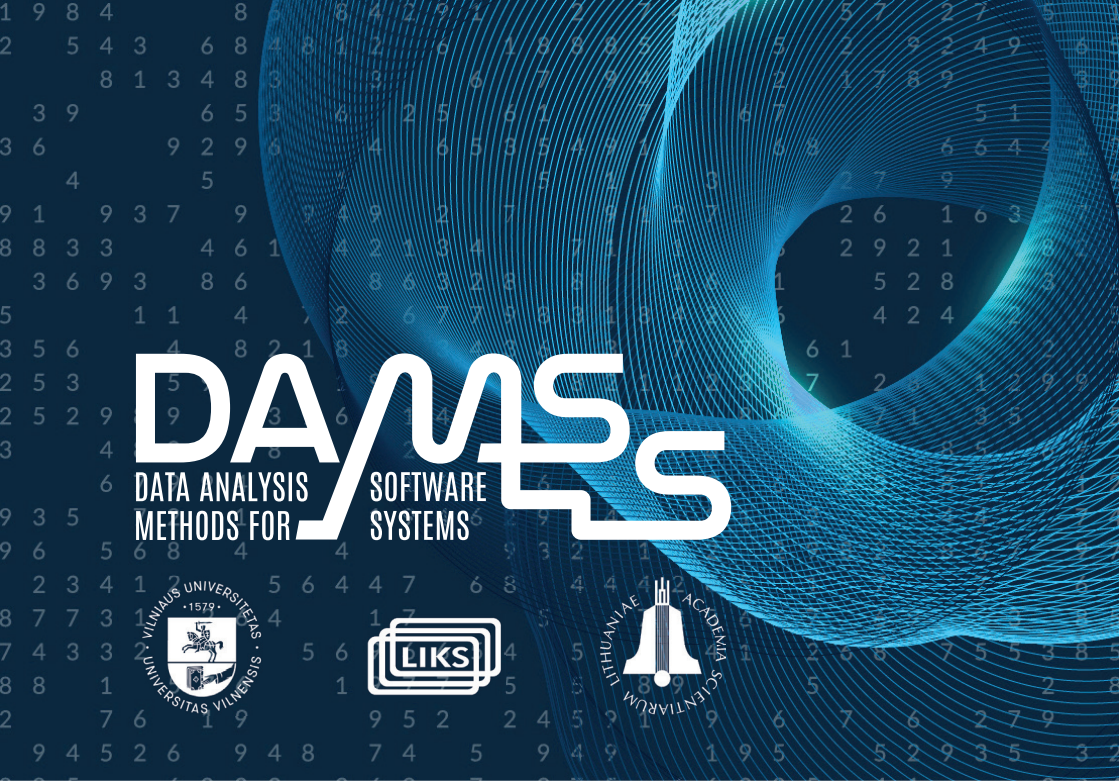


15th Conference on

DATA ANALYSIS METHODS
for Software Systems

November 28–30, 2024

Druskininkai, Lithuania, Hotel "Europa Royale"
https://www.mii.lt/DAMSS

LITHUANIAN COMPUTER SOCIETY
VILNIUS UNIVERSITY INSTITUTE OF DATA SCIENCE AND DIGITAL TECHNOLOGIES
LITHUANIAN ACADEMY OF SCIENCES

15th Conference on

DATA ANALYSIS METHODS
for Software Systems

November 28–30, 2024

Druskininkai, Lithuania, Hotel "Europa Royale"
https://www.mii.lt/DAMSS

VILNIUS UNIVERSITY PRESS
Vilnius, 2024

Contacts:

Dr. Jolita Bernatavičienė
jolita.bernataviciene@mif.vu.lt
Tel. (+370 5) 2109 315
Prof. Olga Kurasova
olga.kurasova@mif.vu.lt

https://doi.org/10.15388/DAMSS.15.2024
ISBN 978-609-07-1112-5 (digital PDF)

Preface

DAMSS-2024 is the 15th International Conference on Data Analysis Methods for Software Systems. It is held in Druskininkai, Lithuania, at the same venue and time every year. The conference successfully achieves its primary goal of lively scientific communication among participants. The conference's history dates back to 2009 when 16 papers were presented. It began as a workshop related to one international project and has evolved into a well-known conference. The initiator of the workshop was the Institute of Mathematics and Informatics which later became the Institute of Data Science and Digital Technologies at Vilnius University. The Lithuanian Academy of Sciences and the Lithuanian Computer Society are partners. The long-standing experience of such partnerships guarantees a high quality of organization and academic relevance.

The conference has grown into one of Lithuania's most important events in computer science and computer engineering, promoting advanced interdisciplinary research and contributing to real innovation. This year's conference includes 87 presentations from 12 countries, with 128 registered participants from Estonia, Latvia, Poland, Croatia, Spain, Portugal, Italy, Serbia, Romania, Israel, Romania, and Lithuania. Such internationality of the conference helps strengthen projects, promote knowledge exchange, and develop innovative ideas globally. Representatives from six Lithuanian universities participated in the conference. So, we have the central annual meeting for Lithuanian computer scientists.

The annual organization of the conference facilitates the rapid exchange of new ideas within the scientific community. DAMSS is also unique in its discussion of opportunities: researchers from Lithuania and abroad, as well as businesses and the public sector, are encouraged to develop joint projects, apply research to practice, and address business and social challenges. The aim is to align research with market needs and business competencies. An essential highlight of DAMSS is the participation of young scientists. Doctoral and Master's students

from Lithuania and other countries can present their research papers, participate in discussions, get acquainted with world-class research, and gain valuable experience. Traditionally, most presentations are posters. Oral sessions are mainly for keynote speakers.

Six IT companies supported DAMSS-2024. This proves the relevance of the conference topics to the business sector. The Lithuanian Research Council and the National Science and Technology Council (Taiwan) support the conference, too.

Topics covered by the conference include Applied Mathematics, Artificial Intelligence, Big Data, Bioinformatics, Blockchain Technologies, Business Rules Software Engineering, Cybersecurity, Data Science, Deep Learning, High-Performance Computing, Data Visualization, Machine Learning, Medical Informatics, Modelling Educational Data, Ontological Engineering, Optimization, Quantum Computing, Signal and Image Processing.

This book covers the presentations from the DAMSS-2024 conference.

PARTNER

International Federation for Information Processing
ifip.org

DAMSS 2024 SUPPORTED BY:

General Sponsors

3RTechnology
3rt.lt

Neurotechnology
neurotechnology.com

Novian
novian.lt

Research Council of Lithuania
lmt.lt

Main Sponsors

National Science and Technology Council
nstc.gov.tw

EPAM Systems
epam.com

Sponsors

Asseco Lithuania
asseco.lt

Visoriai Information
Technology Park (VITP)
vitp.lt

Order in Document Chaos: Logistics Documents Classification

Danylo Abramov[1,2], Eimantas Zaranka[1,2], Monika Zdanavičiūtė[1,2], Nerijus Šakinis[1,2], Tomas Krilavičius[1,2]

[1] Vytautas Magnus University
[2] Centre for Applied Research and Development

danylo.abramov@vdu.lt

Arun Kumar Mishra wrote that every movement of goods from one point to the next must have the attached documents. According to Statista, logistics industry worldwide grows 0.5 trillion dollars each year, meaning more transportation is required the more documents needs to be processed. To efficiently manage huge volumes of documents and automate a decision-making process a classification system is required. This research focuses on logistics documents classification utilizing deep learning and machine learning algorithms. For this study, 50 GB of unlabeled data were presented, and the initial experiments were conducted using 5078 manually selected documents. Manually selected documents were assigned to 4 commonly used logistics document categories: CMRs, invoices, receipts, and others. The dataset was split into train and test sets, where 80% or 4058 of the documents were designated for training and 20% or 1014 of the documents for testing. Five main preprocessing steps were applied: convertion from PDF to JPG, resizing, deskewing, tint and noise removal. Two main methodologies were applied, application of neural networks and traditional machine learning classification techniques. Both approaches utilized pretrained backbone models on ImageNet. For neural networks we used EfficientNet80, VGG16, MobileNet, ResNet50, DenseNet and InceptionV3. The neural network with the ResNet50 backbone outperformed other models achieving 0.9582 accuracy, 0.9593 precision, 0.9582 recall and 0.9585 F1 score. Rest models showed comparable results in performance evaluation: EfficientNet80 achieved 0.9467, VGG16 0.9176, MobileNet 0.9307, DenseNet 0.9387 and InceptionV3 0.9387 F1 scores. In addition, tradi-

tional machine learning classifiers, including Support Vector Machines (SVM), Random Forest, K-Nearest Neighbors (KNN), and XGBoost (XGB), were trained using features extracted from the ResNet50 backbone. The best-performing machine learning model was the Support Vector Classifier, achieving an accuracy of 0.9471, 0.9470 precision, 0.9471 recall and 0.9466 F1 score, while the XGBoost classifier, Random Forest, and K-Nearest Neighbors classifiers achieved F1 scores of 0.9440, 0.9455, and 0.9281, respectively. The research showed that the most promising solution for logistics document classification is the ResNet50 model and that it could be implemented in logistic environments to automate document separation. Future research will focus on dataset expansion utilizing pre-trained ResNet50 model to label the remaining unused documents and further fine-tune models to enhance model F1 score, minimizing the need for human intervention in document classification.

The Potential of Humanoid Robots for Children with Autism Spectrum Disorders: a Preliminary Study

Linas Aidokas

Institute of Data Science and Digital Technologies
Vilnius University

linas.aidokas@mif.vu.lt

Autism spectrum disorder (ASD) is a developmental disorder characterised by impaired social communication, restricted interests, lack of emotional control and repetitive behaviours. Autistic people are impaired in their abilities for social interaction, social communication and imagination. Previous researchers have linked the delay in social development in autistic people to emotional impairment. People with autism spectrum disorders have problems with interpersonal relationships and difficulties building relationships with their environment. Humanoid robotics in combination with developmental therapy can help people build interpersonal and environmental relationships. The causes of ASD are still unknown, but studies have confirmed that autistic children's behavior improves with intervention. It is reported that many people with ASD achieve levels of engagement in tasks through interaction with robots and that robotic systems can be useful for some people with ASD to promote social communication and support interpersonal communication. Based on the evidence that some interventions work, we have developed a curriculum to help autistic children build interpersonal relationships by using robots as a facilitators. Determining the optimal motion sequences of robots when interacting with humans with ASD is important to achieve more natural human-robot interactions and to explore the full potential of robotic interventions. The main goal is to create a learning environment for autistic children in which they can build relationships with other people using verbal and non-verbal interactions. The research focuses on the development of learning activities mediated by a robot to help autistic children communicate with other people. We have

designed and developed a set of robotic movements and a set of verbal expressions that could be used by teachers and therapists to teach communication to autistic children. There is a possibility that visualisation by humanoid robots instead of humans may lead to a higher level of task engagement in people with ASD. With appropriate adaptations, the robot can be used to teach a wide range of autistic individuals.

KATH: A Comprehensive Platform for Integrating DNA Analysis Tools, Genetic Databases, and Open-Data Standards in Genetic Research

Kazimieras Bagdonas

Kaunas University of Technology

kazimieras.bagdonas@ktu.lt

The development of the KATH system represents a novel approach to addressing critical challenges in modern genetic research, particularly in integrating diverse DNA analysis tools and databases. Developed in collaboration with researchers from Harvard University, KATH serves as a comprehensive platform designed to assist geneticists in managing, analyzing, and interpreting DNA data. The system facilitates seamless access to key genetic databases, including Clinvar, gnomAD, and LOVD, while integrating advanced DNA analysis tools such as CADD and REVEL. KATH's architecture allows for the incorporation of additional algorithms, positioning it as a scalable and flexible solution for diverse research needs. A key feature of KATH is its support for open-data publication standards, enabling researchers to share and validate workflows with the broader scientific community.

In this study, we demonstrate the capabilities of KATH by analyzing ancient DNA sequences obtained from three sets of mummified remains published on the NIH website. We applied KATH's integrated tools to identify pathogenic gene mutations and conducted a comparative analysis with the reference human genome. Our results highlight the system's ability to efficiently process and analyze complex genetic data, providing critical insights into the genetic makeup of populations and offering broader implications for evolutionary biology and disease research.

The KATH platform represents a significant advancement in genetic research by combining data integration, workflow sharing, and commercialization opportunities in a user-friendly environment. Its potential for rapid validation and verification of results and the ability to facilitate cross-disciplinary collaborations underscores its value as a transformative tool for geneticists and researchers worldwide.

Self-Organization of Bacterial and Human Populations: Same Model, Different Scale

Romas Baronas[1], Boleslovas Dapkūnas[1], Remigijus Šimkus[2]

[1] Institute of Computer Science
Vilnius University
[2] Institute of Biochemistry
Vilnius University

romas.baronas@mif.vu.lt

The direct movement of organisms toward a chemoattractant and away from a chemorepellent is called chemotaxis. Since the pioneering work of Keller and Segel, published in 1970, the dynamics of chemotactic population and chemoattractant are usually modelled mathematically using a system of nonlinear equations of the reaction-diffusion-chemotaxis type. Although the chemotaxis-type models are mostly applied in biology, they have also been applied to solve problems in various other fields, including the social sciences. Neto and Claeyssen have applied a chemotaxis model to describe economic growth containing capital-induced labour migration. Short et al. proposed a chemotaxis-type system to describe the movement of criminals toward increasing concentrations of an attractiveness value. In this work, the chemotaxis-based self-organization and patterning relationship between bacterial and human populations were investigated computationally. The pattern formation in luminescent suspensions of E. coli was studied as an example of bacterial self-organization. Two aspects of human social behaviour were also analysed: the formation of economic agglomerations, which are regions with higher levels of capital and labour force than their neighbours, and the formation of crime hotspots, when criminals move toward a higher concentration of an attractiveness value. Nonlinear two- and one-dimensional-in-space reaction-diffusion-chemotaxis models were used to simulate the pattern formation in all three chemotactic populations living within a restricted area - a circle. The numerical simulation was carried out using the finite difference technique. Although bacterial self-organization and geographical migration of the human population have large differences, both movements share close similarities except for the difference in scale.

Highly Imbalanced Data Case: Pattern-Guided Feature Selection to Detect Financial Fraud

Dalia Breskuvienė, Gintautas Dzemyda

Institute of Data Science and Digital Technologies
Vilnius University

dalia.breskuviene@mif.vu.lt

Financial fraud represents widespread challenges in the modern financial landscape, necessitating innovative methodologies and technologies to detect and mitigate losses resulting from sophisticated fraudulent tactics.

Fraud detection systems need to respond quickly while maintaining the stability of the financial ecosystem. To address the critical issues identified in fraud detection, we introduce a novel feature selection method, FID-SOM (Feature Selection for Imbalanced Data Using Self-Organizing Maps), specifically designed to overcome the challenges posed by imbalanced data in fraud detection scenarios. Feature selection can significantly enhance classification performance related to accuracy and response time. Given the inherent imbalance in fraud detection data, feature selection must be conducted with high attention. To accomplish this task, we utilize Self-Organizing Maps (SOMs), an artificial neural network designed to organize data into clusters based on similarity, simplifying the complex data landscape typical of financial fraud scenarios. FID-SOM is engineered to address the challenge of high-dimensional data in scenarios characterized by highly imbalanced data. It has been specifically developed to efficiently process and analyze vast, complex datasets commonly encountered in the financial sector. More importantly, it demonstrates adaptability to the dynamic nature of big data environments, thereby making it a practical and reassuring solution for fraud detection.

The proposed method's distinctive aspect lies in forming a new dataset containing the Best-Matching Units of the trained SOM as vectors of attributes corresponding to the initial features. These attributes are

sorted in descending order based on the score calculated using methods like KL-divergence Information Gain and/or Variance. By retaining the required number of attributes, we select features corresponding to those attributes for further analysis. The proposed FID-SOM method has demonstrated its capacity to perform comparably to existing methods and exhibits the potential to surpass them. This potential for superior performance makes FID-SOM a thrilling prospect in fraud detection.

Enhancing Cybersecurity Using Keystroke Dynamics and Data Fusion Techniques

Arnoldas Budžys, Viktor Medvedev, Olga Kurasova

Institute of Data Science and Digital Technologies
Vilnius University

arnoldas.budzys@mif.vu.lt

In response to the growing cyber threats facing critical infrastructure, this research presents a deep learning-based authentication system that uses keystroke dynamics to strengthen security against unauthorised access, including insider threats. Traditional methods often fall short in such sensitive environments, thus advanced solutions are required. Our approach integrates keystroke-based behavioural biometrics with data fusion techniques that transform keystroke dynamics data into image representations. Using a new Gabor Filter Matrix Transformation method, we transform keystroke dynamics into graphical formats, allowing enhanced pattern recognition. A Siamese neural network with a triplet loss function processes these images to accurately distinguish between authorised and unauthorised users. Our extensive experiments on datasets such as Carnegie Mellon University and GREYC-NISLAB , covering over 54,000 password samples, demonstrate that the proposed method achieves higher authentication accuracy comparing to related works. The system achieves an equal error rate value of 0.045, outperforming traditional models and offering scalable adaptability to different password types and user profiles. Empirical studies using publicly available datasets confirm the effectiveness of the approach, as indicated by a reduction in equal error rate and improved user authentication accuracy. This study highlights the need for advanced authentication methods to address insider threats and unauthorised access to critical infrastructure. By integrating deep learning and data fusion, our approach provides a scalable and accurate solution to this pressing security challenge.

DisinfoDetect: A Dashboard-Driven Solution for Identifying Misinformation in Media

Veronika Bryskina[1,2], Bohdan Zhyhun[1,2], Paulius Savickas[1,2], Milita Songailaitė[1,2], Tomas Krilavičius[1,2]

[1] Vytautas Magnus University
[2] Centre for Applied Research and Development

veronika.bryskina@vdu.lt

Fake news is becoming a recognised problem in society, creating large volumes of misinformation and raising questions about media integrity. The spread of misinformation has serious consequences, affecting public opinion, undermining trust in institutions, and sometimes leading to real-world harm. The sheer volume of information produced daily making manual verification of news impractical and time-consuming, leading in the need for automated tools that can assist with the issue. In this work, we present a proof of concept of a tool designed to detect disinformation in English-language media sources. At the core of our solution is the RoBERTa model, fine-tuned on a diverse set of articles from American (mostly for non-disinformation) and Russian (mostly for disinformation) English-language sources. This approach allows us to save time by avoiding the need to train the model from scratch. Additionally, RoBERTa's advanced capabilities, such as its ability to grasp the context and meaning of complex sentences and its bidirectional nature (analysing sentences from both the beginning to the end and in the opposite way), enable it to capture long-range dependencies between words. These features are valuable in identifying complex linguistic structures, such as hyperbole, unverified claims, biased language, and sarcasm, commonly found in news sources. Besides, we are using this approach to collect and analyse texts from Lithuanian media sources, focusing on various domains such as politics, economy, society, and business, to identify the prevalence of disinformation within these outlets. This allows us to gain insights into how misinformation is distributed across

different sectors. To enhance the usability of our solution, we developed a Dash library-based web application that displays the model's evaluation results and delivers an intuitive interface for users to interact with the system. This application allows users to upload news articles or plain text, analyse them for potential misinformation, and view detailed, real-time feedback on the model's predictions. Additionally, the system allows for batch processing of multiple texts at once, providing scalability for larger datasets. The result of our work is a developed and tested version of a deep learning based disinformation detection system, capable of analysing disinformation in the selected national and international news and media sources and presenting the analysis results with an informative dashboard.

Leveraging Phone Sensors for Early Detection of Symptom Changes in Cancer Patients

Gabrielė Dargė[1,2,3], Gabrielė Kasputytė[1,2,3], Ričardas Krikštolaitis[1,3], Tomas Krilavičius[1,2,3], Dovilė Kuizinienė[1,2,3], Adomas Bunevičius[4]

[1] Vytautas Magnus University
[2] Centre for Applied Research and Development
[3] Centre of Excellence of AI for Sustainable Living and Working (SustAInLivWork)
[4] Lithuanian University of Health Sciences

gabriele.darge@vdu.lt

Cancer patients often experience fluctuations in their health, which can be challenging to monitor continuously through traditional methods, such as patient self-reporting during clinic visits. With the growing prevalence of smartphones and their built-in sensors, new opportunities have emerged for real-time, passive monitoring of patient behaviour and health. These phone sensors – tracking factors like physical activity, location, and phone usage – can provide continuous data that may signal subtle changes in a patient's condition even before they become consciously aware of worsening symptoms. This research explores the use of phone sensor data to understand how real-time behavioural changes might be related to the onset of symptoms in cancer patients. By leveraging sensor data alongside patient surveys, the study investigates how patterns in daily activity, such as movement and phone usage, might signal an impending change in symptoms. The study included patients grouped by cancer type, gender, functional status (ECOG scale), and age. Loess regression and statistical analysis were applied to examine the dynamics of sensor data before and after symptom onset. Various variables describing patients' activity, such as distance from home, screen time ratio, and others, were evaluated over time using phone sensors. The study's results revealed that patients' activity levels, or time spent at home, begin to decline even before the patient reports the onset

of symptoms, such as fatigue. The study found that tracking sensor readings can help predict worsening symptoms and facilitate early intervention, which is crucial to the quality of patient care. These findings highlight the importance of using phone sensors for real-time patient monitoring, enabling personalised care for cancer patients.

Acknowledgements: This research was co-funded by the European Union under Horizon Europe programme grant agreement No. 101059903 and by the European Union funds for the period 2021-2027 and the state budget of the Republic of Lithuania financial agreement Nr. 10-042-P-0001.

Transforming Black-Box Models into Explainable AI for Breast Cancer Recognition

Sobia Dastgeer, Povilas Treigys

Institute of Data Science and Digital Technologies
Vilnius University

Sobia.dastgeer@mif.stud.vu.lt

Artificial intelligence (AI) with deep learning (DL) models using medical imaging has sparked a revolutionary wave in healthcare in recent years. DL models often function as black boxes, making their decision-making process difficult to interpret. These considerations get more significant when people's health is at risk, as in the involvement of AI applications used in healthcare. The situation is further worsened by the fact that, unlike conventional machine learning algorithms, state-of-the-art deep learning algorithms consist of complex connected structures, millions of parameters, and a black box mentality that provides no insight into how they operate inside. A recent field of study in machine learning called Explainable Artificial Intelligence (XAI) aims to understand how AI systems make judgement that are hidden from visibility. The idea of explainable AI (XAI) emerged as a response to the issue of AI black boxes, with the goal of offering explanations that are transparent, believable, and sensitive to human understanding. For improved comprehension, a case study regarding the function of XAI in the identification of Breast cancer (BC) with their performances are discussed in this study. BC is a prevalent and extremely deadly disease. It is currently the second largest cause of cancer-related mortality among females globally. This study focuses on the integration of several XAI techniques like Grad-CAM, LIME, SHAP, PDPs and other deep learning models used for breast cancer classification and detection and how these explainability strategies are the better way to establish trustworthiness when utilizing AI systems in the healthcare industry. We will examine several key ideas in XAI, outline various challenges related to XAI in the healthcare industry, and explore

whether XAI can actually progress the field by fostering greater understanding and trust among other things.

Acknowledgements: This research is funded under the Programme "University Excellence Initiatives" of the Ministry of Education, Science and Sports of the Republic of Lithuania (Measure No. 12-001-01-01-01 "Improving the Research and Study Environment").

Red Team Tactics Against Malware Detection Using Adversarial Attacks

Juozas Dautartas, Arnoldas Budžys, Haroldas Jomantas, Olga Kurasova, Viktor Medvedev

Institute of Data Science and Digital Technologies
Vilnius University

juozas.dautartas@mif.stud.vu.lt

Static and dynamic malware analysis has been used for a while among cybersecurity professionals and researchers. While static analysis can be used to gather useful information about file features such as strings, hash values, creation date, import address tables, sections and many more, attackers have adapted to these analysis methods quite easily. Dynamic analysis, on the other hand, offers a much deeper insight into what the program does as it monitors the process itself. However, this analysis method requires an isolated virtual environment and can slow down the workflow of the system as it requires additional resources. By combining these two methods, security researchers and security products started using machine learning and deep learning algorithms to detect and mitigate known and unknown cyber threats. Application of these technologies allows antivirus and EDR (Endpoint Detection and Response) systems make decisions faster regarding a file or process is malicious or not. It helps to save some computational resources as well because deeper inspection and classification can take place in a centralised remote server instead of on local computer resources. It comes with no surprise that threat actors started to investigate and search for weaknesses in these detection algorithms as well. Therefore, a deeper insight regarding weaknesses in these deep learning and machine learning models is necessary. Furthermore, deep learning algorithms such as Generative Adversarial Neural Networks, Variational Autoencoders can be used to generate adversarial malware samples that could evade detection. In our research, we aim to design a Command and Control (C2) framework that would use deep learning algorithms to make it more

evasive than standard C2 frameworks. This will help red team members to better train blue teams and notice anomalies by not being over-reliant on automated tools.

Acknowledgements: This project has received funding from the Research Council of Lithuania (LMTLT), agreement No S-MIP-24-116.

Cointegration of Functional Time Series

Urtė Deinoravičiūtė[1], Jurgita Markevičiūtė[1], Povilas Treigys[2]

[1] Institute of Applied Mathematics
Vilnius University
[2] Institute of Data Science and Digital Technologies
Vilnius University

urte.deinoraviciute@mif.stud.vu.lt

Cointegration is a method used to estimate the existence of a long-term equilibrium relationship between two or more time series. While this method is widely utilized for traditional time series, some research is still needed to make cointegration fully usable in the context of functional time series (FTS). Such progress could provide new applications in different fields, from finance and economics to natural sciences. Even though the tests for stationarity or unit root of functional time series are well-defined, with some of them practically implemented, cointegration tests are mostly in the theoretical development phase. It is an active field of research with contributions from M. Franchi and P. Paruolo (2020), M. Ø. Nielsen, W.-K. Seo and D. Seong (2023), W.-K. Seo (2024) and others. This poster presentation provides an overview of the literature about integration (different stationarity and unit root tests) and cointegration testing for functional time series. Furthermore, some examples of real-world data suitable for FTS cointegration testing are provided. In addition, the poster introduces some studies of cointegration testing on traditional time series in the field of natural sciences.

Acknowledgements: Publication / Research is funded by the Research Council of Lithuania under the Programme "University Excellence Initiatives" of the Ministry of Education, Science and Sports of the Republic of Lithuania (Measure No. 12-001-01-01-01 "Improving the Research and Study Environment"). Project No.: S-A-UEI-23-11.

References

Franchi, M., & Paruolo, P. (2020). Cointegration in functional autoregressive processes. Econometric Theory, 36(5), 803–839. doi:10.1017/S0266466619000306

Nielsen, M. Ø., Seo, W.-K., & Seong, D. (2023). Inference on common trends in functional time series. https://arxiv.org/abs/2312.00590

Seo, W.-K. (2024). Functional principal component analysis for cointegrated functional time series. Journal of Time Series Analysis, 45, 320-330. https://doi.org/10.1111/jtsa.12707

From Data Science to Decision Support and Back!

Boris Delibašić

Faculty of Organizational Sciences
University of Belgrade, Serbia

boris.delibasic@fon.bg.ac.rs

Data Science (DS) and Decision Support Systems (DSS) have long existed as both complementary and competing technologies. While DS has recently assumed a leading role by promising the automatic construction of DSS models, the foundational aim of DSS has been to make the decision-making process more informed and consistent without seeking to replace the human decision-maker. DSS models are traditionally designed to be correct, complete, consistent, comprehensible, and convenient – attributes that DS models do not always possess and whose potentials are often not fully understood within the DS community. In this talk, I will demonstrate how DS models can be integrated into DSS frameworks and how leveraging the strengths of DSS can enhance the effectiveness of DS models in achieving their objectives.

Case Study on Small-Scale Dynamic Neural Networks Explainability

Martynas Dumpis, Dalius Navakauskas

Vilnius Gediminas Technical University

martynas.dumpis@vilniustech.lt

The understanding of dynamic neural networks, particularly those that handle time-variant data, is a key challenge in the development of interpretable AI models. As the application of AI in areas such as time-series analysis, sensor data processing, and real-time decision-making expands, the need for understanding these networks becomes even more critical. Dynamic neural networks are known for their ability to capture temporal dependencies in data, yet their inner workings often remain unclear, raising concerns about trust and reliability. This research focuses on small-scale dynamic neural networks – specifically Long Short-Term Memory (LSTM) and Finite Impulse Response (FIR) Neural Network (NN) – and their ability to provide transparent insights into their decision-making processes. The study addresses a growing need to understand not just how these models perform but why they arrive at specific classification results. A binary classification task using accelerometer data was conducted to assess the neural network's explainability. The accelerometer data, collected from wearable devices, is commonly used for detecting human activities such as running and walking. The challenge lies not only in achieving accurate classification but also in understanding which features of the data are most significant for the model's decisions. We utilise Layer-wise Relevance Propagation to interpret NNs, assigning relevance scores to input features to highlight the signal components that significantly contribute to the model's outputs. By examining the adaptation of relevance over time, this approach uncovers the most critical features driving the classification. Moreover, we use the Short-Time Fourier Transform for time-varying data characterisation and comparison with frequency responses of FIR filters in FIRNN synapses, providing a complimentary means to interpret the results. This study's results enhance the understanding of small-scale dynamic NNs operations by providing valuable insights into their decision-making processes.

Analysis of Event and Human Factor-Based Decision-Making in Cybersecurity Exercises Using MCDM

Karina Čiurlienė[1,2]

[1] Vilnius Gediminas Technical University
[2] Vilnius University

karina.ciurliene@vilniustech.lt

The number of cyberattacks continues to grow steadily and has become more sophisticated in recent years. While organizations are adopting innovative cyber defense technologies and automating processes, detecting and responding to these attacks often remain reactive and event-driven, where human-centric decision-making plays an important role. Also, it must be pointed out that among the most frequent types of attacks are those that deal with human factor vulnerabilities. Cybersecurity is an interdisciplinary field encompassing platforms, systems, technologies, and humans. During cyberattacks or incident response, the behavior and decision-making of cybersecurity professionals are shaped not only by their technical skills and experience but also by psychological and social factors such as emotional states, stress, and fatigue. Cybersecurity professionals tend to act predictably and rationally, however, innate reasoning abilities and emotions often influence their decisions and people make irrational decisions when they are highly stressed. Therefore, recent researches suggest that a holistic approach instead of technical solutions alone is required to contrast cyberattacks.

This research aims to analyze the event-based decision-making of cybersecurity professionals during cybersecurity exercises, emphasizing the human factors. Data for this research were collected through surveys during the international cybersecurity defense exercise „Locked Shields 2024" organized in Vilnius. The user profile, competence assessment data as well as emotional data framed by Plutchik's model of emotions were collected. Criteria and decision-making options were identified. AHP method was used to calculate weighting coefficients and prioritize the criteria. To deepen the analysis of decision-making, MCDM meth-

ods including SAW and TOPSIS were employed. The finding revealed the importance of human factors in decision-making and offered valuable insights for the enhancement of cybersecurity training programs.

Acknowledgements: This research was funded by Research Council of Lithuania under the Programme "University Excellence Initiatives" of the Ministry of Education, Science and Sports of the Republic of Lithuania (Measure No. 12-001-01-01-01 "Improving the Research and Study Environment"). Project No.: S-A-UEI-23-11".

Analysis of Blended Course Quality Metrics: Research, University and Student Perspectives

Beata Gancevska, Simona Ramanauskaitė

Department of Information Systems
Faculty of Fundamental Sciences
Vilnius Gediminas Technical University

beata.gancevska@vilniustech.lt

The quality of education is an important component in providing effective learning and supporting student development. Nowadays, learning delivery includes online and blended formats, making it crucial to maintain high standards in course quality. However, there is a problem in finding a single answer to the question of which metrics define the quality of courses. Different opinions exist regarding quality criteria, with existing research studies identifying various methods and universities having their own standards and metrics. These differing opinions make it difficult to assess what course quality is and how it should be measured. This research discusses different perspectives: first, an analysis was conducted on the solutions used and recommended by existing research papers. Second, it examines the course quality assessment methods and metrics used by higher education institutions. Finally, it analyzes what aspects are most important to students. Students' opinions are very important because the courses are created for them and should meet their needs. Students' feedback is a valuable resource for assessing course quality, as it provides insights about course materials, teaching methods, and other aspects that affect learners' needs. For this reason, this research focuses on analyzing feedback from students in a university study program over the past two academic years. It concentrates on an analysis that utilizes context extraction from students' open-ended responses about completed courses rather than direct survey answers. By examining these responses, the research aims to understand student preferences and needs for blended courses and identify areas for improvement. Text analysis and clustering methods

are used for this task. The research results presented in this poster contribute to the identification of quality criteria and the automation of the e-course quality assessment process. Additionally, the research results also provide suggestions for e-course developers to improve the learning experience.

Optimising Deep Vision Models for Eyeglasses Detection in Low-Power Devices

Henrikas Giedra, Dalius Matuzevičius

Department of Electronic Systems
Vilnius Gediminas Technical University

henrikas.giedra@vilniustech.lt

The rapid advancement of deep learning has driven the integration of artificial intelligence (AI) into edge computing, enabling models to run on resource-constrained devices such as smartphones, wearables, and IoT systems. This shift is crucial for applications requiring real-time processing without relying on cloud infrastructure, which can introduce latency and security risks. One such application is AI-based virtual eyewear measurement, which relies on accurate eyeglass detection from 2D images for virtual try-ons or facial recognition. Deploying deep learning models on low-power devices, however, presents challenges due to limitations in processing power, memory, and energy, necessitating careful optimisation of model architectures to balance accuracy and computational efficiency.

This study investigates optimising convolutional neural networks (CNNs) for eyeglass detection on low-power devices by training various architectures on large datasets and systematically exploring model size, input image resolution, and post-training optimisations to assess their impact on accuracy and inference time.

The experiments reveal the trade-offs between accuracy and inference time across different model complexities. Post-training optimisations reduced inference time with minimal accuracy loss, underscoring their importance for deployment on low-power devices.

These findings provide valuable insights into optimising CNN architectures for low-power AI applications. Beyond eyeglass detection, this research offers general guidelines for developing efficient deep learning models in resource-constrained environments, highlighting the role of

architectural design and post-training optimisation techniques in achieving high performance in edge computing systems.

Acknowledgements: This project has received funding from the Research Council of Lithuania (LMTLT), agreement No S-ITP-24-12.

Investigating Post-hoc Explainability Techniques for Image Segmentation

Rokas Gipiškis, Olga Kurasova

Institute of Data Science and Digital Technologies
Vilnius University

rokas.gipiskis@mif.vu.lt

Post-hoc explanations refer to a class of explainable AI (XAI) methods where explanations about the model are generated after the model has been trained. This subclass of XAI techniques plays a dominant role in interpretable computer vision, with most of the applications falling under image classification and object detection tasks. Here, we evaluate an extension of post-hoc XAI methods to image segmentation tasks based on two criteria: 1) their suitability for generating reliable explanations and 2) their potential for improving models outside the XAI domain. We investigate different types of gradient-based and perturbation-based post hoc explanations to evaluate their effectiveness in image segmentation tasks. We also discuss the trade-off between computation time and the reliability of the generated explanations. The explanations are further evaluated using quantitative deletion curves to measure their sensitivity to various XAI method parameters, such as occlusion type in perturbation-based methods. We also investigate potential application areas beyond XAI, ranging from adversarial attacks to continual learning. Although post-hoc explanations are primarily used to better understand the model in addition to standard model evaluation metrics, their use cases can extend further. We discuss how post-hoc explanations can be applied in adversarial contexts and how they can be used to distil and compress the most important information by identifying the most important regions in an image.

Enhancing 3D Map Generation from Satellite Imagery Using R-CNN

Kęstutis Girnius[1], Aušra Gadeikytė[2], Eglė Butkevičiūtė[1]

[1] Department of Software Engineering
Kaunas University of Technology
[2] Department of Applied Informatics
Kaunas University of Technology

ausra.gadeikyte@ktu.lt

Satellite imagery is used for applications like environmental monitoring, urban planning, climate change studies, infrastructure development, and disaster management. However, converting two-dimensional satellite images into realistic three-dimensional models remains a challenging task due to factors such as data quality variability, limited resolution, atmospheric distortions, and the computational limitations of current methods. The aim of this work is to investigate a region-based convolutional neural network model for object detection and classification of satellite images. Specific problems such as inconsistent quality of input data and the substantial computational requirements for processing large datasets are considered. The study involves training the enhanced R-CNN model on two datasets: a manually annotated dataset of cities in Lithuania and the "Manually Annotated High-Resolution Satellite Image Dataset of Mumbai for Semantic Segmentation." These datasets provide a variety of urban landscapes and structural features, essential for developing a model that can generalise different geographic regions and architectural styles. The training process included 30 epochs, where the model learns to accurately detect features of various buildings. In order to evaluate the performance of the model, metrics such as the Precision-Recall (PR) curve and Average Precision (AP) score were considered, focusing on an Intersection over Union (IoU) threshold of 50%. The trained model achieved an AP score of 0.974, indicating a high level of accuracy in object detection and classification tasks. The investigation of satellite images demonstrates that with enhanced algorithms and

improved processing techniques, satellite imagery can be utilised to create highly accurate, large-scale maps more efficiently. Therefore, there is great potential for future development. Further investigation should focus on refining the proposed model to handle higher-resolution data, integrating additional data sources, and constructing 3D images.

Classification of Cardiovascular Diseases Using Machine Learning Methods and PQRST Intervals Segmentation of Electrocardiogram Records from Holter Monitoring

Uladzislau Hadalau

Vilnius Gediminas Technical University

uladzislau.hadalau@vilniustech.lt

According to data from the World Health Organization, cardiovascular diseases are the most common cause of human death. One of the main reasons for this statistic is the lack of possibilities for mass long-term diagnostics of heart activity using a Holter monitor due to the complexity of processing each record lasting from 24 to 48 hours by cardiologists. Another reason is the absence of online monitoring tools for high-risk patients with automated emergency call triggering.

To solve this problem, a decision was made to implement a fully automated high-precision classifier of cardiovascular system pathologies in Holter monitoring electrocardiogram records using machine learning methods.

Within the framework of the master's thesis, a hypothesis was confirmed about the possibility of accurately identifying the basic set of cardiovascular pathologies, according to the Association for the Advancement of Medical Instrumentation – Electrocardiography Standards – Part 57 (AAMI EC57). By utilizing a public dataset, it was possible to compare the obtained results with existing works and, as of May 2024, the implemented model achieved the best accuracy indicator of 0.994 on a test sample comprising 20% of the total number of records in the dataset, with the original distribution of class sample objects.

However, the pathologies of the AAMI EC57 standard are extremely limited and do not represent significant medical interest. Therefore, a decision was made to significantly expand the set of classified pathologies.

The research problem is the hypothesis testing about the implementation of a classifier for 54 classes of heart rhythm, compiled in cooperation with practicing cardiologists. This includes disturbances of the heart's electrical axis and electrical position, hypertrophy of various parts of the heart, disturbances of heart automatism and excitability, and conduction disturbances (blockades).

Methodology: This work uses a closed dataset, which includes 20,000 collected anonymized Holter monitoring electrocardiogram records from individuals aged 6 to 82 years and lasting from 12 to 48 hours. To evaluate the quality of PQRST interval segmentation methods, a metric based on Intersection over Union (IoU) is used. The evaluation of classification methods is carried out using metrics such as accuracy, recall, precision, F-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

The aim of the research is to confirm the proposed hypothesis by implementing a classifier based on the segmentation of PQRST complex intervals using the range of intervals of existing medical standards for the specified classes. The resulting classifier should comply with international medical standards. In particular, the metrics for the used methods should be comparable to the work of real cardiologists in order to undergo medical certification.

Advanced Ensemble Techniques for IoT Cybersecurity: A Performance Comparison on the CICIoT2023 Dataset

Simran Kaur Hora, Antanas Čenys, Nikolaj Goranin

Vilnius Gediminas Technical University

simran-kaur.hora@vilniustech.lt

The exponential growth of the Internet of Things (IoT) across sectors such as healthcare, smart cities, and industrial systems has introduced substantial cybersecurity challenges. IoT devices, with limited resources and exposure to open networks, are increasingly vulnerable to sophisticated cyberattacks. Traditional security measures and single-model machine learning approaches often fail to handle the complex and dynamic threat landscape associated with IoT. This study evaluates advanced ensemble machine learning techniques, particularly Gradient Boosted Trees (GBT) and XGBoost, for enhancing IoT cybersecurity. Using the CICIoT2023 dataset, which includes seven attack categories (DDoS, DoS, Recon, Web-based, brute force, spoofing, and Mirai), the study compares the effectiveness of these ensemble models against traditional models like Decision Trees (DT) and Random Forest (RF) in both binary (benign vs. malicious) and multi-class classifications (benign and specific attack types). Experimental results show that XGBoost achieved superior performance in both binary and multi-class classification, with a 99.4% accuracy, 76.4% precision, and 78% F1-score in multi-class classification, significantly outperforming DT and RF in distinguishing both general anomalies and specific attack types. These findings highlight the significant potential of ensemble learning in strengthening IoT cybersecurity and underscore the need for robust, adaptive models in IoT-based attack detection systems.

Tracking Equilibrium Departures in Output with O-F SVM and Soft Data

Iulia Igescu

National Bank of Romania, Modelling Department
Romania

iulia.igescu@bnro.ro

Survey data (soft data) offer insights into how believes orient animal spirit, defined by Keynes as "a spontaneous urge to action" - hence sensitive to uncertainty. The example studied here is the industry sector of Romania. The superstar firms dominating this sector had started a major restructuring process in 2018. The Virus Crisis of 2020-2022 added additional shocks to this process: moments of panic at the outbreak of the Virus Crisis prolonged then into a Confidence Crisis. These unusual shocks succeeding in a short time-span plagued hard data (industrial production index) with idiosyncratic anomalies: outliers or novelties. Forecasts on such data could conceal equilibrium stability. As beliefs regarding the state of the economy manifested also as anomalies in soft data (soft data are reported faster than hard data), a policymaker could instead resort to machine learning (One-Factor Support Vector Machines) to monitor soft data anomaly patterns for signals of animal spirit-driven stability changes in hard data.

Application of Image Recognition Methods to Determine Land Use Classes: Lithuanian Case Study

Julius Jancevičius, Diana Kalibatienė

Vilnius Gediminas Technical University

julius.jancevicius@stud.vilniustech.lt

AI application in satellite data is becoming increasingly accessible to interested parties, especially regulatory agencies. This has led providers to improve the quality of satellite imagery and the frequency of data availability. However, the application of satellite data in cloudy areas remains a problem that does not offer many solutions. This research employs an advanced approach to land use classification in Lithuania through satellite image recognition utilising Sentinel-2 data. In consideration of the region's seasonal variability and frequent cloud cover, we apply a structured methodology comprising pre-processing, classification, and post-processing steps to enhance classification accuracy for various land types. Pre-processing methods include image merging, background cleaning, and spectral index calculations (such as NDVI and NDWI) to refine image clarity, while cloud interpolation techniques address cloud cover issues. The developed models are validated with statistical accuracy metrics, such as Cohen's Kappa and F1. This work addresses a methodological gap by focusing on the specific geographic and climatic challenges faced by Lithuania and contributes tools that are useful for environmental monitoring and land management. The prototype developed has the potential to provide invaluable assistance to Lithuanian regulatory agencies by offering precise tools for the sustainable management of land, facilitating the monitoring of environmental changes and ensuring compliance with local regulations.

Integration of Wind Power Generation and Nord Pool Electricity Market Data

Mindaugas Jankauskas, Artūras Serackis

Vilnius Gediminas Technical University

m.jankauskas@vilniustech.lt

The integration of renewable energy sources into the electricity grid has become more important as the world population seeks sustainable and eco-friendly alternatives to fossil fuels. This integration requires accurate energy production forecasts and precise financial estimates to guarantee the economic feasibility of renewable energy projects. For wind farm operators, understanding the technical and financial aspects of energy production is crucial to optimal performance and profitability. Differences in wind patterns and power market pricing create a complexity that standard forecasting approaches, which often focus exclusively on energy output, are unable to effectively capture. This paper presents an innovative method for predicting the profits of a wind park by integrating wind power production data with electricity prices from Nord Pool, Europe's leading power market. By integrating these two essential data sources, we want to present wind farm owners with a more precise and comprehensive perspective of their revenue potential. In contrast to conventional techniques that often overlook the economic aspect of energy production, our methodology employs an integrated prediction model that considers both the total amount of energy produced and its market value. The model employs historical wind power production data and real-time electricity prices to provide a complete forecasting framework that captures the dynamic interaction between energy supply and market demand. We employ advanced time-series analysis and machine-learning methodologies to forecast future earnings with considerable precision. To optimise our predictive models, we utilise TPOT (Tree-based Pipeline Optimization Tool), an automated machine learning framework that uses genetic programming to efficiently explore the parameter space and select optimal model configurations. This approach systematically addresses uncertainty and enhances the performance

of the model, guaranteeing resilience and flexibility in response to fluctuating market conditions, seasonal changes, and unexpected events that could affect both wind energy generation and power prices. The accuracy of the model is evaluated using performance indicators that include the mean absolute error (MAE), the mean squared error (MSE), and the coefficient of determination (R^2). These metrics offer a quantitative evaluation of the model's forecasting ability. We validate the model using actual information obtained from the functioning wind farm and historical pricing records from Nord Pool. Our tests indicate that our integrated forecasting method considerably improves traditional energy-centric models. The model gives wind farm owners more accurate financial estimates, helping them plan their investments, pricing strategies, and market participation. Our methodology not only improves operational decision-making but also provides strategic advantages by improving financial management techniques in the renewable energy sector. Provides investors with the skills they need to expertly negotiate the intricacies of competitive power markets. This study highlights the economic feasibility of renewable energy and helps in the overarching objective of incorporating sustainable energy sources into the global power framework. We provide a holistic solution that supports the continued growth and success of wind farms in a dynamic and increasingly competitive market by addressing both the technical and commercial aspects of wind energy production.

Adversarial Attacks on AI: Understanding and Securing Machine Learning in Cybersecurity

Andrius Januta

Nord Security

a.januta@gmail.com

Artificial Intelligence (AI) and Machine Learning (ML) have become essential tools in modern cybersecurity, but these models themselves are vulnerable to a wide range of attacks. One of the most serious threats is adversarial attacks, where malicious actors manipulate the inputs of ML models to produce incorrect or harmful outputs. The presentation will explore the primary vulnerabilities of AI and ML models that pose a threat to cybersecurity. The main focus will be on understanding how adversarial attacks work, the consequences they can have for cybersecurity applications, and how these attacks can be used to facilitate malicious activities such as manipulating machine learning outcomes. We will also examine the threats of prompt hacking, where AI models become susceptible to deceptive queries.

It's About Time: Revisiting Reciprocity and Triadicity in Relational Event Analysis

Rūta Juozaitienė[1], Ernst-Jan Camiel Wit[2]

[1] Vytautas Magnus University
[2] Università della Svizzera Italiana, Italy

ruta.juozaitiene@vdu.lt

Societies are intricate systems comprising interdependent social actors interconnected through diverse relationships. It has long been recognized that reciprocity and triadic closure are two fundamental components of this interdependence and have, as such, been included in social network models. However, computational limitations and modelling complexity have meant that reciprocity and triadic closure statistics included in such models have necessarily been simplified. Using novel computational and modelling approaches in relational event analysis, our aim is to explore a spectrum of endogenous network effect definitions, ranging from straightforward binary variables to complex, temporal functions accounting for the diminishing relevance of past events. Through simulation studies and real-world dataset analyses, it highlights the importance of comprehensively considering temporal dynamics and subtle assumptions in defining network effects. Neglecting these aspects can lead to significant pitfalls in the analysis. Fundamentally, this research highlights the time-varying nature of reciprocity and triadic closure effects evident in empirical datasets. While exponential decay functions sometimes capture their temporal structure, more complex continuous functions of time often describe the intricate structure of both effects more precisely.

Towards AI-assisted/Powered Smart Environments: from Techno-Centric to Human-Centric and Value-Centric Approaches

Janusz Kacprzyk

Systems Research Institute
Polish Academy of Sciences, Poland

Janusz.Kacprzyk@ibspan.waw.pl

The concept of the so-called smart environments is recently attracting much attention both in science, R&D and even media. The smart environments are basically "small worlds" that constitute a collection of sensors, computers and humans – both individuals and social groups of various sizes - who are synergistically integrated to better complete some tasks and fulfil goals, implement policies, meet expectations by various stakeholders, etc. We are concerned with environments in which the human stakeholders are crucial, and they involve individuals, social groups, enterprises, organisations or even the society as a whole. The problem is to develop and implement some automated agents who can help the human stakeholders develop, plan and finally implement strategies, policies and operations which are crucial for the particular human stakeholders for dealing with problems they face. We assume that the availability and access of various broadly perceived sensors, controllers, etc. is constant and pervasive.

Recently, the use of tools and techniques that are AI (artificial intelligence)- assisted, AI-powered or even AI-enabled is strongly advocated in virtually all fields, also here. Our main interest is the general human-centric direction in smart environments, which is a new way of extending the traditional techno-centric perspective, i.e. just concerning sensors and other inanimate tools and techniques, of the smart environments by assuming that the human being is the key player (stakeholder, actor) as in virtually all complex real-world problems. Such a perspective implies a very wide and complex agenda of research and implementations

combining an active and proactive involvement in the operation of the smart environment, and incentives for the participation of the humans. One of the aspects to be accounted for is a need to find proper relations between the egocentric and (social) value-oriented views. Some inspirations from the human-in-the-loop and society-in-the-loop paradigms can here be employed.

We show how artificial intelligence (AI) can provide tools and techniques to develop new decision-making and reasoning models via the so-called AI-assisted perspective. Then, to deal with more complex decision and reasoning problems, we show the use of the so-called AI-powered/enabled approach, notably in the decision support systems perspective. We emphasise difficulties and challenges faced by the human stakeholders in their role as cognitive partners of automated agents in advanced smart environments, e.g. due to human cognitive biases. Some examples of the use of our approach for critical infrastructure planning are shown.

Detecting Pre-Migraine Night Patterns with Wearable Biosensors and Machine Learning

Viroslava Kapustynska[1], Vytautas Abromavičius[1], Artūras Serackis[1], Šarūnas Paulikas[1], Kristina Ryliškienė[2], Saulius Andruškevičius[2]

[1] Faculty of Electronics
Vilnius Gediminas Technical University
[2] Center of Neurology
Vilnius University

viroslava.kapustynska@vilniustech.lt

Migraine is a prevalent neurological disorder characterized by moderate to severe headaches, often accompanied by disturbances of the autonomic nervous system (ANS). ANS changes may occur during various phases of a migraine. Early detection of these physiological changes could be important in preventing or mitigating migraine episodes.

Objective: This study aims to identify significant patterns of autonomic nervous system (ANS) alterations during pre-migraine nights by analyzing physiological data collected from a cohort of ten individuals using wearable biosensors. We examined 78 features derived from electrodermal activity (EDA), skin temperature (SkinTEMP), metabolic equivalent of task (MET), pulse rate (PR), and accelerometer (Acc) signals to compare pre-migraine and migraine-free nights.

Methods: Data were collected using the Empatica Embrace Plus wearable device and pre-processed to isolate nocturnal sleep periods relevant for analysis. The analysis was carried out over various time frames (ranging from 5 to 120 minutes) to assess how the duration of the frame while training models affect the significance of the characteristics. Machine learning models, including XGBoost, Histogram Gradient Boosting, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), were trained on imbalanced data using cost-sensitive learning to predict migraine occurrences. ANOVA was used to assess differences in physiological features, with F-statistic and P-value heat maps illustrating feature significance across time frames.

Results: Shorter analysis frames of 5 and 10 minutes yielded higher accuracy and recall, effectively capturing physiological changes preceding migraines. The key features of the EDA, pulse rate, accelerometers, activity counts, MET, and skin temperature signals with the highest predictive power included mean, median, standard deviation, max, min, clearance factor, crest factor, impulse factor, peak value, RMS, and shape factor. Importance scores identified skin temperature (0.803) and pulse rate (0.735) as the most significant, followed by EDA (0.509), accelerometer data (0.333), activity counts (0.295), and MET (0.376).

ANOVA confirmed the significance of these features, showing clear differences between pre-migraine and migraine-free nights. Models like XGBoost and Random Forest consistently outperformed others, demonstrating the potential of wearable biosensors for early migraine prediction.

The best performing generalized XGBoost model using a 5-minute frame achieved an accuracy of 0.806, precision of 0.638, recall of 0.595, and an F1-score of 0.607.

Conclusions: These findings suggest that shorter analysis frames (5-10 min) are optimal for detecting significant physiological changes associated with pre-migraine nights, emphasizing the potential of wearable biosensors for early migraine prediction. EDA, SkinTEMP, PR, and Acc signals, with their most significant features, are particularly useful for early detection, paving the way for more effective intervention and migraine management. More research is needed to improve the predictive accuracy for clinical applications.

Applying Reinforcement Learning to Successfully Drive a Car Around a Track

Oskaras Klimašauskas

Institute of Data Science and Digital Technologies
Vilnius University

oskaras.klimasauskas@mif.vu.lt

This work addresses the problem of controlling a car to optimise route traversal using deep reinforcement learning techniques. Two methods are investigated to enable the car to successfully traverse the track. The first method uses radar information used in games - the car uses eight beams to measure the distances from itself to the track boundaries. The second method relies on image analysis, where the car receives information from the environment through in-game visual frames. The aim of this paper is to compare the performance and feasibility of these two different approaches to the problem of car route traversal by testing them in different contexts and using different strategies. In order to increase the generalization ability of the models, additional strategies such as learning on different maps and changing the starting position of the car on the track are applied in both approaches. The experiments allow to evaluate the advantages and disadvantages of each approach and to highlight how the additional strategies influence the learning process and the final performance.

Perception-Driven Augmented Reality and Its Role in Machine Learning

Bożena Kostek

Gdansk University of Technology, Poland

bozenka@sound.eti.pg.gda.pl

Abstract: This study explores the notion of how perception-driven augmented reality (AR), as processed by Artificial Intelligence (AI) systems, can drive advancements in technologies that play a role in the field of machine learning. However, the title of this presentation is only clear when the traditional approach to AR, which refers to an interactive experience that combines the real world and computer-generated 3D content, is reframed from the lens of "looking closer and in more detail into a subject to be processed by AI." Hence, this is a way for AI to "zoom in" on and process the world around us to a granular degree. There are many examples of such approaches, one of them being perceptual computing, allowing human-computer interfaces to understand and respond to inputs such as voice, textual information, image, and gestures enhanced by emotions. The granularity mentioned earlier refers to the expertise derived from human perception. Despite the fact that deep models, including transformers, can achieve high-level quality performance without human intervention, it does not change the fact that the resources on which these models are trained use all this information inscribed in the data. It should also be remembered that what is natural to humans when analyzing, e.g., medical data, such as zooming in on a specific feature or region of interest and then describing it in a natural language, may not yet be possible for an AI model to achieve that, specifically when it concerns underrepresented languages in the medical domain. In this study, it will be shown that an augmented approach to intelligent speech processing may resolve some of the problems involved when medical data are involved. One may envision structured forms and interfaces provided to a physician or medical assistant when a patient's data should be filled in to evolve into improved structures based on contextual awareness and feedback from the process. However, the basic

problem that remains so far unresolved is more common, i.e., when the environment interacts with the healthcare provider during the process of recording the details of a patient visit. In such a case, sensing an environment and deriving its acoustic and noise properties should be analyzed with a granular level of detail based on human expertise and then processed by an AI model to gain human-like perception.

Acknowledgements: This study was supported by the Polish National Center for Research and Development (NCBR) project: "ADMEDVOICE-Adaptive intelligent speech processing system of medical personnel with the structuring of test results and support of therapeutic process," no. INFOSTRATEG4/0003/2022.

Evaluating ML Binary Classification for Predicting Stroke-Related Mortality

Dalia Kriksciuniene, Virgilijus Sakalauskas

Vilnius University

dalia.kriksciuniene@knf.vu.lt

In this study, we evaluated the effectiveness of five binary classification machine learning models – Logistic Regression, Random Forest, XGBoost, Support Vector Machine (SVM), and Neural Networks (MLPClassifier) – for predicting stroke-related mortality using a dataset of clinical and demographic features from the neurology department of Clinical Centre of Montenegro. By comparing these models across key performance metrics such as accuracy, precision, recall, and F1-score, we gained insights into their predictive power and reliability. Additionally, we examined the importance of features to identify the most critical factors driving mortality predictions. Our results indicate that ensemble-based models like Random Forest and XGBoost outperformed other methods, delivering higher accuracy and interpretability. These models consistently identified Health Status, Age and Stroke Symptoms as the most influential predictors, underscoring the importance of these variables in stroke outcome prediction. While Neural Networks showed competitive performance, particularly in terms of precision and recall, the model's lack of interpretability remains a limitation in clinical applications where understanding the driving factors is crucial. Interestingly, simpler models like Logistic Regression, while offering less accuracy, provided clearer insights into feature importance, making them potentially valuable in settings where transparency and ease of interpretation are critical. Conversely, SVM, while delivering good results for certain metrics, struggled with generalizability across different test set sizes. Ultimately, our findings highlight the potential of machine learning models in predicting stroke-related mortality, with Random Forest and XGBoost standing out as the most robust options. These models offer both strong performance and the ability to interpret feature importance, making them suitable for real-world clinical applications. This research underscores the promise of data-driven approaches in improving stroke care by facilitating early identification of high-risk patients and guiding more personalised treatment strategies.

Evidence-Based Network Flow Modelling and Assessment for Cyber Security

Virgilijus Krinickij, Linas Bukauskas

Institute of Computer Science
Vilnius University

virgilijus.krinickij@mif.vu.lt

In the rapidly evolving landscape of cybersecurity, real-time detection and alignment of network incidents and anomalies have become critical. Traditional methodologies for incident detection often rely on offline PCAP file analysis, which is not feasible in the context of modern network environments where real-time assessment is paramount. Additionally, many current techniques suffer from inherent limitations due to data buffer and window size constraints, reducing their effectiveness in capturing and aligning incidents across asynchronous network flows. The increasing complexity of cyber attacks, combined with the volume of network traffic, demands highly efficient, real-time solutions for anomaly detection that are both scalable and responsive to evolving threats. The core challenge addressed in this research is the inadequacy of current algorithms and methods for network flow alignment to process the volume and speed of real-time network flows. Algorithms such as Dynamic Time Warping (DTW), Needleman-Wunsch, and others – commonly used in sequence alignment – are applied in this context to align and detect anomalies in network flows. However, these algorithms are traditionally computationally expensive, requiring significant processing power. This work explores an innovative approach to incident and anomaly detection in real-time network flows using machine learning algorithms coupled with asynchronous alignment techniques. Unlike traditional methods, which struggle with computational overhead and delays due to their reliance on synchronous data windows, the proposed solution leverages asynchronous network flows to simulate cyber incidents in synthetic scenarios. These simulations allow the machine learning models to dynamically adapt to the evolving nature of network traffic without being constrained by static buffer sizes or fixed time windows. The paper dem-

onstrates how applying asynchronous alignment techniques to real-time network flows can provide a robust framework for detecting anomalies and cyber incidents more effectively. We evaluate these algorithms' performance under synthetic, simulated cyber attack scenarios, highlighting their strengths and weaknesses in terms of speed, accuracy, and computational efficiency. By focusing on the dynamic nature of modern network traffic, we present a novel methodology that can significantly enhance the capability of cybersecurity systems to detect and respond to incidents in real time without the need for pre-captured data such as PCAP files. Ultimately, this research addresses the growing need for real-time, asynchronous network flow assessment using machine learning in cybersecurity, providing an effective solution to the limitations of current incident alignment techniques. This approach improves the speed and accuracy of incident detection and offers a scalable model that can be applied in diverse network environments. We demonstrate the effectiveness of our approach through experimental evaluations using real-world network flows. Overall, our work contributes to advancing the field of network traffic analysis by introducing a novel approach that addresses the limitations of traditional synchronous methods and provides a foundation for more robust and adaptive cyber security solutions.

Identification of Key Areas in Histology Images for Detection of Collagenous Colitis: A Deep Learning Approach

Vytautas Kiudelis, Robertas Petrolis, Rima Ramonaitė, Dainius Jančiauskas, Juozas Kupčinskas, Povilas Šabanas, Algimantas Kriščiukaitis

Lithuanian University of Health Sciences

algimantas.krisciukaitis@lsmu.lt

Collagenous colitis (CC) is an inflammatory disease of the large bowel that causes chronic watery diarrhoea, abdominal pain, faecal incontinence, nightly defecation, and weight loss, resulting in significantly impaired quality of life. The diagnosis of CC is rather challenging as it can only be diagnosed upon histological examination of colonic biopsies taken from normal or near normal appearing mucosa. Current routine histological interpretation of biopsies involves subjective evaluation leading to inter-rater variability discrepancies in diagnosis and treatment plan. The aim of this study was to develop an algorithm for robust segmentation of light microscopy images of histological specimen slides identifying key areas containing important diagnostic features of CC. Images of histological specimens from 10 patients (~60 images per patient) were pre-segmented into superpixels using a simple linear iterative clustering algorithm. The areas containing candidate diagnostic features for identification of CC, in particular, the thickened subepithelial collagen layer – the essential diagnostic feature, were marked by the experts. The feed – forward neural network containing three hidden layers with ten neurons in each was trained to identify the superpixels containing sought diagnostic features. The model was tested on 250 images from 5 patients not used for training and showed accuracy of 0.807, sensitivity – 0.801 and specificity – 0.813. The shown neural network's ability to segment histology images could be used for assisted diagnostic process emphasising areas with candidate key features for identification of collagenous colitis.

Change Detection in Satellite Imagery Using Transformer Models and Machine Learning Techniques: A Comprehensive Captioning Dataset

Kürşat Kömürcü, Linas Petkevičius

Institute of Computer Science
Vilnius University

kursatkomurcu@gmail.com

This paper addresses the growing need for automating the caption generation of satellite image pairs, focusing on change detection tasks. The study leverages four major datasets—CLCD, LEVIR-CD, DSIFN, and S2Looking—to create a satellite image caption change detection dataset containing descriptions of 16,753 image pairs. The primary aim is to generate descriptive and accurate captions using the Llama model and evaluate the performance of various machine learning and transformer models in detecting changes between pre-event and post-event images based on these captions. The introduction highlights the importance of automated change detection in remote sensing for applications such as urban planning, environmental monitoring, and disaster management. Traditional manual interpretation of satellite images is time-consuming and requires expertise, underscoring the value of machine learning models in automating this process. The study uses a combination of deep learning techniques, particularly transformer models like BERT, DistilBERT, RoBERTa, and XLNET, and classical machine learning models, including Logistic Regression, Naive Bayes, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN), to evaluate the generated captions. Four datasets are utilised for this task. The CLCD dataset focuses on multi-temporal and multi-sensor analysis for land cover changes, while the LEVIR-CD dataset is centered on building change detection in urban areas. DSIFN enhances high-resolution image change detection with a dual-stream network approach, and S2Looking provides Sentinel-2 images for large-scale change detection studies. Data augmentation

techniques, including random rotations and scaling, are employed to address dataset imbalances and improve the model's ability to generalise across different scenarios. For caption generation, the Llama model is used due to its robust transformer-based architecture, which excels at processing complex visual data and converting it into coherent and contextually relevant natural language descriptions. The generated captions are further refined for clarity and accuracy. The model is trained to maximise the likelihood of the correct word sequences, using a self-attention mechanism to capture dependencies and relationships within the data. The results section presents an in-depth evaluation of various machine learning and transformer models. SVM consistently outperforms traditional machine learning models, achieving the highest accuracy across all datasets. Transformer-based models also demonstrate strong performance, with BERT and RoBERTa leading the pack. BERT excels in training accuracy due to its bidirectional training on masked language modeling, while RoBERTa shows better generalisation on validation datasets, particularly those that benefit from optimisation techniques. DistilBERT offers a faster alternative with slightly reduced accuracy, and XLNET, while competitive, does not outperform BERT and RoBERTa. The study concludes that the Llama model, combined with transformer models, effectively generates accurate and descriptive captions for satellite image pairs, facilitating automated change detection tasks. This research contributes to a new satellite image caption change detection dataset and provides valuable insights into the performance of various models in this domain. The findings underscore the potential of using advanced deep learning techniques, such as transformer models, to translate complex visual data into descriptive language, enabling more informed decision-making across a range of applications.

Acknowledgements: This research was funded by the European Union (project No S-MIP-23-44) under the agreement with the Research Council of Lithuania (LMTLT).

High-Performance Computing in Science

Eduardas Kutka[1], Jolita Bernatavičienė[2]

[1] Information Technology Research Center
Vilnius University
[2] Institute of Data Science and Digital Technologies
Vilnius University

eduardas.kutka@mif.vu.lt

High-Performance Computing (HPC) is the foundation of modern scientific research. HPC enables the solution of complex computations in parallel. Solving these problems on a personal computer can be time-consuming, but the parallel processing power of HPC significantly reduces the computation time. Here, we'll explore the Vilnius University Faculty of Mathematics and Informatics (VU MIF) HPC computational facilities, which serve scientific needs. We'll outline the procedures for obtaining access, ensuring researchers can efficiently use this powerful technology.

Additionally, we'll delve into the opportunities provided by the European Union for large-scale projects. European HPC infrastructures, such as those coordinated under the EuroHPC Joint Undertaking, offer substantial computational power to tackle grand scientific and industrial challenges. These resources are particularly invaluable for researchers working on projects that require extensive computational capacity beyond local capabilities.

Preparing and running calculations on HPC systems is a complex task. European-level HPC training initiatives are organized into HPC competence centres to address this challenge. These centers are crucial in providing specialized training and support to researchers and professionals. Through these initiatives, individuals can access various levels of training, from introductory courses to advanced technical workshops, ensuring that users can effectively utilize HPC resources.

In conclusion, the availability of HPC facilities at Vilnius University MIF, combined with EuroHPC JU HPC resources and comprehensive training initiatives, provides a robust ecosystem for advancing scientific research. Researchers can accelerate their computational projects by leveraging these assets, contributing to significant scientific breakthroughs and innovations.

Advanced Global Optimization Techniques and Their Applications

Dmitri E. Kvasov

University of Calabria, Italy

kvadim@dimes.unical.it

In many simulation-based applications of optimisation, the objective function can be multi-extremal and non-differentiable, thus precluding the use of descending schemes with derivatives. Moreover, the function is often given as a black-box, and therefore, each function evaluation is an expensive operation with respect to the available computational resources. Derivative-free methods can be particularly suitable to address these challenging problems studied in the framework of global optimisation and can be deterministic or stochastic in nature. A numerical comparison of these two groups of methods is interesting for several reasons and has a notable practical importance. In the presentation, the methods of these two groups are considered and their applications (including the field of machine learning) are briefly examined.

Robust Facility Location Under Uncertainty in Customer Behavior

Algirdas Lančinskas[1], Pascual Fernández[2], Julius Žilinskas[1]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Department of Statistics and Operations Research
University of Murcia, Spain

algirdas.lancinskas@mif.vu.lt

Facility location problems are commonly formulated as mathematical optimisation problems, aimed at determining the best locations for facilities such as stores, warehouses, or factories to effectively meet customer demand within a specified geographic region. These problems are important in various industries, from retail to logistics, as they directly influence operational efficiency, customer satisfaction, and profitability. However, in real-world scenarios, environments are dynamic, and customer behavior is often uncertain or subject to change due to various factors. This uncertainty makes a significant challenge to traditional optimisation methods, which are usually used to find a single optimal solution based on a fixed set of assumptions about customer behavior. Given these complexities, usually, it is better to focus on finding robust solutions that may not be the best for any one specific customer behavior model but ensure profitability across multiple customer behavior models and potential future changes. This research addresses the discrete facility location problem for an entering firm, taking into account uncertainties in customer behavior. A multi-objective approach that integrates the concept of a knee point has been used to determine the robust solution(s) from the set of Pareto-optimal or non-dominated solutions. The heuristic algorithm for multi-objective discrete facility location problems has been developed for finding Pareto-optimal solutions. The algorithm dynamically adjusts the ranks of location candidates based on their fitness. The performance of the algorithm was experimentally investigated by solving different instances of the facility location problem using real geographical and population data.

Acknowledgement: This research has received funding from the Research Council of Lithuania (LMTLT), agreement No S-ITP-24-8.

Certain Computational Aspects in Approximation and Harmonic Analysis

Elijah Liflyand

Bar-Ilan University, Israel

liflyand@gmail.com

In applications of approximation and harmonic analysis where numerical methods come into play, sharpness and optimality of relations submitted for calculations are often of crucial importance. Even better if such relations are asymptotic. When convergence of Fourier expansions is studied (just approximation, signal processing, etc.), the so-called Lebesgue constants are an object whose behaviour must be thoroughly analyzed. The situation becomes even more complicated when linear means of Fourier expansions are involved rather than partial sums. For that case, a result due to my late friend Eduard Belinsky has been for decades the best known. However, it was in the form of bilateral estimates, where many additional terms had to be estimated. We have recently succeeded to shape it in an asymptotic way. In particular, if the asymptotics is "genuine" (which may be achieved by the appropriate choice of a function generating summability method), that is, the remainder term is of really better behaviour, one has to concentrate theoretical and numerical efforts only on the leading term. Concrete examples illustrate the latter.

Spectral and Molecular Dynamics Studies of Carotenoids and Stilbene in Complex Systems for Supercomputing

Mindaugas Mačernis, Goda Bankovskaitė, Jonas Franukevičius, Darius Abramavicius

Institute of Chemical Physics, Faculty of Physics
Vilnius University

mindaugas.macernis@ff.vu.lt

Molecular dynamics (MD) can utilize thousands of cores for research studies [1]. However, the Newtonian laws lack quantum physics phenomena, which are essential for understanding molecular properties in solvents, crystals, and amorphous materials [1-4]. Quantum MD methods, such as Car-Parrinello and Atom-Centered Density Matrix Propagation (ADMP) models or AIMD, can be updated to include excited-state dynamics, but understanding remains limited regarding physical properties and computational benchmarks [2-4]. In this study, we present results and discussions on the challenges of Quantum MD for selected carotenoids (Cars), stilbene, and other specific molecules in complex systems.

According to MD and DFT Raman analysis of lycopene and β-cyclodextrin interactions, we demonstrate that the Raman v1 mode shifts to lower or higher frequencies depending on the position of lycopene within the β-cyclodextrin. While MD provides insights into possible structural properties, it remains sensitive to the chosen environment, making it challenging to accurately model the complex dynamics of the lycopene and β-cyclodextrin system. Separate classical and quantum analyses are required.

Stilbene molecules are relatively small compared to carotenoids, but new methods are needed to interpret experimental data [4]. The combination of various Quantum MD approaches provides a deeper understanding, though it makes the required computational resources less predictable.

Moreover, phosphorus-phosphorus through-space indirect spin-spin J-couplings in different conformations, as derived from QMD, present challenges for conformer identification (Fig. 1). Using our proposed Molecular Dynamics Conformation Probability analysis, we identified at least five distinct conformers in CH(PPh2)2.

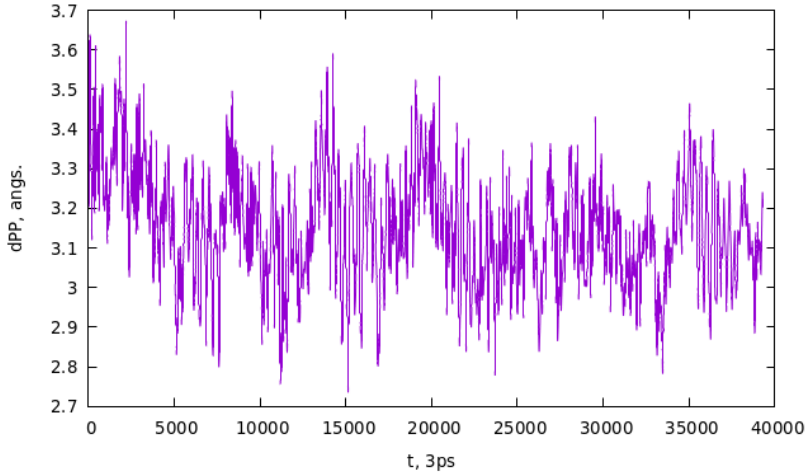


Fig. 1. The P-P distance from Quantum MD is crucial for understanding NMR-based quantum computing

In conclusion, we discuss the limitations of MD and QMD methods in larger scale supercomputers.

Acknowledgements: Theoretical analysis and calculations were supported by the Research Council of Lithuania (Grant No. S-MIP-23-48). Computations were performed using high-performance supercomputer "VU HPC" Saulėtekis of Vilnius University at Faculty of Physics.

References

[1] M. Macernis, J. Franukevicius et al. arXiv preprint arXiv:2210.00934 (2022)

[2] M. Macernis, A.Bockuviene et al. , J. Mol. Stuct. 1226, 129362 (2021)

[3] M. Macernis, S. Streckaite et al. J. Phys. Chem. A126, 6 , 813-824 (2022)

[4] Halimski, I, M. Macernis, D. Abramavicius et al. PCCP 26 (36), 23692-23702 (2024)

On the Road to the Automatic Text Simplification for Lithuanian

Justina Mandravickaitė, Eglė Rimkienė,
Danguolė Kotryna Kapkan, Danguolė Kalinauskaitė

Vytautas Magnus University

justina.mandravickaite@vdu.lt

Text simplification aims to reduce the complexity of the text while preserving essential information. This aspect is important for making information more accessible to a wide range of readers with diverse levels of reading comprehension, such as those with cognitive disorders, non-native speakers, children, and the general public, to name a few. We report experiments on the simplification of Lithuanian administrative texts, seeking to simplify them to a plain language level which would make these texts easier to understand for common people. We chose mT5 and mBART as foundational models and fine-tuned them for the text simplification task. In addition, we tested ChatGPT for simplification of administrative texts as well. Finally, we evaluated the outputs of these models quantitatively using metrics and qualitatively by employing expert evaluation. Our experiments showed that mBART performed the best in simplifying Lithuanian administrative texts, reaching the highest BLEU, ROUGE, SARI, and BERTScore scores. Qualitative evaluation via assessing the simplicity, meaning retention and grammaticality of simplified sentences complemented these results. Meanwhile, mT5 tended to cut longer original sentences in the middle thus losing a part of the information. It also struggled to present sentences in correct Lithuanian grammar, making common spelling mistakes, jumbling the syntactic structure, or, in several cases, getting stuck on generating the same phrase over and over. ChatGPT showed potential, especially as we tested it with zero-shot prompting. However, it was rather difficult to control the desired level of simplification and that information not present in an original sentence would not be added to its simplified version. Our future plans include improving the model (e.g. exploring

different fine-tuning techniques and conducting more comprehensive experimentation in terms of training parameters) and increasing the dataset for better model performance and generalizability. Moreover, we plan a more thorough analysis of the model decision-making process to consider checking for such aspects as factuality or model bias.

From Text to Insight: Enhancing Relation Extraction in Climate Change Research

Sanda Martinčić-Ipšić

University of Rijeka, Croatia

smarti@uniri.hr

Global warming and climate change have profound and far-reaching effects on global ecosystems, weather patterns, sea levels, and human societies, constituting a critical threat to the planet's biodiversity and the prospect of a sustainable future. Simultaneously, the volume of climate change data is rapidly increasing, particularly in published scientific publications. The automated processing of information from unstructured textual data is crucial, with a primary focus on the natural language processing task of information extraction and, more specifically, its subtask, relation extraction. Relation extraction involves identifying relationships between entities within sentences, paragraphs, or larger text units, aiming to automatically generate machine-interpretable data collections that capture entities, their relationships, and associated attributes. This talk addresses the challenge of extracting named entities and relations from scientific publications from renowned journals in the climate change domain. Firstly, the statistics of the collected dataset is presented along the problems encountered in data preprocessing. Secondly, the domain adaptive pretraining of the SciBERT and Climate(Ro)BERT(a) models and from scratch training of CliReBERT and CliReRoBERTa models is elaborated. The discussion is focused on the model architectures and training parameters used, highlighting the advantages and disadvantages of domain adaptive pretraining compared to training from scratch. Thirdly, the task of extracting relations and named entities in the climate change domain is elaborated upon, presenting results on LLM-enabled relation annotation and discovery. These results are used to train all pretrained models (i. e. BERT and RoBERTa) for supervised relation extraction downstream task. Finally, the plan for constructing a knowledge graph from extracted relations in the climate change domain is discussed.

Adapting AI and Tree Growth Models for Sustainable Forestry in Lithuania

Arnas Matusevičius[1,2], Gabrielė Kasputytė[1,2], Anton Volčok[1,2], Martynas Narmontas[1], Gintautas Mozgeris[1], Ljusk Ola Eriksson[3], Tomas Krilavičius[1,2]

[1] Vytautas Magnus University
[2] Centre for Applied Research and Development
[3] Linnaeus University, Sweden

arnas.matusevicius@vdu.lt

Forest management in Lithuania is focused on sustainable development, balancing economic, environmental, and social objectives. Lithuanian forests, covering approximately one-third of the country's land area, are a critical resource for biodiversity, timber production, carbon sequestration, and recreation. To better understand how forests develop over time and how various factors influence their growth, it is important to identify tree growth functions. These functions are mathematical models that describe the relationship between forest growth and other variables such as age, environmental conditions, and management practices. Additionally, AI plays a crucial role in forest planning and management due to its ability to process large datasets, enhance decision-making, and improve sustainability practices. AI and machine learning (ML) are extensively used in forestry, as these methods enable the analysis of large datasets, including climate data, soil conditions, and historical forest management records, to optimise forest operations. For example, AI algorithms can help plan sustainable logging activities by predicting tree growth patterns and yield, while ML models assess the impact of different forest management practices on biodiversity and ecosystem health. AI-driven decision support systems can also assist in the real-time monitoring and management of forest resources, ensuring efficient and sustainable practices. This research aims to develop tree growth functions for Lithuanian forests and extend the GAYA stand simulator to new contexts, adapting it to Lithuanian conditions and reflecting the country's unique compositions of species and ecological conditions. We

are integrating these growth functions into the GAYA framework to ensure accurate projections for Lithuanian forest ecosystems. Moreover, through a literature review, we have identified ways to improve forest planning using AI methods. Initial results show that estimating tree volume growth requires accounting for various forest parameters such as volume, height, age, diameter, and information about cutting activities. By applying ML models, forest planning can be further improved by integrating diverse datasets and evaluating multiple forest management scenarios.

Acknowledgements: This research is funded by the Horizon Europe Framework Programme (HORIZON) under the call Teaming for Excellence (HORIZON-WIDERA-2022-ACCESS-01-two-stage) – Creation of the Centre of Excellence in Smart Forestry "Forest 4.0", No. 101059985. It is co-funded by the European Union under the project FOREST 4.0 – "Ekscelencijos centras tvariai miško bioekonomikai vystyti", No. 10-042-P-0002.

The Rapid Application Development (RAD) in Digital Transformation Era

Jolanta Miliauskaitė, Diana Kalibatienė, Asta Slotkienė

Vilnius Gediminas Technical University

diana.kalibatiene@vilniustech.lt

The needs of business and society are changing rapidly, so the development of software systems should be quickly with the ability to adapt to continuous change of expectations. In response to these quick development needs, the Rapid Application Development (RAD) methodology was proposed in 1991 by Martin James (Martin, 1991), emphasising rapid prototyping and iterative feedback over extensive up-front planning. Since RAD was proposed, new technologies have emerged, and approaches to the software development process have changed. It is therefore important to understand the original concept of RAD and how it has changed over time in order to address the digital transformation of society by developing the digital readiness of higher education institutions (HE) and increasing employment opportunities for students. This research focuses on the RAD concept evolution analysis applying bibliometric analysis of scientific publications taken from the Web of Science digital library. The analysis carried out allows us to better understand and develop a digital transformation by briefly reviewing the main concepts related to RAD, such as low-code, rapid prototyping, iterative development, fast-track development, etc. In order to answer the main research question, "What is the intellectual structure of RAD in Computer Science?" the RAD concept was analysed in terms of the time period covered by RAD, the distribution of scientific papers on RAD across the Web of Science categories, the countries considering the use of RAD, and the main topics studied in RAD. The obtained results show that RAD processes evolve without influencing essential software engineering processes. However, we observed that their sub-processes and activities change in actual RAD processes depending on the main business goal. Summing up, RAD has a short time of application in the field of Computer Science. A number of authors, like (Aveiro et al., 2023; Robal et al., 2024; Agarwal et al.,

2000), argue that further research is needed to understand better the strengths and weaknesses of each RAD and low-code approach and to develop guidelines for choosing the most appropriate method for different types of applications.

Acknowledgements: This research was inspired as a continuation of the Erasmus+ KA220-HED project "Embracing rapid application development (RAD) skills opportunity as a catalyst for employability and innovation" (RAD-Skills).

References

Martin, J. (1991). Rapid application development. Macmillan Publishing Co., Inc.

Aveiro, D., Freitas, V., Cunha, E., Quintal, F., & Almeida, Y. (2023). Traditional vs. low-code development: comparing needed effort and system complexity in the NexusBRaNT experiment. In 2023 IEEE 25th Conference on Business Informatics (CBI) (pp. 1-10). IEEE.

Robal, T., Reinsalu, U., Leoste, J., Jürimägi, L., & Heinsar, R. (2024). Teaching Rapid Application Development Skills for Digitalisation Challenges. In International Baltic Conference on Digital Business and Intelligent Systems (pp. 177-192). Cham: Springer Nature Switzerland.

Agarwal, R., Prasad, J., Tanniru, M., & Lynch, J. (2000). Risks of rapid application development. Communications of the ACM, 43(11es), 1-es.

Advancing NLP for Lithuanian Language at Neurotechnology

Vytas Mulevičius

Neurotechnology

vytas.mulevicius@neurotechnology.com

Less-spoken languages like Lithuanian often receive limited attention from AI researchers, which poses unique challenges for advancing NLP and speech technologies in these languages. This talk will delve into the techniques and strategies we use at Neurotechnology to develop large language models (LLMs) and custom speech solutions, even with restricted datasets and limited prior research. We'll discuss our approach to data collection, model training, and testing, and share insights on overcoming obstacles in building robust language technology for Lithuanian. Finally, we'll invite researchers and developers to collaborate in advancing AI for less-spoken languages, highlighting the importance of inclusive, multilingual AI innovation.

Increasing Availability of Low Latency Stateful Microservices

Kęstutis Pakrijauskas, Dalius Mažeika

Vilnius Gediminas Technical University

dalius.mazeika@vilniustech.lt

Microservice architecture is a widely used approach of application development due to its modularity and loose coupling and facilitate continuous deployment, integration and adaptability to changing workloads. Ensuring high availability and data accessibility is crucial for business success. However, maintaining low-latency stateful microservices presents a greater challenge compared to stateless microservices. Deploying new versions or performing lifecycle management tasks on stateful microservices often requires a graceful failover, which can introduce downtime and impact the microservice's availability budget. To mitigate this issue, we propose a method for graceful failover designed to minimise the availability impact during lifecycle management of low-latency stateful microservices. The method enables the seamless redirection of database requests from one node to another with minimal disruption to client site. This is achieved by monitoring database connection activity and forcefully terminating idle client connections. As a result, the method preserves the availability while allowing essential maintenance activities on stateful microservices. To validate the effectiveness of the proposed method, a series of experiments was conducted to assess the availability of stateful microservices during failover. The results demonstrate that the graceful failover method leads to a negligible impact on availability during maintenance operations, thus ensuring the reliability and continuity of service.

CUDA Implementations for the Acceleration of Microrheology Models

Javier Navarro[1], Gloria Ortega[1], Ester Martín Garzón[1], Antonio Manuel Puertas[2]

[1] TIC-146 Supercomputación-Algoritmos, Dpt. of Informatics
University of Almería, Spain
[2] Department of Applied Physics
University of Almería, Spain

jnl941@inlumine.ual.es

Real-world data often exhibits high dimensionality across various domains, including microrheology, a field of studying fluid properties at microscopic scales through particle motion. The computational demands of processing large datasets, like microrheological simulations, within a reasonable time frame remain significant, requiring dimensionality reduction or acceleration techniques. Among these, GPU-accelerated methods, particularly through NVIDIA's CUDA platform, are increasingly important for efficiency in high-performance computing contexts.

This work involves the development of an accelerated microrheology model using CUDA for GPU platforms, starting with a sequential Fortran model that posed high computational loads. With GPU acceleration, the CUDA-based version of the microrheology model demonstrated substantial performance gains, making high-dimensional simulations more feasible. Various optimization techniques were applied to CUDA to ensure maximum efficiency across GPU architectures, with performance benchmarks conducted on multicore and GPU configurations. To facilitate interdisciplinary collaboration, the project includes a GitLab repository for code version control.

The accelerated CUDA version of this microrheology model provides a significant tool for handling complex datasets, meeting computational requirements while enhancing energy efficiency. This research highlights the value of CUDA-based parallelization in advancing research in computational microrheology and large-scale simulations.

Development of a Large Lithuanian Speech Corpus for Speech Recognition, Artificial Intelligence, and Other Innovative Language Technologies

Gediminas Navickas[1], Gailius Raškinis[2], Danguolė Mikulėnienė[3], Vytautas Kardelis[4], Indrė Makauskaitė[1], Pijus Kasparaitis[5], Margarita Beniušė[5], Laimonas Vėbra[1], Steponas Tolomanovas[1], Asta Kazlauskienė[2], Saulė Milčiuvienė[2], Gražina Korvel[1]

[1] Institute of Data Science and Digital Technologies, Vilnius University
[2] Institute of Digital Resources and Interdisciplinary Research, Vytautas Magnus University
[3] Institute of the Lithuanian Language
[4] Institute of Applied Linguistics, Department of the Lithuanian Language, Vilnius University
[5] Institute of Computer Science, Vilnius University

gediminas.navickas@mif.vu.lt

The lack of robust and accessible speech resources for the Lithuanian language poses a challenge to its digitization, including efforts related to speech recognition, artificial intelligence (AI), and language technologies. This issue is acknowledged in the State Digitization Development Program 2021-2030 of the Ministry of the Economy and Innovation of the Republic of Lithuania. The program emphasizes the need to integrate advanced tools and technological solutions to improve the accessibility, security, and efficiency of e-services at both the national and international levels.

The project "Development of the Large Lithuanian Speech Corpus (LIEPA-3)" contributes to the digitization of the Lithuanian language. During the project, a large corpus of 10,000 hours of annotated speech will be created. The sources will consist of 5000 hours of read speech, 4900 hours of spontaneous speech and 100 hours of four Lithuanian dialects. The corpus will be freely available in open formats on different platforms.

The results of this project will facilitate innovation in AI and digital services, as well as improve public access to e-services in Lithuania. It will simplify interactions with digital platforms, promote digital inclusion and contribute to the broader adoption of AI technologies in various sectors. Ultimately, LIEPA-3 will support a more connected and digitally literate society by providing essential resources for developing user-friendly digital services. Also, the created speech corpus will be a good source for researchers in many fields, primarily contributing to linguistic research and informatics.

Multi-Source Data Merging of Orthoimagery and Satellite Imagery for Deep Learning

Linas Petkevičius

Institute of Computer Science
Vilnius University

linas.petkevicius@mif.vu.lt

The satellite image processing has a lot of new applications in recent years. Open-access satellite data, such as Sentinel-2, of 10 meters, is often not sufficient for computer vision problems. While high-resolution images could be ordered by specialized satellites, the topic of super-resolution images has had a lot of breakthroughs in recent years, too; however, high-resolution data is necessary. Another approach for improving the data labels is the usage of orthophoto images taken by planes or unmanned aerial vehicles (UAV).

In this research, we presenting the results on merging Lithuanian landscape data of orthophoto images and satellite images. The investigated data contains the orthophoto images from the National Land Service under the Ministry of Environment and Sentinel-2 satellite images. The Lithuanian landscape data is analysed from raw TIFF format, which contains RGB colour channels.

The image splitting to patches is done using a sliding window approach. The alignment of orthophoto and satellite images is done by the image registration. The image transformation mapping is done by the OpenCV library. The image transformation mapping is done by the feature matching and homography transformation. The homography transformation is done by the RANSAC algorithm.

The image alignment and mapping problems are presented in this research.

The following merged data sources are prepared for the deep learning model training. The existing segmentation masks are mapped to the high-resolution orthophoto images. The new masks are corrected based

on high-resolution data. The new corrected masks are updated for Sentinel-2 satellite image format.

Acknowledgements: This research was funded by the European Union (project No S-MIP-23-44) under the agreement with the Research Council of Lithuania (LMTLT).

Voxel-Based 3D Object Generation from Single Images Using an Enhanced Deep Learning Architecture

Algirdas Pocius[1], Tomas Blažauskas[2], Eglė Butkevičiūtė[2]

[1] Department of Applied Informatics
Kaunas University of Technology
[2] Department of Software Engineering
Kaunas University of Technology

tomas.blazauskas@ktu.lt

Deep learning has revolutionised the field of 3D modelling by providing powerful tools for generating three-dimensional objects from various input sources, such as images, point clouds, and even textual descriptions. The ability to reconstruct accurate 3D models from limited information is crucial for numerous applications, including computer-aided design (CAD), virtual reality (VR), augmented reality (AR), and game development. However, generating precise 3D objects from single-view images remains a significant challenge due to issues like geometric complexity, occlusion, and computational cost. The aim of this study is to enhance existing computer vision and graphics methodologies by improving and optimising deep learning algorithms that enable the generation of accurate 3D objects, thereby improving the efficiency of computer-aided design. The study utilised the "ShapeNetCore (v2)" dataset, which includes over 51,000 unique 3D models across 55 different categories. A developed deep-learning architecture was used to generate 3D objects in the form of voxels from single-view images. Data processing and model training were conducted using the PyTorch framework, which offers flexibility and efficiency in building and training deep neural networks. To address challenges such as geometric complexity and occlusion, efficient data preprocessing techniques, including data augmentation and normalisation, to enhance the quality and diversity of the training data were incorporated. For evaluating the model's performance, metrics such as Chamfer Distance and Intersection-over-Union (IoU) were applied. The Chamfer Distance quantifies the similarity between the pre-

dicted and ground truth point clouds, while the IoU measures the overlap between the predicted and actual voxel grids. Preliminary experimental results demonstrate that the proposed model effectively generates accurate 3D objects from single images, achieving an overall IoU score of 0.6549. These initial findings suggest that the model performs well across various object categories. This work contributes to the field of 3D object generation by presenting an optimised deep-learning solution that enhances the accuracy of reconstructed objects. The model's adaptability to various object categories and its potential applications in computer-aided design, virtual reality, and game development highlights its significance in advancing 3D modelling technologies.

The Analysis of chatGPT Efficiency for Context Extraction Using Lithuanian News Multi-label Text Dataset

Sergej Pokusajev, Pavel Stefanovič

Faculty of Fundamental Science
Vilnius Gediminas Technical University

sergej.pokusajev@vilniustech.lt

Analysing text data presents various challenges, particularly in text data classification tasks where the text can represent numerous contexts. This complexity often makes it difficult to assign labels unequivocally to newly created datasets. As one alternative to classical multi-label text classification solutions, large language models (LLM) can be used. Considering the LLM's capacities and dependency on the prompt, it is not well known how accurate the multi-label classification task for Lithuanian language texts can be. Therefore, this research experimentally analysed a Lithuanian news multi-label text classification case. The dataset of 12,486 texts was labelled by experts into ten classes, with each text potentially belonging to multiple classes (Collective, Development, Finance, Industry, Innovation, International, Law enforcement, Pandemic, Politics, and Reliability). chatGPT was selected as LLM, and different prompt-engineering options were tested to estimate the Lithuanian written multi-label text classification metrics. Three different directions were used to obtain context for the given text: 1) chatGPT was asked to provide a one-word context list; 2) chatGPT was asked to provide a one-word context list from the 10 classes originally used by experts; 3) chatGPT was asked to provide free-written text about the context reflecting the analysed data. All results were summarised and evaluated, highlighting the advantages and disadvantages of using chatGPT for text data context extraction. The results were compared with previous research, where traditional classification methods were used. This paves a path to understanding how suitable and adjustable, through prompt engineering, the chatGPT is for Lithuanian multi-label text data classification tasks.

Peculiarities of GenAI Usage in Higher Education: Possibilities and Additional Challenges

Urtė Radvilaitė, Birutė Pliuskuvienė, Pavel Stefanovič, Simona Ramanauskaitė

Department of Information Systems, Faculty of Fundamental Sciences
Vilnius Gediminas Technical University

birute.pliuskuviene@vilniustech.lt

The fast-growing popularity of generative artificial intelligence (GenAI) has raised new challenges for educational institutions. Higher education students have wide access to those tools and utilise their opportunities for the study process and individual task implementation. Accordingly, the study and student evaluation process requires changes to adapt to the changes and ensure up-to-date technology usage as well as student skill growth. In this research, we overview the main GenAI-raised challenges at the higher institution and teacher level. The biggest attention is dedicated to student cheating prevention and unfair work identification in text-based tasks. A decision support system for Lithuanian written text generation fact automated identification is proposed. Also, the guidelines for the GenAI-generated text detection model development are provided by reflecting feature engineering and data augmentation's effect on the model accuracy and performance.

The Elastic Slit Membrane: The Concept and the Target Transient Modes

Jūratė Ragulskienė, Kęstutis Pilkauskas, Paulius Palevičius

Department of Mathematical Modelling
Kaunas University of Technology

jurate.ragulskiene@ktu.lt

This article presents the concept and the mathematical model of the elastic slit membrane. The complex structure of the membrane requires a non-standard approach for building the mathematical model and the numerical simulation scheme for the representation of the transient dynamics of the membrane. It is assumed that slits move in accordance with the fundamental axisymmetric mode of the circular-shaped membrane. The mathematical model incorporates the separable variables in time and the solution of a special case of Bessel's differential equation in space. A large scale optimization strategy is used to generate the target transient modes of the membrane.

Automated Course Similarity Estimation, Based on Course Syllabi

Simona Ramanauskaitė[1], Eglė Baradinskienė[2], Cristina Ionela Bulat[3], Umberto Morelli[4]

[1] Vilnius Gediminas Technical University
[2] Exacaster
[3] Universitatea Alexandru Ioan Cuza, Romania
[4] Fondazione Bruno Kessler, Italy

simona.ramanauskaite@vilniustech.lt

Course similarity estimation is an important task for various educational processes, ranking from curriculum design, identification of course optimization or collaborative online international learning (COIL) possibilities, transfer credit assessments and others. While syllabi are presented as text descriptions, and the number of courses might reach high numbers, manual analysis and comparison of course syllabi become time-consuming tasks. In this research, we explore the estimation of course similarity based on course syllabi using both traditional AI methods and Large Language Models (LLMs). We begin by reviewing traditional approaches for text similarity estimation – text vectorization document embedding is done before applying different similarity or distance measurements. The selected courses' syllabi from different universities were processed, and a comparison in the form of a Cartesian product on the course similarity was proposed. The results were compared with human based course similarity estimation results. As an alternative to the course similarity estimation, the selected course syllabi were compared using LLM. As a result of prompt engineering, the comparison results between the two courses were outputted both in a numeric form as well as a justification of the decision provided in the text. The results highlight the strengths and limitations of both traditional and LLM-based approaches, offering valuable insights for educators and institutions looking to automate course-matching solution.

Acknowledgements: The results were achieved in MERIT project (grant agreement no. 101083531), co-funded by the European Union.

Comparative Evaluation of Adjacency Matrix Applications for EEG Signal Classification Tasks

Jonas Mindaugas Rimšelis[1, 2], Gajane Mikalkėnienė[1, 2], Jolita Bernatavičienė[1]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Republican Vilnius Psychiatric Hospital

jonas.rimselis@chgf.stud.vu.lt

Major depressive disorder (MDD) can be defined by distinct abnormal resting-state functional brain connectivity. EEG-based functional brain connectivity can enhance the diagnostics of MDD, especially when combined with machine learning methods. Functional brain connectivity, measured as the correlation between EEG electrodes, is represented as a graph in the form of an adjacency matrix. Different metrics capture unique aspects of this connectivity, making comparative analysis essential for effective classification. In this study, we compare several functional connectivity metrics – Pearson's correlation, phase-locked value (PLV), phase lag index (PLI), and the imaginary part of coherence (iCoherence). These metrics are used to classify two resting-state conditions. The analysis is conducted on two datasets: the EEG Motor Movement/Imagery Dataset, which contains 100 subjects with eyes closed and eyes open recorded with 64 electrodes, and the Republican Vilnius Psychiatric Hospital dataset, featuring 100 control subjects and 100 patients diagnosed with major depressive disorder, recorded with 20 electrodes. Machine learning methods, combined with LASSO for feature selection, produced strong results on the EEG Motor Movement/Imagery Dataset. Bootstrapped mean accuracies using binarized adjacency matrices were: 84.15% (CI: [77.50%, 91.31%]) for support vector machines, 83.40% (CI: [75.00%, 90.00%]) for random forests, and 81.70% (CI: [73.69%, 90.00%]) for XGBoost. Additionally, results will be presented on the Republican Vilnius Psychiatric Hospital dataset.

Acknowledgment: The conference participation is funded by EPAM.

Analysis of the Concept of Reasoning

Darius Sabaliauskas

Institute of Data Science and Digital Technologies
Vilnius University

darius.sabaliauskas@mif.stud.vu.lt

The concept of „reasoning" dates back over 2500 years, beginning with the ancient Greeks, who studied reasoning processes; Aristotle introduced the term „logos," meaning reason, argument, and logic (Anton, 1997). The Romans followed with „ratiocinatio," signifying reasoning or conclusions. In the Middle Ages, Thomas Aquinas used „disputatio" and „argumentatio" (Hoenen, 1997) to discuss reasoning, which Renaissance and Enlightenment thinkers like Descartes, Locke, and Hume further explored (Descartes, 2019; Locke, 1847; Hume, 2000). In the twentieth century, Bertrand Russell and Ludwig Wittgenstein formalized logic (Russell & Whitehead, 1910-1913; Wittgenstein, 2023) as the basis for reasoning structures, while Alan Turing laid the groundwork for computing machines capable of reasoning, establishing reasoning as a core area in artificial intelligence research. In the 21st century, this concept is now used in areas such as artificial intelligence, machine learning, natural language processing, task planning, and more. It is also applied in cognitive sciences and neuropsychology. The concept of reasoning has become interdisciplinary and is used in various fields, from academic disciplines such as mathematics and logic to professional environments where critical thinking and innovative problem-solving are required. This research examines the concept of reasoning as a multifaceted cognitive structure, with particular emphasis on abstract reasoning as a critical cognitive skill that enables individuals and artificial systems to recognize patterns, solve complex problems, and understand complex concepts without relying on concrete experiences. In our research, we conducted a bibliometric study to classify and summarise how the concept of reasoning has transformed over the past millennia, how it is currently applied, and its contribution to science, we used for this advanced bibliometric analysis an extensive Web of Science dataset (Zupic & Čater, 2015). This work

contributes to a broader and comprehensive study aimed at providing a high-level conceptual overview of the concept of reasoning, encompassing various perspectives, methodologies, and innovative approaches.

References

Anton, J. P. (1997). Aristotle on the Nature of Logos. *he Society for Ancient Greek Philosophy Newsletter*.

Descartes, R. (2019). *Discourse on the method of rightly conducting one's reason and of seeking truth in the sciences.* Good Press.

Hoenen, M. J. (1997). The Transition of Academic Knowledge. Scholasticism in the Ghent Boethius (1485) and Other Commentaries on the Consolatio. *In Boethius in the Middle Ages*, 167-214.

Hume, D. (2000). *A treatise of human nature.* Oxford: Oxford University Press.

Locke, J. (1847). *An essay concerning human understanding.* Kay & Troutman.

Russell, B., & Whitehead, A. N. (1910-1913). *Principia Mathematica.* Cambridge: Cambridge University Press.

Wittgenstein, L. (2023). *Tractatus logico-philosophicus.*

Zupic, I., & Čater, T. (2015). Bibliometric methods in management and organization. *Organizational research methods*, 429-472.

Exploring the Distribution of Zeros of the Prime Zeta Function: New Insights Through 3D Visualization and Statistical Analysis

Martynas Sabaliauskas, Igoris Belovas, Rugilė Čepaitytė

Institute of Data Science and Digital Technologies
Vilnius University

martynas.sabaliauskas@mif.vu.lt

This study focuses on the prime zeta function, an under-explored function in analytic number theory. Special attention is devoted to its zero-free region and the distribution of the zeros. Our research identified over 10000 zeros within the complex rectangle $(0.1,1) \times (0,10^4)$. Using statistical analysis, we validated some conjectures regarding zeros' distribution patterns (both within and beyond the critical strip), offering new insights into their underlying structure.

A significant aspect of this study is the utilization of 3D visualizations, which allows us to understand better the behaviour of the prime zeta function and the spatial distribution of its zeros. These visual representations provide a more precise, intuitive grasp of complex surfaces and singularities associated with the prime zeta function. The 3D models depict the intersections of real and imaginary surfaces and highlight zero-plane isolines, aiding in detecting the patterns previously challenging to observe.

The findings reveal a concentration of zeros within the interval $(0.3 < \sigma < 0.7)$, with a prominent peak in the neighbourhood of the line $\sigma = 0.6$. Additionally, the density of zeros diminishes significantly outside the critical strip, $\sigma > 1$. It has been proved that the prime zeta function $\zeta_{\mathbb{P}}(s)$ has no zeros in the half-plane $\sigma > \sigma_0$, where $\sigma = 1.77954465354699\ldots$ is the root of the equation $\zeta_{\mathbb{P}}(\sigma) = 2^{1-\sigma}$. Note that by our calculations,

$$\sigma_{\max} = \max_{|t|<2\cdot 10^5} \{\sigma \mid \zeta_{\mathbb{P}}(s) = 0\} = 1.682628788045196\ldots .$$

Summing up, these results, supported by graphical and numerical data, contribute to the broader understanding of the dynamics of the prime zeta function and pave the way for future explorations through advanced visualization techniques. These insights will be useful for researchers seeking to explore the prime zeta function's intricate behaviour further.

Quantum Machine Learning for Liver Disease Prediction

Laura María Donaire[1], Gloria Ortega[1], Francisco Orts[2], Ester Martín Garzón[1], Ernestas Filatovas[2]

[1] TIC-146 Supercomputación-Algoritmos, Dpt. of Informatics University of Almería, Spain
[2] Institute of Data Science and Digital Technologies Vilnius University

laura.donaire@ual.es

The liver is the body's largest gland, crucial for many vital functions such as processing nutrients, filtering toxins from the bloodstream, and supporting immune defense. Exposure to viruses or harmful chemicals can cause liver damage, which may lead to liver disease—a serious condition that can impair liver function and require immediate medical care. This study explores the early prediction of liver diseases using Quantum Machine Learning (QML), an emerging interdisciplinary field that merges quantum physics with machine learning techniques. Specifically, hybrid QML models combine classical neural networks with quantum layers, utilizing the unique properties of quantum mechanics, such as superposition and entanglement, to enhance data processing and representation. This allows the model to capture complex patterns in the data, potentially improving performance in classification and regression tasks compared to traditional methods. Furthermore, QML has demonstrated higher computational efficiency and performance.

The dataset used for this study, sourced from UC Irvine Machine Learning, contains 583 patient records from India, where 416 cases are diagnosed with liver disease, and 167 are not—resulting in a highly imbalanced dataset. Each record includes 11 attributes, such as patient gender, liver function tests, and a classification label indicating the presence of liver disease.

In this work, we have proposed the 'QML-Liver' model, a hybrid architecture that combines two classical layers (with 2 and 1 neurons, respectively) and a quantum layer containing one qubit. To evaluate its

performance, we utilized four key metrics: accuracy, precision, recall, and F1-score. Our QML-Liver model has demonstrated remarkable performance in predicting liver diseases, achieving an accuracy of 82%, a precision of 81%, a recall of 99%, and an F1-score of 89%. Compared to state-of-the-art models, our model improves these metrics using a simpler architecture and an imbalanced dataset. This approach, which integrates quantum mechanics techniques with classical neural networks, highlights its potential in liver disease prediction by allowing better generalization and handling of data variability.

An Agent Based Simulation and Optimization Model for Minimizing Costs of Renewable Energy Communities

Stefano Sanfilippo[1], Lorenzo Farina[2], Pietro De Vito[1], Mattia Repossi[1], José Juan Hernández-Cabrera[3], José Évora-Gomez[3], José Juan Hernández Gálvez[3], Anna Bolognesi[1], Daniele Domanin[4]

[1] STAM S.r.l, Italy
[2] University of Genoa, Italy
[3] Instituto Universitario de Sistemas Inteligentes y Aplicaciones Numéricas en Ingeniería – Universidad de Las Palmas de Gran Canaria, Spain
[4] C.E.G. Società Cooperativa Elettrica Gignod, Italy

stefano.sanfilippo1992@gmail.com

This paper presents a novel model developed to simulate and optimize Renewable Energy Communities (REC) within the framework of the H2020 PROBONO project, a European initiative aimed at delivering validated solutions for the design, construction, and operation of zero-emission, positive-energy buildings in sustainable green neighborhoods. The model is based on an agent-based and discrete-event simulation approach, allowing the representation of individual Points of Delivery (PODs). Each POD has specific characteristics, such as hourly energy consumption and photovoltaic (PV) production profiles. The model surpasses existing tools by integrating a meta-heuristic optimization process aimed at maximizing REC benefits, which includes not only capital expenditures but also operating expenditures and revenues from the first year of installation. The optimization is driven by a mathematical model based on nonlinear programming and genetic algorithms. The model considers multiple complex variables and constraints, providing optimal design recommendations for PV installations that maximize the Net Present Value of the investment, accounting for factors such as inflation rate and projected cash flows. The model's main strength is its flexibility. It allows users to adjust key parameters like energy prices, installed PV

capacity, and geographical location, making it suitable for different European contexts. While parameterization is important for adapting the model to various scenarios, it's not the most unique feature. What really stands out is how the agents' behavior is driven by dynamic strategies, not just simple parameter adjustments. These strategies are different based on the specific country where the REC is simulated, enabling the agents – representing the PODs – to adapt to varying scenarios. By aligning with national regulations, the agents' interactions reflect the unique conditions of each country. Unlike static approaches, these adaptive strategies allow for more realistic and context-sensitive decision-making processes. The adaptability of the model was validated through a case study, where multiple scenarios were optimized with varying energy prices. These optimizations demonstrated the model's capability to optimize PV capacity installations for each member, delivering significant economic benefits. For example, in a high-energy price scenario, the model optimized the PV capacities to maximize REC benefits. In addition to optimizing installations, the model simulated energy exchanges between PODs and with the grid, factoring in national regulatory constraints. This approach provided a comprehensive framework for REC management, ensuring that the model could handle diverse scenarios and deliver valuable insights for decision-making in applications. In comparison to existing tools such as GSE's Autoconsumo simulator, ENEA's RECON, and commercial products like Energy Community Designer by Maps, this model offers a distinct advantage by not only simulating energy interactions within communities but also optimizing the design of energy production systems. Additionally, it provides support for a wide range of scenarios, making it a powerful tool for investment decisions in RECs across Europe.

A Neural Network for Electricity Demand Modeling: El Espino Case

Stefano Sanfilippo[1], José Juan Hernández Gálvez[2], Christoph Kändler[3], José Juan Hernández-Cabrera[2], José Évora-Gomez[2], Octavio Roncal-Andrés[2]

[1] STAM S.r.l, Italy
[2] Instituto Universitario de Sistemas Inteligentes y Aplicaciones Numéricas en Ingeniería – Universidad de Las Palmas de Gran Canaria, Spain
[3] Europäisches Institut für Energieforschung, Germany

stefano.sanfilippo1992@gmail.com

The design of energy infrastructures, such as microgrids in remote areas presents a number of challenges. The first problem is the lack of reliable, high-quality data on electricity consumption. In these regions, it is difficult to estimate electricity demand due to the difficulty of gathering data, which is constrained by a number of factors including technological, economic and logistical. This poses a significant barrier to a reliable planning and operation of energy systems. A second problem is that daily consumption curves cannot be fully constructed, as periods of data collection may be missing or contain substantial gaps. Traditional methods of demand forecasting are not well-suited to address these difficulties, since they typically rely on consistent and comprehensive datasets. This paper investigates the first problem and addresses the second problem by presenting a neural network-based approach that provides a robust solution for electricity demand estimation in environments where data quality and availability are problematic. The proposed model moves away from relying on hourly demand curves, which are often impractical to obtain in areas with incomplete or poor-quality data. Instead, we propose a causal model that incorporates key variables such as: temperature, humidity, hour of day, month of year, or whether the day is a weekday or weekend. These factors capture the non-linear dynamics that influence electricity consumption. By focusing on these external variables, our model can provide reliable predictions, even in

the presence of missing data. This offers a robust alternative for working with incomplete datasets, which is a key innovation as it reflects the real-world conditions of data collection in remote areas. Additionally, the model is able to handle low-quality data, as it detects and removes outliers, filters incorrect measurements, and ensures more accurate data input. This significantly improves the accuracy of demand estimation. The model was validated using real-world data from El Espino, Bolivia and achieved an accuracy of over 90%. This demonstrates the model's ability to adapt to and compensate for incomplete and unreliable data, making it a valuable solution for demand estimation in remote areas where data limitations are common.

Social Media Analytics for Making Online Platforms Safer!

Rajesh Sharma[1, 2]

[1] University of Tartu, Estonia
[2] IIT Delhi, India

rajesh.sharma@ut.ee

Societal issues such as misinformation and abusive speech have become a menace in recent times. Individuals have started (mis)using online social media platforms to spread misinformation and hate speech, primarily due to the anonymity provided by these platforms and in the name of free speech. These are critical issues as they can create riots in a society, leading to loss of property and lives. We will touch upon various societal issues (mentioned earlier), which can be studied using several data science techniques. The digital traces, if studied responsibly and scientifically, can help mitigate this menace and make these platforms safer for society.

Access Control Approach in Microservices Architecture

Ernestas Serkovas

Kaunas University of Technology

ernestas.serkovas@ktu.edu

Modern and contemporary systems, such as the Internet of Things, are often large-scale and have complex structures. Together with the development of containerization technologies, the use of microservices architecture began. Systems that are developed based on microservices architecture rely on the distribution of functions to the independent and isolated from each other small-scale services. This lets us achieve greater system availability and reliability as well as scalability. The autonomy and isolation of the services of this architecture made it possible to optimize the management and development of the fragmented system. As the number of systems based on microservices architecture grows, especially in the business field, so do challenges related to the security of services and devices. Access control is one of the main security issues faced when designing and developing a system based on microservices architecture. Services are designed to trust requests that come from communicating services. This trust can be exploited to disrupt other services or devices if access control in one of the components of the microservices architecture is compromised. As the Internet of Things technology evolves, it increasingly uses small, constrained devices that also require adequate security and access control. After analyzing the microservices architecture and access management problems related to it, an access management method was proposed and designed, which solves the access management problem in an environment of limited resources. The access control method was designed in microservices architecture, which is made up of three layers – API gateway, fog layer, and end devices layer. The proposed method suggests using an OAuth 2.0 protocol, which is based on the JSON Web Token (JWT), for access control in the API gateway. Furthermore, the mTLS method is proposed for the access control between the API gateway and fog layer servers. Once a

safe communication channel is established, further JWT tokens are used. Fog layer servers will use JWT tokens amongst themselves, which are signed by the ECDSA algorithm public and private key pairs. Most of the devices in the end layer do not have many resources and capabilities to support higher security methods. Due to this reason, lightweight JWT tokens are proposed for the end devices. Moreover, according to the requirements set out in the paper, a prototype was created and tested. The implemented access control method in a microservices architecture was researched with the use of tests and software for analyzing resource usage. When researching the prototype, the aim was to figure out if the chosen access control methods worked as intended and if they were effective. In addition, it was desired to evaluate the resource usage of the proposed access control method. The experimental research on the prototype allowed to compare the proposed method prototype with identical infrastructure without the method and evaluate the proposed methods' suitability, efficiency, resource consumption, and speed. The research showed, that the proposed access control method's increase in processor and random-access memory usage is insignificant, compared to the base measurements, and can be used in devices with limited resources. The speed of the proposed access control method is slower, but the optimization of the method could possibly lower the negative impact on the requests' response time.

Social Factors Affecting Internal Stakeholders' Perceptions of Software Product Quality Characteristics

Jolanta Miliauskaitė[1], Asta Slotkienė[1], Luis Mendes Gomes[2]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Faculty of Sciences and Technology
University of the Azores, Portugal

jolanta.miliauskaite@mif.vu.lt

Software quality models can be categorised into the development process and product quality models based on ISO/IEC 25000. The former focuses on measuring the quality of the software development process and lifecycle. Another is the product quality model, which depends on how deeply internal stakeholders are concentrated and involved in software development processes. Software quality assessment is complex due to the multifaceted nature of software systems and the diverse expectations of different stakeholders (Ndukwe et al., 2023). For this assessment, various parameters can be applied by which software product quality can be evaluated, which was impacted by development process quality, such as software architecture, conformance to functional specifications, ability to scale, and adherence to the development methodology. Since software development is a human-oriented process, it is possible to say that any factor affecting internal stakeholders will directly affect software quality and success (Davis, K., 2014; Guveyi et al., 2020). Internal stakeholder capabilities impact final results, such as individual actions and team interrelation. Thus, once the impact of the social factors has been understood, it is possible to improve software development processes and, at the same time, achieve better values of software quality characteristics. These characteristics could be measured by software quality models, such ISO 25010, defined by the International Organisation for Standardisation (ISO). Subsequently, the ISO 25010 standard emerged, updating the ISO 9126 model by redefining the fundamental characteristics and adding Security and Compatibility, increasing quality dimensions from six to eight.

The paper conducted research to identify and analyse the social factors that influence internal stakeholders' perceptions of the quality of software products using the ISO/IEC 25010 standard. The survey was designed to capture insights from internal stakeholders who participated in software development processes about the key factors impacting software quality characteristics, such as functional suitability, reliability, performance efficiency, usability, security, compatibility, maintainability, and portability. This research allows us to evaluate the social factors influencing internal stakeholders and how they perceive the quality of software products based on the ISO/IEC 25010 standard. Empirical results and statistical analysis show that the biggest impact is the team members' consistency, capability of the team, alignment and interralation communication. By applying these insights, software developers and product owners can make informed decisions that lead to better product quality results and higher customer satisfaction.

References

Ndukwe, I. G., Licorish, S. A., Tahir, A., & MacDonell, S. G. (2023). How have views on Software Quality differed over time? Research and practice viewpoints. Journal of Systems and Software, 195, 111524.

Guveyi, E., Aktas, M. S., & Kalipsiz, O. (2020). Human factor on software quality: a systematic literature review. In Computational Science and Its Applications–ICCSA 2020: 20th International Conference, Cagliari, Italy, July 1–4, 2020, Proceedings, Part IV 20 (pp. 918-930). Springer International Publishing. Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models. ISO/IEC 25010, 2023

Davis, K. (2014). Different stakeholder groups and their perceptions of project success. International journal of project management, 32(2), 189-201.

Hate Speech Detection for Lithuanian Language

Milita Songailaitė, Justina Mandravickaitė, Eglė Rimkienė, Mindaugas Petkevičius, Eimantas Zaranka, Tomas Krilavičius

Vytautas Magnus University

egle.rimkiene@vdu.lt

The rapid increase of online content, which is often coupled with the ease with which people can share their opinions, has contributed to a rise in social issues such as cyberbullying, insults, and hate speech. To mitigate these issues, some online platforms have implemented measures like disabling anonymous comments or completely removing the option to comment on articles the users used to have. Additionally, certain platforms employ human moderators to identify and remove hate speech. However, due to a huge volume of online interactions, manually moderating content requires substantial human resources. Advances in artificial intelligence, particularly in natural language processing (NLP), offer promising results in hate speech identification. Automated hate speech detection systems can facilitate content moderation by efficiently processing and managing large volumes of data. In this study, we present a comparative evaluation of hate speech detection solutions for the Lithuanian language. We used several deep learning models for hate speech detection: Multilingual BERT, LitLat BERT, Electra, open Llama2 for the Lithuanian language, RWKV, BiLSTM, LSTM, CNN and ChatGPT. For the Electra model, we trained ourselves from scratch with Lithuanian texts that made more than 2.5 billion tokens. Multilingual BERT, LitLat BERT, Electra and RWKV were further fine-tuned to classify Lithuanian user-generated comments into three main classes: hate, offensive, and neutral speech. For comparison purposes, we also trained BiLSTM, LSTM and CNN models for the task. Open Llama2 for the Lithuanian language and ChatGPT were used without fine-tuning, and Open Llama2 for the Lithuanian language was then fine-tuned to get better results. To train or adapt the models to the hate speech detection task, we prepared an annotated dataset. It has had 27 357 user-generated comments (hate

speech – 4220, offensive – 7821, neutral – 15 316). All models were evaluated with accuracy, precision, recall, and F1-score metrics. Our future plans include augmentation of our annotated dataset with additional data sources and hate topics as well as experiments in model bias, robustness and output explainability.

Analysing and Identifying Disinformation in Lithuania Using Graph Kernels

Milita Songailaitė, Tomas Krilavičius

Department of Informatics
Vytautas Magnus University

milita.songailaite@vdu.lt

Monitoring social and traditional media requires handling vast amounts of unstructured data, which can be challenging to manage and analyse effectively In such scenarios, techniques that can extract only key information, reducing the data volume and organising it into structured formats, prove to be highly beneficial. The focus of this study was on testing such method, based on the approach of Knowledge Graphs. Knowledge Graphs are a structured way of representing information by connecting entities through relationships, often captured using SVO (Subject-Verb-Object) triplets. These triplets allow for a clearer understanding of how different concepts are related within the analysed data. Using Knowledge Graphs, the occurrences of disinformation in the messaging platform Telegram were investigated. Disinformation was sought by calculating different Graph Kernels between Knowledge Graphs constructed from collected Telegram messages and graphs made from already confirmed disinformation cases published in the EUvsDisinfo database. This method was applied to analyse more than 1 million messages from 30 Russian and Belarusian Telegram channels selected by experts, which, as the initial analysis showed, are also followed by Lithuanian citizens and frequently contain manifestations of disinformation. The disinformation cases were specifically selected only when they were related to the ongoing war between Russia and Ukraine. The results of the analysis showed that a certain amount of disinformation appeared in all the analysed channels, but most channels particularly emphasised disinformation related to biological and nuclear weapons, as well as the views of the residents of the occupied territories. When analysing the methods, it was found that the best way to form Graph

Kernels for disinformation detection was the Shortest Path Kernel. This method, unlike others, allowed distinguishing the graphs most characterised by disinformation from the large number of graphs, which was the primary goal of this study.

Visualising SARS-CoV-2 Phylogenetic Relationships Using Protein Language Models

Brendonas Stakauskas, Virginijus Marcinkevičius

Institute of Data Science and Digital Technologies
Vilnius University

brendonas.stakauskas@mif.stud.vu.lt

Recent advancements have demonstrated the efficacy of Large Language Models (LLMs) based on Transformer architecture for natural language processing tasks, which have been successfully adapted for protein sequence data. Transformer-based models pre-trained on extensive protein databases can predict structures, functions, and other properties from protein sequences alone. These models have shown significant success in understanding the dynamics of viral mutations.

Viruses constantly mutate, and the changes that appear in viral code can lead to different reactions in the host body. Constant changes complicate vaccination, as the new strains can be immune to older vaccines. To monitor the dynamics of mutational changes phylogenetic tree building algorithms are used. Phylogenetic trees represent evolutionary relationships among various biological species based on their genetic information. In this work, we study similarity-based methods for phylogenetic relationship (evolutionary path) visualisation using protein sequence embeddings from protein language models.

The data is gathered from the GISAID Data Science Initiative. This data provider is popular among laboratories gathering sequential data and researchers that study these datasets.

Our used dataset consisted of 41387 SARS-CoV-2 sequences that were collected in Lithuania from February 2020 to March 2023. The dataset spans a comprehensive three-year period (2020-2023), covering key phases of the pandemic in Lithuania. Studying a geographically confined and heavily regulated population may result in lower viral diversity, while the extended timeframe allows us to capture the full extent of the

virus's evolutionary dynamics over time. Data processing was applied using next strain workflow (available on the next strain GitHub page). Data processing included phylogenetic tree building, which was done using IQ-Tree and TreeTime algorithms. Data was filtered to only include virus samples that were collected from human hosts.

Embeddings were built using ESM class models, namely ESM-1b and ESM-2 consisting of 650 million and 3 billion parameters, respectively. Embeddings are built for every token in sequential data of protein. As the data size is huge, we used the mean of outputs from the model. Script for embedding extraction can be found in ESM GitHub repository.

For visualisation purposes the data dimension was shrunk using dimensionality reduction techniques such as TSNE and UMAP. The data points were grouped according to the sequence metadata such as sequencing date or their assigned pango lineage. Although some patterns emerge in produced data plots, the overlapping exists. Internal nodes, acquired from tree-building algorithms, were used in viral evolution visualisations. The phylogenetic tree paths for this dataset were deep thus making it hard to make sense of the information present in the tree. Evolutionary steps acquired from the tree are sparse, and evolutions appear on the different proteins in the SARS-CoV-2 virus. We investigate techniques like node and protein joining to build a dataset from the protein embeddings and phylogenetic tree data, allowing for the visualisation of viral mutation dynamics and the similarity between individual virus samples. The goal of this study is to investigate the consistency of these relationships.

Information Retrieval Tool for Secondary Term Creating: Preliminary Results

Dace Šostaka, Kārlis Sondors, Juris Borzovs, Jānis Zuters

Faculty of Science and Technology
University of Latvia

dace.sostaka@lu.lv

In the spring of 2024 (the University of Latvia, Faculty of Computing), a computer program was developed to review user-specified sources on the Internet related to ICT terms. The program is scalable, and it is possible to add sources of the user's choice in other fields and languages.

The program's primary goal is to significantly reduce the time spent searching for ICT terms and the mechanical steps involved in preparing the material for the Terminology Commission meeting. This will enhance the Terminology Commission's productivity, free up human resources for the more creative aspects of secondary terminology creation, and increase the degree of production of high-quality terms. The research is focused on measuring the program's usefulness and the actual time savings it offers.

The program was tested with 40 terms (the number of secondary terms accepted on average per month by Terminology Commission) from (ISO/IEC/IEEE 24765:2017(E) "Systems and software engineering - Vocabulary". The research was carried out by comparing the time needed when 1) manually searching the selected sources (106 min) and 2) searching with the program (32 min). Thus, the actual time-saving when using the computer program is 74 min per 40 terms.

The Analysis of Synthetic Data Application Peculiarities on Time-Series Forecast Model Selection

Rokas Štrimaitis, Simona Ramanauskaitė, Pavel Stefanovič

Department of Information Systems
Vilnius Gediminas Technical University

rokas.strimaitis@vilniustech.lt

Time-series analysis and forecast fall into the artificial intelligence (AI) model area, where constant model adjustment is needed. While concept shift in classification tasks is relevant too, in most cases, the time-series concept shift is faster than in other AI areas. This requires additional data analyst work on systematic time-series forecast model tuning, or some automated model tuning must be done. In some areas (for example, accounting, collaboration with different partners and their data forecasting), the variety of time-series data is so high manual development of data forecast models becomes not an option. Therefore, foundational models for time-series forecasting are developed. In our previous research, we investigated the possibility of automating time-series forecasting model selection. The results and similar research papers indicate that this task is feasible. At the same time, the foundational time-series forecasting model achieved a forecast error rate that has a place to improve. In this research, we analyse the effectiveness of synthetically generated data applications for more accurate time-series forecasting model selection. The obtained results allow us to estimate the peculiarities of synthetic data application, highlighting its benefits and potential misuse cases.

Automated Coin Classification Using Transformer-Based Deep Learning Models

Mantas Šutas[1], Eimutis Karčiauskas[2], Eglė Butkevičiūtė[2]

[1] Department of Applied Informatics
Kaunas University of Technology
[2] Department of Software Engineering
Kaunas University of Technology

eimutis.karciauskas@ktu.lt

Numismatics, the study of coins and currency, is very important for enhancing our understanding of historical economies, cultures, and societies. Coins serve as records of the past, offering insights into the art, politics, religion, and commerce of different eras. However, the classification and analysis of coins is a complex task that traditionally relies on highly skilled experts. This complexity arises from the intricate details and vast variations among coins. Additionally, many coins suffer from damage or corrosion over time, which eliminates critical features and poses significant challenges for accurate classification. This paper introduces an automated coin classification framework leveraging the Swin Transformer deep learning model. The Swin Transformer's hierarchical architecture and shifted window mechanism enable it to effectively capture intricate features and contextual information within the coin images. This makes it particularly well-suited for detailed numismatic datasets that require high-level feature representation and spatial awareness. By processing images of both the obverse and reverse sides of coins and integrating physical metadata like weight and diameter, the system enhances classification accuracy. The image preprocessing process addresses challenges such as damaged or corroded surfaces, while data augmentation simulates real-world conditions to improve robustness and classification accuracy. The Swin Transformer's hierarchical architecture effectively captures intricate features, making it well-suited for detailed numismatic datasets. Evaluated on a diverse collection of coins, the system demonstrates high scalability and accuracy, offering a valuable tool for numismatic research and cultural heritage preservation.

Rather than replacing human expertise, this tool is designed to assist experts by enhancing the classification process through collaboration between technology and human insight. It automates the initial stages of classification, allowing experts to focus on more nuanced analyses and interpretations that require human judgment.

The Hybrid Model Optimization of the Chiller Infrastructure and Its Preliminary Experimental Results

Rytis Petrauskas, Renaldas Urniezius, Karolis Mickevicius, Ignas Kristutis, Paulius Obolevicius

Dept. of Automation, Electrical and Electronics Faculty
Kaunas University of Technology

renaldas.urniezius@ktu.lt

The control of compressor speed in refrigeration systems is vital for optimizing energy consumption and enhancing HVAC performance. By dynamically adjusting the compressor motor speed, operations can align with current cooling demands, leading to significant energy savings and improved system efficiency. Effective speed control results in stable temperature regulation and reduced noise levels while extending equipment lifespan and lowering maintenance costs.

Conducted in spring 2024 at Kaunas University of Technology, this research explores various types of compressors and their control algorithms. The primary goal is to develop a control algorithm for a laboratory cooling system's compressor to boost efficiency. Tests evaluated compressor performance at predefined constant speeds, revealing a clear correlation between speed, refrigerated space temperature, and compressor efficiency.

The implementation of this control algorithm resulted in energy consumption reductions of 15% to 40% and efficiency improvements of 8% to 20% compared to fixed-speed operations. Additionally, cooling times improved by 11% to 79%, comparing the worst and best fixed-speed results within the studied temperature ranges. The team foresees that the outcome of this study will lead to novel renewable energy infrastructure solutions in the heat pump solutions and their future standardization, improving Heat Pump Keymark certifications' impact.

Exploring Techniques for Diffraction Ring Detection in Molecular Research

Tautvydas Naudžiūnas, Guoda Perminaitė, Povilas Treigys

Institute of Data Science and Digital Technologies
Vilnius University

tautvydas.naudziunas@mif.vu.lt

A substantial amount of visual data is created while conducting single-molecule experiments. Optical tweezers are a more widely used research method, and there is a notable gap in algorithms specifically designed for analysing visual data from magnetic tweezers experiments, which are executed similarly. Existing algorithms, when available, often have significant limitations, such as not being able to track a moving bead accurately. Even though tools are not readily available, some require expensive, specialised equipment purchases, making them hardly accessible.

One of the primary challenges in this study is the low temporal and spatial resolution of the images the authors work with. The visual data generated during experimentation using magnetic tweezers is often of poor quality and noisy, complicating detecting diffraction rings. These rings, critical in interpreting changes in the molecule under study, typically exhibit low contrast against the background, further complicating their detection by conventional computer vision algorithms.

This study aims to investigate and develop an algorithm capable of processing low-quality, noisy images, effectively reducing noise while preserving essential features for diffraction ring detection and parameterisation. At the end of this study, the performance and accuracy of the developed algorithm will be evaluated through experiments utilizing a diverse set of training videos. This will provide an in-depth analysis of the algorithm's effectiveness in varying conditions, particularly regarding its ability to detect and track diffraction rings accurately, which would allow us to be able to detect changes or movement of the molecule connected to the paramagnetic bead.

Acknowledgment: The conference participation is funded by EPAM.

Virtual Reality Meets Mathematics: Innovative Teaching Modules for Geometry. Challenges and Results of the Math3DgeoVR Project

Tatiana Tchemisova[1], Adam Nowak[2], Anna Laska-Leśniewicz[2], Tomasz Kopczyński[3], Jacek Stańdo[2], Nina Szczygiel[1], Ana Breda[1], Beatrix Bačová[4], Mária Kúdelčíková[4], Michaela Holešová[4]

[1] University of Aveiro, Portugal
[2] Lodz University of Technology, Poland
[3] University of Silesia in Katowice, Poland
[4] University of Žilina, Slovakia

tatiana@ua.pt

The Math3DGeoVR project (Mathematical Models for Teaching Three-Dimensional Geometry Using Virtual Reality) is a European initiative (2021-1-PL01-KA220-HED-000030365), under the action Partnerships for Cooperation (Erasmus+ program).

This project is designed for students and academics involved in teaching and studying various Mathematical and Engineering disciplines as well as experts in Education. The project's primary focus is developing an innovative virtual reality technology to enhance students' spatial reasoning skills. A dedicated international consortium of educators and IT developers—experts in Mathematics, Management, Informatics, and Education from five European universities (Poland, Estonia, Slovakia, and Portugal)—has collaboratively developed this cutting-edge teaching method. The technology, delivered through VR glasses, integrates geometric instruction with immersive three-dimensional environments, highlighting the application of Mathematics both in academic settings and real-world situations. Several modules were successfully tested during a Summer School at the University of Aveiro, with over 30 students from partner universities participating.

The Math3DgeoVR project is a significant input into modern, attractive education. Its innovative teaching methods will benefit learners within and beyond the partner institutions.

Constructing Solitary Solutions to Nonlinear Differential Equations Using AI-Assisted Differentiation Operators

Tadas Telksnys, Inga Telksnienė, Romas Marcinkevičius, Zenonas Navickas, Minvydas Ragulskis

Kaunas University of Technology

tadas.telksnys@ktu.lt

Constructing solitary solutions to nonlinear differential equations using AI-assisted differentiation operators By applying inverse balancing techniques in combination with an AI-assisted generalized differential operator method, we have derived deformed kink solitary solutions for a non-autonomous Riccati system that incorporates diffusive coupling. Such systems can be found in many applications, including biological applications, population dynamics modelling, and other areas. The solutions obtained do generalize kink solitary solutions within the classical solitary solution framework: while classical solitary solutions employ an exponential time transformation, deformed solitary solutions can utilize any non-singular transformation function. This allows for a much wider spectrum of solutions, which is not limited to the more traditional soliton framework but a rich variety of analytical forms. Furthermore, the class of differential equations that can be considered using these techniques is also significantly widened due to the introduction of a non-autonomous term that may be present in both the original and image equations. However, the construction of these more general solutions is far from trivial, both analytically and computationally. It requires a multi-stage process that involves analytically solving a specific system of nonlinear equations related to both differential equations and solution parameters. While obtaining a single solution for the aforementioned system is fairly straightforward, listing all possible cases is a process that is not feasible to perform by hand, necessitating the use of more advanced techniques. Standard tools are difficult to apply since there are not many techniques that work smoothly with symbolic rather than

numerical data. For that reason, a custom AI-based tool is employed to analyze the symbolic big data set, enabling the elimination of numerous degenerate cases and significantly enhancing the efficiency of the proposed approach. This tool assists with the pruning of so-called dead-end branches, which do not result in constant or degenerate solutions.

Looking for Simplified Molecular Reaction Coordinates from Computation Data

Stepas Toliautas, Delianas Palinauskas, Laura Baliulytė, Darius Abramavičius

Institute of Chemical Physics, Faculty of Physics
Vilnius University

stepas.toliautas@ff.vu.lt

A range of molecular sensors for measuring properties in micro-scopic environments is based on BODIPY (boron-dipyrromethene) molecule with a rotation-capable chemical group attached. The sensing mechanism can be modeled, as a first approximation, by the evolution of the electronic excitation along the potential-energy curve of the lowest excited state with respect to the rotation angle. However, the nontrivial structure of base and rotation groups presents several challenges for such an approximation.

A reaction coordinate based on the average of opposite dihedral angles, as applied in [1] and later works, is shown to better estimate the actual rotation than a single rotation angle, both for symmetric and asymmetric rotation groups. A similar approach is currently being studied to estimate the mean curvature in TPPS4 (meso-tetra-(4-sulfonatophenyl) porphyrin) oligomers.

Acknowledgements: Part of the presented research is supported by the Research Council of Lithuania (LMT grant no. S-MIP-23-48). High-performance computations were carried out using resources of the supercomputer "VU HPC" of the Vilnius University at the Faculty of Physics location.

References

[1] Toliautas, S., Dodonova, J., Žvirblis, A., Čiplys, I., Polita, A., Devižis, A., Tumkevičius, S., Šulskus, J., Vyšniauskas, A., Enhancing the Viscosity-Sensitive Range of a BODIPY Molecular Rotor by Two Orders of Magnitude. Chemistry – A European Journal, vol. 25, 2019, pp. 10342-10349.

Information System Architecture of the Lithuanian National Biobanking Infrastructure

Justas Trinkūnas[1], Roma Puronaitė[2,3], Laurent Jacotot[4], Mike Woodward[4], Mindaugas Morkūnas[2]

[1] Vilnius Gediminas Technical University
[2] Vilnius University Hospital Santaros klinikos
[3] Institute of Data Science and Digital Technologies Vilnius University
[4] Modul-Bio

Justas.Trinkunas@santa.lt

In 2019, the Human Biological Resource Center project was started to create a modern national biobanking infrastructure in Lithuania and join the BBMRI–ERIC. Project objectives were formulated as follows: to create a National Human Biological Resources Center (HBRC) with a unified and standardized management system for collecting, processing, storing, and managing biological samples and related health information. The Lithuanian National Biobank collaborates with various organizations and institutions to fulfil its mission of supporting health research and public health initiatives. Vilnius University Hospital Santaros Klinikos (VULSK) as the main implementer of the IT project, together with its partners National Center of Pathology (VPC), The National Cancer Institute (NCI), Vilnius University (VU), Innovative Medicine Center (IMC), Lithuanian University of Health Sciences (LSMU), Hospital of Lithuanian University of Health Sciences Kauno Klinikos (LSMUL KK), signed the HBRC Joint Activity Agreement and Lithuania joined the BBMRI-ERIC network as the full member in 2024.

With the rapid advance of multi-omics-based assays and medical decision-support artificial intelligence tools, cutting-edge biomedical research striving to deploy personalized medicine in clinical practice increasingly depends on high-quality, well-annotated samples and datasets. Following the broad participants' consent, the samples in the storage can be linked with clinical information to form biospecimen col-

lections and further serve the development of personalized medicine strategies. However, inconsistent operational procedures cause collections to be scattered across Lithuanian biobanks, making samples and data inaccessible and slowing down request processing.

To achieve this goal, we designed and implemented biobanking platform reflecting the biobanking process in the Lithuanian biobank network. The architecture has three main parts: the hospital infrastructure (VULSK HIS, NCI HIS, VPC HIS), the biobank laboratories software (Modul-Bio's MBioLIMS BioBanking Software) for biospecimen management, and the information technology infrastructure integrating biobank networking operations (biobank participant registry, HIS integrations, The State Data Agency integration).

MBioLIMS BioBanking software manages the complete life-cycle from reception to distribution of biospecimens and associated data. Its multi-site functionality facilitates the connection of different biobanks within the same network. This allows national biobanks to work in the same system autonomously but also share nomenclature and standardized processes. It is possible for a supervising site to view and search all biobanks' data in the cluster.

As the feasibility study of the implemented system, we report here a case study of a VULSK biobank with 12,300 patients enrolled by June 1st 2024. At this date, the biobank database contained more than 101,500 inpatient, 823,200 outpatient, and 6,200 emergency encounters (including historical) and more than 120,000 DICOM images, mostly Ultrasound (US), Computed Radiography (CR), and Computed Tomography (CT) modalities. Patients had cancer and blood diseases (TOP 3 according to ICD-10-AM cancer and blood disease groups, D70-D77 - 33.2%, C81-C96 - 32.2%, D60-D64 - 29.1%), infections (TOP 3 ICD-10-AM infection groups, U00-U49 - 29.2%, A30-A49 - 18.9%, B95-B97 - 8.2%), other diseases (TOP 3 ICD-10-AM other diseases groups, E70-E89 - 38.2%, I10-I15 - 35.0%, J09-J18 - 24.7%).

Visual Representations in Financial Process Mining

Ilona Veitaitė[1], Audrius Lopata[2], Saulius Gudas[3]

[1] Institute of Social Sciences and Applied Informatics
Vilnius University
[2] Faculty of Informatics
Kaunas University of Technology
[3] Institute of Data Science and Digital Technologies
Vilnius University

ilona.veitaite@knf.vu.lt

Process mining, introduced by Aalst in 2004, extracts process-related information from historical event logs and has become an essential tool for data-driven process improvements. Business process mining, an emerging and rapidly growing field of research, aims to analyse business processes by applying data mining and machine learning techniques to event data. Despite its novelty, process mining has quickly gained broad support due to its ability to provide fast, reliable, and ongoing insights for discovering, monitoring, and optimising business processes. Process mining enhances traditional Business Intelligence (BI) tools by offering detailed, micro-level analysis of process behaviour that complements the macro-level insights BI provides on overall business operations. Additionally, it plays a crucial role in digital transformation efforts, delivering deep insights that drive operational excellence. By bridging the gap between process science and data science, process mining has become an indispensable tool for fast-growing and ambitious manufacturing organisations. While modern business systems like CRM, Finance Management Systems and ERP capture extensive event data, visualising this data presents significant challenges. Translating complex, abstract process data into intuitive visual formats is critical but difficult, as it requires balancing clarity, accuracy, and comprehensiveness. Effective visualisations must highlight patterns, detect anomalies, and simplify interpretation for auditors and decision-makers, but the complexity of processes and volume of data can lead to misleading visuals. Aligning

visualisations with diverse stakeholder needs and ensuring they enhance communication across teams is another challenge. Each task is designed based on the specific problem at hand, with a focus on the visual representation and comprehension of data. Process visualisation involves animating process executions in various formats. It can also be depicted through process cube operations, visualising dimensions of financial data, displaying process models as process maps, or using statistical diagrams and other visual techniques. This research explores these challenges by presenting examples of process mining visualisations derived from historical event logs and discusses how visualisations can better support business decision-making.

A Complex Network Approach to Marine Traffic Interaction Modeling: Feature Extraction for Real-Time Traffic Awareness

Julius Venskus[1], Robertas Jurkus[2], Povilas Treigys[1]

[1] Vilnius University
[2] Klaipeda University

julius.venskus@mif.vu.lt

Marine traffic interaction modelling is critical for improving situational awareness and enhancing safety in maritime operations. This paper presents an approach that leverages complex network theory and feature extraction techniques to model and analyze marine vessel interactions in real-time. Using Automatic Identification System (AIS) data, we construct a dynamic maritime network where nodes represent vessels, and edges capture interaction factors such as proximity, speed and heading correlation, and route similarity. Key network features, including degree centrality, clustering coefficient, and betweenness centrality, are extracted to characterize vessel interactions, offering deeper insights into patterns such as congestion, cooperative navigation, and risk-prone behaviours. Through the application of clustering and network analysis algorithms, we identify distinct interaction types and anomalous activities, enabling real-time predictions of traffic patterns and potential risks. The framework is validated using historical AIS data and maritime incident reports, demonstrating its effectiveness in enhancing real-time traffic awareness for vessel traffic service (VTS) operators and autonomous ship systems. This study offers a solution for modelling marine traffic dynamics and improving decision-making and operational efficiency in increasingly congested and complex maritime environments. This research contributes to the development of advanced tools for real-time maritime traffic management, providing actionable insights that enhance navigational safety and optimize marine operations.

Acknowledgements: This project has received funding from the Research Council of Lithuania (LMTLT), agreement No S-MIP-24-117.

Fault Detection in Solar Power Plants Using Energy Production Data

Dominykas Vilčinskas, Lukas Voveris, Jolita Bernatavičienė

Institute of Data Science and Digital Technologies
Vilnius University

dominykas.vilcinskas@mif.stud.vu.lt

Addressing the critical need for timely identification of faults in solar power plants is essential, as even minor malfunctions can lead to substantial electricity losses over time, reducing overall efficiency and profitability. This study focuses on the analysis of energy production data from a solar power plant in Lithuania, consisting of 143 strings distributed across 12 inverters, collected over a 19-month period. To facilitate the analysis, 16 key features were extracted from each string's time series data during the preprocessing phase. These features capture the global structure of the data, transforming each time series into an object with 16 representative attributes. This transformation made the data more suitable for further analysis, allowing for more effective identification of anomalies. Various statistical and machine-learning techniques were applied to detect systems exhibiting abnormal behaviour. The combination of Principal Component Analysis (PCA) and Alpha-Hull methods was used to reduce dimensionality and identify anomalous systems by isolating points outliers. Alongside this, other machine learning algorithms, such as Isolation Forest (iForest) and Local Outlier Factor (LOF), were employed to detect anomalous systems. The results suggest that these methods can potentially identify solar energy generation systems exhibiting abnormal behaviour, with a combined anomaly score offering a comprehensive assessment of string performance. This approach provided valuable insights into which systems were underperforming or exhibiting abnormal behaviour. In addition to the aforementioned techniques, the study also utilised Random Sample Consensus (RANSAC) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) methods to construct detailed fault profiles, which enabled a more in-depth analysis of each

system's performance, offering further confirmation of previously identified outlying systems. This combination of approaches demonstrated its potential for fault detection in solar power plants.

Acknowledgment: The conference participation is funded by EPAM.

Enhancing Keypoint Detection in Thermal Images through Loss Function Optimisation and Model Evaluation

Gabriela Vdoviak, Tomyslav Sledevič

Department of Electronic Systems, Faculty of Electronics
Vilnius Gediminas Technical University

gabriela.vdoviak@vilniustech.lt

Thermal imaging-based activity recognition becomes essential in various domains such as surveillance, healthcare, robotics, augmented reality, autonomous vehicles, behavioural analysis, and sports. These fields often operate under challenging conditions, such as low-light environments, variations in illumination, and heightened privacy concerns. Each challenge demands advanced technological solutions to ensure accurate and reliable performance. The objective of this study is to enhance the performance of keypoint detection by integrating various loss function optimisation methods and comparing different YOLOv8-Pose models in terms of keypoint detection accuracy and processing time. A novel single-person thermal dataset has been introduced as part of this study, which contains a comprehensive collection of 1000 images captured using a thermal camera. The keypoint detection process is carried out using the YOLOv8n-Pose model and different loss functions such as L_1 loss, L_2 loss and L_{OKS} loss that are applied during the training phase to optimise the model's performance. The results show a 1.3% improvement in keypoint detection accuracy when using L_1 loss or a combination of L_1 and L_{OKS} loss functions. In addition to loss function optimisation, this study evaluates the performance of several YOLOv8-Pose models, ranging from Nano to Extra Large, in terms of keypoint detection accuracy and processing capabilities. The findings demonstrate that the Nano model has an exceptional OKS of 95.6%, with a processing time of up to 10.4 milliseconds per frame, making it suitable for real-time applications. The findings of this research underscore the critical importance of both optimising loss function and model selection to improve the accuracy of keypoint detection in thermal images in diverse environments and conditions.

Trapping and Manipulating Drug-loaded Microbubbles by Acoustic Vortex Tweezers

Chih-Kuang Yeh

National Tsing Hua University, Taiwan

ckyeh@mx.nthu.edu.tw

Microbubbles (MBs) can be pushed through blood circulation under radiation force guidance and facilitate the drug adhesion via cavitation. However, the retention and accumulation of MBs on the target site is typically unstable under flow conditions, particularly in the regions of endarterial region and thrombosis. In this study, we propose the acoustic vortex tweezers (AVT) precisely collecting MBs at specific locations under different flow conditions and ultrasound parameters. Owing to the features of long working distance (>10 mm) and single beam configuration, the AVT become feasible in vivo applications. Moreover, the AVT trapping drug-loaded MBs perform drugs accumulation at specific site within blood vessel and B-mode images can see the manipulating process of MBs. The AVT trapping process was optically observed in a 200-μm capillaries mimic tube and acoustically monitored by a 7 MHz B-mode imaging in a tissue mimic phantom. Self-made MBs (sizes of 1.2 μm) were injected with a 1.7 cm/s flow velocity. The AVT was realized by a 4-element customized transducer with 90-degree phase different in each adjacent element (frequency: 3.1 MHz, pressure: 800 kPa, duty cycle: 9.6%). When the AVT was applied, primary radiation force displaced MBs perpendicularly from the center streamline and the secondary radiation force make the MBs aggregation to reach size of 12-14 μm simultaneously. The AVT collected the attached clusters toward the potential-well to form a big cluster 31.1 μm and continued to retain in the flow without lost. The trapping size of 240μm was measured by the maximum pressure gradient. The retaining big cluster can be detected by B-mode imaging resulting to 1.06 mm speckle pattern. These results suggested that AVT are useful in precise retention of MBs under intravascular conditions and the surveillance ability can ensure process safety. Furthermore, MBs signals within

mouse capillaries could be locally improved 1.7-fold and the location of trapped MBs could still be manipulated during the initiation of AVT. The proposed AVT technique is a compact, easy-to-use, and biocompatible method that enables systemic drug administration with extremely low doses.

Quality Assessment of LLM Models Generated Unit Tests: Quality Metrics Completeness from Code-Aware Perspective

Dovydas Marius Zapkus, Asta Slotkienė

Institute of Data Science and Digital Technologies
Vilnius University

marius.zapkus@mif.stud.vu.lt

Unit testing is a fundamental aspect of software testing, which ensures the correctness and robustness of code implementations, but their creation requires considerable time and resources from developers. It has already been proven that special software can generate test cases using conventional methods such as SBST or random testing (Tang et al., 2024). However, the generated test's code reached high code coverage metrics but was highly unreadable by developers. Large Language Models (LLMs) can solve readability issues by learning from training data containing real human-written test code examples. However, another challenge arises, such as unit tests reaching better coverage, but they are independent of functional context (Ryan et al., 2024). Researchers suggest solving this issue by additionally introducing the context of the code fragment into LLM's training set, improving overall results and its quality metrics. Recent research with LLM-generated unit tests focuses on code coverage as a unit test quality measure (Pan et al., 2024; Lops et al., 2024; Bhatia et al., 2024). However, this is not enough, and we suggest involving additional ways in which unit tests will be reliable and understandable. These additional two ways: comparison measures based on abstract syntax trees such as CodeBLEU, RUBY, and measurements based on machine-translation metrics such as ROUGE, METEOR, chrF. According to this, this paper proposed to research and analyze how to measure the quality of the LLM-generated unit tests. In this research, three LLM models were applied, which were used for unit test generation according to the provided source codes. The generated unit tests

were evaluated by test quality metrics such as coverage and machine translation-based metrics. Our research results allow us to highlight several results of generated unit tests with several LLM models. The first observation was that LLM models generated unit test coverage that achieved an average of 76%. The second research result was that semantic and syntactic similarity based on AST was achieved up to 0,99 between LLM-generated unit tests.

Acknowledgment: The conference participation is funded by EPAM.

References

Tang, Y., Liu, Z., Zhou, Z., & Luo, X. (2024). Chatgpt vs SBST: A comparative assessment of unit test suite generation. IEEE Transactions on Software Engineering.

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.

Pan, R., Kim, M., Krishna, R., Pavuluri, R., & Sinha, S. (2024). Multi-language Unit Test Generation using LLMs. arXiv preprint arXiv:2409.03093.

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).

Lops, A., Narducci, F., Ragone, A., Trizio, M., & Bartolini, C. (2024). A System for Automated Unit Test Generation Using Large Language Models and Assessment of Generated Test Suites. arXiv preprint arXiv:2408.07846.

Application of Machine Learning Techniques for Lithuanian Enterprise Clustering

Eimantas Zaranka[1,2], Dovilė Kuizinienė[1,2], Tomas Krilavičius[1,2]

[1] Vytautas Magnus University
[2] Centre for Applied Research and Development

eimantas.zaranka@card-ai.eu

The precise identification of enterprise activity codes stands as a crucial task enabling the rapid and effective establishment or renewal of databases encompassing both public and private companies, which in return helps to make an informative decision about countries' economic tendencies. The research involves combining multi-source datasets, data cleaning, explanatory data analysis, retrieval of embeddings, feature selection, the optimal number of clusters identification, data clustering, and post-clustering analysis. Gathered insights allow for informative decisions about taxes, needed state aid and competition analysis. In both the Republic of Lithuania and the European Union, the enterprise classification system operates under the Nomenclature of Economic Activities (NACE), which employs a six-digit framework. For instance, code 461900 indicates that the business conducts the sales of various goods that involve agents. The initial two digits represent overarching enterprise classifications, in this case, retail trade, while the final four digits delineate specific categorisations within the country's industries. This study aims to apply clustering methods to help in the identification of the economic activities of enterprises using descriptions that could be found in the "Company Description" section of the rekvizitai.lt website. The dataset consists of 28350 business descriptions. Two main themes were observed in the data: (1) the average description lengths are 14, excluding stop-words; (2) the most common activities in the Lithuania economic sector are wholesale, retail, agriculture, and service industry. In this study, 3 embedding methods (BERT, LaBSE and Word2Vec), 4 feature selection methods (PCA, UMAP, SVD, and autoencoders) and 8 clustering methods (K-means, GMM, agglomerative, mean shift, OPTICS, BIRCH, HDBSCAN, DEC) were used for experimentations with 195 mod-

els trained in total. Three main metrics, silhouette score, Davies Bouldin score, and Calinski-Harabasz Index, are evaluated across all clustering algorithms, with adjusted Rand Index and mutual information evaluated for hard-clustering methods. The initial experiments showed that LaBSE and Word2Vec are the most prominent methods for embedding retrieval, while PCA and UMAP are most suitable for dimensionality reduction. The elbow approach was employed in additional experiments to determine the ideal number of clusters. Although these experiments demonstrated that data may be grouped into fewer clusters, the outcomes did not indicate a statistically significant improvement, and adhering to the original NACE space facilitates a more accurate assessment of the current economic landscape situation. Clustering results from K-means, agglomerative, and mean shift methods showed good intra-clustering and slightly above average inter-clustering results. This research demonstrates that enterprise activity sectors can be categorised using Lithuanian descriptions and the K-means, agglomerative, or mean shift clustering algorithms. Future research will focus on all three algorithms hyperparameter optimisation to improve inter-clustering and intra-clustering results.

From Images to Smart Data: Digitization of Logistic Documents

Eimantas Zaranka[1,2], Monika Zdanavičiūtė[1,2], Tomas Krilavičius[1,2]

[1] Vytautas Magnus University
[2] Centre for Applied Research and Development

monika.zdanaviciute@card-ai.eu

According to the Transport Innovation Association estimations, on average, a single lorry driver carries about 50 sheets of paper consisting of only 15 CMR documents. In Lithuania alone, there are about 50,000 active trucks each month, resulting in about 192 tonnes of wasted paper annually. The manual entry of document data into enterprise resource planning (ERP) systems not only is time-consuming but inefficient and could consist of errors. To address these issues, a framework for the digitisation of logistic documents, such as invoices, receipts and CMRs, is proposed that uses object detection, optical character recognition (OCR) and a semi-supervised finetuning pipeline. This study focuses on both the experimentation and implementation phases of research. During the experimentation phase, multiple object detection models like SSD MobileNet, SSD ResNet-50, Faster-RCNN, EfficientDet-D4, and CenterNet HourGlass104 were evaluated. OCR models like Tesseract, EasyOCR, KerasOCR, Kraken, Doctr, and Google OCR were tested. Extensive evaluations showed that using the combination of Faster-RCNN and Google OCR works the best for document digitisation. The object detection model was trained using approximately 1000 images that were equally distributed among three classes of documents. The Faster-RCNN model achieved an average precision (AP) of 0.95 at an Intersection over the Union (IoU) threshold of 0.5, 0.84 AP at an IoU of 0.75, and 0.71 AP IoU ranging from 0.5 to 0.95, with an average recall (AR) of 0.75, within the same range. OCRs performances were manually assessed due to the lack of annotations, with Google OCR proving the best results in the presence of minor inaccuracies in bounding box placement or noise within the bounding box. To further increase the accuracy of the object detec-

tion model, a semi-automated labelling process was introduced, where a trained Faster-RCNN model is used to generate initial bounding boxes and class labels on unseen data, which later are manually adjusted for further finetuning of a pre-trained Faster-RCNN model. The proposed system is an improvement in automating logistics document digitisation, reducing dependence on manual labour and an overall increase of efficiency in the transportation industry.

Are We Able to Model the Human Auditory System in Speech Signal Enhancement?

Daniel Zakševski, Gintautas Tamulevičius

Institute of Data Science and Digital Technologies
Vilnius University

gintautas.tamulevicius@mif.vu.lt

Human hearing has unique characteristics that are actively modeled and integrated into speech signal analysis and processing. Masking, frequency selectivity, individual nature, and adaptivity of the human auditory system are the features we are trying to understand, model, and implement in speech signal processing applications. In the speech enhancement domain, the human auditory system is often modeled by band-pass filters, and now artificial neural networks are being used increasingly. While many new ideas and models are being proposed and developed, most have inherent weaknesses. Some solutions are based on predefined and static filters (e. g., mel frequency cepstral analysis, bark scale filters, and equivalent rectangular bandwidth filters). This is inconsistent with the adaptive and individual nature of the human auditory system, as these static models cannot capture the dynamic changes in auditory filter shapes or the individual differences in hearing abilities. Other solutions are considered data-driven, i.e., a neural network trained on specific data is proposed as a method. A persistent challenge in current speech enhancement methods is their significant performance degradation in low signal-to-noise ratio (SNR) environments. This contrasts with the human auditory system's ability to effectively extract and comprehend speech even under such conditions.

In this study, we are exploring and comparing human auditory models, searching for innovative ideas of dynamic, adaptive, or active properties applicable to speech signal enhancement tasks. Successful modeling of these properties and their integration into non-linear neural models may lead to the development and implementation of a highly efficient, human auditory system-based speech enhancement approach. This will potentially improve the performance of speech enhancement systems in real-world scenarios, leading to more intelligible speech for various applications, such as hearing aids, telecommunications, and voice assistants.

Noise Dataset for Deep Learning-Based Speech Enhancement

Faustas Žiliajevas, Justina Ramonaitė, Daniel Zakševski, Gražina Korvel, Gintautas Tamulevičius

Institute of Data Science and Digital Technologies
Vilnius University

grazina.korvel@mif.vu.lt

In the real world, ambient noise often impeded human speech, such as passing cars, background chatter, or other environmental sounds. While humans have adapted to filter out most of this noise, computers struggle to do the same. Even with advanced noise cancellation algorithms, unpredictable and ever-changing noise remains a significant challenge. Researchers use many types of real-world noise to improve these algorithms to train artificial neural networks. This requires diverse and regularly updated noise datasets. The goal of this work is to create a noise dataset that will help advance this field. This work presents a comprehensive noise dataset recorded in Vilnius, Lithuania, in 2024. The recordings capture high-quality single-channel audio at a sampling rate of 44.1 kHz, stored in WAV format, and quantized at 24 bits. The dataset was produced using a Shure SM78b dynamic microphone and a MiX-Prie 3 audio interface. The dataset consists of a variety of indoor and outdoor sounds. Long-duration sounds, each lasting 30 minutes, include construction noise, a city street, a bus stop, neighborhood sounds, forest ambiance, rainfall, and the sound of a vaporizer. In addition to these longer recordings, the dataset includes shorter audio samples of specific sounds, such as a garbage truck, a kettle, a coffee maker, a car wash, and a microwave oven. The variety of noise recordings and multichannel format of the dataset make it an optimal choice for tasks such as noise classification, speech enhancement, and noisy speech recognition.

Acknowledgements: This project has received funding from the Research Council of Lithuania (LMTLT), agreement No S-MIP-24-118.

An Overview of the Machine Learning Algorithms Used for Music Source Separation

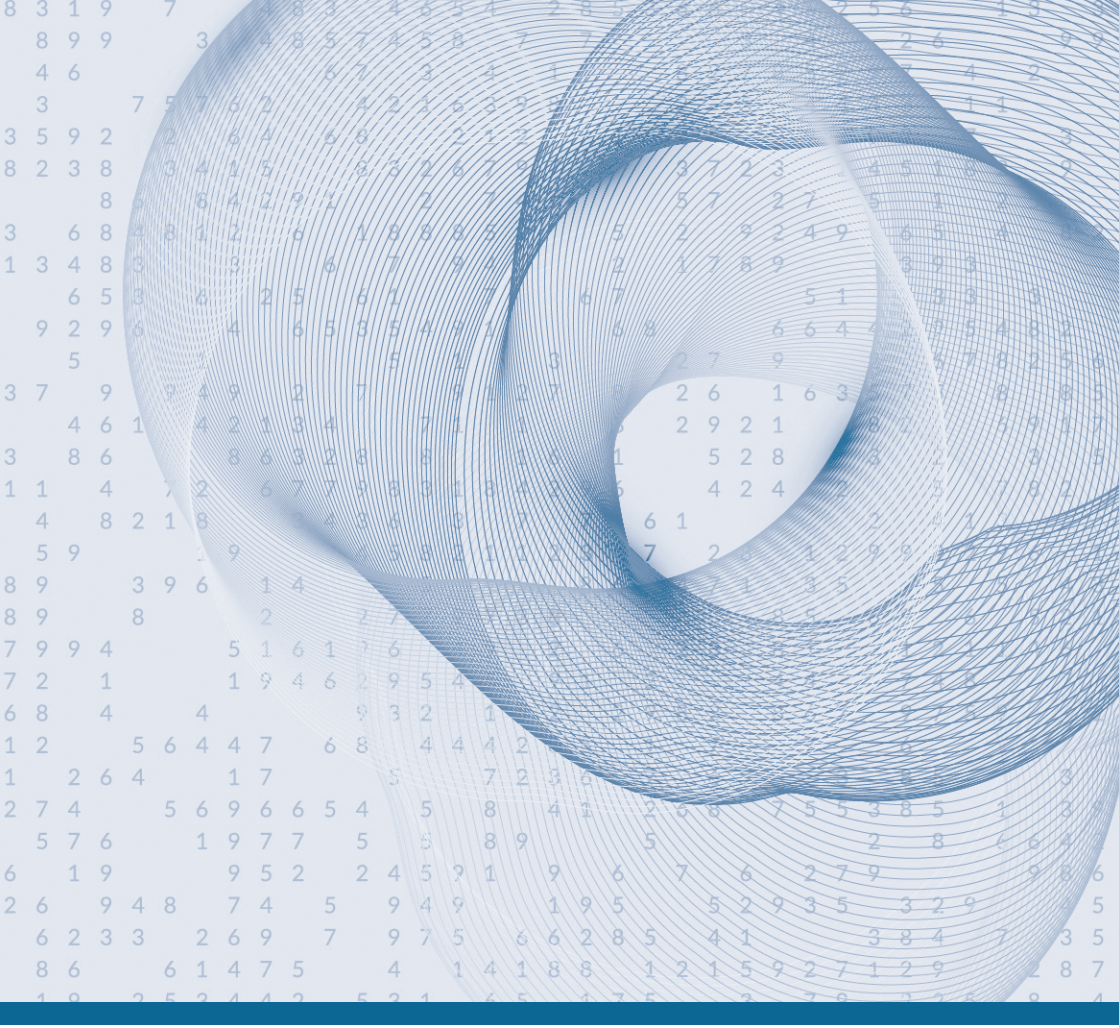
Aidas Žygas, Gražina Korvel

Institute of Data Science and Digital Technologies
Vilnius University

aidas.zygas@mif.stud.vu.lt

Music source separation refers to the process of isolating individual audio sources from a mixed audio signal, allowing for enhanced manipulation and analysis of musical elements. With the growing availability of music in digital formats, the number of research articles on music source separation has increased significantly. Although there are countless different music source separation models available, most of them utilize modified versions of machine learning algorithms, including 1) Convolutional Neural Networks, 2) Recurrent Neural Networks, and 3) Attention-Based Transformers. This work presents and discusses the results of a survey on music source separation approaches. The primary purpose of this analysis is to understand the underlying logic of each algorithm and how they differ from each other. Additionally, it is crucial to compare the performance of the models that adapt machine learning algorithms using a quantitative metric – SDR (signal to distortion ratio) in order to determine which machine learning algorithm is the most efficient for separating stems in a musical recording. Our literature analysis not only answers the question of which models are most commonly used in the existing literature but also provides answers to which datasets are employed for model training and which acoustic features are utilized to represent the acoustic signals of musical instruments. The results of this analysis contribute to a deeper understanding of current research approaches and provide a foundation for future research in this area.

Acknowledgment: The conference participation is funded by EPAM.



15th Conference

DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS

Compiler **Jolita Bernatavičienė**
Prepared for press and published by
Vilnius University
Institute of Data Science and Digital Technologies
4 Akademijos St., LT-08412 Vilnius

Vilnius University Press
9 Saulėtekio Av., III Building, LT-10222 Vilnius
info@leidykla.vu.lt, www.leidykla.vu.lt
Books online bookshop.vu.lt
Scholarly journals journals.vu.lt