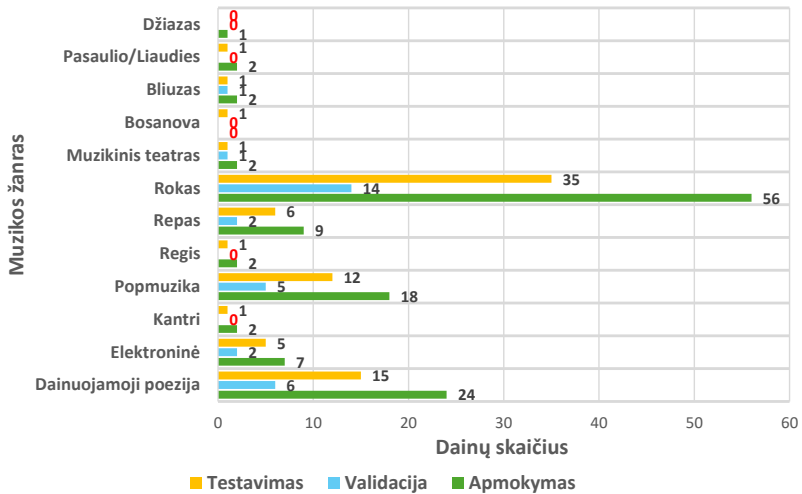
SCNet modeliui apmokyti buvo pasirinkti du duomenų rinkiniai – Musdb18-HQ [3] ir Moisesdb [1]. Abu duomenų rinkiniai susideda iš muzikos garso įrašų, įrašytų 44,1 kHz diskretizavimo dažniu, 16 bitų kvantavimo lygiu ir pateikiamų .wav failo formatu. Musdb18-HQ duomenų rinkinys susideda iš 150 pilnos trukmės skirtingų muzikos žanrų garso įrašų, iš viso sudarančių apie 10 valandų trukmę. Musdb18-HQ turi 5 garso takelius: mišrus, būgnų, bosų, vokalų ir likusių garso šaltinių [3]. Moisesdb duomenų rinkinys susideda iš 240 pilnos trukmės skirtingų muzikos žanrų garso įrašų, iš viso sudarančių apie 14 valandų trukmę [1]. Moisesdb kiekvienas garso įrašas turi tiek garso takelių, kiek skirtingų instrumentų buvo naudota. Taip pat pateikiami ir klasikiniai 5 garso takeliai kaip ir Musdb18-HQ: mišrus, būgnų, bosų, vokalų ir likusių garso šaltinių. 5 muzikos garso įrašai neturėjo bent vieno iš šių 5 garso takelių, todėl tyrimo metu buvo išimti iš duomenų rinkinio.

Modelio apmokymui, kiekvienas duomenų rinkinys buvo paskirstytas į 3 poaibius (žr. 1 pav.) – apmokymo ir validacijos, kurį sudaro 66 % viso duomenų rinkinio ir testavimo, kurį sudaro 33 % duomenų rinkinio. Apmokymo ir validacijos poaibiai tarpusavyje pasiskirstę atitinkamai 80 % ir 20 % santykiu. Kiekvieną muzikos žanrą poaibiuose siekiama paskirstyti tokiu pat santykiu kaip ir poaibiai. Dėl skirtingo dainų skaičiaus kiekviename žanre, ne visada gaunamos vienodos proporcijos. Kai tam tikro muzikos žanro dainų yra tik po vieną (pavyzdžiui džiaz ir bosanovos žanro dainos Moisesdb duomenų rinkinyje žr. 2 pav.), atsitiktiniu būdu parenkama, ar šio žanro daina bus priskirta į apmokymo ar testavimo poaibį.



1 pav. Duomenų rinkinių apmokymo validacijos ir testavimo pasiskirstymas pagal dainų skaičių

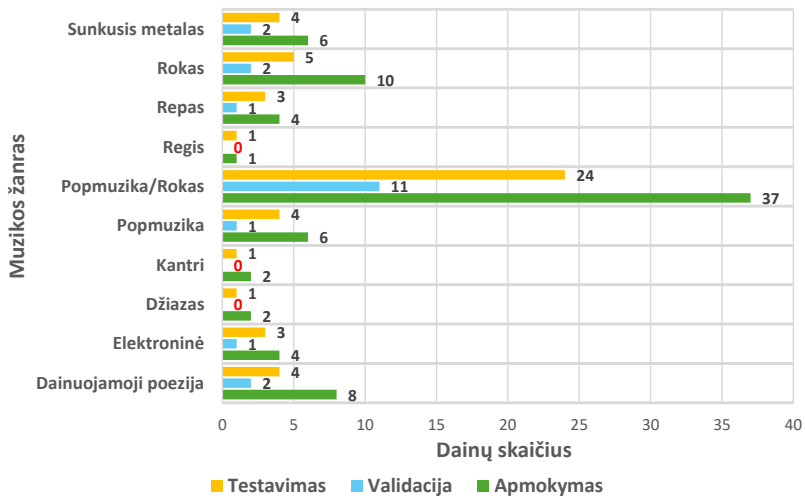
Apžvelgus Moisesdb dainų pasiskirstymą pagal žanrus (žr. 2 pav.), matoma, kad daugiausia duomenų turima roko muzikos žanro. Apmokymo validacijos ir testavimo poaibiuose yra atitinkamai 56, 14, 35 dainos. Moisesdb susideda iš vos 1 džiazo ir bosanovos žanro dainų. Tyrimo metu džiazo daina buvo priskirta apmokymo, o bosanovos – testavimo poaibiams. Dainų skaičius tarp žanrų yra nevienodas – roko žanras turi žymiai daugiau dainų, nei visi kiti.



2 pav. Dainų skaičiaus pasiskirstymas Moisesdb duomenų rinkinyje pagal apmokymo, validacijos ir testavimo poaibius.

Apžvelgus Musdb18-HQ dainų pasiskirstymą pagal žanrus (žr. 3 pav.), matoma, kad daugiausia duomenų turima popmuzikos/roko muzikos žan-

ro. Apmokymo validacijos ir testavimo poaibiuose yra atitinkamai 37, 11, 24 dainos. Musdb18-HQ turi vos 2 regio žanro dainas. Tyrimo metu regio žanro dainos buvo paskirstytos po 1 dainą į apmokymo ir testavimo poaibius. Dainų skaičius tarp žanrų yra nevienodas – popmuzikos/roko žanras turi žymiai daugiau dainų, nei visi kiti, tačiau tarp likusių žanrų pasiskirstymas yra tolygesnis, nei Moisesdb.



3 pav. Dainų skaičius pasiskirstymas Musdb18-HQ duomenų rinkinyje pagal apmokymo, validacijos ir testavimo poaibius.

4 Rezultatai

Modeliui apmokyti su Musdb18-HQ skirta 100 epochų, Moisesdb – 65 epochos. Abiem duomenų rinkiniams mokymosi greitis nustatytas $5e-4$, partijos dydis 32. Modelio ir neuroninio tinklo konfigūracijos parametrai nebuvo optimizuojami ir keičiami. Apmokius modelį su vienu duomenų rinkiniu, buvo įvertinta jo SDR metrika (toliau – SDR) pagal skirtingus garso šaltinius ir visų garso šaltinių SDR pagal skirtingus muzikos žanrus. SDR parodo santykį tarp atskaitos signalo ir garso šaltinių atskyrimo metu susidariusių triukšmų ir išreiškiama decibelais (dB) [7]. Įvertinimui naudojamas to paties duomenų rinkinio testavimo poaibis ir nekeičiant apmokymo duomenų rinkinio naudojamas kito duomenų rinkinio testavimo poaibis. Žemiau pateikiami gauti rezultatai (žr. 1 lentelę, žr. 2 lentelę).

1 lentelė. Visų žanrų vidutinis garso šaltinių SDR pagal skirtingus duomenų rinkinių poaibius.

Apmokymo (AP) ir įvertinimo (IV) duomenų rinkinys	Bosai	Būgnai	Kiti garso šaltiniai	Vokalai	Visi garso šaltiniai
AP: Mudb18-HQ, IV: Musdb18-HQ	8,90	9,90	6,46	9,74	8,75
AP: Musdb18-HQ, IV: Moisesdb	10,87	10,93	7,07	10,17	9,76
AP: Moisesdb, IV: Moisesdb	11,99	11,90	7,88	10,74	10,63
AP: Moisesdb, IV: Musdb18-HQ	8,31	9,27	5,88	9,70	8,29

2 lentelė. Skirtingų duomenų rinkinių, naudotų apmokymui ir įvertinimui visų garso šaltinių vidutinė SDR pagal žanrus.

Žanras	AP: Mudb18-HQ, IV: Musdb18-HQ	AP: Musdb18-HQ, IV: Moisesdb	AP: Moisesdb, IV: Moisesdb	AP: Moisesdb, IV: Musdb18-HQ
Bliuzas	-	13,14	14,57	-
Elektroninė	6,18	7,22	8,10	6,14
Kantri	12,85	9,97	10,79	13,57
Rokas	8,71	9,09	9,71	8,92

Naudojant Musdb18-HQ duomenų rinkinį apmokymui/testavimui, geriausiai išskirti garso šaltiniai yra būgnai ir vokalai, kurių SDR yra 9,9 dB ir 9,74 dB atitinkamai. Naudojant Moisesdb duomenų rinkinį apmokymui/testavimui, geriausiai išskirti garso šaltiniai yra bosai ir būgnai, kurių SDR yra 11,99 dB ir 11,90 dB atitinkamai. Visais duomenų rinkinių poaibių atvejais, modelis prasčiausiai išskiria kitų garso šaltinių garso takelį ir dažniausiai prasčiausiai išskiria elektroninės muzikos žanrą. Turint skirtingus apmokymo ir testavimo poaibius, naudojant Musdb18-HQ rinkinį apmokymui, o Moisesdb testavimui, gaunama aukštesnė SDR, nei apmokymui naudojant Moisesdb ir testavimui Musdb18-HQ. Kai įvertinimui naudojamas Moisesdb duomenų rinkinio testavimo poaibis, geriausiai išskiriami bliuzo žanro garso šaltiniai, nors bliuzo apmokymo ir validacijos poaibiuose buvo vos po vieną dainą, o testavimo – 2 dainos. Kai apmokymui naudojamas Moisesdb duomenų rinkinys, o testavimui Musdb18-HQ, geriau atskiriami kantri muzikos žanro garso šaltiniai, nei apmokymui naudojant Musdb18-HQ, nors

visų žanrų skirtingų garso šaltinių vidutinė SDR visada žemesnė, naudojant skirtingus apmokymo ir testavimo duomenų rinkinius. Daugiausia abiejuose duomenų rinkiniuose apmokymo poaibiuose yra roko muzikos žanro dainų, tačiau SDR šiame žanre niekada nėra aukščiausia.

5 Išvados

SCNet modelį apmokius pasirinktais dviem duomenų rinkiniais ir įvertinus modelio garso šaltinių atskyrimo muzikoje rezultatus, gautos šios išvados:

- SCNet modelis negali gerai apibendrinti skirtingų apmokymo ir testavimo duomenų rinkinių, nes SDR visada būna žemesnė, nei modelį apmokant ir testuojant su tuo pačiu duomenų rinkiniu. Modelio, apmokyto ir testuoto su Musdb18-HQ duomenų rinkinio poaibiais visų garso šaltinių SDR yra 8,75 dB. Tuo tarpu naudojant Musdb18-HQ testavimo poaibį ir modelį apmokius su Moisesdb apmokymo poaibiu, SDR sumažėja iki 8,29 dB. Panaši tendencija matoma ir naudojant Moisesdb testavimo rinkinį. Modelį apmokius su Moisesdb apmokymo poaibiu, visų garso šaltinių SDR yra 10,63 dB, o su Musdb18-HQ SDR yra 9,76 dB.
- Roko muzikos žanras yra sudėtingas jį sudarančių garso šaltinių atžvilgiu ir didelis duomenų kiekis (56 dainos Moisesdb ir 10 dainų Musdb18-HQ apmokymo poaibiuose) nelemia geriausių garso šaltinių atskyrimo rezultatų tarp skirtingų muzikos žanrų. Roko muzikos žanro visų garso šaltinių SDR yra nuo 8,71 dB iki 9,71 dB, priklausomai nuo skirtingų apmokymo ir testavimo duomenų rinkinių poaibių.
- Muzikos žanrai, turintys nedidelį apmokymo duomenų kiekį kaip bliuzas ir kantri, pasiekia aukštą SDR. Bliuzo SDR yra 14,57 dB, naudojant Moisesdb apmokymo ir testavimo rinkinių poaibius, o kantri SDR yra 13,57 dB, naudojant Moisesdb apmokymo ir Musdb18-HQ testavimo poaibius. Tai rodo, kad didelis apmokymo duomenų kiekis nebūtinai užtikrina geresnius garso šaltinių atskyrimo rezultatus ir signalizuoja, kad panašūs bliuzo ir kantri muzikos žanrai tarpusavyje turi artimas akustines savybes dėl kurių modelis gali pasiekti aukštesnę visų garso šaltinių SDR.
- Moisesdb duomenų rinkiniui reikėjo mažiau epochų (65), nei Musdb18-HQ (100), kad pasiekti panašius SDR rezultatus modelį įvertinant su Musdb18-HQ testavimo poaibiu. Moisesdb duomenų rinkinys yra di-

desnis ir susideda iš 235, o Musdb18-HQ iš 150 dainų. Tai rodo, kad Moisesdb duomenų rinkinio didesnė duomenų įvairovė lemia modelio greitesnį mokymąsi per epochas. Taip pat, didesnis modelio apmokymo epochų skaičius ir mažesnis duomenų kiekis lemia geresnį modelio apibendrinimą, kadangi Musdb18-HQ visų garso šaltinių SDR buvo aukštesnė, kai testavimo poaibis buvo pakeistas iš Musdb18-HQ į Moisesdb (9,76 dB > 8,75 dB). Moisesdb atveju buvo atvirkščiai, SDR buvo žemesnė (8,29 dB < 10,63 dB).

Padėka. Dėkojame Vilniaus universiteto ITAPC padaliniiui už suteiktus IT išteklius (HPC), kurie leido greičiau apmokyti modelius bei supaprastino viso tyrimo atlikimą. Tyrimas finansuojamas pagal LR Švietimo, mokslo ir sporto ministerijos programą „Universitetų ekselencijos iniciatyvos“ (LR ŠMSM mokslo plėtros programos pažangos priemonė Nr. 12-001-01-01-01 „Gerinti mokslo ir studijų aplinką“).

Literatūra

- [1] I. Pereira, F. Araújo, F. Korzeniowski, and R. Vogl, “Moisesdb: A dataset for source separation beyond 4-stems”, arXiv.org, 2023, doi: <https://doi.org/10.48550/arXiv.2307.15913>.
- [2] W.-T. Lu, J.-C. Wang, Q. Kong, and Y.-N. Hung, “Music Source Separation with Band-Split RoPE Transformer”, arXiv.org, 2023, doi: <https://doi.org/10.48550/arXiv.2309.02612>.
- [3] Z. Rafii, A. Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and R. Bittner, “MUSDB18 - a corpus for music separation”, Zenodo (CERN European Organization for Nuclear Research), 2017, doi: <https://doi.org/10.5281/zenodo.1117372>.
- [4] Y. Luo, Z. Chen, and T. Yoshioka, “Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation”, arXiv.org, 2019, doi: <https://doi.org/10.48550/arXiv.1910.06379>.
- [5] X. Mei, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang, “Audio Captioning Transformer”, arXiv.org, 2021. <https://arxiv.org/abs/2107.09817>.
- [6] W. Tong et al., “SCNet: Sparse Compression Network for Music Source Separation”, arXiv.org, 2024, doi: <https://doi.org/10.48550/arXiv.2401.13276>.
- [7] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation”, IEEE Transactions on Audio, Speech and Language Processing, vol. 14, no. 4, pp. 1462–1469, Jul. 2006, doi: <https://doi.org/10.1109/tsa.2005.858005>.

Viršelio dailininkė *Jurga Tėvelienė*
Maketuotoja *Vida Vaidakavičienė*

Vilniaus universiteto leidykla
Saulėtekio al. 9, LT-10222 Vilnius
info@leidykla.vu.lt, www.leidykla.vu.lt
Knygos internete: www.knygynas.vu.lt
Mokslo periodikos žurnalai: www.zurnalai.vu.lt

