DAMSS
DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS



17th Conference on

DATA ANALYSIS METHODS for Software Systems

ABSTRACT BOOK

November 26–28, 2026

Druskininkai, Lithuania,
Hotel "Europa Royale"

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



17th Conference on

DATA ANALYSIS METHODS
for Software Systems

November 26–28, 2026

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

VILNIUS UNIVERSITY PRESS
Vilnius, 2026

Co-Chairmen:

Dr. **Saulius Maskeliūnas** (Lithuanian Computer Society)
Prof. **Gintautas Dzemyda** (Vilnius University, Lithuanian Academy of Sciences)

Programme Committee:

Prof. **Jolita Bernatavičienė** (Lithuania)
Prof. **Juris Borzovs** (Latvia)
Prof. **Janusz Kacprzyk** (Poland)
Prof. **Ignacy Kaliszewski** (Poland)
Prof. **Bożena Kostek** (Poland)
Prof. **Tomas Krilavičius** (Lithuania)
Prof. **Olga Kurasova** (Lithuania)
Assoc. Prof. **Tatiana Tchemisova** (Portugal)
Assoc. Prof. **Gintautas Tamulevičius** (Lithuania)
Prof. **Julius Žilinskas** (Lithuania)

Organizing Committee:

Prof. **Jolita Bernatavičienė**
Prof. **Olga Kurasova**
Assoc. Prof. **Viktor Medvedev**
Laima Paliulionienė
Assoc. Prof. **Martynas Sabaliauskas**
Prof. **Povilas Treigys**

Contacts:

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

Copyright © 2026 Authors. Published by Vilnius University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

https://doi.org/10.15388/DAMSS.17.2026
ISBN 978-609-07-1297-9 (digital PDF)

© Vilnius University, 2026

Preface

This book contains the abstracts of the presentations delivered at the 17th Conference on Data Analysis Methods for Software Systems (DAMSS 2026). The conference takes place on 26–28 November 2026 at the Europa Royale Hotel in Druskininkai, Lithuania.

Over seventeen editions, DAMSS has become the place where Lithuanian researchers and developers working in informatics, data analysis, and complex software systems meet every year. Most participants come from universities and research institutes, and they are joined by industry professionals working on practical problems. The conference brings together people with very different experience. Some develop new methods, others apply them in real systems, and others work on problems for which suitable methods do not yet exist. This community grows every year. This year, 135 participants registered for the conference and 95 presentations are included in the programme, both more than last year. Participants come from nine countries: Czechia, Greece, Lithuania, Poland, Portugal, Spain, Switzerland, Taiwan, and the United Kingdom. The conference creates new opportunities for future partnership.

Artificial intelligence is rapidly changing how software systems are built, how data are analysed and how research is carried out. These changes are happening faster than they can be evaluated or settled into established practice. For this reason, one of the most important goals of DAMSS today is the open sharing of experience. Participants discuss what works and what has not worked, the challenges that arise when implementing AI solutions, and the questions that remain unanswered. Such experience is rarely reflected in publications, but it is precisely this kind of experience that enables colleagues to streamline their work and identify the best solutions more quickly.

It is especially important that many students and doctoral candidates take part in the conference. For them, DAMSS is a chance to see in a few days what has been achieved in their field over the past year, which directions are most relevant today, and where the research community

sees its greatest challenges. When young researchers present their work, they receive questions, critique, and advice from experienced colleagues. This often helps them define their research problem more precisely, choose suitable methods, or discover an unexpected application of their work. These discussions benefit both sides, as questions from young researchers often prompt experienced colleagues to look at their own work in a new light.

The abstracts collected here offer an overview of current research directions and results. Yet much of the value of DAMSS comes from the discussions that grow out of them, whether it is a question asked at a poster, a method adopted from a related field, or a dataset shared after a conversation over coffee. We hope this abstract book will help readers find authors whose work is close to or complements their own research, and encourage people to start working together.

DAMSS 2026 would not have been possible without its sponsors. We sincerely thank Visoriai Information Technology Park (VITP), Vinted, Neurotechnology, Novian, 3RTechnology, and the Research Council of Lithuania (LMT). Their support shows that the questions discussed at the conference matter beyond academia, and their representatives bring practical problems to the conference that give research a clear direction.

Finally, we thank the conference organizers, the Vilnius University Institute of Data Science and Digital Technologies, the Lithuanian Academy of Sciences and the Lithuanian Computer Society, as well as the conference partners, the International Federation for Information Processing (IFIP) and the IEEE Lithuania Section. We also thank all authors, reviewers, session chairs, and members of the Programme and Organizing Committees for their work and contribution.

DAMSS 2026 IS SUPPORTED BY:

Organizers

Vilnius University Institute of Data Science and Digital Technologies
Lithuanian Academy of Sciences, Lithuanian Computer Society

Partners

International Federation for Information Processing

IEEE Lithuania Section

General Sponsors

Neurotechnology

Novian UAB

3RTechnology

Research Council of Lithuania

Vinted UAB

Main Sponsor

Visoriai Information Technology Park (VITP)

A Reproducible Framework for Implementing and Benchmarking Quantum Comparator Circuits

Fernando J. Alonso, Ester M. Garzón, Francisco Orts

University of Almería, Spain

danijaime17@gmail.com

Quantum comparator circuits are fundamental components of quantum arithmetic and appear in applications such as image thresholding, database search, state preparation, and oracle-based algorithms. Although numerous comparator designs have been proposed, their practical comparison remains difficult. Published circuits frequently use different conventions for operand order, output polarity, ancillary qubits, garbage outputs, gate decomposition, and resource accounting. Moreover, some proposals are described only through figures or composite gates whose implementation is not fully specified. These differences hinder reproducibility and may lead to resource estimates that cannot be directly compared. This paper presents QCMP, an open-source and extensible framework for implementing, validating, and benchmarking quantum comparator circuits with Qiskit. QCMP defines a common plug-in interface that separates comparator construction from input encoding, simulation, result decoding, and metric calculation. Each implementation declares its operands, output qubits, restored auxiliary qubits, garbage outputs, and relevant assumptions. This makes it possible to execute heterogeneous designs through the same command-line and Python interfaces while preserving their original computational semantics. The framework provides a structural resource-analysis module that operates before input preparation and hardware-dependent transpilation. It reports quantum cost, critical-path delay, total qubits, auxiliary qubits, garbage outputs, T-count, and T-depth. Configurable profiles allow composite gates to be evaluated using alternative decompositions. QCMP can additionally sample circuits for different operand widths and infer symbolic resource expressions as functions of the input size, including their valid domain and exceptional

small-size cases. The repository currently integrates representative architectures based on reversible ripple networks, temporary logical-AND gates, quantum Fourier transforms, comparator trees, and circuits with operands encoded as compile-time gate sequences. The latter require special support because one input is incorporated into the circuit topology instead of being loaded into a quantum register. QCMP handles this distinction explicitly, allowing both conventional two-register designs and constant-operand comparators to share the same execution interface. A layered validation methodology is used. Implementations are checked against exhaustive truth tables for feasible input sizes, boundary and random cases for larger sizes, expected output polarity, restoration of input and work registers, and published gate or resource counts. Where meaningful, coherent behavior on superposed inputs is also examined. Circuits containing measurement or reset operations are identified as non-unitary rather than being presented as coherent comparison oracles. When a paper leaves a gate, recurrence, or generalization underspecified, the adopted implementation is isolated and documented separately from the operations explicitly provided by the source. QCMP does not claim that a single comparator is optimal for every architecture. Instead, it provides a reproducible experimental basis for studying the trade-offs among qubit count, circuit depth, fault-tolerant cost, garbage management, and input assumptions. The framework supports more reliable comparison of published proposals and facilitates the incorporation of future quantum comparator designs by combining executable implementations, explicit provenance, automated validation, and unified metrics.

Auditable Authority and Recoverable Execution in Distributed Scholarly Publishing

Kazimieras Bagdonas

Kaunas University of Technology

kazimieras.bagdonas@ktu.lt

Decentralized publishing proposals address the delays and biases of peer review with distributed records, reviewer reputation and shared governance. Their implementations must keep a signed review bound to its manuscript, assignment and policy, a payment intent distinct from a completed transfer, and a record's meaning unchanged after a crash. We use FNDTN, a prototype publishing system for pseudonymous authors, reviewers and evaluators, to ask where weaker evidence can acquire authority and what prevents it. A node admits signed commands into a SQLite store, materializes submissions in Git, and keeps signed profiles for synthetic bonds and multi-round review. The intended Bitcoin anchoring and conditional payments are not evaluated.

Seven adversarial cases each state a failure mode, principle, mechanism and measured evidence. A process-crash experiment interrupted a signed submission at eight boundaries, with a normal-completion control, under exact and re-signed retries; after recovery, all 108 runs held one target operation and one Git effect, and no deferred intent was executed. In the ordering comparison, a baseline reducer's result depended on arrival order in all four conflict sets, and a hash-ordered log gave one result per set without deciding legitimacy. In the signed economic profile, every tested conflict froze without usable totals, and separately implemented checkers reconstructed both profiles across conflicts, restarts and process termination. Mutation analysis killed 125 of 132 mutants, with 5 equivalent and 2 surviving; the sanitized suite and the fuzzing harnesses reported no error; and a clean container on the same host reproduced every compared outcome. Benchmarks showed that editorial ingest cost grows with the square of the command journal length.

The cases apply established principles and yield three design obligations: preserve signed meaning through every representation change; give all consumers of a fact the same acceptance rule, with explicit conflict and unavailable states, and bind each decision to its exact context; and define completion and recovery separately for each persistence domain. These results hold for one pinned issuer, one operator and static keys, under process termination on one platform. In a simple payoff model, only identity cost deters undetected dishonesty, and pseudonymous enrollment keeps it low. Distributed agreement, funded settlement, identity history, privacy and incentives remain unestablished. The approach provides a non-exhaustive review procedure applicable to other systems, links each mechanism to measured evidence, tests whether that evidence detects missing mechanisms, and separates established from unestablished claims.

Lossy Compression of Simulated High-Energy Physics Data Using Vector Quantization

Paulius Balčiūnas

Faculty of Mathematics and Informatics
Vilnius University

paulius.balciunas@mif.vu.lt

Modern high-energy physics increasingly depends on large-scale simulations that produce data faster than storage capacity can affordably grow. When a controlled reconstruction error is acceptable, lossy data compression can greatly reduce storage needs while preserving the information relevant to analysis. This work examines vector quantization (VQ), a lossy compression approach widely used for images, audio and other signals. As a case study, we apply two VQ methods to simulated high-energy physics data from the CMS experiment at CERN, namely pixel and strip tracker data, which differ in structure and sparsity. Residual Vector Quantization (RVQ) is a classical method that iteratively refines the reconstruction by quantizing the residual error at each stage. The Vector-Quantized Variational Autoencoder (VQ-VAE) is a deep learning model that jointly learns a compact representation and a codebook. We ask whether a neural-network-based VQ method can outperform a simpler classical method, such as RVQ. Both methods are evaluated using compression ratio, root mean square error, precision and recall, together with example reconstructions. The results show that VQ-VAE outperforms RVQ on pixel data but has slightly higher reconstruction error than RVQ on strip data. Overall, the findings suggest that learned compression is a promising direction for scientific data and motivate further research toward its integration into physics analysis workflows.

Developing a Lithuanian Dataset of Malicious Social Media Communications

Edgaras Baranauskas, Jevgenijus Toldinas

Faculty of Informatics
Kaunas University of Technology

edgaras.baranauskas@ktu.edu

Creating a dataset of malicious social media interactions in Lithuanian necessitates considering privacy, linguistic variety, and annotation quality. When collected messages contain personal data, they must be processed in accordance with the GDPR, which includes an acceptable legal basis and protections for scientific research. The Lithuanian dataset containing more than 600 labelled messages, approximately 30% of which were neutral was created by Čepulionytė [1]. Four non-exclusive labels represented aggressive, insulting, toxic, and malicious content. The dataset description lacks a detailed account of source selection and annotator agreement. The Common Language Resources and Technology Infrastructure CLARIN-LT [2] brings together language-oriented resources, advanced tools, and their metadata at both national and European levels, focusing primarily on resources related to the Lithuanian language. Lithuanian is a language characterized by extensive inflection and morphological complexity, which complicates text classification compared to English. Non-standard orthography, mixed Cyrillic/Latin characters, or internet vernacular that evades traditional morphological tools like Morfuoklis [3] are frequently used in malicious communication on Lithuanian social networks. To address the issues mentioned above, we suggest using OpenAI to generate a corpus of synthetic datasets in Lithuanian. We created several prompt scripts that then were used to ask OpenAI to generate dataset of ham and malicious social media Lithuanian messages in accordance with the GDPR. In our experiments GPT-4o, Claude, Gemini OpenAI tools were used. We generate a consolidated dataset by comparing datasets from the corpus and including only the different messages. We utilize machine learning algorithms and MATLAB to develop and assess models on a per-dataset basis from the corpus. The final experiment utilized

a consolidated dataset. The models were evaluated and compared using accuracy, precision, recall and F1-score. The results obtained were visualized using ROC curves. The most significant difficulties that we have previously faced in the process of collecting, processing, and developing the Lithuanian dataset of malicious social media communications are discussed.

References

[1] Čepulionytė, A.; Toldinas, J.; Lozinskis, B. A Multilayered Preprocessing Approach for Recognition and Classification of Malicious Social Network Messages. Electronics 2023, 12, 3785. https://doi.org/10.3390/electronics12183785

[2] CLARIN-LT repository. Hosted at CLARIN-LT, the Republic of Lithuania. Available online: https://clarin-repo.lt/info/legal (accessed on 7 10 2026).

[3] An online tool for morphological analysis and synthesis of the Lithuanian language. Available online: https://sitti.vdu.lt/morfuoklis/lt/apie (accessed on 7 10 2026).

Computational Optimization of Enzyme-Based Biosensors with Cyclic Signal Amplification

Romas Baronas, Linas Litvinas, Karolis Petrauskas

Institute of Computer Science
Vilnius University

romas.baronas@mif.vu.lt

Biosensors are analytical devices that convert biological recognition events into measurable signals. In catalytic electrochemical biosensors, enzyme-catalyzed reactions generate electroactive species, producing electrical signals related to analyte concentration. Among these, amperometric enzyme biosensors are widely employed owing to their simplicity, cost-effectiveness and reliable analytical performance, with applications spanning clinical diagnostics, environmental monitoring, and industrial bioprocesses [1]. Their sensitivity can be enhanced through cyclic reaction schemes that amplify the electrochemical signal. In such systems, the analyte initiates a reaction cycle involving repeated regeneration and conversion of reaction products, a process known as the triggering mechanism [2].

Computational modeling enables the prediction and optimization of biosensor responses under steady-state and transient conditions. By accounting for sensor geometry, enzyme-catalyzed reaction kinetics, and mass transport, numerical simulations provide valuable tools for evaluating biosensor performance, optimizing design and operating conditions and developing novel biosensing systems [2].

This study investigates and optimizes amperometric bienzyme biosensors employing catalysis–catalysis–electrochemistry (CCE) triggering mechanisms, accounting for internal and external diffusion, interfacial partitioning and nonlinear Michaelis–Menten kinetics. Biosensor operation is described by a two-compartment reaction-diffusion model comprising an enzyme layer and an external diffusion layer. Exact analytical solutions for steady-state concentration profiles and current are derived for selected cases, while transient responses are simulated using a finite difference method implemented in Java [3].

Constrained multi-objective optimization is performed using the parallel HybAlg framework, combining Bayesian global search with Hooke–Jeeves local refinement [4]. The optimization aims to maximize sensitivity and signal gain while minimizing enzyme utilization. These competing objectives are governed by the interplay between enzyme kinetics and mass transport, with the effects of reaction intensity and external diffusion resistance strongly dependent on the kinetic regime. The framework enables systematic exploration of a computationally demanding seven-dimensional design space and identification of physically feasible Pareto-optimal configurations. In particular, Pareto analysis reveals fundamental trade-offs between sensitivity, signal amplification and enzyme utilization, demonstrating that maximizing CCE-induced signal gain alone does not ensure an optimal balance of biosensor performance characteristics. These findings provide a basis for the rational design and optimization of CCE-based amperometric bienzyme biosensors.

References

[1] C.M. Pandey, B.D. Malhotra. Biosensors: Fundamentals and Applications, 2nd ed., De Gruyter, 2019.

[2] R. Baronas, F. Ivanauskas, J. Kulys. Mathematical Modeling of Biosensors, 2nd ed., Springer, 2021.

[3] R. Baronas, K. Petrauskas. Effects of diffusion limitations and partitioning on signal amplification and sensitivity in bienzyme electrochemical biosensors employing cyclic product conversion. Appl. Sci., 16(3):1171, 2026.

[4] L. Litvinas. An open-source parallel algorithm of Bayesian-based global search with Hooke-Jeeves local refinement for multi-objective optimization problems. Nonlinear Anal. Model. Control, 31(2):463-476, 2026.

Human-Language Model Alignment in Native and Foreign Language Processing: An Eye-Tracking Study of Cognitive Effort

Błażej Białczak, Kamil Sterniuk, Bożena Kostek

Faculty of Electronics, Telecommunications and Informatics
Multimedia Systems Department and Audio Acoustics Laboratory
Gdańsk University of Technology, Poland

bozena.kostek@pg.edu.pl

Understanding how cognitive effort differs between native (L1) and foreign (L2) language processing is important for developing adaptive language-learning technologies and cognitively informed natural language processing (NLP) systems. Although Transformer-based language models achieve remarkable performance across linguistic tasks, the extent to which their computational representations align with human reading behavior remains poorly understood. This study investigates the relationship between eye-tracking measures of cognitive effort and computational indicators derived from deep language models, focusing on reading in Polish (L1) and English (L2).

An experimental framework was developed to collect and analyze eye-movement data during language-processing tasks. The study involved 26 Polish-speaking participants aged 19–25, who read Polish and English texts of varying complexity and completed reading-comprehension tasks. A dedicated Python-based application integrated with a Tobii eye tracker enabled synchronized stimulus presentation, gaze recording, word-level fixation analysis, and visualization of reading patterns. The experimental material comprised 20 texts, equally divided between Polish and English. Cognitive effort was assessed using oculomotor indicators, primarily fixation counts and durations, complemented by word-level gaze statistics and regression measures.

The recorded behavioral data were compared with computational measures from Transformer-based models, including BERT, HerBERT, mDeBERTa, and GPT-2. Five analytical approaches were examined: token masking, natural language inference, language-model surprisal, attention-based predictors, and question-answering attribution. Correlation analy-

ses assessed whether model-derived indicators of word importance, predictability, and contextual processing aligned with observed human gaze behavior.

The strongest agreement with eye-tracking measurements was observed for language-model surprisal, with mean Spearman correlations of 0.643 across all texts, 0.810 for Polish, and 0.498 for English. These findings suggest that words deemed less predictable by language models require greater visual attention during reading. Attention-based predictors also showed meaningful, though language-dependent, associations with oculomotor measures. In contrast, token-masking indicators exhibited less stable relationships, while question-answering attribution showed limited correspondence with human gaze patterns. The evaluated question-answering model correctly identified exact answers in only two of eight examples, yielding a mean token-level F1 score of 0.321.

The results indicate that computational measures of linguistic predictability can provide useful approximations of selected aspects of human reading effort, though model attention should not be equated directly with human cognitive attention. The observed language-dependent differences further underscore the importance of language-specific validation and calibration. The study contributes an integrated experimental and computational framework for investigating human–model correspondence and demonstrates its potential to identify difficult text fragments, support adaptive learning environments, and improve the selection of educational materials. Further validation with larger and more diverse participant groups is required before these findings can be generalized.

Ethics statement: On July 1, 2026, the Rector's Committee for Research Ethics Involving Human Participants at Gdańsk University of Technology issued a favorable opinion on the study. All participants provided informed consent, and research data were collected without personal identifiers.

Visual Programming Framework for Learning Data Analysis

Tomas Blažauskas, Eimutis Karčiauskas

Kaunas University of Technology

tomas.blazauskas@ktu.lt

Spreadsheet applications such as Microsoft Excel are currently the starting point for data analysis in schools. Moving towards text-based programming adds additional syntactic and conceptual demands. Research on the transition from block-based to text-based programming highlights difficulties with syntax and data types. This motivates creating a learning environment that lets learners develop algorithmic thinking before learning a programming language's syntax. Block-based programming lets learners build algorithms with fewer syntactic challenges. However, some students perceive these environments as introductory tools with limited applicability. As programs grow, existing block-based development environments can also become cumbersome to navigate and modify. Therefore, we should create compact representations suited to meaningful data-analysis tasks. Building on earlier work related to visual programming and data science, we present a Blockly-based framework for school-level data analysis. It combines general programming constructs with task-oriented blocks for selected Python operations and data-analysis libraries. Supported operations include importing CSV files, selecting columns, filtering and sorting rows, grouping data and calculating summary statistics. We implemented interactive tables and plotting to support exploration and interpretation. Separate workspaces for reusable components were introduced to help organize larger programs. The framework generates readable Python code, making the correspondence between visual operations and textual expressions explicit. Its design supports algorithmic thinking in data analysis before learners master Python syntax. This approach should enable a gradual pathway towards text-based programming. The created educational development environment combines compact data-processing operations, modular program organization, and access to established data-analysis libraries. We conducted two experiments to

evaluate the visual programming language's compactness and the size of the generated Python code. We divided nine distinct school-level data analysis tasks into 27 computational subtasks. In the first experiment, two solutions were constructed for each subtask: one using general-purpose blocks from the same environment and another additionally using specialized blocks from the "Data" category. We compared block count, visual footprint, structural depth, and whether solutions fit within a predefined workspace. In the second experiment, we examined Python code generated from solutions using specialized blocks. We compared this code with separately written, concise Python and pandas reference implementations. The metrics included lines of code, lexical tokens, and abstract syntax tree nodes. The specialized data blocks reduced the total block count from 1,632 to 1,080 (33.8%) and decreased the aggregate visual footprint under the default layout by 46.0%. The number of solutions fitting within the predefined workspace increased from 5 to 13 out of 27. Reductions occurred in 11 subtasks, while the remaining 16 solutions were unchanged. The generated Python code contained 2,136 lexical tokens, compared with 1,509 in the reference implementations, making it 41.6% larger. The generator introduced no helper functions. These results show how specialized blocks influence program compactness within this benchmark. However, they do not establish improvements in student learning or code comprehensibility. In the future, we will address operations that currently require lengthy block sequences, expand support for data visualization, and reduce implementation-specific constructs in the generated Python code. Additional experiments will examine execution time, memory consumption, and scalability across datasets of different sizes. We will also evaluate the framework's impact on student learning and code comprehensibility.

From Obfuscation to Defence: Adversarial Malware in Controlled C2 Environments

Arnoldas Budžys, Juozas Dautartas, Viktor Medvedev, Olga Kurasova

Institute of Data Science and Digital Technologies
Vilnius University

arnoldas.budzys@mif.vu.lt

Malware obfuscation changes the appearance of malicious software while aiming to preserve its functionality. These changes can make malware detection and analysis more difficult. This work reviews common obfuscation techniques, their limitations, and possible countermeasures. Particular attention is given to adversarial machine learning (AML), which can be used to modify malware features and investigate the robustness of machine learning-based detectors [1].

The work forms part of a research project aimed at developing a secure command and control (C2) framework for conducting realistic yet controlled cyber exercises [2]. The project explores the use of artificial intelligence and AML methods to generate and modify malware samples in an isolated laboratory environment. The C2 infrastructure is designed to coordinate their execution, ensure control over the experiments and collect information for the evaluation of defence systems.

In our previous study [3], we investigated a method of adversarial modification of Windows API import features using a conditional variational autoencoder (CVAE). The experiments used a dataset comprising six classes, containing features extracted from Windows executable files. Adding 20 selected API import features resulted in a reduction in the malware detection rate achieved by the evaluated classifier from 87.5% to approximately 30%. These results demonstrate a limitation of static, API-based malware detection and motivate further investigation of modified samples in controlled environments.

The analysis takes into account both detection results and the functionality of the modified samples. A sample that is no longer detectable but is unable to perform its intended task does not indicate

the successful application of obfuscation techniques that preserve functionality. Ongoing research investigates how to effectively integrate these methods into C2-based experiments, combining detection measurements with independent observation of program behaviour. The overall goal is to ensure the reproducibility of research, enhance the realism of cyber exercises, and identify vulnerabilities in existing defence systems.

Acknowledgments. This project has received funding from the Research Council of Lithuania (LMTLT), agreement No. S-MIP-24-116. Project website: https://amberc2.projektas.vu.lt/.

References

[1] Dautartas, J.; Čypas, J. R.; Kurasova, O.; Medvedev, V. (2026). Evasion of Malware Classifiers by Injecting Category-Specific Benign Features // 2026 international conference on Advances in Artificial Intelligence and Machine Learning (AAIML), 20–22 March 2026, Tokyo, Japan: IEEE, 2026. DOI:10.1109/AAIML67890.2026.11498090

[2] Budžys, A.; Dautartas, J.; Jomantas, H.; Medvedev, V.; Kurasova, O. (2025). AMBER C2: Enhancing Cyber Defence With Ethical Adversarial Machine Learning // 16th Conference on Data Analysis Methods for Software Systems (DAMSS), 27–29 November 2025, Druskininkai, Lithuania, 2025. DOI:10.15388/DAMSS.16.2025

[3] Dautartas, J.; Kurasova, O.; Čypas, J. R.; Medvedev, V. (2026). Learning to Look Benign: Targeted Evasion of Malware Detectors via API Import Injection. arXiv preprint arXiv:2605.18624, 2026. DOI:10.48550/arXiv.2605.18624

Not Every Peak Is a Beat: R-Peak Detection Errors in Single-Lead Holter ECG

Nika Bukolovaitė[1], Zigmantas Kęstutis Juškevičius[1, 3], Žygimantas Abramikas[1,2], Viktor Medvedev[1], Jolita Bernatavičienė[1]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Clinic of Cardiovascular Diseases, Institute of Clinical Medicine
Faculty of Medicine, Vilnius University
[3] UAB Zive

nika.bukolovaite@mif.vu.lt

Accurate detection of R-peaks is extremely important when analysing ECG signals from portable devices, as RR intervals, heartbeat classification, and heart rhythm assessment all depend on it. Even a single incorrectly detected R-peak can skew the results of the signal analysis.

During the analysis of single-lead Holter recordings, including noise removal and beat annotation, we observed recurring R-peak detection errors, sometimes even in relatively clean signals. These included detections on T waves and small spurious peaks, missed QRS complexes and incorrectly positioned peaks in predominantly negative QRS complexes. Challenging signal morphologies, such as inverted T waves and altered QRS shapes, were also observed.

This poster presents an initial investigation of these errors using examples from real Holter recordings. We examine how QRS polarity and signal filtering affect detection, and discuss the possible role of T-wave morphology and noise. Several widely used open-source R-peak detectors are compared with the detector currently in use, focusing on differences in detected peak positions and agreement between methods. We also explore how the frequency band used for detection and recording-level polarity correction influence the detected peak locations. As verified reference annotations are not yet available, agreement between detectors is analysed without treating it as a measure of detection accuracy.

Before changing the detection algorithm, three questions need to be answered: where to mark the peak when the QRS complex has no clear positive wave, what timing error is acceptable, and how to ensure fixes don't break recordings that already work well.

Acknowledgements. This project has received funding from the Research Council of Lithuania (LMTLT), agreement No. S-ITP-25-9.

Cross-Institution Reliability of Multi-Label Chest X-Ray Classification Software: Discrimination, Calibration and Subgroup Fairness under Domain Shift

Paulius Bundza, Justas Trinkūnas

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

paulius.bundza@vilniustech.lt

Automated chest X-ray interpretation software is often selected from in-distribution leaderboards, yet accuracy on a source hospital is a weak proxy for behaviour in a new institution. Differences in acquisition protocol, equipment vendor, patient mix and labelling conventions induce domain shift that can silently degrade a deployed classifier. This contribution studies that software-reliability problem for seven pre-trained multi-label convolutional models from TorchXRayVision: six DenseNet-121 variants (224x224) trained on NIH ChestX-ray14, CheXpert, MIMIC-CXR (CheXpert or NegBio labelers), PadChest, or a multi-source mix, and one ResNet-50 (512x512) trained on the multi-source mix. A reproducible evaluation pipeline scores each (model, label, uncertainty-policy) combination with AUROC, average precision, Brier score and expected calibration error (ECE). Uncertainty is quantified by 500-iteration patient-level cluster bootstrap 95% confidence intervals. Uncertain labels follow an ignore-uncertain policy. Fairness is measured on four axes available in the target cohort: sex, age band, radiographic view (AP versus PA) and equipment manufacturer. In-distribution baselines use NIH ChestX-ray14 (112,120 frontal images for the DenseNet models; a 1,000-image stratified subset for ResNet-50). Out-of-distribution evaluation uses a study-level BRAX v1.1.0 frontal cohort (18,424 images) that was absent from every training set. Aggregate comparison is restricted to seven labels shared by all models: atelectasis, cardiomegaly, consolidation, edema, pleural effusion, pneumonia and pneumothorax. The same seven classifiers are served bit-identically by a containerized Django, Celery and PostgreSQL prototype (MCADS),

so the reported numbers describe the deployed software rather than a detached research table. On BRAX, macro AUROC ranged from 0.652 to 0.741. Multi-source or richly annotated training separated the top three models (PadChest DenseNet 0.741, multi-source DenseNet 0.738, ResNet-50 0.733) from four single-source North-American models (0.652-0.677). Discrimination and calibration were largely independent: ResNet-50 achieved the best ECE (0.028) while ranking only third in AUROC, whereas a MIMIC CheXpert-labeler DenseNet reached ECE 0.392. Relative to NIH, overlapping models lost 3.4-8.7 percentage points of macro AUROC under domain shift. Subgroup gaps were large: AP versus PA AUROC differed by up to 21.0 percentage points; for the multi-source DenseNet, pediatric (ages 0-4) and geriatric (85+) bands fell to 0.563 and 0.420. In-distribution-only benchmarks therefore systematically overstate software robustness. A deployable chest X-ray stack should combine external validation, post-hoc calibration and explicit subgroup warnings for portable AP views and extreme age, rather than selecting a model by a single AUROC. Limitations include NLP-derived labels, extreme rarity of some BRAX findings (edema n=9) and a single external hospital. Ongoing work extends the protocol to additional external cohorts and couples calibrated probabilities with severity-aware triage rules.

Trust Issues: AI Inside Command and Control Frameworks

Antanas Čypas, Juozapas Rokas Čypas, Juozas Dautartas, Viktor Medvedev

Institute of Data Science and Digital Technologies
Vilnius University

antanas.cypas@mif.vu.lt

The use of artificial intelligence (AI) is increasingly explored in command and control (C2) infrastructures, including LLM-assisted construction of C2 infrastructure [1] and adversarial modification of malware features [2]. However, because these efforts typically target isolated tasks, a systematic view of how AI is applied across the different stages and roles of a C2 framework is still lacking. It is also often unclear whether a claimed AI capability actually works or is only reported.

Within this work, C2 frameworks are understood as infrastructures used to coordinate, control, and manage software agents deployed on target machines during cybersecurity operations and exercises. The term has no agreed definition. It originates in military doctrine, where its definition "is still being debated" even within the institutions that coined it [3]. The same lack of consensus persists in its cybersecurity usage, which this work addresses by using "C2 framework" exclusively in the narrower sense defined above.

This study reviews published studies and documented implementations of AI within C2 frameworks. Specifically, we examine at which stage of the C2 framework lifecycle it is used, focusing on three main areas: infrastructure and agent generation [1], agent coordination and task planning, and detection evasion using adversarial machine learning [2, 4]. For each area, we note whether each AI capability was actually shown to work or only reported, and we organize the results into a structured map of AI roles across the C2 lifecycle. This review will help shape decisions for a C2 framework currently in development [4].

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

References

[1] Hasui, K.; Matsugaya, S.; Shimamura, M.; Hashimoto, M. (2026). LLM-assisted Generation of Pseudo-C2 Servers for IoT Malware Dynamic Analysis // 21st Asia Joint Conference on Information Security (AsiaJCIS 2026), EPiC Series in Computing, vol. 111, pp. 1–16, 2026. DOI:10.29007/7hdm.

[2] Dautartas, J.; Čypas, J. R.; Kurasova, O.; Medvedev, V. (2026). Evasion of Malware Classifiers by Injecting Category-Specific Benign Features // 2026 international conference on Advances in Artificial Intelligence and Machine Learning (AAIML), 20–22 March 2026, Tokyo, Japan: IEEE, 2026. DOI:10.1109/AAIML67890.2026.11498090.

[3] Curts, R.; Campbell, D. (2005). Building an Ontology for Command & Control // 10th International Command and Control Research and Technology Symposium (ICCRTS), 2005.

[4] Budžys, A.; Dautartas, J.; Jomantas, H.; Medvedev, V.; Kurasova, O. (2025). AMBER C2: Enhancing Cyber Defence With Ethical Adversarial Machine Learning // 16th Conference on Data Analysis Methods for Software Systems (DAMSS), 27–29 November 2025, Druskininkai, Lithuania, 2025. DOI:10.15388/DAMSS.16.2025.

Mask Off: Adversarial Training Against Malware Disguised with Benign-Category APIs

Juozapas Rokas Čypas, Olga Kurasova, Viktor Medvedev, Juozas Dautartas

Institute of Data Science and Digital Technologies
Vilnius University

juozapas.cypas@mif.vu.lt

Machine learning malware detectors that use static features of Windows executables, such as imported API functions, can be tricked when an attacker adds imports typical of legitimate software, masking the malware as a benign program. Our previous work [1, 2] showed that such an attack can be targeted when benign software is split into functional categories (office, development, multimedia, security, administration): each malware sample is matched to the category it most resembles and receives that category's most characteristic APIs. A related analysis [3] used the same categories to show which benign patterns are linked to missed malware and suggested testing whether such modified malware can improve training.

This study tests that idea by turning the attack into a defence. Half of the malware in the training split receives APIs drawn at random from the top of its assigned category's ranked list, with higher-ranked APIs more likely to be chosen, and keeps its malware label. The aim is to teach the detector that benign-looking imports do not make a file benign. The procedure needs no access to the detector's internals and can be applied to any classifier.

We test four types of classifier on the WinAPI-AdvMal dataset [4]. Our experiments have shown that adversarial training can significantly improve the detection of adversarially perturbed malware. There is a tradeoff, though, between improved detection of adversarial malware and the detection of unmodified malware. Detection of unmodified malware falls by 5–10 percentage points, gains are smaller against additions chosen differently from the training set, and malware with randomly added APIs is detected less often. How APIs are chosen for

adversarial training matters: ranked selection outperforms uniform selection from the same lists, and a single benign API list without categories does not reproduce the improvement.

In practice, the method can be repeated whenever a detector is retrained. Comparing predictions on the same samples before and after adversarial training shows, for each benign category, which errors were corrected and which new misses appeared. Its settings should be tuned on separate validation data and checked against unmodified malware, false positives, and other attack strategies. All experiments modify feature vectors, not executables. The next step is testing on real modified executables, including command and control (C2) malware studied in our ongoing project.

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

References

[1] Dautartas, J.; Čypas, J. R.; Kurasova, O.; Medvedev, V. (2026). Evasion of Malware Classifiers by Injecting Category-Specific Benign Features // 2026 international conference on Advances in Artificial Intelligence and Machine Learning (AAIML), 20–22 March 2026, Tokyo, Japan: IEEE, 2026. DOI:10.1109/AAIML67890.2026.11498090.

[2] Čypas, J. R.; Dautartas, J.; Kurasova, O.; Medvedev, V. (2025). Improving Malware Detection by Analyzing Similarities of Multi-Category Benign Software // 16th Conference on Data Analysis Methods for Software Systems (DAMSS), 27-29 November 2025, Druskininkai, Lithuania. DOI:10.15388/DAMSS.16.2025.

[3] Medvedev, V.; Dautartas, J.; Čypas, J. R.; Kurasova, O. (2026). Poster: Beyond a Single Benign Class: What Functional Categories of Benign Software Reveal About Static Malware Detection. Proceedings of the 2026 ACM SIGSAC Conference on Computer and Communications Security (CCS'26). DOI:10.1145/3830454.3846425.

[4] Dautartas, J.; Kurasova, O.; Čypas, J. R.; Medvedev, V. (2026). WinAPI-AdvMal: A Six-Class Windows API Import Dataset for Adversarial Malware. DOI:10.5281/zenodo.20208958.

Composing Independently Issued Grants for Cross-Organization LLM Agent Workflows

Andrius Daranda

Kauno kolegija Higher Education Institution

andrius.daranda@go.kauko.lt

Large language model (LLM) AI agents now call tools across different organizations. One organization may grant access to data. Another may control computation, storage, or publication. A workflow that spans them must combine grants that were issued independently, and it must not break any issuer's restrictions. This paper calls this operation a cross-domain authorization join. The joined authorization must fulfill all contributing grants and remain enforceable at the precise point where a tool exerts its effect. The paper introduces JoinGuard, a gateway-enforced architecture for such joins. Its primary contribution is an operational composition contract for independently managed grants. Notably, it does not introduce a new capability primitive. The threat model treats AI agents as untrusted. They may be malicious, and they may collude. Authority services are trusted, and every participating domain trusts the gateway. A tool executes an effect only when it receives a valid permit issued by the gateway. Permit verification establishes authenticity. Exact effect and dependency binding fixes the arguments, targets, holders, and provenance. Single-use enforcement blocks replay. Revocation fencing and activation barriers ensure that no permit is issued and no effect runs after a revocation is acknowledged. Restrictions carry over to derived artifacts. Safety is defined against the latest acknowledged authorization state, not against real-time revocation. A task and trace oracle, independent of the gateway's decision logic, checks each run for enforcement violations, policy gaps, and availability failures. The gateway sees an AI agent only through the requests it submits. Therefore, the evaluation models adversaries as request schedules. It covers 11 configurations: intact JoinGuard, nine variants with seeded defects, and a deny-all control that marks the utility floor. Three adversary methods

run under three deterministic seeds. The first is a fixed library of 16 hand-crafted attack traces. The second is a seeded random schedule generator. The third is a bounded search over schedule mutations. Random generation and search each receive 48 schedules per seed. This gives 99 method-system-seed combinations and 3,696 episodes. Intact JoinGuard produced no enforcement violations. Each seeded defect exposed a corresponding failure. Without permit verification, it accepted forged permits. Without dependency binding, mutated arguments, targets, and holder aliases passed unnoticed. Without single-use protection, direct-sink replay succeeded. Without revocation fencing, effects occurred after revocation acknowledgment. Without activation barriers, permits were issued across recorded barriers even though consume-time fences stayed active. Without artifact-restriction retention, prohibited downstream publications went through. One ablation targets composition itself by removing a single issuer's restriction. The result was an intent-level gap, not a checker-level violation. The missing restriction never reached the composed artifact, so the gateway correctly enforced the weaker policy it was given. The independent oracle still flagged the resulting publication as forbidden. This result separates enforcement correctness from policy-composition correctness. Crash experiments separate safety from availability in the same way. With recovery or durable decision handling disabled, workflows stalled or stayed incomplete. These outcomes are classified as availability losses, not as unauthorized effects. The adversaries in this evaluation are scripted, random, and bounded-search procedures. Adaptive LLM-driven adversaries are left to future work. Within the stated trust assumptions and adversary classes, JoinGuard maintained its specified enforcement invariants. The targeted ablations show each mechanism's contribution: removing them produced enforcement failures, an authorization-composition gap, an activation-barrier breach, and recovery-related availability losses.

Patient-Specific 3D Drilling Aimer Generation for Mandibular Ramus Fractures

Romas Daugnora, Jonas Domarkas, Vilius Šarkinas, Eimantas Zaranka, Monika Katilienė

MB Skaitmeninė verslo komunikacija
UAB DRE Design & Consulting
Vytautas Magnus University

monika.katiliene@card-ai.eu

Mandibular fractures are common maxillofacial injuries that can impair jaw function and require careful treatment planning. In conventional miniplate fixation, surgeons determine plate positions and screw trajectories from imaging and intraoperative assessment, manually adapting plates to the bone where required. Variations in fracture geometry and the proximity of the mandibular canal make it challenging to achieve stable fixation while protecting nearby anatomical structures. To support these decisions, this study presents an AI-based platform that combines anatomical segmentation with constrained fixation planning and quantitative geometric and biomechanical assessment. The selected plan is used to generate compact, patient-specific 3D drilling aimers that guide screw-hole placement. Computed tomography (CT) volumes are segmented using the nnU-Net-based DentalSegmentator model. Six point cloud architectures – RandLA-Net, Point Transformer V3, PointNet, PointNet++, 3DMedPoint, and SuperPoint3D—were evaluated on 372 annotated mandibular point clouds. In the three-class task, RandLA-Net achieved the highest overall test accuracy of 87.0%, compared with 83.2% for Point Transformer V3. Its classwise accuracies were 87.4%, 86.1%, and 98.4% for permitted, avoidance, and fracture regions, respectively. After geometric refinement and fragment reconstruction, class proportions within screw corridors serve as anatomical scoring metrics. Candidates must satisfy constraints on plate placement across both fragments, plate–bone intersection, screw-tip position within bone, and clearance from the mandibular canal, fracture region, and neighbouring screws. Plate fit is quantified using root mean square error and a fracture-

weighted variant. Finite element analysis estimates interfragmentary displacement and von Mises stress. These measures are combined with geometric and anatomical metrics to rank feasible candidates. Compact, patient-specific drilling aimers are constructed to fit the bone fragments, with guide channels aligned with the selected screw trajectories. The resulting models are prepared and sent for 3D printing.

Acknowledgments. These results were obtained as part of the project "Development of Innovative Products at DRE Design & Consulting UAB" (Project No. 02-019-K-0066), co-funded by the European Union through the European Regional Development Fund.

Mitigating Hallucinations in Large Language Models

Konstantinos Diamantaras

International Hellenic University, Greece

kdiamant@ihu.gr

Large Language Models (LLMs) have transformed natural language generation, yet they continue to suffer from hallucinations – outputs that are linguistically fluent but factually incorrect, ungrounded, or logically inconsistent. This talk offers a comprehensive synthesis of state-of-the-art hallucination mitigation techniques, organized into six methodological categories: Training and Learning Approaches, Architectural Modifications, Input and Prompt Optimization, Post-Generation Quality Control, Interpretability and Diagnostic Methods, and Agent-Based Orchestration. The presentation will examine the underlying causes of hallucinations – from noisy or outdated training data to probabilistic decoding biases – and review cross-stage mitigation strategies spanning pre-generation, generation, and post-generation phases. Emphasis will be placed on hybrid paradigms such as retrieval-augmented generation, contrastive and reinforcement learning from human or AI feedback, and emerging self-reflective multi-agent systems. The talk will address persistent challenges—such as scalability, latency–accuracy trade-offs, and benchmark inconsistencies – while highlighting underexplored research directions including knowledge-grounded fine-tuning, factual–creative balance, and context-aware verification pipelines. Ultimately, it will focus on designing more reliable, transparent, and context-sensitive LLMs for high-stakes domains such as healthcare, law, and defense.

Budget-Constrained Competitive Facility Location-Design with Sequential Customer Choice Rules

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

[1] University of Murcia, Spain
[2] Vilnius University, Lithuania

algirdas.lancinskas@mif.vu.lt

This research addresses a discrete competitive facility location and design problem in which an entering entity tries to maximize captured customer demand by selecting facility locations and designs under a limited budget. Customer behavior is modeled using two sequential choice rules: a proportional rule, applied when facilities satisfy a minimum attractiveness threshold, and a binary rule, applied otherwise, with demand shared equally among facilities with maximum attractiveness.

We propose a population-based metaheuristic that combines evolutionary search with a learning-based ranking mechanism. The algorithm evaluates candidate locations based on their contribution to solution quality and uses the learned ranks, together with geographical proximity, to guide the generation of new feasible solutions. An important feature is the ability to retain and transfer accumulated knowledge between related problem instances.

Computational experiments are conducted using geographical and population data from Spanish municipalities, considering different investment budgets, facility designs, and competitive scenarios. The proposed algorithm is evaluated against exact optimal solutions obtained by solving the problem with a deterministic optimization algorithm. Experimental results demonstrate that the metaheuristic identifies good solutions and transferring learned ranks between problem instances significantly improves the probability of obtaining optimal and near-optimal solutions.

The results highlight the potential of learning-based metaheuristics for solving competitive facility location problems and demonstrate the benefits of transferring optimization knowledge across related problem instances.

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

Quantum Optimization of Constrained Vehicle Routing Problems

Ernestas Filatovas[1], Dmitry Podkopaev[1, 2], Remigijus Paulavičius[1]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Systems Research Institute, Polish Academy of Sciences

ernestas.filatovas@mif.vu.lt

Vehicle routing problems (VRPs) are central to logistics and transportation, where route planning must account for numerous operational constraints, including customer time windows, heterogeneous fleets, vehicle capacity restrictions, working-time limits, and other practical considerations such as travel times and costs. These interdependent requirements create tightly coupled routing and scheduling decisions, increasing computational complexity and making efficient optimization challenging. Quantum annealing offers a promising approach to combinatorial optimization, but its practical applicability depends on effective mathematical formulations and rigorous computational evaluation. This work investigates quantum-oriented optimization of constrained vehicle routing problems, with a focus on time windows. We study the transformation of a mixed-integer linear programming (MILP) model into a quadratic unconstrained binary optimization (QUBO) formulation suitable for quantum annealing. Representing time-related constraints and inequalities requires additional binary variables and penalty terms, which can substantially increase the size and complexity of the resulting model. We systematically compare alternative QUBO formulations, focusing on variable encoding, constraint representation, and penalty calibration. Since changes to individual components can affect interactions throughout the model, we evaluate their impact on the complete QUBO structure and solution quality rather than considering isolated constraints. In particular, we examine how different constraint groups contribute to model size and interaction structure, helping identify the main sources of formulation complexity. Computational experiments are conducted on vehicle routing instances of different sizes. The MILP

model provides reference solutions, while simulated quantum annealing is used to assess the feasibility and quality of solutions obtained from alternative QUBO formulations. Structural results reveal that the effects of alternative QUBO formulations depend on problem size, with reductions in quadratic interactions becoming more pronounced for larger instances. These findings provide insights into the development of more effective quantum and hybrid optimization approaches for constrained vehicle routing problems.

Acknowledgements. This research was funded by the Research Council of Lithuania under Grant Agreement No. S-ITP-25-5. The authors are also thankful for the HPC resources provided by the IT Research Center of Vilnius University.

Modelling Before Solving with Metaheuristics

Carlos M. Fonseca

University of Coimbra, CISUC/LASI, DEI, Portugal

cmfonsec@dei.uc.pt

To a large extent, research in metaheuristics continues to pursue the development of dedicated algorithms for specific problems. This is in stark contrast to the field of mathematical programming, where solvers have become so effective that, in most cases, solving a given optimisation problem amounts to developing a mathematical model and passing it to an existing solver to obtain the desired solution. In the context of metaheuristics, however, such a clear separation between problem models and solvers is widely considered unworkable. This presentation explores the potential impact of placing a greater focus on problem modelling on the future development and practical adoption of metaheuristics. It also reports on ongoing work, carried out in the scope of the ROAR-NET COST Action, to address some of the challenges posed by this approach.

Acknowledgements. This presentation is based upon work from COST Action Randomised Optimisation Algorithms Research Network (ROAR-NET), CA22137, supported by COST (European Cooperation in Science and Technology). Part of the work presented is funded by national funds through FCT – Foundation for Science and Technology, I.P., within the scope of the research unit UID/00326 – Centre for Informatics and Systems of the University of Coimbra, https://doi.org/10.54499/UID/00326/2025.

Public EV Charging under Changing Participation: A Stochastic Agent-Based Simulation

Jose J. H. Galvez, Mathias Osorio, Alexis Lozano, Jose Evora, Jose J. Hernandez Cabrera

Universidad de Las Palmas de Gran Canaria, Spain

jose.galvez@ulpgc.es

Planning a public electric vehicle charging network requires deciding where to install stations, how much capacity to provide, and when to expand it. Demand evolves as vehicles enter the fleet and users gain access to private charging. This study proposes a stochastic agent-based simulation method, calibrated with historical charging records from Gran Canaria, to explore these changes and support phased infrastructure planning.

The records comprise account identifiers, session-start timestamps, station locations, and delivered energy. Each account will be represented by an agent, without equating accounts with individual people or vehicles. Behavioural profiles will describe charging frequency, preferred days and times, spatial preferences, and energy variability. Dynamic states will record recent sessions, elapsed time since public charging, usage regime, and domestic charging access where observed or specified by a scenario.

Sessions will be generated sequentially. A recurrence model will determine charging probability from elapsed time, calendar context, and usage regime. Conditional on a session, a spatial choice model will select a station and an energy distribution will determine delivered energy. Routine behaviour will reproduce familiar time–location combinations, while an occasional-use component will permit atypical visits under plausible spatial and temporal constraints.

Spatial choice will use charging potential fields: continuous attractiveness surfaces interpreted as behavioural utilities. Activity anchors will generate attraction that decays with travel distance or time, with weights varying by weekday and time of day. Station utility will combine this attraction with familiarity and documented attributes such as charging speed, price, and availability. Normalised utilities over feasible stations will yield choice probabilities. Where external anchor information is unavailable, latent preference centres will be estimated without labelling them as homes or workplaces.

Calibration will fit recurrence and energy models while accounting for observation exposure and censored waiting times. Potential-field parameters will be estimated from conditional station-choice likelihoods using feasible alternatives available at each session. Hierarchical estimation will pool accounts; chronological validation will control model complexity and assess simulated behavioural patterns. Final evaluation will use later held-out observations with parameters fixed, comparing the method with independent attribute sampling and empirical station-choice models without potential fields.

Candidate stations will enter feasible choice sets through their coordinates and documented attributes, without invented personal usage histories. Repeated simulations will estimate session counts and energy demand, propagating parameter, anchor, and stochastic uncertainty. Paired runs with and without a candidate will assess redistribution from existing stations. Additional public charging will be estimated only where evidence supports an outside option or a recurrence response to improved access; otherwise, it will remain a sensitivity scenario.

Population scenarios will combine new-account arrivals with changes in existing users' participation. Historical entry cohorts will inform initial profiles, but first observed use will not be equated with vehicle acquisition. Agents may become occasional users, inactive, or reactivated. Home-charging adoption may reduce routine public use while retaining occasional visits. Its rate and effect will remain scenario assumptions unless independently supported, since absence from public records cannot identify domestic installation. Experiments will cross low, central, and high vehicle entry with slow, intermediate, and rapid home-charging adoption, including uneven domestic access and alternative spatial profiles of incoming users.

Planning experiments will compare retaining the network, expanding existing stations, and adding feasible sites. Demand distributions will identify the probability and timing of exceeding planning thresholds. Where connector counts, power, durations, and availability support capacity modelling, occupancy, waiting, and unmet demand will also be assessed. Site priorities and expansion timing will reflect service, coverage, implementation lead times, and uncertainty, alongside land, grid, and cost constraints where available. The intended contribution is a calibrated behavioural framework connecting individual routines and spatial attraction to infrastructure decisions under changing public-network participation.

Identifying Unusual Topic Alignment Patterns in E-Learning Materials

Beata Gancevska, Simona Ramanauskaitė

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

beata.gancevska@vilniustech.lt

The quality of e-learning materials is an important part of course quality. One question is whether the materials used in a course match the topics that students are expected to study. This is not always easy to assess, particularly in courses with many learning resources. Some materials cover several topics, while others may be difficult to assign to a single topic. Automated methods can help by giving each resource a score for its relationship with the course topics. However, the scores themselves do not always provide a clear answer. In some cases, the way scores are distributed across topics may be as interesting as the highest score. This study examines unusual patterns in the topic alignment scores of e-learning materials. Semantic alignment scores are obtained using large language models (LLMs) to evaluate presentation slides against the course topics. The main purpose is to examine the resulting alignment patterns rather than focus primarily on evaluating the accuracy of the models used to generate the scores. Exploratory data analysis is used to examine score distributions and identify patterns that stand out. These may include resources with low scores for most or all topics, as well as resources with strong alignment to multiple topics. Selected cases are examined in more detail to see how they relate to the learning resources and the topics defined in the course. The study does not treat unusual scores as proof that a resource is irrelevant or that a course is poorly organized. Instead, they are considered a starting point for closer examination. The results are summarized using statistical measures and visualizations that show differences between common and less common scoring patterns. This makes it possible to examine the results from different perspectives rather than relying only on a single alignment score. The aim of this research is to find out what can be learned from

unusual topic alignment results and whether they can help identify learning resources that need further review. The study provides a basis for understanding how alignment scores can be interpreted in practice. It may also help in developing approaches for reviewing the organization of e-learning materials based on existing semantic alignment scores.

When Malware Changes: Assessing Detection Robustness with Generative AI

Mantas Globys, Viktor Medvedev

Institute of Data Science and Digital Technologies
Vilnius University

mantas.globys@mif.stud.vu.lt

Evaluating malware detection systems requires understanding how detection changes after code modification. Generative artificial intelligence (AI) supports research on this problem through code variant generation and detection rule development. Huang et al. [1] and LLMalMorph [2] examine variants generated using large language models (LLMs). Rule-LLM [3] studies YARA and Semgrep rule generation using models such as GPT-4o and Claude-3.5-Sonnet. These applications require separate consideration of code structure, functionality, and detection outcomes.

Building on these research directions, we propose a methodology for assessing malware detection robustness using generative AI. The first question concerns how modifications that preserve relevant functionality affect malware detection. The second asks whether AI-assisted rules detect additional cases missed by the baseline tool. We will examine both under controlled conditions, without assuming that AI use improves detection.

The assessment covers sample selection, variant preparation and validation, baseline measurement, and evaluation of additional rules. Robustness means retaining detection after modifications that preserve relevant functionality. We will measure the proportion still detected among variants that pass predefined behavioural checks and whose original samples were detected. Originals and their variants will be compared under the same detection conditions. Rule development and final evaluation will use separate data. Each original and its variants will be kept in the same data partition. False positives, invalid variants, and unsuccessful generation or analysis attempts will also be recorded. This will help relate conclusions to the conditions examined.

For a limited pilot, we will assess the suitability of data sources and detection tools. Potential sources include MalSource [4] for malware source code. Assemblage [5] offers an alternative source of candidate benign control files. We will check their project purpose, provenance, and platform compatibility. The presentation will describe the proposed approach and its practical limitations. The contribution at this stage is a structured evaluation protocol for subsequent experiments. Empirical validation and conclusions about possible detection improvements remain future work.

References

[1] Huang, Z.; Pa Pa, Y. M.; Yoshioka, K. (2024). Exploring the Impact of LLM Assisted Malware Variants on Anti-Virus Detection // 2024 IEEE Conference on Dependable and Secure Computing (DSC), pp. 76–77. DOI:10.1109/DSC63325.2024.00027.

[2] Akil, M. A.; Li, A. S.; Karim, I.; Iyengar, A.; Kundu, A.; Parla, V.; Bertino, E. (2025). LLMalMorph: On the Feasibility of Generating Variant Malware Using Large-Language-Models // 2025 IEEE 7th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA), pp. 160–171. DOI:10.1109/TPS-ISA67132.2025.00026.

[3] Zhang, X.; Du, X.; Chen, H.; He, Y.; Niu, W.; Li, Q. (2025). Automatically Generating Rules of Malicious Software Packages via Large Language Model // 2025 55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), pp. 734–747. DOI:10.1109/DSN64029.2025.00072.

[4] Calleja, A.; Tapiador, J.; Caballero, J. (2019). The MalSource Dataset: Quantifying Complexity and Code Reuse in Malware Development // IEEE Transactions on Information Forensics and Security, vol. 14, no. 12, pp. 3175–3190. DOI:10.1109/TIFS.2018.2885512.

[5] Liu, C.; Saul, R.; Sun, Y.; Raff, E.; Fuchs, M.; Pantano, T. S.; Holt, J.; Micinski, K. (2024). Assemblage: Automatic Binary Dataset Construction for Machine Learning // Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track, pp. 58698–58715. DOI:10.52202/079017-1871.

A Multi-Criteria Decision-Making Framework for Cybersecurity Incident Severity Evaluation

Indre Grigaraviciute, Nikolaj Goranin

Department of Information Systems
Vilnius Gediminas Technical University

indre.grigaraviciute@vilniustech.lt

Cybersecurity incident response requires organizations to identify, assess, and evaluate security incidents according to their potential operational, informational, and organizational impact. Conventional incident severity evaluation often relies on predefined rules or expert judgment, which may provide limited transparency when multiple and potentially conflicting factors must be considered simultaneously. This study proposes a Multi-Criteria Decision-Making (MCDM) framework for cybersecurity incident severity evaluation. The framework evaluates cybersecurity incidents according to five criteria: Operational Impact, Data and Information Impact, Affected Asset Scope, Response Urgency, and Recovery Complexity. Expert judgments are incorporated through the Analytic Hierarchy Process (AHP) to determine the relative importance of the evaluation criteria. The resulting criteria weights are then applied to five representative cybersecurity incident scenarios, which are evaluated and ranked independently using the VIKOR and PROMETHEE methods. The use of two complementary MCDM ranking approaches enables the consistency of incident severity evaluation results to be examined across different decision-making methodologies. Comparative analysis using Spearman's rank correlation is employed to assess the agreement between the resulting rankings, while sensitivity analysis examines the robustness of the evaluation results under variations in criteria weights. The proposed framework provides a transparent and reproducible decision-support mechanism for evaluating the relative severity of cybersecurity incidents based on multiple operational, informational, asset-related, and response-related factors. The approach is intended to support more systematic and consistent cybersecurity incident severity evaluation in organizational incident-response environments.

Where Does Extraction Fail? Ambiguity-Stratified Evaluation of Security Control Extraction from Regulatory Text

Luis-Carlos Herrera, Justinas Janulevičius

Vilnius Gediminas Technical University

luis.herrera@vilniustech.lt

Compliance automation converts regulatory text into machine-readable security controls, and every subsequent stage, from knowledge graphs to configuration verification, operates on that representation; extraction error therefore bounds what those stages can establish. Published extraction pipelines report an aggregate score, cascade designs escalate low-confidence inputs, and hybrid designs reserve language models for ambiguous clauses. Each of these choices presupposes, without testing it, that extraction error concentrates in ambiguous clauses. Within a doctoral project on knowledge bases for information security governance, this study evaluates the extraction stage against an ambiguity-stratified gold standard. It asks how performance on spans and obligation tuples varies across a six-category taxonomy of regulatory ambiguity (lexical, syntactic and semantic ambiguity, vagueness, incompleteness and referentiality) (RQ1), and what share of a per-clause oracle's gain is captured when escalation from a smaller to a larger model is conditioned on annotated ambiguity, decoder confidence or disagreement between prompt variants (RQ2). A corpus of 3,215 clause units was segmented from the NIS2 Directive and the GDPR with PyMuPDF 1.28.2 and from the CPRT export of NIST SP 800-53 Release 5.2.0. Of these, 380 clauses are annotated with obligation tuples over nine roles and with multi-label ambiguity categories and trigger spans, aligned to character offsets by a validating script. An 80-clause development sample, drafted with language-model assistance and reviewed by the first author, fixed fourteen annotation conventions, the prompt and the GLiNER settings, and was excluded from evaluation. The test sample comprises 300 unseen clauses: all remaining in-scope NIS2 and GDPR units and 99 SP 800-53 units stratified by unit type

and ambiguity cue words. Three configurations are compared: GLiNER (modern-gliner-bi-large-v1.0) zero-shot; SecureBERT 2.0 fine-tuned for token classification under five-fold cross-validation grouped by article or control family; and Qwen3-8B emitting JSON under grammar-constrained decoding (vLLM 0.8.5, xgrammar 0.1.18, temperature 0), escalating to Qwen3-32B, quantized to 4 bits with AWQ so that it fits one NVIDIA L40 virtual GPU with 24 GB. Confidence is recovered by re-scoring outputs without the grammar, tuples are matched by Hungarian assignment, and intervals come from a bootstrap over 45 article and control-family clusters. On the 285 clauses that impose a duty, tuple F1 is 0.786 [0.740, 0.826] for Qwen3-8B and 0.771 [0.702, 0.834] for Qwen3-32B, and SecureBERT 2.0 attains the highest span F1, 0.794 [0.757, 0.825], against 0.682–0.705 for the instruction models and 0.293 for GLiNER. Ambiguous clauses (33%) are not measurably harder: the core-tuple difference is −0.073 [−0.197, 0.043] for the 8B and −0.030 [−0.162, 0.084] for the 32B model. Error follows source and structure instead: core-tuple F1 ranges from 0.667 to 0.701 on the GDPR and from 0.788 to 0.832 on SP 800-53, and the number of duties is wrong in a quarter of clauses. No ambiguous clause is flagged as indeterminate, although the flag probability separates ambiguous clauses above chance (AUC 0.66). At a 20% budget the oracle gains 0.116 in core-tuple F1 over random escalation; ambiguity-conditioned escalation captures 6% of that gain, disagreement 16%, and confidence falls below random, because the larger model improves 59 clauses and degrades 48. The results do not support the assumption that extraction error concentrates in ambiguous clauses. Extraction error tracks source, clause structure and duty segmentation rather than ambiguity, the models register indeterminacy in their token probabilities without reporting it, and escalation requires a signal of relative model competence rather than of clause difficulty; evaluation of regulatory extraction should therefore be stratified by structure as well as by ambiguity. Retrieval-augmented extraction and an independent annotator are future work.

Clinical Transferability and Institution-Specific Adaptation of a prostateCA Foundation Model for Prostate MRI Zonal Segmentation

Chia-Chi Hsiao[1], Kuan-Jung Chen[2], Chun-Chieh Wang[1], and Da-Chuan Cheng[2]

[1] Department of Radiology, Kaohsiung Veterans General Hospital, Kaohsiung, Taiwan
[2] Department of Biomedical Imaging and Radiological Science, China Medical University, Taichung, Taiwan

dccheng@mail.cmu.edu.tw

Deep learning models developed using public medical imaging datasets often exhibit reduced generalizability when applied to clinical data from different institutions because of domain shift. This study investigated the clinical transferability and institution-specific adaptation of a prostateCA foundation model for peripheral-zone (PZ) and transition-zone (TZ) segmentation in prostate magnetic resonance imaging (MRI). Two foundation-model construction strategies were investigated using the PROSTATEx (n = 204) and Prostate158 (n = 158) datasets, followed by evaluation on an independent institutional cohort of 100 patients. Before institutional adaptation, the two models achieved Total Dice scores of 0.7377 and 0.7229, respectively, indicating limitations in direct cross-institutional transfer. To address these limitations, the models were fine-tuned using institution-specific annotated data under two training configurations involving 40 and 70 patients, with performance assessed through 10-run cyclic holdout validation. Fine-tuning improved Total Dice scores to 0.8010–0.8055, demonstrating the benefit of local adaptation. Notably, the performance differences between the two foundation-model construction strategies diminished after fine-tuning, while the two training configurations yielded comparable aggregate segmentation performance. However, case-level analysis revealed heterogeneous responses to adaptation, including substantial improvements, limited gains, and newly introduced segmentation

errors. These findings demonstrate that publicly trained prostate segmentation models provide a useful starting point for clinical transfer, but their successful application requires institution-specific adaptation and careful evaluation beyond aggregate performance metrics.

Financing Green Solutions: Asset Returns and Tail Risks

Eugenijus Gabrielius Ivanauskas, Liepa Urbonaitė, Saulius Jokubaitis

Institute of Applied Mathematics
Faculty of Mathematics and Informatics
Vilnius University

saulius.jokubaitis@mif.vu.lt

The translation of green technologies from theoretical potential to industrial adoption hinges on the stability of the financial mechanisms that fund them. The poster will present recent results on the risk dynamics and dependence structures of the Green financial ecosystem, comprising Green Bonds, Clean Energy ETFs, and Carbon Credits. Employing a rolling-window ARMA-GARCH-Vine Copula framework we map the evolving topology of sustainable finance and test the "decoupling hypothesis". The analysis reveals that while Green Bonds have successfully decoupled and act as effective portfolio diversifiers, Clean Energy equities remain deeply integrated with broad market risks, functioning as a centralized "hub" for volatility transmission. We further identify a distinct structural shift where with the fade of the post-pandemic volatility the market network stabilizes into an R-Vine structure that transmits shocks more efficiently. Finally, we assess the viability of clean cryptocurrencies, finding them structurally incompatible with institutional hedging strategies due to extreme tail-risk dependence.

Bridging Corpus Linguistics and Relational Lexical Databases Through Machine Learning: The SERMIS Approach

Dalia Jakulytė, Vitalij Denisov, Arturas Garbuz

Klaipėda University

dalia.jakulyte@ku.lt

Historical language research employs two complementary paradigms. Corpus linguistics relies on annotated texts and automated language-processing tools, whereas lexical and morphological research often requires highly structured databases that support complex linguistic queries and data enrichment, e. g. SLIEKKAS (https://titus.uni-frankfurt.de/sliekkas/) [1]. The SERMIS (System of Morphological Information on Historical Texts, https://sfc.shmf.ku.lt/sermis/) project explores ways to connect these paradigms through a workflow linking corpus-based annotation environments with a relational SQL-based linguistic information system [2,3]. SERMIS is an established web-based platform for the study of historical Lithuanian texts. Its data model represents texts, lexemes, word forms, morphological features, derivational structures, and semantic attributes as interconnected entities stored in a relational database. The system is currently used for linguistic research and already contains fully structured data from the seventeenth-century text Knyga Nobažnystės (1653), while data from Morkūnas' Postilė (1600) are being incorporated. To date, linguistic data have largely been entered and annotated manually. The next stage of development aims to integrate corpus-linguistic and machine-learning technologies into the data acquisition workflow. Several annotated historical texts are available as pilot datasets for investigating how existing corpus annotation tools may support the semi-automatic identification of lemmas, word forms, and grammatical categories. Rather than creating linguistic information directly within database tables, researchers would first exploit corpus-processing technologies and then transform the resulting annotations for import into SERMIS. Within this framework, automatically generated annotations could be reviewed, corrected, enriched, and linked to

additional lexical, morphological, semantic, and derivational information already maintained in the database. The proposed approach would combine the scalability of corpus-based processing with the structured knowledge representation provided by relational databases. The project presents SERMIS as a platform for integrating corpus-linguistic methods with relational lexical resources and discusses future development of data-transfer mechanisms that would facilitate the conversion of annotated corpora into structured linguistic knowledge.

References

[1] Gelumbeckaitė, J., Šinkūnas, M., & Zinkevičius, V. (2012). Old Lithuanian Reference Corpus (SLIEKKAS) and Automated Grammatical Annotation. Journal for Language Technology and Computational Linguistics (JLCL), 27(2), 83–96. https://doi.org/10.21248/jlcl.27.2012.165

[2] Jakulytė, D. (2004). Knygos nobažnystės (1653) morfologijos tyrimo galimybės naudojant reliacinę duomenų bazę [Possibilities of studying the morphology of "Knyga nobažnystės" (1653) using a relational database]. Acta Linguistica Lithuanica, 51, 1–14. https://journals.lki.lt/actalinguisticalithuanica/article/view/1126/1206

[3] Jakulytė, D., & Pakalniškienė, D. (2018). Senųjų raštų morfologijos duomenų bazės taikymas semantikos tyrimams: veiksmažodžių semantika [Application of the morphological database of Old Writings to semantics research: Verb semantics]. Vilnius University Open Series, 1, 373–388. https://doi.org/10.15388/Proceedings.2018.15

Evaluating GAN-Based Augmentation Across Deep Learning Architectures for Endoscopic Image Classification

Laimonas Janutėnas, Dmitrij Šešok

Vilnius Gediminas Technical University

laimonas.janutenas@vilniustech.lt

Data scarcity remains a major challenge in medical image classification, where obtaining sufficiently large and accurately annotated datasets is often difficult and costly. Generative Adversarial Networks (GANs) provide a potential solution by generating synthetic images that can be incorporated into training datasets. This study investigates whether the addition of GAN-generated synthetic images improves endoscopic image classification performance across different deep learning architectures. Experiments are conducted using the Kvasir v2 dataset, consisting of 8000 endoscopic images distributed equally across eight gastrointestinal classes. Synthetic images are generated using an enhanced StarGAN-based augmentation approach and incorporated into the training dataset. Five deep learning architectures - VGG-16, ResNet-50, DenseNet-121, InceptionNet-V3, and EfficientNet-B7 - are evaluated. The results are compared with previously reported baseline accuracies for the corresponding architectures. GAN-based augmentation results in improved classification accuracy across all five evaluated models, with gains ranging from 0.298 to 0.704 percentage points. The largest improvement is observed for VGG-16, while EfficientNet-B7 achieves the highest overall classification accuracy of 95.25%. The magnitude of improvement varies between architectures, indicating that the effectiveness of GAN-based augmentation is architecture-dependent. Overall, the results demonstrate that GAN-generated synthetic images can improve endoscopic image classification performance across diverse deep learning architectures and support the use of generative augmentation for medical imaging tasks with limited training data.

Weak and Strong Tie Dynamics: A Continuous-Time Relational State Approach

Rūta Juozaitienė

Vytautas Magnus University

r.uzupyte@gmail.com

Researchers across diverse domains increasingly study systems as dynamic networks, with nodes connected by relationships that evolve over time. Many established modelling approaches, however, represent ties as either present or absent. Such binary representations overlook the substantively distinct states that characterize many real-world relationships. For example, social ties may develop from acquaintanceship to friendship or close friendship, with each state reflecting a relatively stable relational condition rather than an isolated interaction. Reducing these states to a binary indicator obscures how relationships form, strengthen, weaken, and dissolve. We introduce a continuous-time framework for modelling dynamic networks in which each directed dyad occupies one of several relational states and evolves through transitions between them. Building on the relational event modelling perspective, we treat state transitions as events and specify transition-specific intensities that depend on exogenous covariates and state-dependent endogenous network statistics. This formulation allows mechanisms such as reciprocity, degree-related tendencies, and triadic closure to operate differently across relational states and transition types. To enable inference from discrete panel observations, we develop a data-augmentation approach that accounts for unobserved intermediate transitions and the resulting changes in network structure. A simulation study evaluates the recovery of transition-specific mechanisms, and an application to longitudinal adolescent friendship data examines the dynamics of absent, weak, and strong ties. The framework accommodates both ordered and unordered relational states, providing a flexible approach to studying network dynamics while retaining the substantive distinctions among relationships.

Multi-Level Reinforcement Learning for Dynamic Assortment Control in a Cosmetics Retail Network

Robertas Jurkus, Vitalij Denisov, Darius Drungilas, Mindaugas Kurmis

Klaipeda University

robertas.jurkus@ku.lt

Retail assortment optimization is difficult in large store networks because stores differ in demand, capacity, and product performance. A single global model may not capture these local differences, while training independent store-specific models from scratch can require substantial data and computational resources. Training a single model across all stores and items would also create a very large state and action space, making the learning process computationally expensive. This study evaluates a multi-level reinforcement learning approach for dynamic assortment adaptation. A Proximal Policy Optimization (PPO) policy is first trained at the global level, then adapted to store-size groups, and finally fine-tuned for individual stores. The method is evaluated using anonymized commercial retail data containing approximately 35,000 stock-keeping units (SKUs) and 26 stores. The performance of global, group-level, and store-level policies is compared using historical profit recovery under a counterfactual replay setting. Recovery increased from 55.53% for the global policy to 63.73% at the group level and 93.94% for store-specific policies across the evaluated store-size groups. Additional experiments were conducted for the large-store group and its four stores using repeated PPO training runs with different random seeds. The same global-to-group-to-store performance pattern was observed across the repeated runs, while store-level results showed low variability, with an average recovery of 94.9% ± 0.5%. The results indicate that multi-level reinforcement learning can combine the scalability of shared policies with the benefits of store-specific adaptation for large-scale retail assortment control.

Acknowledgments. This work was supported by the R&D project "Developing a coherent system for promoting innovative activities" (No. 02-020-K-0040), implemented by UAB "Inherent Simplicity Baltic" in partnership with Klaipeda University (KU) during 2024-2026.

A Privacy-Preserving Pipeline for Dialect-Aware Transcription and Evaluation of Admissions Interviews

Diana Kalibatienė, Julius Jancevičius, Kęstutis Normantas, Algirdas Laukaitis, Mindaugas Jankauskas, Dovilė Jodenytė

Vilnius Gediminas Technical University

julius.jancevicius@vilniustech.lt

University admissions processes, especially for foreign students, increasingly incorporate asynchronous video interviews to evaluate applicants at scale. However, manual review and transcription are still labour-intensive processes that are prone to errors, especially when interviewees exhibit a variety of linguistic and dialectal characteristics. Therefore, reliable automated transcription is essential for enabling scalable and equitable analysis of interviews in higher education. This study presents a reproducible, privacy-preserving pipeline for processing and evaluating authentic university admissions interviews. This pipeline encompasses secure data preparation, automated detection of personally identifiable information (PII) with human verification, dialect identification and a dialect-conditioned assessment of automatic speech recognition (ASR) performance. Dialect labels are inferred using an ECAPA-TDNN accent identification model from SpeechBrain, which was trained on the CommonAccent corpus. This enables the systematic grouping of interview recordings according to dialectal variation. The quality of the transcriptions is evaluated using the Word Error Rate (WER) across three transformer-based automatic speech recognition (ASR) architectures: Whisper, HuBERT and Wav2vec 2.0. Experiments are conducted on a real-world dataset of applicant interviews, and results are reported for dialect groups containing more than fifteen recordings, ensuring meaningful comparative analysis. The findings of comparing Whisper, HuBERT, and wav2vec2 models, reveal significant differences in their performance and determining dialect groups. Whisper achieves the lowest WER consistently, ranging from 20% to 27%. In contrast, HuBERT records error rates between 31% and 46%, and wav2vec 2.0 between 44% and 64%. The results also show

that transcription accuracy is strongly affected by dialect variation. This highlights the limitations of evaluating models based solely on aggregate performance measures. The study introduces a dialect-conditioned evaluation framework to provide a more nuanced assessment of ASR robustness and identify model-specific sensitivities to linguistic diversity. The proposed approach promotes transparency, privacy protection and fairness in automated interview processing, thereby contributing to the development of trustworthy AI solutions for educational applications. Furthermore, enhanced transcription quality can facilitate downstream analytics, such as applicant evaluation and the identification of early indicators of student support requirements and potential dropout risks. The results emphasise the importance of selecting dialect-aware models in high-stakes educational settings, and show that incorporating dialect-specific evaluation criteria can lead to more reliable and equitable ASR deployment in university admissions workflows.

Benchmarking Large Language Models in Lithuanian

Danguolė Kalinauskaitė

Vytautas Magnus University

danguole.kalinauskaite@vdu.lt

Benchmarks play an important role in evaluating the capabilities of large language models (LLMs). These standardized tests enable researchers to assess various aspects of a model's language proficiency, such as text comprehension, reasoning ability, and generalization skills. While such benchmarks have provided valuable insights into the capabilities of LLMs in high-resource languages, similar comprehensive evaluation frameworks are lacking for many low-resource languages, including Lithuanian.

A common approach to creating LLM benchmarks for languages with small speaker populations or low resource availability is through manual or automatic translation. However, translated benchmarks introduce two major issues: translation noise and reduced cultural relevance. Even with high-quality human translations, benchmarks may still fall short of providing reliable evaluations unless they are properly localized. Research has shown that localized benchmarks that are adapted to the culture and context of the target language correlate better with human preferences than those that are only translated.

Examples of benchmarks for other languages and the experience of their development highlight the need for a representative benchmark for the Lithuanian language. Our work addresses this gap by aiming to create a comprehensive set of benchmarks for evaluating LLMs on Lithuanian-language tasks, covering subjects across different areas. To our knowledge, this is the first systematic effort to develop a toolkit for measuring LLMs' multitask accuracy for the Lithuanian language. Moreover, this work emphasizes a broader challenge common to small, morphologically rich, low-resource languages: how to evaluate and improve LLMs when data and resources are limited. In our presentation, we delve into a deeper conceptual framework of this evaluation.

The core research approach is the construction of a modular benchmark, where each benchmark targets a different task, testing

different abilities of language models. The methodology integrates approaches from NLP, linguistics, education science, and AI ethics. Such resources will serve as universal comparison instruments for assessing LLM reliability in areas relevant to Lithuania, identifying performance gaps, and guiding targeted model development. In addition, the research will produce a comparative analysis of different LLMs compared on the tasks from our benchmark. The outcomes of this work are expected to enable deeper and more systematic research into LLMs' ability to understand the Lithuanian language and national context.

Acknowledgements. We thank Prof. Tomas Krilavičius and Assoc. Prof. Andrius Utka for their valuable ideas, insights, and fruitful discussions, which greatly enrich this research. We also thank Milita Songailaitė for her technological support.

Mutual Knowledge Distillation Between Heterogeneous Architectures for Continuous Sign Language Recognition

Michał Kalinowski, Bożena Kostek

Multimedia Systems Department and Audio Acoustics Laboratory
Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology, Poland

bozena.kostek@pg.edu.pl

Continuous sign language recognition (CSLR) remains a challenging task, partly because sufficiently large and diverse annotated datasets are scarce. This scarcity makes training new architectures from scratch difficult and motivates methods that can exploit knowledge already captured by well-performing models. In this work, we investigate mutual knowledge distillation between two structurally different CSLR architectures, CorrNet and SlowFastSign, selected for their strong performance on the PHOENIX14 dataset. The underlying hypothesis is that heterogeneous architectures learn complementary representations from the same visual input and can therefore improve each other through bidirectional knowledge transfer.

A common training framework was developed in which CorrNet and SlowFastSign are optimized simultaneously. In addition to their original task-specific objectives, each model receives a distillation signal derived from the other network's output probability distribution. This enables soft predictions from one architecture to guide the learning of the other without requiring identical internal representations. The study also considers Transformer-based architectures as an alternative architectural family; however, in our experiments, they did not outperform the selected CorrNet and SlowFastSign baselines on the dataset under investigation.

Particular attention was paid to training stability, which proved essential for a reliable assessment of mutual distillation. Batch-normalization parameters were frozen after loading the pretrained model weights, as batch-size-one training otherwise caused progressive degradation of the stored statistics. Gradient accumulation was used to

achieve a larger effective batch size and reduce update variance, while independent gradient scaling was applied to the two networks to control numerical stability separately. Finally, the distillation loss was normalized by sequence length and batch size so that its magnitude remained comparable to the remaining loss components. These measures provide a stable framework for evaluating whether bidirectional knowledge transfer between heterogeneous CSLR models can improve recognition performance without requiring additional annotated training data.

Separating Clinical Reasoning from Termination Authority in Agentic LLMs

Lina Kankevičienė, Julija Daranda

Kauno kolegija Higher Education Institution

lina.kankeviciene@kaunokolegija.lt

Agentic large language models (LLMs) used in clinical discourse must determine the appropriate questions to ask and the optimal time to stop asking. Most evaluations focus on diagnostic accuracy. However, researchers have placed less emphasis on assessing whether these models can reliably determine when they have obtained sufficient information. This study investigates whether an LLM's self-assessment of information sufficiency can safely stop data collection compared with an explicit evidence-coverage criterion. CLARITY constitutes a pipeline involving a History–Safety–Coordinator–Critic framework. At each stage, the coordinator elects to ASK, ASSESS, or ESCALATE. We conducted comparisons among prompted stopping, post hoc self-gating, and deterministic authorization. We derived the necessary evidence from author-designed, domain-specific red-flag checklists that have not undergone independent validation. The Safety agent possessed the complete checklist and supplied the coordinator with a list of potentially missing items estimated by a large language model (LLM). Conversely, the deterministic approach employed programmatically verified coverage. Coverage-violating termination (CVT) was defined as terminating the process while at least one required item remained unaddressed. CVT rates encompassed all valid evaluations within each method and information level, including escalations. Urgency accuracy was assessed based on concordance with rule-generated reference labels on a scale from 1 to 3, where 3 denoted the highest urgency. The escalation received an urgency rating of 3. We evaluated 300 synthetic acute-care cases across six complaint categories and four levels of information availability using qwen3.5:9b. Further testing was also performed with gemma4:e4b. Prompted CLARITY produced CVTs in 15.0% of cases at 25% information availability and 48.0% at 75%. The model's final sufficiency score demonstrated only a partial capacity to distinguish

complete from incomplete evidence states. The AUROC decreased from 0.82 at 25% availability to 0.70 at 50% and 0.62 at 75%. In a post hoc terminal analysis, a sufficiency threshold of 0.7 modestly reduced CVT to 12.7%–45.7%. At a threshold of 0.95, CVT declined to 1.7%–3.0%, while conclusion rates also diminished to 6.3%–20.7%. Following correction of a terminal-transition bypass, we reran the deterministic implementation across 1,200 evaluations. It yielded zero CVTs, consistent with the rule, while concluding 72.7%–98.3% of cases with degraded information. Escalation reached 27.3% at 25% information availability. Regarding Qwen, urgency accuracy remained above the single-pass baseline at information availability of 50% or lower. With Gemma, deterministic authorization caused a 10-percentage-point drop in urgency accuracy compared to prompted CLARITY when only 25% information was available (95% confidence interval, −19 to −1). The model's sufficiency score provided valuable information; however, as a terminal filter, it did not align with the deterministic policy's combination of low CVT and high conclusion rates. This analysis did not evaluate a real-time self-gating policy capable of ongoing questioning. Explicit coverage rules can establish a clear, verifiable stopping criterion; however, they may also increase escalation and affect overall task performance. These findings demonstrate adherence to specified criteria rather than clinical safety. Clinical application of this model would require independent validation of both the evidence requirements and the reference labels.

Unsupervised Learning of Player Roles and Lineup Archetypes in EuroLeague Basketball

Gabrielė Kasputytė, Tomas Krilavičius

Vytautas Magnus University

gabriele.kasputyte@vdu.lt

Basketball player evaluation is shifting from traditional position-based classifications to data-driven methods for identifying player functionalities. As players become more versatile and teams use positionless tactics, standard positional labels no longer fully describe players on the court. Although several studies have investigated player role identification, important limitations remain. Many approaches rely on box-score statistics, which may miss possession-level impact and contextual tactical information. Existing work also tends to emphasize offensive roles, while defensive archetypes are less frequently analyzed. In addition, most studies focus on the National Basketball Association, whereas European basketball has received less attention despite its different tactical systems, pace, player utilization, and game structure. To address these limitations, our study proposes a multi-stage probabilistic framework for EuroLeague basketball analytics. First, we build a player-role model that describes functional responsibilities, using possession-adjusted box-score statistics, play-type and shooting features, defensive indicators, and a regularized adjusted Plus-Minus estimate. We use Gaussian Mixture Models to identify player roles and retain role probabilities, so a player can be represented as a hybrid profile rather than a single category. Moreover, we use these roles as inputs for lineup analysis. For each EuroLeague lineup, we construct role-composition variables such as the number of creators, shooters, rim finishers, rim protectors, on-ball defenders, and stretch bigs, along with lineup-level indicators of spacing, creation, size, and defensive structure. PCA is used to analyze the main structural dimensions of the lineup space, while clustering methods compare alternative lineup archetypes. Preliminary results suggest that role-based lineup representations produce more stable and interpretable archetypes than using the full feature set alone. This provides a practical link between player-role modeling and lineup-level decision-making in European basketball analytics.

MORPHO-NN: Adaptive Growth and Pruning for Information Structure Recovery in Multichannel Signals

Giedrė Keršulytė, Artūras Serackis

Vilnius Gediminas Technical University

arturas.serackis@vilniustech.lt

Efficient neural network architectures are essential for deploying artificial intelligence in systems with limited memory, computation, and energy resources. Fixed dense architectures may contain redundant filters and connections, while reducing model size alone does not ensure that the resulting topology reflects the informative structure of the data. This study investigates whether MORPHO-NN, a dynamic architecture formation strategy, can grow a compact and accurate convolutional neural network from a small initial structure while recovering informative channels and cross-channel interactions. Synthetic one-second, 12-channel signal windows were generated at a sampling rate of 256 Hz. Three binary information components defined an eight-class classification task: the amplitude of a 3–6 Hz sinusoid, the strength of a 40–60 Hz burst, and the correlation between a channel pair within the 14–20 Hz band. In the S1 condition, information was concentrated in channels 0–2, while the remaining channels contained noise. In S1_reloc, the same signals were relocated to channels 7–9 through a channel permutation. This controlled setup provided known reference structures for assessing both classification performance and structural recovery. MORPHO-NN was trained directly on standardized signals, starting with two convolutional layers containing four active filters each and allowing up to 32 filters per layer. During training, important filters were split, inactive connections were activated using gradient information, and low-importance connections and filters were removed using first-order Taylor estimates. Comparisons included a dense CNN, magnitude pruning, RigL-like, GradMax, and NORTH-Select implementations, and scheduled Net2Wider expansion. Each method–condition combination was evaluated over 10 random seeds and 40 training epochs; variability

is reported as standard deviation across seeds. Evaluation covered classification accuracy, macro F1, active parameter counts, theoretical multiply–accumulate operations (MACs), channel-recovery ROC-AUC, the fraction of connections assigned to noise channels, and the fraction of filters connected to both members of an informative channel pair. Paired accuracy comparisons used Wilcoxon tests with Holm correction. In S1, MORPHO-NN achieved 92.62 ± 0.87% accuracy, compared with 90.28 ± 0.63% for the dense CNN. It significantly outperformed the dense model, GradMax, and scheduled Net2Wider expansion (adjusted $p = 0.0156$). MORPHO-NN used an average of 3,823 active parameters and 0.532 million MACs, approximately 29% and 26% of the dense model's respective budgets. RigL-like and NORTH-Select achieved 92.05 ± 1.37% and 91.52 ± 0.80%; their differences from MORPHO-NN were not significant after correction. Magnitude pruning achieved the highest accuracy, 93.67 ± 0.66%, at the same connection budget. MORPHO-NN achieved a connection-based channel-recovery AUC of 1.00 in both conditions. In S1, only 13.3% of first-layer connections were assigned to noise channels, compared with 35.0% for magnitude pruning and 70.8% for RigL-like. The corresponding informative-pair connectivity fractions were 0.85, 0.37, and 0.16. Following information relocation, MORPHO-NN achieved 91.17 ± 4.32% accuracy, indicating reduced stability. Magnitude pruning reached 93.64 ± 0.62% and outperformed MORPHO-NN in all 10 runs (adjusted $p = 0.0156$). These findings indicate that MORPHO-NN can form compact architectures that selectively allocate connections to informative channels and channel pairs. In S1, it improved accuracy over a dense network while substantially reducing theoretical computational requirements. However, superiority over budget-comparable magnitude pruning was not established, and relocation exposed training instability. Further work should address the interaction between structural adaptation and weight optimization and evaluate generalization to broader signal-processing tasks.

SAR-to-Optical Image Translation Using Wavelet Based GAN Networks

Kürşat Kömürcü[1,2], Vytautas Paura[2], Valentas Gružauskas[2], Linas Petkevičius[2]

[1] IRISA Université Bretagne Sud
[2] Vilnius University

vytautas.paura@mif.vu.lt

In this article, we propose a modified generative adversarial network (GAN) for synthetic aperture radar (SAR)-to-optical image translation that uses a single-pass generator based on Haar-wavelets and Fourier phase locking. Preserving Fourier phase is often assumed to help image translation because phase is said to encode structure. For SAR-to-optical translation that assumption needs qualification. To test the assumption a collection of Sentinel-1/2 imagery over 150 areas of the Baltic states (consisting of more than 30,000 SAR and 120,000 optical images) was gathered to create training and testing data. The amplitude-weighted phase agreement between the SAR input and the aligned optical image is on average negative: dominant scene structures are in anti-phase, since C-band backscatter and visible reflectance have opposite contrast polarity: forests are bright in SAR and dark in optics, and SAR phase additionally carries speckle and layover. We therefore treat phase preservation as an inductive bias on intermediate features rather than a constraint on outputs. The newly created GAN network with the single-pass generator has about 63% fewer parameters than the baselines (3.5M vs 7.9 - 13.1M parameters compared to existing models). Additionally, a layer of spectral transform blocks rewrite the amplitude spectrum of bottleneck features freely while holding their phase to within ~20° per frequency. Two findings follow. First, the trained blocks keep the phase nearly fixed in any case, and zeroing them at inference moves no metric by more than ~3%, identifying amplitude modulation under an (almost) fixed phase as the effective mechanism. Second, removing the blocks degrades realism (FID up ~61%) and edge fidelity (GSSIM from 20.4% to 15.8%) while slightly lowering pixel error through darker, flatter images. GAN still attains the closest match to the aligned reference in mean

absolute error (13.4% of full intensity) among the compared methods. We further show why no single metric ranks these methods: the best-FID model is not the most faithful, two baselines degenerate near the untranslated input, and a random, unrelated optical image already reaches 14.4% mean absolute error and 30.7% SSIM. Our method, when tested on five standard unpaired benchmarks evaluated on their test splits, preserves layout (e.g., FID 64.0 for map-to-satellite) but is weak where objects must change shape. This new proposed method greatly reduces inference times compared to baseline models. On one NVIDIA V100 GPU it translates a 256 by 256 image in 16.7 ms. because generators built only from residual convolution blocks are faster (9-10 ms), whereas other tested methods are up to 2.8 times slower.

Acknowledgements. This project has received funding from the Research Council of Lithuania (LMTLT), agreement No S-ITP-25-3.

Computational Methods for Investigating the Microscopic Mechanisms Underlying the Impact of Microplastics on Public Health

Jakub Kraus, Piotr Mika, Mateusz Wiśniewski, Paulina Trybek, Zbigniew Dendzik

Institute of Physics, Faculty of Science and Technology
University of Silesia in Katowice, Poland

paulina.trybek@us.edu.pl

Nanoplastics, ubiquitous environmental contaminants arising from the fragmentation of plastic waste, pose a growing potential threat to biological systems and public health. Their nanoscale dimensions enable interactions with subcellular structures. This study explores the use of molecular dynamics simulations to elucidate the microscopic mechanisms by which nanoplastics in aqueous environments may alter the structure and properties of biomolecules, particularly proteins and lipid bilayers. The proposed computational approach can be used to investigate whether hydrophobic nanoplastics promote peptide adsorption through templating effects, destabilize α-helical structures in favor of β-sheet-like conformations, facilitate β-hairpin unfolding, or partition into hydrophobic membrane cores, thereby affecting bilayer thickness, lipid diffusion, and molecular conformations. Such molecular changes may be relevant to protein misfolding and aggregation processes implicated in disorders including prion diseases and Alzheimer's disease. By complementing epidemiological observations and evidence of environmental exposure, including bioaccumulation along aquatic food chains, this framework could help identify mechanisms through which chronic, low-dose nanoplastic exposure influences protein-folding landscapes and cellular integrity. These simulations offer a route toward mechanistic hypotheses that can subsequently be tested experimentally.

Single-Reference Identity Preservation for Dogs Through Large-Scale Adapter Training

Tomas Krilavičius, Justina Mandravickaitė, Monika Katilienė, Tadas Žiemys, Hande Safiye Erdal, Eimantas Zaranka

Vytautas Magnus University
Dogo App UAB

monika.katiliene@vdu.lt

Personalised image generation for animals presents a harder problem than the equivalent task for human faces. No large annotated corpus of individual animals exists, no dedicated identity encoder is available, and the visual cues that separate one animal from another lie in coat pattern and body proportion rather than in the facial geometry that face-specific methods exploit. This work examines whether an image-conditioned adapter trained at scale on pet photographs can preserve individual identity from a single reference image, without per-subject fine-tuning and without an identity encoder. Breed labels are present for part of the training corpus but are not required at inference, where the reference photograph is the only input describing the animal. A corpus of 1,255,805 photographs collected through a commercial pet application was reduced to a training set of 601,360 images. Of these, 653,155 met a minimum resolution of 1024 by 1024 pixels and entered the pipeline. Subjects were located with a detection model and a second detector applied to the residual failures, segmented with a boundary-refinement network, and filtered through connected-component analysis of the resulting masks, detection-based validation of the segmented output, and perceptual hashing for duplicate removal. Captions were generated by a vision-language model selected from seven candidates. Three background treatments were compared under otherwise identical conditions, and complete background removal was found to outperform both blurring and no modification. Adapter training on Stable Diffusion XL failed to preserve coat colour or individual features. Balanced subsets

constructed from colour labels, alternative caption sources, reduced effective batch size and initialisation from the released ViT-H adapter rather than from scratch all left this failure unchanged, establishing that it was not attributable to data composition. The failure was resolved by a variant employing a resampler with sixteen conditioning tokens, trained on the full corpus at 512 by 512 pixels resolution on Stable Diffusion 1.5. Thirty-two tokens brought no further improvement. The selected adapter was then combined with components it had never been trained alongside. Substituting the base checkpoint at inference across twelve community derivatives of Stable Diffusion 1.5 changed rendering style while leaving identity intact, provided the architecture was unchanged. Textual inversion embeddings gave no useful improvement. LoRA and ControlNet both combined well with the identity adapter without interfering with subject preservation, though neither is required for it. Quantitative evaluation follows the DreamBooth protocol. Eighty references were drawn from eight breeds, four among the most numerous categories in the corpus at roughly twenty-one thousand images each and four among the least numerous at around three hundred. At an adapter scale of 0.5 the system reached a DINO score of 0.8607, a CLIP-I of 0.9231 and a CLIP-T of 0.3608. Breeds represented by twenty-one thousand images scored no higher than those represented by three hundred, the difference between the two groups being smaller than the variation within either. Both subject fidelity figures are raised by the uniform dark background and identical framing shared by reference and output. Varying the adapter scale across three settings shows that raising it past a certain point stops being worthwhile. Moving from 0.5 to 0.7 improved subject fidelity by 0.0226 on DINO and cost 0.0530 of prompt fidelity. Moving from 0.7 to 0.9 improved subject fidelity by only 0.0140 while costing 0.0469, almost the same as before. The identity gains shrink while the prompt losses stay the same, so 0.7 is the point beyond which the exchange no longer pays. The benefit is also uneven across references. Those whose identity was preserved least well at 0.5 improved by up to 0.0542 across the full range, while the best-preserved improved by 0.0219. A higher scale rescues the references the adapter struggles with and adds almost nothing where it already performs well. Taken together the results show that identity preservation for animals

is achievable at production quality from a single photograph, without per-subject training, provided the conditioning carries enough capacity to represent the subject and the corpus is large enough to cover the variation within it.

Acknowledgements. These results were obtained as part of the project "Development of Innovative Products at Dogo App UAB" (Project No. 02-048-K-0006), co-funded by the European Union through the European Regional Development Fund.

A Tour of the Separate Islands of Model-Based Optimization

Jakub Kůdela

Brno University of Technology, Czech Republic

Jakub.Kudela@vutbr.cz

Expensive black-box optimization has produced several mature research traditions, including DIRECT-type partitioning methods, deterministic response-surface optimization, model-based derivative-free optimization, Bayesian optimization, and surrogate-assisted evolutionary algorithms. Although these approaches address closely related problems, they have developed largely within separate communities, with distinct terminology, software ecosystems, benchmark suites, and experimental conventions. This talk reviews their common principles, methodological differences, and current research directions, then examines the surprisingly limited evidence from direct cross-family comparisons.

Democratizing In-Database Machine Learning at Scale: BigQuery ML Adoption at Vinted

Gytis Kulaitis

Vinted

gytis.kulaitis@vinted.com

Recent industry trends in artificial intelligence have heavily focused on increasingly complex, compute-intensive architectures. While modern tabular foundation models like TabPFN have shown remarkable zero-shot performance on smaller datasets [1], for large-scale production workloads and routine predictive tasks, classical machine learning methods—such as gradient boosted decision trees (XGBoost), linear models, and time-series decomposition—often remain the most computationally efficient and robust solutions [2, 3]. Traditionally, operationalizing these classical models requires substantial MLOps infrastructure: exporting data from analytical warehouses into dedicated computing clusters, writing custom Python pipelines, containerizing environments, and maintaining separate orchestration stacks. This workflow introduces heavy developer friction, slows time-to-production, and poses a steep technical barrier for SQL-focused data professionals.
To mitigate these architectural overheads, in-database machine learning paradigms have emerged, bringing model training and inference directly to where the data resides [4]. In this talk, we present the practical adoption of Google Cloud's BigQuery ML [5] within Vinted, Europe's leading online C2C marketplace. BigQuery ML enables the declarative specification of model training, hyperparameter tuning, and evaluation directly via standard SQL syntax, encompassing standard algorithms (e.g., XGBoost, logistic regression, k-means, ARIMA+) as well as automated machine learning (AutoML) capabilities [6].

We describe the end-to-end integration of BigQuery ML into Vinted's production data stack. By coupling declarative SQL model statements with the data build tool (dbt), we established a version-controlled, reproducible workflow where feature transformations, model retraining,

and batch inference are governed under the same CI/CD pipelines and lineage graphs as core analytics marts. We discuss real-world use cases—ranging from buyer gross merchandise value forecasting to fraud pattern similarity—and compare them against equivalent custom Vertex AI pipelines. Our findings demonstrate that in-database models can be prototyped and deployed into batch production in a fraction of the time while achieving comparable predictive performance. Finally, we present how decoupling model training from serving in Vinted's infrastructure allows BigQuery ML artifacts to be deployed as real-time inference services with minimal engineering overhead.

References

[1] N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter, "TabPFN: A transformer that solves small tabular classification problems in a second," in *International Conference on Learning Representations (ICLR)*, 2023.

[2] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 507–520, 2022.

[3] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep neural networks and tabular data: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 6, pp. 7494–7510, 2024.

[4] J. M. Hellerstein et al., "The MADlib analytics library: or MAD skills, the SQL," *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 1700–1711, 2012.

[5] Google Cloud, "BigQuery ML overview and reference architecture," *Google Cloud Documentation*, 2024. [Online] Available: https://cloud.google.com/bigquery/docs/bqml-introduction.

[6] X. He, K. Zhao, and X. Chu, "AutoML: A survey of the state-of-the-art," *Knowledge-Based Systems*, vol. 212, p. 106622, 2021.

Neuro-Fuzzy Prediction of Decision Making in Cyber Defense Exercises

Karina Kulikauskaitė

Vilnius Gediminas Technical University

karina.kulikauskaite@vilniustech.lt

Decision-making in cyber defense exercises (CDX) is shaped by a complex interplay of technical, emotional, and collaborative human factors operating under uncertainty and time pressure. This study proposes a human-centred Adaptive Neuro-Fuzzy Inference System approach for predicting Counterfactual Decision Reflection (CDR) outcomes among CDX participants. A dataset capturing technical, emotional, and teamwork-related dimensions was collected during the international cyber defense exercise Amber Mist 2024 and analyzed using ANFIS-based prediction to examine the relationship between human factors and CDR. Results indicate that teamwork, communication, and collaboration exert a stronger influence on decision stability than isolated technical competencies. Gaussian, generalized bell, trapezoidal and triangular membership functions were analyzed used to evaluate prediction errors. To enhance model robustness, six different synthetic data augmentation techniques were applied. ANFIS models trained on augmented data achieved substantially improved predictive performance, with significantly reduced testing error. These findings suggest that neuro-fuzzy modeling can offer interpretable, human-centric insight into decision-making processes that purely technical performance metrics fail to capture. The proposed framework provides cyber defense trainers with a data-driven tool for identifying non-technical factors that undermine decision stability under operational stress. Future work will focus on validating the model for real-time applications in adaptive training systems.

A Lithuanian Image-Caption Dataset: Combining News Media and Translated Benchmarks

Tautvydas Kvietkauskas, Ūla Šveiterytė

Vilnius Gediminas Technical University

tautvydas.kvietkauskas@vilniustech.lt

Vision-Language Models (VLMs) are widely used in image captioning tasks, but their development for less common languages is limited by a lack of high-quality multimodal datasets. This poster presents a hybrid approach for creating Lithuanian image-caption datasets, combining two different data sources: articles from Lithuanian news portals with associated photographs, and international image-caption datasets: NoCaps, VizWiz, MS COCO, TextCaps, Flickr30k, SBU, DOCCI and Winoground. Data from the news portal underwent multi-stage filtering; articles, images, and text were cleaned and standardized. Following this processing, a dataset of Lithuanian news portals was created, comprising 272 964 articles and 364 122 images. Descriptions of international datasets were automatically translated into Lithuanian and evaluated using a quality control methodology based on English and Lithuanian evaluation metrics. English descriptions were assessed according to indicators of fluency, grammar, spelling, relevance, and semantic anomalies, using GPT-2, LanguageTool and pyspellchecker tools, whilst Lithuanian translations were evaluated using metrics based on LaBSE, Sentence-BERT MPNet and Hunspell. Based on these indicators, a composite quality score of 0.6 was calculated, enabling the automatic identification and removal of low-quality image-caption pairs and semantically inaccurate translations. After cleaning and filtering, a final bilingual dataset was formed from all the international sets, consisting of 1 162 252 unique images and 2 009 222 captions. By applying the proposed methodology, high-quality Lithuanian multimodal datasets were prepared, which can be used for the training, adaptation, and evaluation of vision-language models in the context of the Lithuanian language.

Disentangling Sources of Variability in Sensor-Channel Importance Rankings for Human Activity Recognition

Maksym Lapin, Dalius Navakauskas

Vilnius Gediminas Technical University

maksym.lapin@vilniustech.lt

Studies of human activity recognition with wearable inertial sensors often end by naming the sensor channel that matters most, and that name guides sensor selection and device design. Yet the ranking also depends on the random seed, the architecture, and the operator that defines what it means to remove a channel. These choices are typically fixed and rarely reported, leaving it unclear how much of the result comes from the data rather than the setup. To quantify this, disagreement between networks that differ only in the random seed serves as the baseline for all other choices. A total of 49 680 small gated recurrent networks, within a wearable budget of 16 384 trainable parameters, were trained on UCI-HAR and PAMAP2: 69 architectures, 30 seeds, and 12 splits per dataset, each assigning whole participants to training, validation, and test (16/5/9 on UCI-HAR, 5/1/2 on PAMAP2). Rankings came from gradient × input and channel occlusion under four removal operators: physical zero, the training set mean, the within-window mean, and an unrelated recorded window, plus a random control. They were compared with Kendall τb and assessed through deletion curves under each operator. Reseeding sets the floor, with $(1 - \tau_b)$ of 0.18 to 0.40 on UCI-HAR and 0.33 to 0.41 on PAMAP2 across ranking methods, against about 1.0 for unrelated rankings. Changes in width and depth add only 2 to 5 percent of the floor, and the three operators that replace a channel with a fixed value or an unrelated window stay at or below it. The operator that replaces a channel with its own window mean, keeping its level but flattening its shape, is different. It disagrees with each of the other three by more than the floor, adding on average 1.9 floors on UCI-HAR and 0.3 on PAMAP2. In 83 of 96 cases (four operators, 24 splits), the ranking produced by an operator also scored highest under that same operator, so deletion

cannot identify a preferred ranking and faithfulness is not a property of a ranking alone. One fixed ranking (gradient × input) gains 1.9 times as much over a random control under one operator as under another on UCI-HAR, and 2.6 times as much on PAMAP2. Thus, a channel's importance is not determined by the data alone, but also depends on the operator through which it is measured

.

Building a Digital Twin of University Admissions from DreamApply Data

Algirdas Laukaitis, Diana Kalibatienė, Arturas Serackis, Kęstutis Normantas, Mindaugas Jankauskas, Julius Jancevičius, Dovilė Jodenytė

Vilnius Gediminas Technical University

algirdas.laukaitis@vilniustech.lt

Student admission is one of the most data intensive activities performed by universities. Every year, admission systems collect documents, video interviews, grades, communication records, and thousands of workflow events. Yet most AI solutions ignore this process context and operate as isolated prediction tools. In this presentation, we show how data from the DreamApply admission management platform can be used to reconstruct the real admission workflow and create a digital twin of the process. The framework combines process mining, multimodal AI agents, and interpretable decision models to analyze applicant documents, video interviews, and structured application records while keeping every processing step transparent and auditable. Rather than allowing AI to make admission decisions directly, we use AI to extract and organize evidence that is evaluated through transparent decision rules embedded within the reconstructed workflow. The framework is implemented as a collection of reproducible notebooks that provide a practical path from research prototypes to organizational systems. The project is fully open source and consists of several GitHub repositories covering process mining, document analysis, video analysis, workflow orchestration, and deployment components. A companion web site available at https://fair-vid.github.io/ connects the presentation with the project resources, notebooks, repositories, and implementation examples. The talk presents the architecture, the process mining approach, and the lessons learned from transforming operational DreamApply data into a transparent AI enabled digital twin for university admissions. It also demonstrates how organizations can introduce AI into high trust decision processes while preserving transparency, auditability, and human oversight.

Data Augmentation for Multi-Temporal Rooftop Solar Panel Segmentation in Lithuanian Orthophotos

Justinas Lekavičius, Valentas Gružauskas

Vilnius University

justinas.lekavicius@mif.vu.lt

Monitoring the growth of rooftop photovoltaic (PV) installations requires building-level information, which in Lithuania is not openly available. Open data report electricity-producing consumers only in aggregated 1 x 1 km statistical grid cells. Very high-resolution aerial orthophotos can help fill this gap, but deep learning segmentation of small rooftop PV objects faces two data problems. Labelled training data are scarce outside a few open datasets from other countries such as China or France, and orthophotos acquired in different years differ in illumination, shadow direction, season, and geometry. This work studies data augmentation as a way to address both problems, and applies the resulting segmentation model to detect new PV installations across acquisition years, as a first step towards a multimodal pipeline for grid-level solar assessment. The study uses the national orthophoto archive (0.25 m resolution, 2018-2023). Each annual acquisition covers only part of the country, so many locations are imaged two or more times at irregular intervals. The archive is filtered to grid cells from the Lithuanian Open Data Portal that contain electricity-producing consumers, and image patches from these cells were pseudo-labelled with a DeepLabV3-ResNet-50 model trained on open French and Chinese PV datasets. A manually reviewed subset forms a verified Lithuanian test set, while the remaining pseudo-labels serve as weak labels for model adaptation. Orthophotos of the same location from different years are first co-registered and radiometrically normalised (e.g., by histogram matching) to reduce pixel-level differences before segmentation. The segmentation model is then adapted to Lithuanian imagery under several augmentation strategies: classic geometric and photometric transformations, and synthetic image-mask pairs generated by a conditional generative adversarial network

(pix2pix). In earlier work on open datasets, pix2pix-based augmentation improved IoU by 2% over classic augmentation, and here the aim is to test whether generative augmentation also improves robustness to the illumination and seasonal differences between acquisition years. This robustness is measured on roofs that did not change between acquisitions, where a reliable model should produce the same mask in every year. Segmentation masks from consecutive acquisitions are then compared to detect newly installed panels, with filtering of false positives caused by roof displacement or seasonal changes, and the results are aggregated per grid cell. The poster will present the dataset construction, the effect of registration and normalisation on alignment error, a comparison of augmentation strategies in terms of segmentation accuracy and multi-year consistency, and change detection results with grid-level maps for selected regions of Lithuania. The work's aim is to contribute as an open multi-temporal PV dataset for Lithuania and a basis for multimodal remote sensing analysis, in which image-derived indicators are combined with meteorological and other in-situ data.

Acknowledgements. This project has received funding from the Research Council of Lithuania (LMTLT), agreement No S-ITP-25-3.

Graphemes, Stressed Graphemes, or Phonemes? Text Input Representations for Lithuanian VITS Speech Synthesis

Vytautas Lėveris, Gražina Korvel

Institute of Data Science and Digital Technologies
Vilnius University

grazina.korvel@mif.vu.lt

Lithuanian spelling is close to phonemic but leaves out lexical stress and pitch accent. Both are free and contrastive, and they shift across the inflected forms of a word (kasà 'braid' vs kãsa 'digs'). A neural text-to-speech model that reads plain letters must therefore infer stress from context. Earlier Lithuanian studies showed that marking stress on letters improves the output of Tacotron 2 and VITS. Phonemic transcriptions designed by linguists, however, have not been compared with letter inputs on a modern end-to-end model. The only VITS experiment with phonemes relied on eSpeak NG, whose Lithuanian rules are not maintained by phoneticians.

We train VITS from scratch on the neutral subset of the Lithuanian Emotional Speech Corpus (EMO), created in 2026 (66.5 h, 88 speakers). In our study, we compare four input representations: plain graphemes, graphemes with separate accent marks, a 59-phoneme inventory without stress or accent information, and a 92-symbol phonemic inventory encoding stress position, accent type, and vowel length within vowel symbols. All four representations are derived from the same transcripts through a consistent pipeline that includes rule-based stress assignment and grapheme-to-phoneme conversion, so the stressed variants share identical stress decisions. Additionally, multi-character phonemes are re-encoded as single symbols. The experiments use the same dataset, random seed, hyperparameters, and training duration, and differ only in the input alphabet.

The models are compared using mel-cepstral distortion (MCD), F0 error, and ASR-based character error rate (CER) on a selected subset of validation utterances, as well as through a pairwise listening test. The study shows whether explicit stress information or phonemic precision matters more for Lithuanian neural speech synthesis.

From Bits to Qubits: Bridging Classical Reliability and Fault-Tolerant Quantum Algorithms

Ugnė Liaubaitė

Institute of Data Science and Digital Technologies
Vilnius University

ugne.liaubaite@mif.vu.lt

Quantum computing introduces a fundamentally different computational paradigm, exploiting superposition and entanglement to perform tasks that can be challenging for classical algorithms. However, the sensitivity of quantum states to environmental noise and imperfect operations presents a major obstacle to scalable quantum computation. This talk explores the transition from classical to quantum information processing, with an emphasis on computational reliability. Starting from the principles of classical error-correcting codes, quantum error correction is introduced, including logical encoding, syndrome measurements, and the concept of fault-tolerance thresholds. The talk further discusses quantum error mitigation as a complementary approach to handling noise and examine how these techniques influence the feasibility and resource requirements of quantum algorithms. The talk highlights the interplay between algorithmic complexity, physical noise, and error suppression in the development of practical quantum computing.

Concept of a Cybersecurity Management Model for Higher Education Institutions

Tatjana Liogiene, Asta Danileviciute, Vaidas Valukonis

Vilniaus kolegija Higher Education Institution

t.liogiene@eif.viko.lt

International and national requirements governing the activities of higher education institutions (HEIs) drive the digitalization of teaching, research, international cooperation, and data management processes, as well as intensive information exchange. These processes create conditions for the emergence of various cybersecurity (CS) risk factors, making it essential to systematically ensure data confidentiality, integrity, and availability, as well as regulatory compliance. A literature review revealed a fragmented approach in CS research related to HEIs, with primary attention given to individual technological, organizational, or human aspects. Integrated organizational-level CS management models are addressed only to a limited extent, highlighting the need for a comprehensive model that combines the key CS domains and their interrelationships. To address this gap, a conceptual CS management model for HEIs was developed based on the NIST Cybersecurity Framework, ISO/IEC 27001, ISO/IEC 27002, and ISO/IEC 27005 standards, the NIS2 Directive, the General Data Protection Regulation, the Artificial Intelligence Act, and the Cyber Resilience Act. The model is structured into four interrelated blocks: (1) a strategic block encompassing CS governance, legal compliance, and strategic risk management; (2) an operational block focused on the implementation of security processes, incident management, and business continuity assurance; (3) a technological block covering IT infrastructure protection and the application of technical security controls; and (4) a human-factor block addressing competency development, security awareness enhancement, and the cultivation of a CS culture. The model identifies nine core components: legal compliance, competency development, risk management, supplier and third-party management, incident management, auditing and maturity control, IT infrastructure security, resilience, and CS governance. Components are assigned to one or more blocks according to their purpose and level of application, while some

components connect multiple blocks to ensure horizontal and vertical integration of CS management throughout the institution. The suitability of the model was evaluated through expert assessment. The evaluation confirmed the relevance, adequacy, and validity of the standards, regulations, and good practices used in the model for CS management in HEIs. The logical structure of the strategic, operational, technological, and human-factor blocks, the adequacy of the identified components, the consistency of their assignment to different blocks, and the model's ability to cover all essential CS domains relevant to HEIs were also confirmed.

Structuring Cybersecurity Hygiene Indicators: Automated Assessment Complexity, Data Privacy and Indicator Prioritisation

Faustas Litvinaitis[1], Simona Ramanauskaite[2]

[1] Fontys University of Applied Sciences, the Netherlands
[2] Vilnius Gediminas Technical University

faustas.litvinaitis@gmail.com

Cybersecurity hygiene becomes a mandatory skill. However, the problem of wide sciety edification is this area is complicated and can not be implemented in tradictional education - interactive tools are needed to guide the user during his or her work with computer. To implement such a system, clear indicators are needed. Those encompasses a wide range of user behaviours and endpoint conditions, but these practices differ substantially in how they can be observed and assessed automatically. While existing cybersecurity-hygiene research provides multidimensional descriptions of secure behaviour, it does not directly specify which indicators can be reliably derived from an end-user endpoint, what data would be required, or what privacy implications such monitoring would introduce. This study develops a structured framework for analysing cybersecurity-hygiene indicators from the perspective of automated endpoint assessment. Indicators derived from the Security Behavior Intentions Scale, its behavioural validation, and the Cyber Hygiene Inventory are classified according to their nature, observable evidence, assessment complexity, and privacy implications. A distinction is made between endpoint-environment indicators, which describe measurable device states or configurations, and behavioural indicators, where endpoint observations may provide only indirect evidence of user actions or intentions. The proposed framework evaluates each indicator along several dimensions: cybersecurity relevance, observability from an endpoint, technical assessment complexity, reliability of inference, required data, privacy sensitivity, and explainability. An initial prioritisation of indicators is obtained using multi-criteria decision-making techniques, providing preliminary importance weights for subsequent technical validation. The resulting framework provides a basis for selecting cybersecurity-hygiene indicators that balance security relevance, automated assessment feasibility and privacy.

Quantum Molecular Shape Similarity: A Variational Circuit for Three-Dimensional Ligand Comparison

Marco Marcozzi[1], Savíns Puertas-Martín[2], Ana Cámara-Artigas[2], Ernestas Filatovas[1], Juana L. Redondo[2]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] University of Almería, Spain

marco.marcozzi@mif.vu.lt

Ligand-based virtual screening ranks candidate molecules by their similarity to a known active compound, and three-dimensional shape is one of its main criteria. Quantum computing offers a different way of formulating this comparison, but current hardware limits it to small problems, so new methods must first be studied on simulators. We present Quantum Molecular Shape Similarity (QMSS), a proof of concept in which the comparison of two molecules is carried out by a variational quantum circuit. Each atom is assigned to one qubit. Atomic coordinates are mapped from the SO(3) spatial representation to SU(2) single-qubit rotations. A linear chain of two-qubit entangling gates correlates neighbouring positions in the atom-to-qubit ordering. The circuit follows a compute–uncompute construction inspired by quantum kernels: the query molecule is encoded, a shared variational transformation aligns it to the target, and the inverse target encoding drives the register back towards the reference state when the two shapes agree. The average Pauli-Z expectation value of this return signal is mapped to a bounded Tanimoto-style similarity score. The variational layer has only three parameters whatever the molecular size, so larger molecules widen the register without enlarging the optimization space. The parameters are found in two stages: quasi-random Sobol exploration followed by local COBYLA refinement. QMSS was evaluated on the 20 natural amino acids, which include six pairs of chemical analogues that are also mutual nearest neighbours in an independent descriptor space (the T-scale). Molecular flexibility is taken into account with conformer sets, and a

classical pre-filter limits the number of circuit evaluations. The task is to recover the analogue of each of the 12 paired amino acids among the other 19. A molecule compared with itself always reaches the maximum score. The analogue is ranked first for 7 of the 12 queries (58%, against about 5% for a random ranking) and appears among the five best candidates in all 12 (against about 26% at random). The runtime of the experiments is dominated by the quantum simulation and grows with an empirical exponent of 1.02 per qubit, in agreement with the O(2^n) cost of statevector simulation, while atom assignment and data preparation take only a few milliseconds. A single circuit evaluation takes 0.21 s at 13 qubits and 9.15 s at 27 qubits. This growth belongs to the simulator and not to the circuit, whose gate count grows linearly with the number of atoms. QMSS is not intended to replace classical shape-similarity methods, which are faster and more mature. However, this proof of concept shows that variational quantum circuits can be use to accurately perform ligand comparison. QMSS's linear scaling for gates number and potential enhancements of the algorithm, for instance a more complex entanglement scheme and a full-scale state tomography, seem to point towards promising results. Evaluation on larger and more diverse sets of drug-like molecules, and tests under finite sampling and hardware noise, are included as the next steps.

Acknowledgements. This research was funded by the Research Council of Lithuania under Grant Agreement No. S-ITP-25-5. The authors are also thankful for the HPC resources provided by the IT Research Center of Vilnius University.

High-Resolution cfDNA Sequencing Data Simulation for the Development of a Minimal Cancer Residual Disease Monitoring Model

Augustas Markevičius, Ramunė Bružaitė, Mindaugas Morkūnas

Institute of Data Science and Digital Technologies
Vilnius University

augustas.markevicius@mif.vu.lt

Minimal residual disease (MRD) monitoring via liquid biopsy is a low-risk alternative to invasive tissue profiling for tracking the progression of cancer. Modern clinical pipelines increasingly rely on analyzing cell-free DNA (cfDNA) in the bloodstream to discern the subclonal architecture of tumors during diagnosis and to monitor the mutational footprints of said tumors at subsequent clinical checkups.

With the increasing accuracy and accessibility of high-resolution DNA sequencing technologies, developing and validating MRD tracking algorithms requires datasets that faithfully capture the characteristics of high-throughput sequencing data with regards to signal-to-noise ratios, DNA fragmentation, etc. However, such datasets are difficult to create from real-life clinical data, due to high experimental costs, patient privacy concerns, and lengthy clinical timelines.

To address the challenge of data scarcity, we present a computational framework aimed at synthesizing high-resolution targeted cfDNA sequencing profiles by leveraging publicly accessible whole-genome sequencing (WGS) data from solid tumors. Our methodology utilizes a statistical resampling pipeline to convert widely available WGS tissue sequencing data into synthetic targeted cfDNA data. For our reference, we employ a public benchmark from the ICGC-TCGA DREAM challenge [1] to establish baseline clonal architecture maps. Additionally, we consider critical aspects such as position-specific sequencing noise, tumor-specific fragment-length truncation [2], and the ultra-low somatic variant frequencies that are characteristic of cfDNA samples [3]. We propose

that this methodology will enable the systematic evaluation of an MRD tracking algorithm designed to reconstruct the subclonal hierarchy of the primary tumor from highly diluted cfDNA signals.

References

[1] H. Carrie et al., "Comprehensive benchmarking of methods for mutation calling in circulating tumor DNA," Nature Communications, vol. 17, no. 1, p. 1082, Dec. 2025, doi: 10.1038/s41467-025-67842-x.

[2] G. Caravagna, T. A. Graham, and A. Sottoriva, "A guide to understanding tumour evolution through the lens of population genetics," Nature Reviews. Cancer, Sep. 2026, doi: 10.1038/s41568-026-00973-5.

[3] A. M. Newman et al., "An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage," Nature Medicine, vol. 20, no. 5, pp. 548–554, Apr. 2014, doi: 10.1038/nm.3519.

Beyond Price History: Stability and Sensitivity of the News-Augmented GARCH Model

Jurgita Markevičiūtė, Diana Vereškaitė

Institute of Applied Mathematics
Vilnius University

jurgita.markeviciute@mif.vu.lt

Financial market volatility is shaped not only by past price movements but also by expectations and reactions to news. Yet most approaches to incorporating sentiment into volatility models rely on additive effects, potentially overlooking nonlinear interactions and asymmetric responses to positive and negative information. The News-Augmented GARCH model addresses these limitations through a nonlinear, asymmetric, multiplicative sentiment component. Its flexibility, however, raises important questions about mathematical stability, parameter sensitivity, and the reliability of estimation. This presentation investigates these questions through theoretical analysis and controlled simulations. We establish the existence of a unique causal solution under a sufficient stability condition that links volatility persistence to the maximum amplification induced by news. This provides a theoretical foundation for evaluating the model beyond conventional GARCH restrictions. The simulation study examines hyperparameter recovery and forecasting accuracy, while partial dependence, Sobol indices, and SHAP values reveal the individual contributions of hyperparameters and their interactions. The results identify the baseline level of the news multiplier as the most influential hyperparameter and reveal substantial interactions among the tuning parameters. Estimates near the lower and upper ends of the investigated ranges require particular caution. Forecast errors are also lower when evaluated against realized rather than historical volatility in the simulated settings, highlighting the importance of the volatility proxy used for model assessment. Together, these findings provide practical guidance for hyperparameter selection and interpretation, supporting more informed use of news-sensitive volatility models. They also establish a foundation for further empirical validation and potential applications in financial risk assessment.

Segmented or Pooled Models? Machine Learning Estimation of Stem Volume from Lithuanian NFI, LiDAR and Sentinel-2 Data

Arnas Matusevičius, Gabrielė Kasputytė, Monika Papartė, Gintautas Mozgeris, Tomas Krilavičius

Vytautas Magnus University

arnas.matusevicius1@gmail.com

Lithuanian forests cover approximately one-third of the country's land area, and forest management requires reliable stem volume estimates for the whole forest area. The Lithuanian National Forest Inventory (NFI) measures volume accurately, but only on sparse sample plots, while airborne laser scanning (ALS) and Sentinel-2 imagery cover the whole country. Because forests are heterogeneous, a common approach is data segmentation: training separate models for different plot types or tree species. Previous studies comparing segmented and pooled models show that segmentation often improves accuracy, but mainly when the segment label is known without error and each segment contains enough plots. This research evaluates whether data segmentation improves machine learning estimates of plot-level stem volume from Lithuanian NFI, ALS and Sentinel-2 data. In this research, 10,877 NFI plots measured in 2020–2024 were combined with 81 ALS metrics and 69 Sentinel-2 features, and linear regression was compared with ten machine learning methods, including Random Forest, XGBoost and neural networks. Accuracy was assessed with repeated hold-out and cross-validation grouped by NFI plot clusters, using the root mean square error (RMSE) and the RMSE relative to mean volume. XGBoost reduced the RMSE by about 15% compared with linear regression, while tree-based and neural network methods differed by less than 5%. On forested plots, the relative RMSE was about 26%, comparable with studies in Norway, Finland and Poland (24–28%). Segmenting plots into forested and non-forested with a two-stage model reduced false volume predictions on non-forested plots from 0.6% to 0.4%, but left the relative RMSE unchanged. Holding

out each measurement year showed that a model trained on 2020–2023 under-predicted forested plots measured in 2024 by 3.7% of mean volume, which is consistent with forest growth after the ALS acquisition. Segmentation by tree species group is examined further, comparing separate models with a pooled model that uses species as a predictor. These results suggest that the choice of algorithm and the separation of non-forested plots have little effect on accuracy, while the time gap between field and ALS data introduces a systematic bias.

Evaluating Information Retrieval Effectiveness of RAG Architectures Across Diverse LLM Scales Using HotpotQA and MuSiQue

Natalija Matyjaškoit, Pavel Stefanovič, Simona Ramanauskaitė

Vilnius Gediminas Technical University

natalija.matyjaskoit@vilniustech.lt

In contemporary technological landscapes where artificial intelligence permeates diverse operational workflows, large language models (LLMs) play a significant role in tasks ranging from routine information retrieval to complex scientific research assistance. Because pretraining novel LLMs requires difficult infrastructure, Retrieval-Augmented Generation (RAG) frameworks are increasingly deployed to cost-effectively enhance external knowledge grounding and factual accuracy. This study investigates the comparative effectiveness of diverse RAGs in document retrieval and multi-hop reasoning, utilizing the complex HotpotQA and MuSiQue benchmark datasets. We systematically evaluate six models across distinct parameter scales (industry-leading large models (Claude, ChatGPT, and Gemini) and compact models (Gemma-3-4B, Qwen3-4B, and LLaMA-3.2-3B)) integrated across multiple retrieval strategies: traditional BM25, Dense retrieval, Hybrid configurations (BM25 + Dense), and advanced setups incorporating neural rerankers (Dense + Reranker, Hybrid + Reranker), alongside a proposed RAG search modification. Retrieval performance is evaluated using all-gold Top-k, Recall, and MRR, while answer quality is assessed using Exact Match (EM), Token F1, and BERTScore F1. Our empirical findings identify the optimal cost-to-performance trade-offs for robust document retrieval and critically illuminate the architectural boundaries, latency profiles, and practical suitability of low-resource, compact language models in knowledge-intensive environments.

A Diagnostic Framework for Assessing Social Bias in Lithuanian and English Datasets and Language Models

Emilija Mazūraitė[1,2], Deimantė Zaikauskaitė[1,2], Justina Mandravickaitė[1,2], Milita Songailaitė[1,2], Tomas Krilavičius[1,2]

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

emilija.mazuraite@vdu.lt

Large language models (LLMs) are used in decision-making contexts such as recruitment and healthcare, where systematic biases towards social groups can lead to unfair outcomes. Most bias-detection methods, however, were developed for English-speaking contexts and do not transfer directly to morphologically rich, low-resource languages such as Lithuanian. We present a diagnostic bias assessment framework for Lithuanian and English that evaluates both textual datasets and language models. Data bias is assessed through sentence-level co-occurrence: sentences containing expert-defined social-group targets are extracted from lemmatised text and searched for expert-defined, potentially adverse attributes. Model bias is evaluated using two distinct methods. The first is a probabilistic, template-based model-bias evaluation method that builds on the approach of Kurita et al. [1]. We adapt the method to Lithuanian, extend it from masked to causal language models, and use it to compare model preferences across gender, religion, and age groups. The second is a generative model-bias evaluation method that assesses whether model-generated content reinforces harmful stereotypes related to age, gender, and religion. A generator model completes prompts adapted from the HONEST benchmark introduced by Nozza et al. [2], and a separate judge model evaluates the completions against stereotypes specified for each bias category. Differences in verdicts between groups within a category (e.g., young adults vs. seniors) are assessed with a chi-square test. The data-bias method was applied to Lithuanian and English general-domain and hate-speech corpora. It identified target groups with potentially

adverse associations, with the highest rates observed in the Lithuanian hate-speech dataset. The template-based model-bias method captures differences in group associations across 12 models (spanning several model families, parameter scales, and both multilingual and Lithuanian-specific models) and is validated on gender-debiased models, all of which associated more attributes with women than their biased counterparts. In Lithuanian, a stronger masculine preference is observed; however, it is confounded by the default use of masculine gender forms. In the generative model-bias evaluation, in which five instruction-tuned models were paired as generators and judges, significant group differences were found in 6 of 14 Lithuanian and 2 of 10 English experiments, and results varied with the generator-judge pairing. Mean agreement between judge verdicts and expert annotations was 61.64-73.33% for Lithuanian and 72.04-74.67% for English across categories. Lithuanian generations were also far more often incoherent (mean 21.18-38.59% vs. 2.95-5.90% for English), which undermines judge-based evaluation because incoherent sentences are labelled non-biased. The framework can flag social bias in both datasets and models, but our Lithuanian results show that bias-assessment methods require language-specific adaptation to account for grammatical features, such as grammatical gender, and highlight the importance of considering generated-text quality when interpreting bias evaluations.

References

[1] K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov, "Measuring bias in contextualized word representations," in Proc. First Workshop on Gender Bias in Natural Language Processing, Florence, Italy, 2019, pp. 166–172.

[2] D. Nozza, F. Bianchi, and D. Hovy, "HONEST: Measuring hurtful sentence completion in language models," in Proc. 2021 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Online, 2021, pp. 2398–2406.

Hybrid Autoscaling Algorithm of Containerized Microservices for QoS Assurance in SaaS Systems

Dalius Mažeika, Povilas Lajus

Vilnius Gediminas Technical University

dalius.mazeika@vilniustech.lt

Software as a Service (SaaS) platforms grow in scale and complexity but maintaining Quality of Service (QoS) within Service Level Agreement constraints remains a significant challenge. QoS assurance in Kubernetes-based microservice architectures typically rely on either horizontal (HPA) or vertical (VPA) pod autoscaling, but each autoscaler has inherent limitations: HPA offers elasticity at the cost of resource overhead and cold-start delays, while VPA is resource-efficient but constrained by node capacity. This trade-off becomes particularly critical for workloads with unpredictable traffic, where a static scaling policy can lead to either SLA violations or unnecessary resource waste. Existing research provides limited knowledge on selecting the most effective approach for a given workload pattern, typically addressing only a single scaling dimension or relying on complex machine-learning predictions that hinder practical adoption. A hybrid autoscaling algorithm is proposed and evaluated. It dynamically identifies incoming load patterns using a sliding-window regression model and selects the most suitable pods autoscaling method – HPA or VPA – accordingly. The regression model operates on lightweight, real-time metrics rather than historical training data, allowing the algorithm to adapt to workload shifts without the latency. The approach was implemented and validated on a Kubernetes cluster using a CPU-intensive factorial service under five distinct load profiles. Results show that the hybrid autoscaler reduces average response time by up to 90% and decreases error rates by up to 99.7% compared to traditional Kubernetes autoscalers, while sustaining high throughput. This improvement comes at the cost of increased memory consumption and moderate CPU overhead. The proposed method is particularly effective for bursty and unpredictable workloads, offering a practical trade-off between performance and resource efficiency. These findings point to a viable path toward autonomous, SLA-aware scaling in production SaaS systems without reliance on prior workload forecasting.

Grad-CAM Explainability in Chest Radiography: A Comparison of Correct and Incorrect Predictions

Edita Mažonienė, Dmitrij Šešok

Vilnius Gediminas Technical University

edita.mazoniene@vilniustech.lt

Explainability has become an important aspect of deep learning applications in medical imaging. Grad-CAM is widely used to visualize image regions associated with model predictions, however, a correct or high-confidence prediction does not necessarily correspond to activation in an anatomically relevant region. This study investigates Grad-CAM activation patterns in correctly and incorrectly classified chest X-ray images. A DenseNet-121 based model trained on the NIH ChestX-ray14 dataset is analyzed across several thoracic findings selected for their differing morphological and spatial characteristics. Correct and incorrect predictions are examined separately. Particular attention is given to three configurations: correct classification accompanied by activation in the expected anatomical region, correct classification accompanied by activation outside it and incorrect classification that nevertheless produces a clinically plausible activation pattern. Near-threshold predictions are considered separately, in order to distinguish genuine differences in model attention from outcomes determined primarily by the selected operating threshold. Grad-CAM visualizations are analyzed qualitatively in relation to prediction outcome and lesion morphology. Activation maps are examined to identify recurring differences between correct and incorrect predictions, and between focal and diffuse abnormalities, where lesion extent is expected to influence the specificity of the resulting maps. Potential confounding features are also noted, as anatomically plausible activation may reflect co-occurring findings or imaging artifacts rather than the target pathology. The study highlights the importance of examining model behavior beyond classification metrics alone. Grad-CAM can provide useful insight into image regions associated with individual predictions, but visually plausible activation

does not by itself establish that a model has learned clinically meaningful features. Predictive confidence and visual explanation should therefore not be treated as equivalent indicators of model reliability. The presented analysis does not constitute clinical validation, and most importantly, the final assessment of the clinical relevance of activation patterns requires evaluation by qualified medical professionals.

Investigation of BIM Model Development for Cultural Heritage Buildings Using Artificial Intelligence and Other IT Solutions

Darius Migilinskas, Birutė Končė

Vilnius Gediminas Technical University

darius.migilinskas@vilniustech.lt

Contemporary academic literature emphasises that the primary challenge of Heritage Building Information Modelling (HBIM) arises from a fundamental conflict between standardised parametric design tools adapted for new construction and the unique, changing forms of historical architecture. Traditional automated solutions designed for modern construction fail to accurately interpret the geometric imperfections, complex spatial configurations, and structural deformations caused by age-related natural changes in historical buildings. Existing artificial intelligence algorithms often encounter systematic errors when attempting to classify such non-standard heritage elements; consequently, the creation of 3D models faces significant manual labour and financial costs, relying heavily on the professional interpretation of the modeller. The application of Building Information Modelling (BIM) for heritage buildings is further complicated by a lack of standardised component libraries, the vast volume of data generated, and the need for specialised, advanced, and costly data-analysis technologies, all of which require exceptional specialist expertise and substantial resources to manage effectively. Throughout the building's lifecycle, design solutions, collected historical data and structural information are typically preserved in a highly fragmented manner, resulting in significant information gaps and losses within the asset management process. For this reason, the application of BIM technologies is critical for modelling the existing conditions of cultural heritage sites, enabling the collection, systematisation, and secure preservation of relevant building data for future use throughout the structure's lifecycle. The study clearly demonstrates that the optimal creation of 3D models for

cultural heritage sites depends directly on the targeted application of advanced digital technologies, customised to the level of detail required for the specific heritage building. In this context, an AI-integrated "Scan-to-BIM" methodology emerges as the most effective tool, ensuring the unparalleled micro-geometric precision needed to capture deformations in unique architectural details and millimetre-scale structural deviations. Photogrammetry and orthophoto technology play a vital supporting role by effectively mapping hard-to-reach heritage structures such as complex roof planes or intricate façade ornamentation, and preserving their valuable characteristics. Traditional BIM modelling remains an essential integrative axis, allowing analogue historical and archival data to be consolidated into a unified digital environment. The research addresses the comprehensive evaluation of BIM technology efficiency in the management and renovation of cultural heritage sites, emphasising in-depth analysis of existing conditions and the collection, preservation, and continuous use of project information throughout the building's lifecycle, while examining the advantages, disadvantages, and practical applicability of various data collection and design methods. A synthesis of the literature review and practical research insights indicates that BIM technology holds immense potential for cultural heritage sites; however, its actual effectiveness depends directly on the rationality of the chosen data collection methods, the selected level of model detail, the competence of the specialists involved, and the specific practical application of these technologies to particular heritage sites. The study's findings suggest that the digitalisation of heritage preservation is evolving toward hybrid systems as a synergy of combined laser scanning and photogrammetry data, thereby radically reducing the costs associated with manual modelling and eliminating errors caused by purely automated solutions, failures in interpreting complex spatial geometry, and the non-professional human factor. The methodological solution analysed in this study enables a fundamental optimisation of the heritage object digitisation process by employing advanced IT tools and artificial intelligence algorithms for precise element recognition, while simultaneously ensuring the comprehensive, millimetre-level authenticity of cultural assets based on supervision by HBIM professionals for the long-term information storage and accessibility of data for future asset management needs.

On the Arithmetic Optimization Algorithm for the Quadratic Assignment Problem

Alfonsas Misevičius, Dovilė Verenė, Gintaras Palubeckis

Kaunas University of Technology

dovile.verene@ktu.lt

In our work, the main goal is to demonstrate the performance of an arithmetic optimization algorithm (AOA) for solving the quadratic assignment problem (QAP). The basic innovation of the proposed algorithm is the introduction of solution transformation functions that establish a mapping between the continuous search space of AOA and the discrete space of the QAP. These functions allow the bidirectional mapping, that is, they enable not only the transformation of real values into permutation-based solutions, but also maintain the reverse mapping from permutations to real-valued representations. The experimental results confirm the high efficiency of AOA for the QAP. Several new best known solutions were obtained for hard QAP instances.

Natural Language Processing in the Era of Artificial Intelligence: The Wind of Change is Blowing

Ruslan Mitkov

Lancaster University, UK

r.mitkov@lancaster.ac.uk

Natural Language Processing (NLP) is currently undergoing a profound and unprecedented transformation. While it has long been acknowledged that NLP technologies are inherently imperfect and far from achieving complete accuracy, recent advances have significantly reshaped the field of language and translation technologies. First, Deep Learning methods, and now Large Language Models (LLMs), have revolutionised the field and captured widespread attention across both academia and industry. This entertaining and accessible talk will shed light on the future of NLP in the era of Artificial Intelligence (AI). The talk will begin by outlining the history of NLP and Machine Translation, reviewing the advances driven by Deep Learning and LLMs. The speaker will then offer a critical evaluation of the use of LLMs in NLP and Machine Translation, focusing on his recent comparative studies involving LLMs, traditional Deep Learning architectures, and rule-based methodologies across selected NLP tasks and applications. Finally, the presentation will turn to the future of NLP. Generative AI is certainly continuing to improve, but are there any attendant dangers? While the speaker makes no claim to prophetic insight, he will draw on his extensive experience and use linguistic intelligence as a benchmark to offer predictions on how artificial intelligence may evolve relative to human intelligence. The talk will also consider how language professions themselves are likely to evolve in the years ahead.

Forecasting German Day-Ahead Electricity Market Prices with Regularized Linear Models

Mykolas Motiejūnas[1,2], Lukas Janušauskas[1,2], Loann Brahimi[2], Baptiste Quidet[2], Gaël Pagès[2]

[1] Vilnius University
[2] Improm AI

mykolas.motiejunas@mif.stud.vu.lt

Electricity markets play central roles in matching supply with demand and setting electricity prices. The German electricity market runs a daily day-ahead auction allowing participants to bid on prices of 15 minute delivery intervals of the next day. Accurate forecasts of day-ahead prices allow producers, consumers and traders to manage risk and allocate assets according to market dynamics. This task is greatly complicated by a strong influence from renewable sources, solar and wind, which cause price volatility and even negative prices. We forecast all 24 hourly prices with regularized linear models (Lasso, Ridge and Elastic Net), simulating a production environment by using only information available before 11:00 on the previous day, one hour ahead of auction closure. As price drivers, we incorporate country-wide meteorological variables (solar radiation, wind, temperature and cloud cover) and forecasted load. Using separate XGBoost models to forecast solar, wind, fossil and other electricity generation, we feature engineer two key features: the renewable share of total generation and the ratio of total generation to forecasted load. All models are evaluated with a weekly walk-forward backtest where hyperparameters are tuned every four weeks. Performance is measured using standard metrics such as MAE, MdAE, RMSE and R2. To evaluate the linear models, we compare them against an exponentially weighted moving average model.

Geometry-Informed Comparison of Quantum and Classical Molecular Generative Models

Glauco Endrigo Moura de Lima, Remigijus Paulavičius

Institute of Data Science and Digital Technologies
Vilnius University

glauco.de@mif.stud.vu.lt

Comparing quantum and classical generative models requires criteria that connect their representations to the task being learned. Equal numbers of trainable parameters do not describe how strongly, or in which directions, those parameters affect molecular predictions. We present a geometry-informed framework for comparing photonic quantum and classical models for molecular graph generation, combining local sensitivity analysis with evaluation under common training and sampling conditions. The framework follows the response to changes in optical controls from the quantum state through photon counting to molecular predictions. A classical network is calibrated to approximate photonic moment features, and the concentration of its task-Fisher spectrum is compared with photonic references to select classical control dimensions. These dimensions guide the construction of two classical comparators based on scale-only Feature-wise Linear Modulation (FiLM), which rescales groups of features produced by a fixed multilayer perceptron. The resulting models, C-FiLM7 and C-FiLM9, use seven and nine trainable group scales, respectively. The generation experiment compares them with a simulated photonic model, Photonic Q, whose nine optical angles control features derived from the complete photon-count distribution. All three models perform discrete graph denoising using separately trained prediction heads, which are then frozen during control optimization. Experiments use 973 distinct molecular graphs with two to four heavy atoms and a fixed training-validation split. Evaluation covers ten paired training seeds, with three generation runs of 1,000 candidates combined into 3,000 candidates per model and seed. Mean validity is 63.70% for Photonic Q, 76.97% for C-FiLM7 and 77.11% for C-FiLM9. Mean uniqueness among valid molecules is 41.49%, 33.90%

and 33.91%, respectively. The mean yield of novel unique structures is 7.19%, 6.17% and 6.22% of all generated candidates, respectively. This yield counts distinct valid structures absent from the full 973-structure reference dataset, after removing repetitions within each combined sample, and divides that count by 3,000. All percentages are averaged across training seeds. Photonic Q therefore produces a higher mean yield of novel unique structures despite its lower validity, and its novel unique count exceeds each classical comparator in eight of ten paired runs. These preliminary results identify a trade-off between validity and the diversity of valid outputs. The framework connects a common mathematical description of local prediction sensitivity with controlled generation experiments, providing a basis for assessing representation choices across quantum and classical models.

Automated R&D with AI Agents

Vytas Mulevičius

Neurotechnology

vytas.mulevicius@neurotechnology.com

Most of a researcher's time goes into the inner loop: change something, run an experiment, check a number, decide whether to keep the change. AI agents are now good at exactly this loop. They can work on a task for hours rather than minutes, and leading labs openly say they are building automated AI researchers.

The talk uses Andrej Karpathy's open-source project autoresearch (March 2026) as a case study. In about 630 lines of code, an AI agent edits a model's training code, trains for five minutes on one GPU, and keeps a change only if a fixed validation metric improves. That comes to roughly 100 experiments a night. The human doesn't write Python. Instead they write program.md, a plain-English protocol that sets goals, constraints and research taste. In two days the agent tried about 700 changes, kept about 20, and cut the nanochat "time to GPT-2" benchmark by 11%.

From this we draw a recipe for automating research: a cheap, trusted metric, a fixed compute budget, a limited set of files the agent may change, version control as memory, and checks that results hold at full scale.

Mathematics shows what happens when automation moves faster than the research community can absorb it. In October 2026, OpenAI released more than 700 manuscripts addressing about 370 open problems, and three were retracted within a day. Terence Tao warned about problems "solved" by AI prompters with no interest in the field. He also noted that researchers now hold back ideas for fear of being scooped. Mathematicians worry about four more things: journals and arXiv flooded with proofs no one can check, results nobody understands, unclear credit, and students put off from the field.

For a data-analysis audience, the risks are also statistical: metric gaming, noise-chasing, and overfitting the validation set after hundreds of trials.

The Impact of Stated Source Reliability on Face Verification with Multimodal Large Language Models

Artūras Nakvosas, Virginijus Marcinkevičius

Institute of Data Science and Digital Technologies
Vilnius University

arturas.nakvosas@mif.stud.vu.lt

Previous work on face verification has shown that multimodal large language models (MLLMs) may follow textual claims, respond in the opposite direction, or retain their image-based decisions. These findings do not establish whether models weight such claims according to the stated reliability of their source. We investigate this question by separating the effect of claim content from the decision shift shared by opposing claims and comparing the observed updates with a Bayesian benchmark.

We test three MLLMs (Gemma-3-12B, Qwen3-VL-8B, InternVL3.5-8B) on 150 genuine and 150 impostor Labeled Faces in the Wild (LFW) pairs with three verification prompts, comparing an image-only baseline against prompts that add a textual claim ("same person" or "different people") from a source with stated reliability r in {0.5, 0.6, 0.75, 0.9, 0.95, 0.99}, expressed in the prompts as source accuracies of 50%, 60%, 75%, 90%, 95% and 99%. The claim is attributed to a witness, an automated matcher or an analyst, or accompanied by a source-reliability tag; for example, "A witness who is right about 90% of the time says the two images show the same person". For each pair of images, prompt, source wording and r, we record the shifts in the model's log-odds from the image-only baseline under the "same person" and the "different people" claims. We decompose these shifts into a claim effect (half their difference) and a common part (their mean). Under a Bayesian benchmark, the expected claim effect equals the log-odds of r, and the common part is zero.

The experiments show that none of the three models consistently adjusts its decisions according to the stated source reliability. Gemma-3 shows a substantial increase in claim effect as reliability rises, but only

on genuine pairs; Qwen3-VL and InternVL3.5 show little sensitivity to reliability. All three models also exhibit a shift toward uncertainty shared by opposing claims. In Qwen3-VL and InternVL3.5, a source-attributed statement that makes no identity claim reproduces much of this shift. Source wording also matters: at 50% stated reliability, where the Bayesian target is zero, Gemma-3's claim effect varies substantially across phrasings. False "same person" claims increase false-accept rates in Gemma-3 and Qwen3-VL; notably, Qwen3-VL is most affected at the lowest stated reliability. For InternVL3.5, changes in error rates are not consistently supported across thresholds targeting false-accept rates of 1% and 10%, given bootstrap uncertainty.

These findings indicate that the tested MLLMs do not reliably use stated source reliability to weight auxiliary evidence. The paired protocol provides a basis for evaluating progress toward calibrated evidence use, while generalization to other models and tasks remains to be established.

Beyond Benchmark Accuracy: Cross-Dataset Generalization and Open-Set Safety in Gastrointestinal Endoscopy AI

Mesut Erkin Ozokutgen, Adebayo Adewumi Abayomi-Alli, Sema Nur Alan

Vilnius Gediminas Technical University

adebayo-adewumi.abayomi-alli@vilniustech.lt

Artificial intelligence models for gastrointestinal endoscopy often perform well on benchmark datasets, but their reliability on data from other sources and on unfamiliar clinical findings remains poorly understood. This study evaluates the cross-dataset generalization, calibration, and open-set safety of gastrointestinal endoscopy classifiers. Five architectures (EfficientNet-B1, EfficientNet-B3, ConvNeXt-Tiny, DenseNet121, and Swin-T) were trained on the eight-class Kvasir v2 dataset with five random seeds each and externally evaluated on HyperKvasir. Overlapping HyperKvasir categories were harmonized to the Kvasir v2 label space for closed-set evaluation (6,759 images), while 3,228 images from non-overlapping classes were retained for open-set analysis. Performance was assessed using accuracy, Macro-F1, AUROC, and expected calibration error (ECE), and confidence-based abstention (threshold 0.8) was used to examine open-set behaviour. Internal Macro-F1 ranged from 0.918 to 0.936, and external Macro-F1 ranged from 0.911 to 0.934, with ConvNeXt-Tiny achieving the best external performance (accuracy 95.1%, Macro-F1 0.934, AUROC 0.984). Mean Macro-F1 differed by less than 0.01 between internal and external evaluation for every model, and ECE remained low (0.023 to 0.048). However, class-level results showed that aggregate scores concealed weaker behaviour for specific findings: normal-cecum was the most difficult class externally (F1 0.78 to 0.80, recall about 0.68), followed by esophagitis (F1 0.84 to 0.90) and normal z-line (F1 0.85 to 0.89). In open-set analysis, unseen classes were often assigned to clinically related known classes with high confidence. Barrett's images were

predicted as esophagitis in 88.5% of cases (mean confidence 0.91) and ileum images as ulcerative colitis in 96.5% of cases (mean confidence 0.94), yet only 17.3% and 11.1% of these images, respectively, were rejected by abstention. These findings indicate that strong closed-set and cross-dataset accuracy does not by itself establish clinical readiness, and that confidence-based abstention alone is insufficient to prevent overconfident errors on unfamiliar findings. Explainability analysis is part of ongoing work.

Making Invisible Colleges Visible: Multidimensional Scaling and Interactive 3D Exploration of the Vilnius University Co-Authorship Network

Karolis Pempė, Martynas Sabaliauskas

Faculty of Mathematics and Informatics
Vilnius University

karolis.pempe@mif.stud.vu.lt

Research at a large university is increasingly a collective activity, yet scalar bibliometric indicators such as publication counts or the h-index describe individual researchers and ignore the relations between them. Co-authorship networks make this relational layer observable and can reveal whether actual research groups, the "invisible colleges" of the sociology of science, coincide with the formal organizational structure. Two practical obstacles arise at this scale: bibliographic repositories present records intended for human reading rather than network analysis, and a network of thousands of researchers is hard to display so that both global patterns and individuals remain accessible. We reconstructed the co-authorship network of Vilnius University with an automated pipeline. The university's organizational hierarchy was traversed recursively to obtain 4,103 staff members, emeriti and affiliated researchers in 15 faculty-level units, and their publication records were harvested from the Lithuanian Academic Electronic Library (eLABa). After deduplication, 75,671 unique publications yielded a weighted graph of 3,221 authors and 20,512 co-authorship ties. Tie strength was computed by fractional counting: a publication with N authors, external ones included, adds 1/N to every pair of its co-authors from the university, so that large consortium papers do not dominate the network. Two complementary views were developed. Static maps were obtained by metric multidimensional scaling with the SMACOF algorithm in two and three dimensions, using the dissimilarity $\delta = 1/(1 + \ln(1 + w))$, where w is the tie strength. This transformation is bounded in (0, 1], assigns the maximal dissimilarity to author pairs without joint papers and damps

the influence of a few exceptionally strong ties (maximum strength 159.8 versus a median of 0.25); a PCA-based initial configuration makes the maps reproducible. After a general-purpose plotting library froze the browser at the university scale, an interactive 3D explorer was built on WebGL (Three.js) with a force-directed layout. It runs in a web browser without installation or server-side components and offers faculty filters, search, node volumes proportional to publication counts, multicolored nodes for 107 authors affiliated with more than one faculty, author cards listing positions, publications and co-authors ranked by tie strength, and shareable URL-encoded views. The network is sparse (density 0.004) yet strongly clustered (average clustering coefficient 0.428), has a heavy-tailed degree distribution (median 8, maximum 165) and small-world properties, with a mean shortest path length of 4.13 within the largest connected component, which holds 89.0% of authors. Collaboration largely follows organizational boundaries: 82.8% of ties connect authors sharing a faculty, and the faculty partition reaches a modularity of $Q = 0.694$. The Louvain algorithm, blind to organizational data, attains only a moderately higher modularity ($Q = 0.813$) but divides the largest faculties among several smaller communities (the largest community has 306 authors, whereas the largest faculty has 893). The strongest interfaculty axis joins Medicine and the Life Sciences Center (898 ties); Mathematics and Informatics is most strongly linked to Medicine (346). Most humanities and social science units are only weakly connected, both to each other and to the natural sciences. In the SMACOF maps, organizational clusters sharpened with scale: absent for one institute (91 authors), emerging for one faculty (248) and evident for the whole university (about 2,900), presumably because individual collaborations weigh less in the global stress function as the number of points grows. The two views answer different questions rather than compete: SMACOF maps suit the assessment of global cluster tendencies, have interpretable distances and are deterministic, whereas the interactive explorer is the only practical way to locate a particular researcher and follow their collaborations. Future directions include the temporal dimension, external collaborators and centrality measures that identify researchers bridging otherwise separate parts of the network.

Acknowledgements. This research was funded by the Research Council of Lithuania (LMTLT), agreement No S-SV-26-165.

Lithuanian Medical Decision Models

Linas Petkevičius

Faculty of Mathematics and Informatics
Vilnius University
linas.petkevicius@mif.vu.lt

The three models target Lithuanian biomedical typed decisions on MEDLINE-derived data was created. laya-multilingual-lt-med is a non-autoregressive fine-tune of the multilingual Laya model that returns calibrated option probabilities in one forward pass. On the MEDLINE-S1-LT test set it reaches 92.4% decision accuracy and 62.8% full-record accuracy after temperature scaling. Qwen3.5-4B-LT-med is a LoRA adapter plus Clef head trained only on English BioDecision examples and then transferred to the Lithuanian test split. It achieves 85.1% decision accuracy and 39.9% full-record accuracy, well above a random head on the same base. medgemma-1.5-4b-it-decitions-lt-med is a LoRA fine-tune of MedGemma 1.5 4B trained directly on MEDLINE-S1-LT. It scores 94.6% decision accuracy and 72.3% full-record accuracy after calibration, the highest of the three. All three models emit probabilities rather than generated text and are explicitly research-only, not medical devices, which enables type save deployment capabilities. They share the same evaluation protocol on 4,857 abstracts and 29,654 decisions. The results for first time illustrate the benefit of in-distribution fine-tuning versus cross-lingual transfer for Lithuanian medical decision tasks.

Acknowledgements. This project was funded by the European Union (project No S-ITP-26-11) under the agreement with the Research Council of Lithuania (LMTLT).

Artificial Intelligence-Based Automated Real Estate Valuation

Monika Petrulevič, Simona Ramanauskaitė

Vilnius Gediminas Technical University

simona.ramanauskaite@vilniustech.lt

The rapid development of artificial intelligence (AI) and machine learning (ML) technologies creates new opportunities for automating traditionally expert-driven processes, including real estate valuation. Conventional valuation methods often require expert knowledge and can be time-consuming when processing large volumes of heterogeneous and geographically distributed data. This research investigates the application of ML techniques for automated real estate valuation and aims to develop and evaluate a data-driven method for predicting property values. An extended dataset was created by combining real estate transaction data with information from a real estate advertisement portal. A methodology was developed for distributing transaction prices among multiple properties involved in a single transaction. Four machine learning regression models were evaluated: K-Nearest Neighbors (KNN), Ridge Regression, Support Vector Regression (SVR), and Multilayer Perceptron (MLP) neural networks. The selected models represent different regression approaches, including instance-based learning, regularized linear regression, kernel-based regression, and nonlinear neural network modelling. The models were compared using regression evaluation metrics, with particular attention given to the coefficient of determination (R^2) and mean absolute error (MAE). The experimental results showed that the Support Vector Regression (SVR) model achieved the best performance among the evaluated approaches. The model obtained $R^2 = 0.715$ and a mean absolute error of 382.84 €/m². These results demonstrate the potential of machine learning techniques for automated property valuation and indicate that nonlinear models can effectively capture relationships between property characteristics, spatial factors, and market prices. The research also demonstrates that model performance is strongly influenced by the quality and preparation of input data. The proposed approach provides a foundation for

developing scalable, data-driven real estate valuation solutions. Future research could focus on expanding the dataset, incorporating additional temporal and spatial characteristics, and evaluating more advanced machine learning and ensemble methods across different geographical areas and real estate market segments.

Data-Driven Imputation of Missing Economic Indicators for EU Regions: A Multi-Metric Accuracy Assessment

Irmantas Pilypas, Irma Šileikienė, Asta Slotkienė

Vilnius Gediminas Technical University

irmantas.pilypas@vilniustech.lt

Regional economic analysis in the European Union relies on complete statistical data at the NUTS 2 level, yet Eurostat regional datasets frequently contain missing values, particularly for less-developed regions and earlier periods. Mean or median substitution ignores inter-variable relationships, whereas Multiple Imputation by Chained Equations (MICE) preserves them but is computationally demanding for large datasets. Tree-based machine learning methods offer a scalable alternative. However, their accuracy and sensitivity to hyperparameter selection on heterogeneous regional data with varying missingness patterns remain insufficiently studied. This study compares XGBoost and Random Forest for imputing missing regional economic indicators. A dataset of eight economic indicators for 260 regions over the period 1990-2023 was compiled from the Eurostat database (6,984 observations, with 10.01-53.81% missing values per indicator). A separate model was trained for each indicator, with the remaining indicators, region and year as predictors. For each indicator, 20% of the observed values were randomly withheld as a test set. Imputation accuracy was assessed using nRMSE, nMAE, R^2, and sMAPE with five-fold cross-validation on the remaining values for the baseline and hyperparameter-optimised configurations (randomised search over 30 configurations). XGBoost outperformed Random Forest for all eight indicators and all four metrics in both configurations, achieving a mean R^2 of 0.9433 compared with 0.8826. The largest difference was observed for population density, an indicator weakly correlated with the others, with R^2 increasing from 0.7229 to 0.9652 (a relative improvement of 33.51%). Both models achieved the highest accuracy for the strongly correlated employment indicators (XGBoost R^2 values of 0.9882 and 0.9885 for total and female

employment rates, respectively). Hyperparameter optimisation did not systematically improve imputation accuracy. It reduced R^2 for six of eight indicators with Random Forest and for all eight with XGBoost, lowering the mean R^2 by 0.066 and 0.013, respectively. These results highlight the importance of model parsimony for heterogeneous regional datasets.

Identifying Emotion-Sensitive Acoustic Features in Italian Speech: A Comparative Analysis of Anger and Neutral Expression

Michele Pizzi[1], Gražina Korvel[1], Bożena Kostek[2], Povilas Treigys[1]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technologies, Poland

grazina.korvel@mif.vu.lt

Understanding how emotions shape acoustic speech characteristics is essential for developing interpretable and robust speech emotion analysis methods. This study investigates acoustic features associated with anger in Italian speech, providing an initial step toward cross-linguistic comparisons of emotional expression. An automated analysis pipeline was developed to extract eGeMAPS and Mel-frequency cepstral coefficient (MFCC) features from the Emozionalmente corpus. The analysis included 280 recordings from 28 speakers, each contributing five matched sentences produced under angry and neutral conditions. For each speaker, identical sentence content was compared across emotional conditions. For each acoustic descriptor, mean values and standard deviations were computed, and Pearson correlations with binary emotion labels were used to identify features most strongly associated with anger. Of 388 extracted acoustic features, the 50 with the strongest correlations with emotion labels were selected for further analysis. The results indicate that spectral flux, loudness variability, and selected MFCC descriptors show the strongest associations with emotional expression, suggesting increased spectral dynamics and acoustic variability in angry speech. These preliminary findings demonstrate the potential of interpretable acoustic descriptors to characterize emotion-related speech variations. Future work will extend the analysis to additional emotional categories and languages to assess the consistency and generalizability of the identified acoustic markers.

Acknowledgments. 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.

Analyzing Group HRV Complexity Matching via Attractor State-Space Geometry

Naseha Wafa Qammar[1], Kristina Poškuvienė[1], Mantas Landauskas[1], Kiran Shahzadi[1], Minvydas Ragulskis[1], Alfonsas Vainoras[2], Nachum Plonka[3], Mike Atkinson[3], Rollin McCraty[3]

[1] Kaunas University of Technology
[2] Lithuanian University of Health Sciences
[3] HeartMath Institute, USA

kiran.shahzadi@ktu.edu

We investigate group-level complexity matching by evaluating the geometric organization of heart rate variability (HRV) dynamics within a reconstructed delay-coordinate state space. Photoplethysmography (PPG)-derived interbeat interval (RR-interval) time series were recorded from 20 adult participants during a structured, six-interval guided heart-focused session. Each participant's HRV recording for each interval was embedded into a two-dimensional state space using an optimal time delay determined independently for that participant and interval. To compare the resulting attractors systematically, the reconstructed two-dimensional attractor images were processed with a convolutional neural network (CNN) autoencoder, which performed unsupervised feature extraction and dimensionality reduction into latent geometric representations. K-means clustering applied to the first two principal components of the latent space grouped the attractors into three recurrent morphology types: circular or loop-like (Type 1), distorted circular (Type 2), and distorted concentrated (Type 3). Repeated-measures analysis with Cochran's Q test showed a significant redistribution of attractor morphology across the session. The prevalence of Type 1 increased from 10.00% in the first interval (baseline) to 60.00% in the final, compassion-focused interval (Cochran's $Q(5) = 19.46$, Holm-adjusted $p = 0.0047$), while Type 2 decreased from 55.00% to 25.00%. This pattern suggests increasing similarity in HRV state-space organization, consistent with group-level complexity matching. As a conceptual illustration, diffusively

coupled Thomas cyclically symmetric and Rössler chaotic systems showed how interaction between structurally different nonlinear systems can produce increasing similarity in attractor geometry; this does not establish that the same mechanism operated in the HRV data. The protocol had a fixed order and no control condition, so interval order is confounded with elapsed time and instruction repetition. The increase in Type 1 should therefore be read as an association, not as evidence of a causal effect of the appreciation- or compassion-focused content. Respiration was not monitored, and the morphology types were not compared with conventional HRV indices. Within these limits, attractor reconstruction combined with CNN-autoencoder clustering can reveal collective physiological organization that is not visually apparent in the raw HRV recordings.

Detecting GenAI-Generated Open-Ended Exam Answers Across Multiple Computer Science Subjects

Urtė Radvilaitė, Birutė Pliuskuvienė, Simona Ramanauskaitė, Pavel Stefanovič

Vilnius Gediminas Technical University

urte.radvilaite@vilniustech.lt

The widespread adoption of Generative Artificial Intelligence (GenAI) tools has created new challenges for the assessment of open-ended exam questions in higher education. While systems such as ChatGPT, Copilot, and Gemini can support learning, their use also raises concerns regarding academic integrity and the authenticity of student work. As a result, there is an increasing need for reliable methods capable of distinguishing between student-written and AI-generated responses. This study investigates the applicability of deep learning and transformer-based approaches for detecting GenAI-generated answers in Lithuanian-language open-ended assessments across multiple computer science subjects. The experimental dataset consisted of authentic student responses collected from midterm and final examinations in five computer science courses, complemented with answers generated using ChatGPT, Copilot, and Gemini. Two deep learning architectures (CNN and LSTM) and four BERT-based transformer models were evaluated for both binary classification (human-written versus AI-generated) and multiclass classification (identification of the specific GenAI tool). The results show that transformer-based models can achieved accuracy of 89-95% in in detecting AI-generated responses while CNN and LSTM for this task reach accuracy of 91% and 90% respectively. All models perform very similarly when identifying the specific GenAI tools used to generate the answers demonstrating the accuracy of 68-85%. The findings reveal that models trained on a single subject do not generalize well to other subjects, indicating strong domain-specific characteristics and the need for more diverse training datasets. Furthermore, the results suggest that responses generated by ChatGPT and Copilot exhibit greater similarity,

whereas Gemini-generated responses contain more distinctive features. The findings demonstrate the potential of deep learning methods to support academic integrity and assessment practices in higher education. However, GenAI detection systems should be used as decision-support tools rather than definitive evidence of student misconduct.

Perfect Coverings with Sparse Observation Windows for Secure Steganography

Jurate Ragulskiene, Loreta Saunoriene, Minvydas Ragulskis

Kaunas University of Technology

jurate.ragulskiene@ktu.lt

Image steganography aims to conceal information within a carrier image while preserving its visual appearance and preventing unauthorized extraction of the hidden message. This study considers a steganographic approach in which a dichotomous secret image is embedded into a grayscale carrier image using a system of overlapping observation windows. In contrast to conventional rectangular observation windows, only selected cells of each window participate in the evaluation of pixel values. Such a structure is referred to as a sparse observation window (SOW), where active cells are involved in determining the number of different grayscale levels, while passive cells are excluded from this operation. A central problem addressed in this study is the construction of perfect coverings for different architectures of sparse observation windows. A covering is considered perfect when a pixel modification performed at the current position of the observation window does not alter the information already encoded at its previous positions. The admissibility of a particular SOW therefore depends on the mutual arrangement of its active cells, the location of the modifying cell, and the selected covering strategy. The analysis demonstrates that the modifying cell does not necessarily have to coincide with a corner of the rectangular window; nevertheless, its position cannot be chosen arbitrarily if conflicts with previously processed windows are to be avoided. The sparse architecture introduces an additional security mechanism because successful reconstruction of the hidden information requires knowledge of the actual distribution of active cells. Even a single permutation of these cells may prevent correct decoding. Consequently, both the number and spatial configuration of active cells can effectively

serve as additional components of a secret key. For example, a 4 × 4 sparse observation window containing nine active and seven passive cells admits 10,564 distinct effective layouts; even when the modifying cell is fixed, 6,106 admissible arrangements remain. Computational experiments with different carrier and secret images demonstrate favorable statistical properties of the resulting stego images in terms of mean absolute error, mean square error, structural similarity, and peak signal-to-noise ratio. The approach is also investigated with respect to partial image blocking, additive noise, and brightness transformations. The results indicate that sparse observation windows provide a flexible combinatorial mechanism for increasing confidentiality while retaining the principal advantages of image-based steganographic encoding.

Evaluation of Multi-Model Translation Strategies for Lithuanian Medical Multimodal Dataset Construction

Augustė Raišytė[1], Dalius Matuzevičius[2], Tadas Žvirblis[3], Linas Petkevičius[1]

[1] Vilnius University
[2] Vilnius Gediminas Technical University
[3] National Cancer Institute

auguste.raisyte@mif.vu.lt

This research investigates methods for adapting existing vision–language models (VLMs) to medical visual question answering in Lithuanian, with particular emphasis on the influence of dataset translation quality on model fine-tuning and downstream performance. Most contemporary multimodal large language models and medical datasets are primarily developed for English, which limits their applicability to lower-resource languages such as Lithuanian. This limitation is particularly relevant in the medical domain, where translation inaccuracies, inconsistent terminology, and semantic ambiguity can directly affect the quality of model training and the correctness of generated answers. Therefore, the study addresses not only model adaptation but also the development of reliable methods for constructing high-quality Lithuanian medical datasets from existing English-language resources. The experiments are based on the PathVQA, VQA-RAD, and SLAKE medical visual question answering datasets. Their textual components, including questions and answers, are translated into Lithuanian using multiple large language models and translation approaches. Rather than relying on a single translated version, the research investigates whether combining independently generated translations can reduce translation errors and improve semantic, linguistic, and medical consistency. Several dataset construction strategies are examined: datasets translated by individual language models, combinations of translations produced by different models, automated merging of multiple translation candidates, and translation merging with human supervision. A particular focus is placed on automated translation consolidation. Multiple independently

generated Lithuanian translations of the same source instance are analysed and merged into a single candidate intended to preserve the original medical meaning while reducing grammatical, terminological, and semantic errors introduced by individual translation systems. This approach is compared with human-supervised merging, in which a reviewer evaluates alternative translations and selects, corrects, or combines them. The resulting dataset variants provide a basis for analysing the relationship between translation methodology, dataset quality, and the performance of subsequently fine-tuned multimodal models. MedGemma and Gemma-family vision–language models are fine-tuned on the resulting Lithuanian datasets using the Unsloth framework, enabling parameter-efficient training while reducing computational and memory requirements. The models are subsequently evaluated on their ability to interpret medical images, understand questions expressed in Lithuanian, and generate medically relevant and linguistically correct answers. The proposed evaluation methodology considers several complementary dimensions rather than relying exclusively on conventional answer accuracy. These include semantic correspondence with reference answers, correct use of medical terminology, preservation of clinical meaning, Lithuanian grammatical quality, and robustness to translation-related variation. By systematically comparing different translation and merging strategies, the study aims to determine how translation quality propagates into fine-tuned model performance and whether multi-model translation consolidation can provide a scalable alternative to fully manual dataset localization. The results are expected to contribute to the development of more reliable methodologies for adapting medical multimodal models to Lithuanian and other low-resource languages.

Acknowledgements. This project has received funding from the Research Council of Lithuania (LMTLT), agreement No S-ITP-26-11.

Large Language Models for Conceptual Modelling: From Automated Generation to Human-AI Collaboration

Jolita Ralytė

University of Geneva, Switzerland

jolita.ralyte@unige.ch

Large Language Models (LLMs) are rapidly changing the way software and information systems are engineered, including activities traditionally requiring substantial human expertise. Conceptual modelling is a particularly promising yet challenging area: LLMs can transform natural-language requirements, domain descriptions, and other textual sources into representations such as Entity-Relationship models, UML class diagrams, ontological structures, and relational schemas. But how reliable are these generated models, and what role should LLMs play in the modelling process? This talk presents an overview of the emerging use of LLMs for conceptual modelling, drawing on recent literature and empirical investigations. It reviews current applications and findings, highlighting both the capabilities of LLMs and their limitations in dealing with abstraction, semantics, relevance, consistency, and domain-specific knowledge. Particular attention is given to the role of prompting and human intervention in improving the quality of generated models. Based on these observations, the talk proposes a research roadmap for LLM-supported conceptual modelling. The central argument is that the key research challenge is no longer simply whether LLMs can generate conceptual models. Rather, it is how to design reliable and explainable human-AI collaboration in which LLMs support modelling while human experts remain responsible for abstraction, relevance, validation, and domain interpretation.

Towards Noise-Adaptive Training of GANs: Adding Noise Information to the Objective Function

Justina Ramonaitė, Gražina Korvel

Institute of Data Science and Digital Technologies
Vilnius University

justina.ramonaite@mif.vu.lt

Speech enhancement is a broad research field that covers the improvement of speech signals by reducing or completely removing various distortions and improving the intelligibility and quality of the signal. Different approaches have been proposed, varying from traditional methods based on filtering or other mathematical and statistical operations to deep learning. The latter includes generative adversarial networks (GANs). The first GAN-type models used for speech enhancement were introduced in 2017 and have improved greatly since then. They have demonstrated significant improvement in the most common quality evaluation metrics, such as PESQ and STOI. This work focuses on 2 denoising models: SEGAN, which is one of the first GANs used for speech enhancement, and CMGAN, which was introduced in 2022 and has one of the highest reported results. This study investigates the possibility of improving the performance of GANs by including noise-related data into the training objective. The main focus of this research is on developing a loss function that would adjust according to the selected properties of initial noise. This work extends the previous study, which showed that adding scale-invariant signal-to-noise ratio component to the objective function improves the results of the model, and investigates the effect of including additional information. An acoustic analysis of different noise types is conducted to investigate their characteristics. The experiments show promising results, maintaining the improvement of the models and also revealing possible further directions toward more elaborate noise evaluation and profiling.

Acknowledgements. This project has received funding from the Research Council of Lithuania (LMTLT), agreement No S-MIP-24-118. The authors are also thankful for the high-performance computing resources provided by the Information Technology Research Center of Vilnius University.

Impact of Diffusion-Denoising-Based ECG Augmentation on the Robustness of Beat-Level Arrhythmia Classification

Jonas Mindaugas Rimšelis[1], Žygimantas Abramikas[1, 2], Zigmantas Kęstutis Juškevičius[1, 3], Jurgita Markevičiūtė[1], Viktor Medvedev[1], Simonas Šimėnas[1], Povilas Treigys[1], Jolita Bernatavičienė[1]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Clinic of Cardiovascular Diseases, Institute of Clinical Medicine
Vilnius University
[3] UAB Zive

jonas.rimselis@mif.vu.lt

Undiagnosed cardiac arrhythmias can lead to serious cardiovascular complications, while continuous Holter monitoring generates large volumes of electrocardiogram (ECG) data that are impractical for exhaustive manual analysis [1]. Deep-learning-based beat-level arrhythmia classification can reduce this burden [2]; however, ambulatory ECG recordings are frequently contaminated by baseline wander (BW), muscle artefacts (MA), and electrode-motion (EM) artefacts, which distort cardiac morphology and reduce classification performance [3]. Diffusion-based denoising, such as DeScoD-ECG, can partially recover classification performance from contaminated ECG, but its iterative inference procedure limits its applicability as a real-time preprocessing step [4].

We investigate whether diffusion restoration can instead be exploited during training to improve robustness without requiring diffusion inference during deployment [5]. Using the MIT-BIH Arrhythmia Database under an inter-patient evaluation protocol with AAMI normal-like (N), supraventricular ectopic (S), and ventricular ectopic (V) beat classes, each clean training beat was paired with a randomly sampled counterpart from a diffusion-restoration trajectory comprising noise-contaminated, partially denoised, and fully denoised ECG states. The classifier was trained on both clean and transformed beats, with an additional consistency loss between their predictions. A separate model was trained for each noise type. Robustness was evaluated under BW, MA, and EM across increasing

contamination intensities and compared with a conventionally trained classifier evaluated on clean, noisy, and diffusion-denoised ECG.

Diffusion-based augmentation demonstrated class-dependent effects on classification robustness. For S beats, augmented classifiers achieved the highest F1 scores in all 12 evaluated noise conditions. At the highest contamination intensity (1.5–2.0), S-class F1 increased from 0.591 to 0.697 under MA and from 0.486 to 0.668 under EM, compared with conventional training evaluated on noisy ECG. In contrast, conventional training combined with inference-time denoising outperformed augmentation-based approaches for V beats in 10 of 12 conditions.

These findings indicate that diffusion-derived augmentation improves beat-level ECG classification robustness, particularly for S beats, without requiring diffusion-based preprocessing at deployment. However, V-beat classification generally benefited more from inference-time denoising, suggesting that augmentation may adversely affect ventricular beat discrimination. This class-dependent trade-off suggests potential limitations of augmentation as a standalone alternative to inference-time restoration.

Acknowledgements. This project has received funding from the Research Council of Lithuania (LMTLT), agreement No. S-ITP-25-9.

References

[1] Z. Kotsialou, N. Makris, and S. Gall, "Fundamentals of the electrocardiogram and common cardiac arrhythmias," Anaesth. Intensive Care Med., vol. 25, no. 3, pp. 219–222, Mar. 2024, doi: 10.1016/j.mpaic.2023.11.014.

[2] F. Zhou, Y. Sun, and Y. Wang, "Inter-patient ECG arrhythmia heartbeat classification network based on multiscale convolution and FCBA," Biomed. Signal Process. Control, vol. 90, p. 105789, Apr. 2024, doi: 10.1016/j.bspc.2023.105789.

[3] Y. Jia et al., "Preprocessing and Denoising Techniques for Electrocardiography and Magnetocardiography: A Review," Bioengineering, vol. 11, no. 11, p. 1109, Nov. 2024, doi: 10.3390/bioengineering11111109.

[4] H. Li, G. Ditzler, J. Roveda, and A. Li, "DeScoD-ECG: Deep Score-Based Diffusion Model for ECG Baseline Wander and Noise Removal," IEEE J. Biomed. Health Inform., vol. 28, no. 9, pp. 5081–5091, Sep. 2024, doi: 10.1109/JBHI.2023.3237712.

[5] C. S. Sastry, S. H. Dumpala, and S. Oore, "DiffAug: A Diffuse-and-Denoise Augmentation for Training Robust Classifiers," in Advances in Neural Information Processing Systems, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., Curran Associates, Inc., 2024, pp. 20745–20785. doi: 10.52202/079017-0654.

An Ordinal Transport Framework for Paired Likert-Scale Responses

Virgilijus Sakalauskas, Dalia Kriksciuniene

Kauno kolegija Higher Education Institution

virgilijus.sakalauskas@go.kauko.lt

Likert-scale questionnaires are commonly analysed using response frequencies, mean or median scores, rank correlations, reliability coefficients, and latent-variable models. Although these methods provide useful information, they do not directly distinguish between systematic differences in item response distributions and disagreement occurring at the level of individual respondents. This paper proposes Paired Ordinal Transport Analysis (POTA), a framework for analysing Likert-scale items answered by the same respondents. POTA combines respondent-level pairing with one-dimensional optimal transport. For each pair of questionnaire items, the method first measures the observed ordinal difference between the answers provided by each respondent. It then determines the minimum transport cost required to transform the marginal response distribution of one item into that of the other. Comparing these two quantities separates the total observed item dissimilarity into two interpretable components. The first represents a systematic shift between the response distributions, such as a general tendency for one item to receive higher or lower ratings. The second represents respondent-level discordance that cannot be explained by differences between the marginal distributions. This decomposition makes it possible to distinguish item pairs that appear similar according to their frequencies, average scores, or correlations but differ substantially in how individual respondents evaluate them. Item-level indicators can be obtained by averaging the pairwise components across the questionnaire. They can be used to detect items with unusual response profiles, heterogeneous interpretation, possible thematic misclassification, or redundant content. Alternative category-boundary weights can also be applied when equal spacing between adjacent Likert categories is considered unrealistic. An experimental survey illustrates the interpretation and diagnostic value of the

proposed framework without relying on domain-specific assumptions. The results indicate that POTA complements conventional questionnaire methods by revealing whether item differences arise primarily from systematic distributional shifts or from heterogeneous respondent-level evaluations. Therefore, the framework provides an interpretable and flexible tool for questionnaire diagnostics, item comparison, and scale development.

A Lithuanian-Specific E5 Text Embedding Model

Lurda Songailaitė[1, 2], Justinas Juozas Dainauskas[1, 2], Milita Songailaitė[1, 2], Tomas Krilavičius[1,2]

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

lurda.songailaite@card-ai.eu

In recent years, natural language processing has made some big leaps, mainly thanks to the development of high-quality text embedding models. One downside is that the progress has been uneven across languages. Lithuanian is an official European Union language with over three million speakers worldwide. Despite this, it remains understudied in the NLP field. Currently, no high-quality semantic embedding model specifically adapted to Lithuanian exists, meaning Lithuanian language information retrieval, semantic search, or other applications need to rely on general multilingual models. These models are trained on dozens of languages at once, with only a small share of Lithuanian data, and therefore capture it only partly. The Lithuanian language being morphologically rich, where words frequently change form depending on grammatical context, is another problem. This means that models trained mostly on English or other morphologically simpler languages, even if Lithuanian is a part of their training data, will have a hard time capturing all of the Lithuanian word forms or their meanings. In this work, this gap between the widely used languages and Lithuanian will be covered by training and evaluating an E5 embedding model specifically for the Lithuanian language. We propose a new Lithuanian E5 model which is fine-tuned from a multilingual E5 base model [1]. The first step uses weak supervision over a large number of Lithuanian query and passage pairs, which were taken from multiple datasets, including The General Lithuanian Language Corpus (BLKT) [2]. This dataset spans over 8.4 million Lithuanian texts covering news portals, legal and administrative documents, internet texts, speech transcriptions, and fiction. The second step fine-tunes using hard negatives taken from a subset of higher-quality pairs made from the same datasets, including a

Lithuanian question-answer corpus of over 10 million machine-generated pairs [3] and a human-written summarisation corpus covering IT, law, medicine, and news domains [4]. The biggest challenge faced in this work was the lack of Lithuanian evaluation benchmarks, like Lithuanian STS (Semantic Textual Similarity) datasets. Given this, the model was assessed with a retrieval-based evaluation set made from the human-written summarisation corpus [4], using document abstracts as queries and full texts as passages. By providing a dedicated embedding model and an evaluation set for Lithuanian, this work aims to help Lithuanian become a more equal participant in the modern NLP field.

References

[1] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual E5 text embeddings: A technical report," arXiv preprint arXiv:2402.05672, 2024.

[2] State Digital Solutions Agency, "General Lithuanian language corpus," Hugging Face, 2026. [Online]. Available: https://huggingface.co/datasets/VSSA-SDSA/LT_AI_BLKT.

[3] UAB TETRAGRAMA, "Question-answer corpus," State Digital Solutions Agency, CLARIN-LT Repository, Apr. 2026. [Online]. Available: https://hdl.handle.net/20.500.11821/99.

[4] Vytautas Magnus University and Vilnius University, "Abstract corpora for artificial intelligence," Hugging Face, 2026. [Online]. Available: https://huggingface.co/datasets/VytautoDidziojoUniversitetas/LT_Summarisation_Corpus.

How Well Do LLMs Detect Disinformation Narratives? Retrieval vs. Zero-Shot on Familiar and Unseen News

Milita Songailaitė[1,2], Tomas Krilavičius[1,2]

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

milita.songailaite@vdu.lt

Analysts monitoring the information environment face more content than they can review manually, and large language models (LLMs) are increasingly proposed as a first screening step. Coordinated disinformation campaigns also adapt quickly to breaking events, so new narratives often circulate widely before any analyst has had the chance to catalogue them. This puts pressure on a common assumption behind LLM-based disinformation screening, that grounding a model in catalogued disinformation cases improves detection, while a second assumption, that large proprietary models are needed, also remains largely untested. This study tests both on binary disinformation classification of English news article bodies with all metadata removed. Two pipelines are compared: a zero-shot classifier that judges an article by its claims, tone, sourcing and framing, and a retrieval-augmented generation (RAG) narrative detector that scores the article against 15 cases retrieved from 3,676 EUvsDisinfo records (2015 to 2025). Twenty-five open-weight and proprietary model variants were evaluated on 100 articles from 2018 to 2025, and 15 models with pre-2026 training data on 77 articles published in 2026. Six human annotators labelled a 30-article subset. On familiar news, RAG scored higher for 18 of 25 models (mean F1 0.87 against 0.79 for zero-shot), mainly by catching more disinformation. On post-2026 news, RAG F1 fell for all 15 pre-2026 models, from a mean of 0.83 to 0.55, while zero-shot remained stable (0.74 to 0.73). The best open-weight model, run on one workstation with 32 GB of VRAM, stayed within 0.06 F1 of the best proprietary model in every condition. In the annotation study, the best human reached an F1 of about 0.97 and the best model run about 0.90, while the average human (0.79) and average model run (0.74) were close. Justifications

differed: models cited emotive wording in 78% of explanations against 17% for humans, whereas humans more often checked outside sources. These results suggest that case-based retrieval is effective for recognising known narratives but does not transfer to new events, that zero-shot screening is the more robust option for emerging narratives, and that local open-weight models are a viable choice where sensitive data must stay in-house. Model output appears best suited to supporting, not replacing, analyst review, although the benchmarks are small and the annotation results are preliminary.

On Enhancing Virtual Machine Placement Balancing Algorithms for Highly Overbooked Cloud Infrastructures

Mikita Syravatnikau

Department of Mathematical Modelling
Vilnius Gediminas Technical University

mikita.syravatnikau@vilniustech.lt

Effective load balancing helps maintain application performance and use resources efficiently, even with many virtual machines sharing the same physical servers. In cloud environments, resource overbooking helps reduce infrastructure costs by allowing total resource allocations to exceed physical capacity, since virtual machines do not usually need all their allocated resources at the same time. However, traditional load balancing algorithms that do not account for overbooking can become less effective as competition for resources increases. Sudden or simultaneous increases in demand may cause long periods of server overload, resource shortages, and reduced application performance. These risks make it important to consider how resource demand changes over time when making placement decisions. This presentation focuses on proactive load balancing under high resource overbooking. It examines how past load measurements can improve decisions about where to place new virtual machines. Load history provides information about how resource use changes over time and can support estimates of upcoming demand. The aim is to use this information to choose placements that reduce the risk of future overload. Reactive balancing, which responds to the current system state and detected overload, provides a point of comparison. The presentation introduces a model of resource overbooking to support this approach. The model separates allocated resource limits, actual demand, and the resources that virtual machines receive under physical capacity constraints. This distinction helps explain why measured resource use may hide unmet demand when a server is already working at full capacity. The model provides a basis for

evaluating how placement decisions based on load history affect resource availability as workloads change. The proposed evaluation considers resource use, load distribution, total server overload time, and the share of resource demand that cannot be met. Total overload time is calculated by adding up, across all servers, the time during which demand for at least one resource exceeds the server's capacity. Together, these metrics describe the duration and severity of resource shortages that can cause performance degradation. Repeatable simulation experiments provide a basis for this evaluation. Preliminary results from an experiment with synthetic workloads suggest that using past load measurements can increase load balancing efficiency compared with using only the current load. The benefit depends on workload characteristics.

Hidden Market Regimes in Daily Cryptocurrency Data: HMM-Based Forecasting and Trading Evaluation

Raigardas Tautkus, Mantas Vaitonis

Institute of Social Sciences and Applied Informatics, Kaunas Faculty
Vilnius University

raigardas.tautkus@knf.stud.vu.lt

This experiment investigates whether Hidden Markov Models (HMMs) can identify latent market regimes in daily cryptocurrency data and whether these regimes provide useful information for return forecasting and trading. Cryptocurrency markets are characterized by high volatility and changing market conditions, making them a suitable setting for investigating models that explicitly account for latent regime changes. Daily OHLCV data for Bitcoin, Ethereum and Litecoin covering the period from January 2019 to January 2026 are transformed into four features: log return, intraday range, volume change and 20-day rolling volatility. The resulting features are standardized and used to estimate a multivariate Gaussian HMM, while preserving the chronological structure of each cryptocurrency time series. The data are divided chronologically into training, validation and test sets using a 60/20/20 split. HMMs with different numbers of hidden states and covariance structures are evaluated, model selection is based on the Bayesian Information Criterion (BIC). The selected model is a six-state Gaussian HMM with diagonal covariance, identifying latent regimes with distinct characteristics in terms of return, volatility, persistence and frequency. The most frequently observed regime accounts for 34.06% of observations and shows positive average returns and an average duration of approximately 9.4 days. A rare regime representing 7.38% of observations shows strongly negative average returns in contrast and much higher volatility. Another low-frequency regime shows the highest average return and the longest average duration, finding that regime frequency and persistence can differ considerably. The identified regimes are evaluated from both statistical forecasting and economic perspectives. Out-of-sample one-step-ahead return forecasts are

generated using filtered state probabilities, the estimated transition matrix and state-conditional expected returns. Forecast performance is evaluated using root mean squared error, normalized RMSE and directional hit ratio,Directional accuracy remains close to 50% for all three cryptocurrencies. A separate trading experiment evaluates a pre-specified HMM-based long/cash strategy against a buy-and-hold strategy, incorporating transaction fees and slippage. The trading results are asset-dependent: the HMM strategy outperformed buy-and-hold for Ethereum and produced a 17.54 percentage-point excess return, but underperforms buy-and-hold for Bitcoin and Litecoin. These results show that identifying persistent and economically distinct market regimes does not necessarily translate into better return forecasts or consistent trading performance. Overall, the study demonstrates that HMMs can provide a useful probabilistic characterization of changing cryptocurrency market conditions, while also highlighting the distinction between successful regime identification, predictive accuracy and economic profitability.

Data Balance Analysis for Lithuanian and English NLP

Deividas Valiuška, Justina Mandravickaitė, Milita Songailaitė, Tomas Krilavičius

Department of Informatics
Vytautas Magnus University

deividas.valiuska@vdu.lt

Dataset defects can introduce shortcuts and leakage that inflate evaluation results and obscure model performance. We present a data balancing framework that detects data quality issues before training, including strong associations between words and labels, systematic differences in text length between classes, and overlap between training and test items. The framework reports each finding with its evidence and a recommended fix. The data balancing framework covers label balance (imbalance ratio, minority-class size, split consistency, and multi-label measures), distribution balance (length as a label predictor, near-duplicates, train–test leakage, and conflicting labels), lexical leakage, and subgroup balance (disparate impact, statistical parity, and Cramér's V). For lexical leakage, we propose a length-aware enrichment test, a length-weighted binomial test that, like Fisher's exact test, conditions on each word's overall frequency. It's designed to avoid both missed shortcuts in minority classes and false findings caused by differences in text length. The test is also applied to lemmas to account for Lithuanian inflection and expose shortcuts distributed across multiple word forms. Each threshold records its basis, of 25 verdict thresholds, 8 follow an external standard (one legal) and 16 are provisional. On 80 clean synthetic datasets that were generated by the framework itself (every class draws made-up words from the same Zipf distribution, so no word, length or subgroup is tied to a label), only 3 of 320 method runs raised a finding, all warnings, in line with the 5% significance level. With six defect types planted at increasing strength, verdicts never weakened as the defect grew (mean Spearman $\rho = 1.00$, minimum 0.94). The widely used Monroe log-odds test missed a majority-class shortcut in every imbalanced trial. Our test found all planted shortcuts with no false ones. Applied to a Lithuanian

hate-speech corpus of 27,357 comments, it identified several previously unnoticed issues, including scraped website-interface text in 9.4% of hate comments versus under 0.6% of other classes, 46 evaluation items with an exact or near copy in training, near-identical comments with conflicting labels, lemma-level lexical shortcuts, and an oversampled "balanced" corpus variant that is 43% exact copies. All findings were manually confirmed. The interface artefact drives a fine-tuned classifier: appending it to non-hate comments raised "hate" predictions from 1.2% to 99.2% (neutral control text: 0.7%). In a Lithuanian fake-news corpus, nearly a third of the "true" class are articles from a single fact-checking column. These results show that inexpensive, interpretable pre-training checks can expose dataset defects that might otherwise lead to misleadingly high evaluation scores.

Unsupervised Graph Neural Network Anomaly Detection on Enterprise Firewall Traffic

Deividas Valiuška, Nerijus Šakinis, Eimantas Zaranka, Justinas Juozas Dainauskas, Rimantė Povilavičienė, Tomas Krilavičius

Department of Informatics
Vytautas Magnus University

justinas.dainauskas@vdu.lt

In this research, we adapted and compared several unsupervised graph-based anomaly detection methods on real enterprise firewall traffic and Sysmon logs. Our main contribution is the adaptation of existing methods, originally designed for offline, static settings, to run without labels in an online, batched-window manner on streaming logs, where malicious or unusual behaviour is a tiny fraction of otherwise ordinary flows. We represented network traffic as a graph in which hosts are nodes and communications are edges, letting the models learn from structural context rather than isolated flow features. We applied four methods to firewall traffic (Anomal-E, DOMINANT, CONAD and OCGNN) and one to Sysmon user activity logs (Log2Vec). Anomal-E adapts the edge-oriented E-GraphSAGE architecture, trained self-supervised with Deep Graph Infomax (DGI), to embed individual flows. The embeddings are scored by an ensemble of Histogram-Based Outlier Score (HBOS), PCA and Isolation Forest detectors, and a threshold based on the interquartile range (IQR) is set once and frozen for all later days. DOMINANT and CONAD are reconstruction-based: DOMINANT combines structure and attribute errors via a dual decoder, while CONAD adds contrastive learning with synthetic anomaly injection. OCGNN takes a one-class approach, scoring nodes by their distance to the centre of a latent hypersphere. These three methods use a contamination threshold of 0.1%. Log2Vec builds a heterogeneous graph of user log events, generates behavioural sequences by weighted random walks, embeds them with a skip-gram Word2Vec model, and flags events that deviate from dominant behaviour, ranking hosts by their count of flagged events. We evaluated all methods

over two-week periods. Anomal-E was trained self-supervised on the first week, which served as the baseline, and then scored both weeks day by day. DOMINANT, CONAD and OCGNN required no pre-training and were fitted on each 5-minute window. Log2Vec was rebuilt for each day. We ran the full workflow on Apache Spark on Kubernetes to handle volumes exceeding a single machine's capacity, and analysed detected anomalies per host to support expert review. Experts reviewed the 10 most anomalous hosts from Log2Vec and the 4 most anomalous hosts from each of DOMINANT, CONAD and OCGNN. Anomal-E was excluded from expert review because its input data resembled that of the other three methods and expert time was limited. Log2Vec revealed no real intrusions, although 3 hosts appeared anomalous and required deeper inspection. DOMINANT and CONAD flagged the internal attacker host of a previously known attack but missed the attacked host, and additionally flagged one non-malicious host whose behaviour experts judged genuinely anomalous. Experts rated these two methods the best. OCGNN flagged the attacked host but missed the attacker host. Because the firewall-based methods detected complementary parts of the known attack, we expect an ensemble to give a more complete view of anomalous events.

From Compliance to Control: Governing GenAI Through the CISO Function

Ilona Veitaitė, Rasa Brūzgienė,
Ieva Andrijauskaitė, Donatas Misiūnas

Vilnius University

ilona.veitaite@knf.vu.lt

Organizations that use generative AI (GenAI) are facing governance gaps that traditional cybersecurity frameworks were never designed to address. This is because large language models can produce convincing but incorrect outputs, can be manipulated by adversaries, and can degrade in reliability over time. This research argues that supervising such systems requires a permanent supervised socio-technical architecture, not just simple periodic technical audits, and we investigate how the position of the Chief Information Security Officer (CISO) should be redefined to meet this requirement in the face of a converging set of EU rules – the AI Act, GDPR, the Cyber Resilience Act (CRA) and the NIS2 Directive. The research concerns the Matrix-Plus-Map (M_and_M) approach, a two-level governance model developed within the GAISO research project. The "Matrix" part plots twelve fundamental CISO functions – from threat intelligence and compliance mapping to incident response, LLM-specific threat detection, access governance and workforce training – against provisions from NIS2, the AI Act, CRA, ISO 37000:2021 and ISO/IEC 27002:2022 controls, demonstrating that each function has traceable obligations under multiple overlapping instruments simultaneously. The 'Map' element offers a decision-logic overlay to distinguish the tasks artificial intelligence can safely perform from those that require human authority, against three tests: whether consequences are reversible; whether personal or organizational liability attaches under regulation; and whether the decision requires contextual or ethical judgment that is beyond current AI capability. The research places human-AI interaction in four phases of governance throughout the system lifecycle: pre-deployment identification, two active-operation stages (contextualization and layering), and post-incident execution and validation. A companion procedural workflow shows the CISO defining use cases and GenAI aggregating regulatory context, decomposing tasks and flagging review checkpoints, so compliance is continually re-verified,

rather than at initial deployment. The same logic applied across all 12 functions yields three levels of AI augmentation: threat-feed analysis and training-scenario generation are in the 60-80% range, policy mapping and access-control functions are in the 40-60% band, and incident response / recovery functions are human-dominant at 2-40% automation because of their real-time, high-stakes nature. The framework is applied to two contrasting sector cases, with differing regulatory density and operational risk. A university case on GenAI-assisted classification of research-data sensitivity engages a modest set of CISO functions and human decision points, governed mainly by NIS2 and GDPR, with AI taking on a substantial share of routine tasks such as drift monitoring and report generation. A FinTech example of real-time fraud detection and credit-risk scoring, which is a high-risk use case under the AI Act and also regulated by the CRA and DORA, triggers a wider range of CISO functions and decisions. Here AI automates the bulk of the fraud scoring work, but mandatory human sign-off is still required for large blocks of transactions, responses to model drift and notifications to multiple authorities for incidents. A comparative analysis of the two cases shows that the heavier oversight burden in the FinTech case is a result of denser regulation and cumulative reporting obligations, and not a deficiency of the AI's performance. The analysis yields four takeaways. First, no CISO function is fully automatable, irrespective of how data-intensive it may appear. Second, overlapping EU regulations generate compound accountability, whereby a single decision may simultaneously trigger obligations under multiple legal instruments. Third, ISO 37000, though not explicitly naming CISO duties, offers the governance scaffolding necessary to translate legal text into operational practice. Finally, the level of required oversight by humans is determined by sector-specific risk and regulatory density, not technological limitations. The authors note that their control mapping is specific to ISO/IEC 27002:2022 (other frameworks such as NIST or CIS Controls might produce different mappings), that automation percentages are a snapshot in time and likely to change as AI capabilities improve, and that both use cases are illustrative and require longer-term field validation.

Acknowledgments. The research was funded by the project "Research on Cyber Resilience Through Application of Generative Artificial Intelligence in Chief Information Security Officer Operations" (GAISO), which has received funding from the Research Council of Lithuania (LMTLT), under agreement No S-ITP-24-13.

Challenge-Based Learning in Cyber Security Education Within a Federated Education Environment

Algimantas Venčkauskas[1], Raimundas Matulevičius[2], André Zúquete[3], Csaba-Zoltán Kertesz[4]

[1] Kaunas University of Technology
[2] Tartu Ülikool, Estonia
[3] Universitatea Transilvania din Brașov, Romania
[4] Universidade de Aveiro, Portugal

algvenc@ktu.lt

The rapidly evolving cyber threat landscape demands cybersecurity professionals equipped not only with theoretical knowledge, but also with advanced practical problem-solving skills, adaptability, and cross-institutional collaboration capabilities. Traditional instructional methods often fail to mirror the complex, distributed, and real-time nature of modern cyber security problems. Cybersecurity studies often face severe operational barriers: institutions cannot safely share raw network logs, proprietary security data, or live threat telemetry due to privacy laws and security risks. A federated infrastructure solves this through decentralized architecture. Applying Challenge-Based Learning (CBL) within a Federated Education Environment (FEE) for a cybersecurity study addresses the core pedagogical, technical, and regulatory challenges of training modern cybersecurity professionals. Combining these two methodologies creates a collaborative, practical, and privacy-compliant learning model. Project CYBERCHALLENGE (https://www.cyberchallenge-erasmus.eu/) focuses on innovative learning by using CBL methodology, MOOC and FEE and its implementation into practice, which brings a collaboration aspect with business organizations on solving challenges raised in the business sector. CYBERCHALLENGE results will be developed: R1 – A four-module Cybersecurity study program based on the MOOCs methodology and uses CBL and sandbox approaches: Fundamentals of Cybersecurity, Identity and Access Management, Cyberattacks and Cyber Defenses: Tools and Methods, and Cyber Security Management. The modules were implemented in the partners' Moodle environments.

R2 – Federated Cyber security lab based on tools sandbox. Federated Cyber Security Lab combines virtual labs of partners. They are connected through a common portal - Federated Lab Capital. R3 – CBL scenarios are created for each module: Fundamentals of Cybersecurity: The security of smart home systems; Identity and Access Management: Secure Access and Identity in a Federated Ecosystem; Cyber Attacks and Cyber Defences: Improving the infrastructure of a partner company by simulating various types of attacks and proposing cyber defence mechanisms to mitigate the identified issues; Cyber Security Management: Scooter Ride-Hailing System. The conducted case study and the voluntary anonymous survey of students with little to moderate experience show that the synergy of CBL and the FEE strengthens practical cybersecurity, identity management, threat detection and incident response capabilities and optimize partners' resources. The presentation also discusses the challenges of task standardization and assessment and provides guidelines for the modernization of cybersecurity study programs.

Acknowledgements. This work has received funding from the European Union's Erasmus+, KA220-HED – Cooperation partnerships in higher education Programme under project Challenges Solving in Cybersecurity Study Program (Grant Agreement No. 2024-1-LT01-KA220-HED-000252504).

Manoeuvre Compatibility Networks to Support Autonomous Ship Decisions

Julius Venskus, Robertas Jurkus

Vilnius University

robertas.jurkus@mif.vu.lt

When considering their movements, autonomous maritime vessels should think about the effect that these have on other nearby vessels; a move which may seem appropriate if each neighbour is looked at individually might not end up forming a viable plan when all the ships are taken into account. In order to examine this issue the study makes use of network theory and Automatic Identification System (AIS) data in the case of three-vessel encounters. We looked at 112 days' worth of AIS observations and, after applying the data-quality and traffic filters, only kept a small number of encounters, namely 11 decision moments. We constructed networks in which the nodes stand for possible manoeuvres and the triangles show the compatible plans for all three ships. Then we compared the decisions made on the basis of individual ship pairs with those based on the entire network. Once the first manoeuvre had been decided, both approaches chose the same actions at the first possible opportunity during each encounter. When the available manoeuvres had been improved at a later stage, the pairwise method opted for a 30-degree change in course and had no compatible three-ship plan. The network method, on the other hand, chose a 22.5-degree change which was part of nine compatible plans. Thirty seconds later, however, none of the nine manoeuvre sets tested offered a common plan under the model assumptions. Another experiment looked at the case of no communication, communication between all three ships, and communication only between a pair that had been selected at random. When reliable and immediate communication was assumed, agreement among all three ships led to the adoption of a common plan in seven out of the nine manoeuvre scenarios at an earlier time. Communication only between a pair resulted in a joint plan in fewer cases and this depended on which ships were in communication and on how the actions of the other ship were predicted. The findings, which relate to two specific

instances during a single interaction, are still in an exploratory phase and illustrate the way in which network analysis can show the limitations of assessing pairs of elements and allow one to tell the difference between a jointly available plan and the coordination required to implement it. The communication scenarios have been simulated, but more extensive testing involving realistic vessel behaviour and data uncertainty is necessary before any conclusions can be drawn regarding practical safety.

Acknowledgements. The project has been funded by the Research Council of Lithuania (LMTLT) under agreement number S-MIP-24-117.

Assessing Maritime Interaction Networks under Prediction Uncertainty with a Complex Network Theory Approach

Julius Venskus[1,2], Robertas Jurkus[1,2], Povilas Treigys[1], Kasparas Kaukas[1], Clinton Iwuoha[1]

[1] Vilnius University
[2] Klaipeda University

julius.venskus@gmail.com

Complex Network Theory offers a framework for analyzing maritime traffic as a system of interacting vessels, finding patterns of relation and identifying traffic important ships. However, the conclusions made from those networks relates on how vessel interactions are represented. When interactions are inferred from trajectory predictions, the prediction uncertainty may introduce additional related connections and change the vessels prioritized for monitoring. This build on our earlier work with maritime interaction modelling and network feature extraction. This study research the robustness of network-based assessments of trajectory prediction uncertainty. The focus shifts from describing vessel interactions to evaluating whether different uncertainty representations produce consistent vessel rankings and how greater network connectivity improves monitoring priorities. Analysis compares three networks constructed from identical trajectory predictions. A point-based network, a network based on Gaussian prediction ellipses, and a network based on conformally calibrated ellipses. Degree centrality, clustering, connected components, and the stability of priority vessel sets are analyzed. Reference networks reconstructed from subsequently observed AIS positions provide a basis for evaluating predicted interactions. Comparisons with matched edge counts help distinguish changes in network structure from the effects of increased network density. In this scenario, conformal prediction regions increased interaction recall from 60% to 82%. The top 10% of vessels ranked by predicted degree accounted for 57% of observed interactions, compared with 48% for point forecasts. After edge counts were matched, this proportion decreased to 49%. This suggests that increased network

density explains much of the supposed improvement. The key result is that accounting for prediction uncertainty may substantially change network topology and vessel rankings without yielding a comparable improvement in monitoring performance. The proposed contribution extends earlier interaction modelling by researching the stability and practical relevance of the conclusions derived from maritime networks.

Acknowledgements. The project has been funded by the Research Council of Lithuania (LMTLT) under agreement number S-MIP-24-117.

Design Pattern Specification Templates for Coordination Methods of Collectives: A Critical Analysis

Tadas Verbyla, Audronė Lupeikienė

Faculty of Mathematics and Informatics
Vilnius University

audrone.lupeikiene@mif.vu.lt

Collectives consist of numerous agents working together to achieve common objective, relying on communication, cooperation, and decentralised coordination, resulting in self-organization, what is also called emerging behaviour. The coordination method here is highly significant because it forms the foundation for the creation of collectives and collective intelligence. This means that its selection and implementation constitute one of the most important stages in the engineering of systems conceptualised as collectives.

Design patterns are the blueprints for solving *common problems in systems development. Reuse technology in the form of using patterns* helps save time, avoid errors as they have been tested in various situations, making them generally reliable and reusable solutions. The use of design patterns is ensured by the ability to completely describe and find them; that is, a necessary and sufficient template for their specification – in other words, a design pattern meta-model – is required. *The use of design patterns is a well-established practice in software development [1]. However, this is not the general rule in developing other types of systems. In addition, s*uch reuse of solutions across various levels and stages of system development is still an emerging discipline.

This is the critical analysis paper that summarizes the results of the assessment of **Pattern-Oriented Software Architecture** (POSA) and Gang of Four Design (GoF) patterns. The main objective of this work is to answer the question: "Are the proposed design pattern specification templates sufficient to describe the coordination of systems conceptualized as collectives?". The need to find such answer is argued by the two facts. First, although state-of-the-art literature reviews have been conducted, they focus on specific sub-fields or on solving problems

related to software systems [2], i. e. on specific paradigms. Second, it was proven that there exist paradigm-independent design problems, described as design patterns [3]. Thus, this article contributes to the theory of creating systems as collectives by formulating requirements for design template in the context coordination problem, by defining templates for different categories of POSA and GoF patterns and by determining a correspondence between requirements and identified templates.

References

[1] E. Gamma, R. Helm, R. Johnson, J. Vlissides. Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley Professional, 617 p., 1994.

[2] E. Kesavan. The Evolution of Software Design Patterns: An In-Depth Review. International Journal of Innovations in Science Engineering and Management, vol. 4, no. 1, 2025, 163–167.

[3] Z. Vaira, A. Caplinskas. Software Engineering Paradigm Independent Design Problems, GoF 23 Design Patterns, and Aspect Design. Informatica, vol. 22, no. 2, 2011, 289–317.

Machine Learning for Laryngeal Cancer Detection Using Endoscopic Images, Voice Signals, and Multimodal Data

Adriana Voišvilo

Institute of Data Science and Digital Technologies
Vilnius University

adriana.sirokyte@mif.stud.vu.lt

Machine learning can support laryngeal cancer detection by analysing endoscopic images and voice signals. Combining both data sources in a multimodal model may provide additional diagnostic information. However, developing and evaluating such models is challenging because publicly available datasets rarely contain matched images and voice recordings from the same patients. This work combines a review of scientific literature with an experiment using two public datasets: CE-NBI endoscopic images and the Saarbrücken Voice Database. Separate machine learning models will be trained and evaluated using the available data, while previous studies combining images and voice signals will be reviewed to explore the potential of multimodal approaches. This presentation covers:

- Literature review: previous studies using image-based, voice-bases, and multimodal machine learning approaches for laryngeal cancer detection.
- Datasets and data preparation: how the image and audio datasets differ and how are they prepared before training the models.
- Experiment: training and testing separate machine learning models using the available image and voice data. The experiment will also consider the importance of data quality, class balance, and appropriate training and testing procedures. Since multiple images may come from the same patient, patient-level data splitting will be important to reduce the risk of data leakage and obtain more reliable results.
- Model evaluation: comparing model performance using metrics, such as accuracy, sensitivity, specificity, F1-score, and ROC-AUC.

- Main challenges: limited data, differences between datasets, the lack of matched image and voice recordings.
- Future directions: potential for monitoring patients before and after treatment. The aim is to investigate the performance and limitations of image-based and voice-based models on their own datasets, review existing multimodal approaches, and identify what data would be needed to develop and evaluate models that combine both sources. Future research could explore collecting endoscopic images and voice recordings from the same patients before and after treatment to investigate whether these data could help monitor changes in their condition.

Edge AI in Data Processing

Joachim Włodarz

Faculty of Science and Technology
University of Silesia, Poland

joachim.wlodarz@us.edu.pl

Edge AI enables the deployment of artificial intelligence algorithms and the respective machine learning models directly on edge hardware devices like IoT sensors, cameras, etc., rather than relying mainly on cloud resources. In this approach only the resource-hungry model training takes place in the cloud, while the real-time inference and also the decision making occur right where the data is generated. Processing data locally on the edge devices has many operational advantages, such as bandwidth reduction, near-zero latency, energy efficiency, offline reliability, and also enhanced privacy. Moreover, by eliminating the round-trip lag of cloud servers, it enables it to react almost instantly, allowing critical applications to function safely and continuously, even without network connectivity. In the present contribution our recent experiences in this field are briefly summarized and discussed.

Assessment of Answer Reliability of a Cybersecurity Knowledge Graph Under Software Vulnerability Abstraction Perturbation

Dovydas Marius Zapkus, Asta Slotkienė

Institute of Data Science and Digital Technologies
Vilnius University

marius.zapkus@mif.vu.lt

The prevention of and defense against cybersecurity threats increasingly rely on structured, openly maintained security knowledge. Knowledge about software vulnerabilities (CVE), their root-cause weaknesses (CWE), and the associated attack patterns (CAPEC) is increasingly represented as a knowledge graph. The reliability of these graphs is, however, fundamentally constrained by the quality of the CVE-to-CWE assignments upon which they are constructed. Şimşek et al.'s research demonstrated that a large proportion of NVD records exhibit invalid, placeholder, or insufficiently specific CWE mappings [1]. Further studies [2,3] show that, contrary to MITRE's mapping guidance, vulnerabilities are frequently assigned to overly general (higher-level) weaknesses, with classification errors concentrated among hierarchically adjacent parent–child pairs. According to Mahmud et al. [3], the absence or ambiguity of descriptive CWE properties degrades knowledge-mining results; notably, their evaluation treats a same-branch substitute as approximately correct, implicitly presupposing those annotations are inherited along the CWE hierarchy. We empirically evaluated this assumption on 207 CWE chains (class→base→variant) across 54 sibling groups, measuring semantic proximity with sentence embeddings. A same-branch variant was closer to the true weakness than its parent in about 60% of comparisons, and in 36% after excluding a single group of near-duplicate weaknesses (CWE-41). The decisive factor was intra-group semantic homogeneity: siblings outperformed the parent in 74% of comparisons within homogeneous groups, but only in 20% within heterogeneous groups. Same-branch substitution should therefore not be treated as approximately correct by

default. It is justified only for semantically homogeneous sibling sets, and under classifier uncertainty, the parent constitutes the safer prediction.

References

[1] Şimşek, Ş., Xia, H., Gluck, J., Medina, D. S., & Starobinski, D. (2025, July). Fixing Invalid CVE-CWE Mappings in Threat Databases. In 2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC) (pp. 950-960). IEEE.

[2] Schwengber Kelm, A., Bockermann, C., & Frochte, J. (2026). Multi-Class vs. Multi-Label BERT for CVE-to-CWE Mapping: How Taxonomy Structure Shapes the Errors. arXiv e-prints, arXiv-2607.

[3] Mahmud, A., Rawajfih, Y., Shahriar, H., & Wu, F. (2025, December). Contamination-aware, taxonomy-driven vulnerability classification for CPS: Evidence and insights at scale. In 2025 IEEE International Conference on Big Data (BigData) (pp. 4298-4304). IEEE.

Language Models in Human-Robot Collaboration: Past, Present and Future

Eimantas Zaranka[1,2], Roel Pieters[2], Tomas Krilavičius[1]

[1] Vytautas Magnus University
[2] Tampere University, Finland

eimantas.zaranka@vdu.lt

As robotics shifts towards the Industry 5.0 paradigm emphasizing sustainability and human-centricity. The demand for seamless Human-Robot Collaboration (HRC) has grown alongside a global fleet exceeding 4.6 million active units. Nevertheless, the safe integration of robots into the shared human environment remains the critical bottleneck. Here the main challenge is a high-level semantic reasoning connection with physical execution in a timely manner. This review aims to (1) classify how language models (LMs) and vision-language-action (VLA) models have been integrated into HRC; (2) identify bottlenecks and (3) outline future research directions. We survey research primarily between 2020 and 2026, drawing on peer-reviewed venues such as CoRL, RSS, ICML, NeurIPS, ICLR, AAAI, ICRA, alongside key robotics journals and arXiv preprints. We selected sources relevant to LMs and VLA usage in HRC while excluding human-robot interaction, which yielded corpus of 48 sources. The integration of language models in HRC distinguishes five core domains: (1) high-level task planning and decomposition, with main frameworks being SayCan, LLM+P, TAMP and Voyager; (2) intent estimation and anticipation – GG-LLM, HOI4ABOT and LIT; (3) multimodal communication and explainability – REFLECT, SayPlan; (4) shared autonomy and assistive control – VoicePilot and Assistron; (5) reward design and learning from human feedback – Eureka, ADAPT. Furthermore, since 2023 two main architectural philosophies dominate the answer to the question – how to connect the discrete, semantic reasoning of language models to the continuous, physical requirements of robot control. These architectures are End-to-End VLA, where main models are RT-2, PaLM-E, OpenVLA, Octo, π0 and Gemini robotics, and LLM-as-planner models – SayCan, Inner Monologue, VoxPoser. Across multiple studies three main lineages of datasets are observed for

benchmarking: (1) the manipulation-data, with an Open X-Embodiment dataset; (2) conditioned evaluation, with CALVIN, LIBERO, ManiSkill2, RoboCasa, BEHAVIOR-1K datasets that provide procedural task suites with language annotations; (3) interaction-centric, with Habitat 3.0, TEACh, Watch-And-Help and PARTNR datasets. While analysing the literature multiple bottlenecks are observed: (1) the physical grounding problem, that prevents large models from translating a vast semantic knowledge of abstract reasoning into a precise, physically viable actions; (2) real-time latency, where inference delays prevent autonomous robots navigation and interaction with dynamic environments safely and fluidly with immediate reaction; (3) hallucinations in safety-critical contexts, which makes model confidently generate incorrect or ungrounded outputs creating unacceptable physical risk in safety and precision oriented environments; (4) catastrophic forgetting under continual adaptation, occurring in real world environments where robot is forced to learn new tasks or adapt to a changing surroundings; (5) evaluation and verification gaps that shows a distinct lack of standardized testing frameworks and metrics which would allow to certify the safety and robustness of the models before deploying them in real-world. However, addressing these bottlenecks five future research directions are outlined: (1) neuro-symbolic HRC with runtime assurance; (2) calibrated uncertainty as the foundation for intent alignment and help-seeking; (3) parameter-efficient continual personalization for edge deployment; (4) interaction-centric world models for anticipatory collaboration; (5) evaluation science for LLM-mediated Human-Robot Collaboration. Thus, this review provides a structured roadmap to overcome current limitations, ensuring a safe, robust, and truly collaborative robotic systems in the human-centric environments.

Acoustic Water Leak Detection Using Machine Learning Methods

Nojus Zaronskis, Milita Songailaitė, Rūta Juozaitienė, Tomas Krilavičius

Vytautas Magnus University

nojus.zaronskis@vdu.lt

As societies around the world strive to adopt more sustainable and environmentally friendly practices, water loss has become a significant issue. Leaking pipelines waste a valuable natural resource, along with the energy and costs of treating and distributing it. Traditional leak detection methods, such as periodic inspections and specialized sensing equipment, often require extensive labor, many devices, or service interruptions. Acoustic monitoring is a promising alternative – water escaping through a crack produces a characteristic sound signal that a microphone can capture continuously, without direct access to the pipe. The challenge is distinguishing these sounds from water flow, machinery and other background noise. This project develops a machine learning-based system for detecting water pipe leakages from microphone recordings. The audio dataset is analyzed to identify differences between leak signals and background noise, then preprocessed, and acoustic feature representations are extracted. Multiple machine learning models are trained and compared using standard classification metrics. The best performance is achieved by a convolutional neural network trained on spectrograms, which reaches 97% accuracy. Error analysis shows that the main difficulty is separating background noises that closely resemble water leaks while training on a relatively small dataset. The work demonstrates that low-cost, automated microphone-based leak detection is feasible and could help reduce water losses in distribution systems.

17th 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

