13th Conference on

DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS

December 1–3, 2022

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

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

13th Conference on

DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS

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

December 1–3, 2022

VILNIUS UNIVERSITY PRESS
Vilnius, 2022

Co-Chairmen:

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

Programme Committee:

Prof. Juris Borzovs (Latvia)
Dr. Jolita Bernatavičienė (Lithuania)
Prof. Robertas Damaševičius (Lithuania)
Prof. Janis Grundspenkis (Latvia)
Prof. Janusz Kacprzyk (Poland)
Prof. Ignacy Kaliszewski (Poland)
Prof. Bożena Kostek (Poland)
Prof. Tomas Krilavičius (Lithuania)
Prof. Olga Kurasova (Lithuania)
Prof. Julius Žilinskas (Lithuania)

Organizing Committee:

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

Contacts:

Dr. Jolita Bernatavičienė
jolita.bernataviciene@mif.vu.lt
Tel. +370 5 2109 315

https://doi.org/10.15388/DAMSS.13.2022
ISBN 978-609-07-0794-4 (print)
ISBN 978-609-07-0795-1 (digital PDF)

Preface

DAMSS-2022 is the 13th International Conference on Data Analysis Methods for Software Systems, held in Druskininkai, Lithuania. Every year at the same place and time. The exception was in 2020, when the world was gripped by the Covid-19 pandemic and the movement of people was severely restricted. After a year's break, the conference is back on track. 2021 conference was successful, and the main objective of lively scientific communication was again achieved. The conference focuses on live interaction among participants, but there is also some scope for a limited number of virtual presentations. For better efficiency of communication among participants, most of the presentations are poster presentations. This format is really effective.

The history of the conference goes back to 2009, when 16 papers were delivered. It started as a workshop and has now grown into a well-known conference. The idea of such workshop was conceived at the Institute of Mathematics and Informatics, which is now the Institute of Data Science and Digital Technologies of Vilnius University. The Lithuanian Academy of Sciences and the Lithuanian Computer Society supported this idea. This idea has been welcomed by the Lithuanian scientific community and abroad. The number of this year's presentations is 81. The number of registered participants is 121 from 10 countries. This is significantly more than in 2021. The conference brings together researchers from six Lithuanian universities. This makes the conference the main annual meeting point for Lithuanian computer scientists. The main goal of the conference is to introduce the research undertaken at Lithuanian and foreign universities in the fields of data science and software engineering. The annual organization of the conference allows the fast interchanging of new ideas among the scientific community. Seven IT companies supported the conference this year. This means that the topics of the conference are actual for business, too.

Topics of the conference cover Artificial Intelligence, Big Data, Bioinformatics, Blockchain Technologies, Business Rules Software Engineering, Data Science, Deep Learning, Digital Technologies, High-Performance Computing, Machine Learning, Medical Informatics, Modelling Educational Data, Ontological Engineering, Optimization in Data Science, Signal Processing, Visualization Methods for Multidimensional Data. This book gives an overview of all presentations of DAMSS-2022 conference.

Partner

International Federation for Information Processing
ifip.org

DAMSS 2022 supported by:

General sponsors

Neurotechnology
www.neurotechnology.com

Novian UAB
novian.lt

Vinted UAB
vinted.com

Sponsors

Asseco Lithuania
asseco.lt

NetCode UAB
www.netcode.lt

Baltic Amadeus
www.ba.lt

Visoriai Information Technology Park (VITP)
vitp.lt

Lithuanian Speech in Humanoid Robots

Linas Aidokas, Gintautas Tamulevičius

Institute of Data Science and Digital Technologies
Vilnius University

linas.aidokas@mif.vu.lt

Speech-based technologies are becoming more feasible in portable devices, such as smartphones, smart speakers, and social robots. Given the importance of communicating in everyday situations, there is a big research interest in the human-robot interaction field or HRI in short. HRI focuses on speech-driven systems used for communicating and interacting between humans and robots. Currently, the successful applications of social robots in business do not use speech. In mission-critical tasks, almost everything is usually done through touch screens.

There are multiple challenges with social robots when the Lithuanian language is being considered. Lithuanian language has a small number of speakers compared to most popular languages, and it presents difficult challenges when Lithuanian speech-driven systems have to be developed. Spoken language interactions are very poor, and user expectations are usually very high. In HRI communication, typical users are generally not cooperative. When users want to exchange information with a robot, it has to be instant and real-time. Most robot users are usually children or old people who do not understand data exchange protocols, and they do not follow a strict interaction scenario. When people are talking without a microphone, and colloquial language is used, it becomes more challenging to identify keywords or phrases. Often there are problems related to «false start», unclear endings or continuation of dialogue or sentence. The current word recognition accuracy is not enough - it should be close to 100%, and the context of what is being talked about in the interaction is also important.

Because of all the drawbacks of the current HRI, current dialogue systems for human-robot interactions must be designed with poor language recognition quality in mind. This also presents new possibilities for improving current HRI systems through research.

Target Class Classification Recursion Preliminaries

Levon Aslanyan[1], Karen Gishyan[2], Hasmik Sahakyan[1]

[1] Institute for Informatics and Automation Problems of the National Academy of Sciences, Armenia
[2] International Scientific Educational Center of the National Academy of Sciences, Armenia

lasl@sci.am

In Machine Learning, the Supervised Classification scenario supposes a number of geometrically "compact" sets of elements (clusters) given by the sets of class representatives (elements of the so-called learning set), asking for universal procedures that may correctly classify new objects into the classes. After the learning / training stage, recognition / classification of the trial object is, in general, a static, one-step process. In our application, which is fundamentally different from the traditional learning model, one of the classes is marked, especially, and the aim is in learning, allocating all objects by a step-by-step procedure to this special class. In an individual step of the recursion, when the temporal class of the object is determined, a predetermined class action is applied that transfers this object to the same or to some other class. The chain of such transformations has to converge to the marked special class. The reinforcement Learning concept is quite close to this scenario. It acts over the set of classes / sites, and these classes, with their elements, compose an environment. The agent (algorithm, recognizer) learns how to effectively transfer among these classes in a way to optimize the target function about the final allocation to the special / target class. In this model, we deal with a dynamic, not one-step, classification task. This paper considers a novel / specific classification problem that does not fit into either of the well-known basic learning scenarios – supervised, unsupervised, or reinforcement. For the outlined special class, we equivalently use the names "normal" and "target". The novelty is in recurrent classifications over the set of objects, in a coincidence of classes and class actions, and in the convergence of classifications to the target class

during the minimal possible steps. Model, algorithm, and evaluation of this target class problem is our objective. We differentiate the white-box and black-box cases. The last one is closer to the model of inverse reinforcement learning and especially to the apprenticeship learning concept. Planning the logic-combinatorial analysis of the inverse target class models, currently, we stay at the Markov chain interpretation.

Application of Merkelized Abstract Syntax Trees to the Internet of Things Location-Based Cybersecurity Solutions

Kazimieras Bagdonas, Algimantas Venčkauskas

Kaunas University of Technology

kazimieras.bagdonas@ktu.lt

In this paper, we present a novel method for geolocation data integration into a multimodal Internet of Things (IoT) security solution using Merkelized Abstract Syntax Trees (MAST). The proposed method has been developed for the IoT devices operating in the IoT Fog, communicating with the Edge devices. The proposed method allows the exploitation of the IoT Networks Node's (NN) localization solution for data source authentication and data validation.

Localization solutions can be obtained via external systems (e.g. GNSS) or can be derived from IoT localization techniques. The least significant bits of the localization solution are masked to provide the desired geographical Zone of Validity (ZoV). The amount of masked bits defines the size of the ZoV. No other information except the number of masked bits needs to be communicated from the Edge node to the NN.

In cases where improved security is required, additional parameters, such as IDs and location data of neighboring IoT NNs, can be incorporated into the extended MAST structure to enhance security. This information must be present in the Edge nodes in order to validate the signature solution. A proposed novel approach to coordinate verification allows the transmission of hashed values without the need to reveal either the information on the ZoV by the Edge node or the exact coordinates by the NN.

The simulation of the algorithm is analyzed and discussed. The proposed method is numerically investigated in regard to the uncertainties introduced by the expected compounded localization errors in IoT ad-hoc networks. The impact on computational and bandwidth requirements is analyzed in relation to the desired level of security. The obtained results provide insight into the possible envelope of application for the proposed method.

Exploring PGAS-Based Gossiping Algorithms for Knödel Graphs

Vahag Bejanyan, Hrachya Astsatryan

Institute for Informatics and Automation Problems of
the National Academy of Sciences of the Republic of Armenia

bejanyan.vahag@pm.me

Knödel graphs of even order n and degree $1 \leqslant \Delta \leqslant \lfloor \log_2(n) \rfloor$, $W_{\Delta,n}$, are regular graphs that have an underlying topology that is time optimal for algorithms gossiping among n nodes. Because of their distinctive properties, Knödel graphs act as a time-optimal topology for broadcasting and gossiping, thus arising in many settings, including social and communication networks or agent-based modelling simulations. Experimentation, often based on the extensive generation and analysis of complex networks, relies on high-performance computational resources to efficiently simulate the flow of information. The efficacy of such processing commonly depends on parallel processing and proper provisioning of distributed resources. The study aims to develop a complete approach devoted to the experimental research of Knödel graphs and to evaluate the memory usage pattern based on global memory address space abstraction. The sequential and parallel generation of synthetic datasets and simulation of push-pull gossiping algorithms with detailed analysis of resource usage and runtime has been studied.

Visualisation of 2D Fractal Structures Associated with the Riemann Zeta Function

Igoris Belovas, Martynas Sabaliauskas, Lukas Kuzma

Institute of Data Science and Digital Technologies
Vilnius University

martynas.sabaliauskas@mif.vu.lt

The Riemann Hypothesis is one of the seven Millennium Prize Problems - greatest unsolved problems in mathematics - selected by the Clay Mathematics Institute. Currently, the only problem to have been solved is the Poincaré conjecture mastered by Grigori Perelman in 2010. The Riemann zeta function is extremely important, playing a central role in analytic number theory. The properties of the function are essential in determining the distribution of primes. Prime numbers, in turn, have a broad spectrum of applications, ranging from quantum mechanics to information security. It is known that non-trivial zeros of the Riemann zeta function belong to the critical strip. The Riemann hypothesis asserts that all the non-trivial zeros of the Riemann zeta function lie on the critical line. We present the Fractal heuristic approach, which is based on applying Mandelbrot sets associated with the Riemann zeta function for visual investigation of its underlying nature. We aim to visualize the fractal geography of the Riemann zeta function using special non-linear transformations and propose an algorithm for the generation of the corresponding Mandelbrot sets. We seek to get the most precise fractal image based on the yellow-black-blue colour palette. We present visualizations of Mandelbrot sets received by the Fractal heuristic approach. These complex plane fractal structures offer additional information on the properties of the Riemann zeta function, such as the arrangement of the non-trivial zeros.

Autoencoder for Fraudulent Transactions Data Feature Engineering

Dalia Breskuvienė, Gintautas Dzemyda

Institute of Data Science and Digital Technologies
Vilnius University

dalia.breskuviene@mif.stud.vu.lt

Fraud is an illegal action by someone who wants to gain financial benefit from another person or institution. It has evident economic consequences on private enterprises, public services, and individuals' financial situation. Fraudulent activity is constantly evolving - it has no persistent patterns. Machine learning is one of the ways to solve fraud detection problems. However, using machine learning algorithms for these issues requires careful data preparation, feature engineering, and feature encoding.

This paper aims to improve machine learning classification quality on fraudulent cases, focusing on enhancing Recall. It is a continuation and improvement of the approach suggested in the article "D. Breskuvienė, G. Dzemyda (2023). Imbalanced data classification approach based on clustered training set. In: Dzemyda G., Bernatavičienė J., Kacprzyk J. (Eds.), Data Science in Applications. Studies in Computational Intelligence. Springer (accepted)".

We use a Synthetic database with the simulated United States customers' credit card transaction information for the experimental evaluation. Many features of this data set are categorical (nominal), with no intrinsic ordering to the categories. In our research, we are especially interested in such features that obtain many different values (categories). We need to use encoders to convert them into a numerical format suitable for machine learning. The two most popular ways to encode categorical data are LabelEncoder, which assigns numeric values to the categorical values, and OneHotEncoder, which creates a binary column for each category. However, when dealing with fifty different values under a categorical variable, LabelEncoder creates random weights for the categories. While OneHotEncoder brings inefficiency in the machine learning model performance by increasing dimensionality. To solve the abovementioned issue, we suggest empowering an artificial neural network. Namely, use an autoencoder, which learns patterns of a dataset and learns to ignore the noise, to reduce dimensionality.

Problems and Solutions of Autonomous Exploration of Unknown Indoor Environments for Micro Aerial Vehicles with Onboard Stereo Camera

Mantas Briliauskas

Institute of Data Science and Digital Technologies
Vilnius University

m.briliauskas@gmail.com

The popularity of using micro aerial vehicles (MAV) in autonomous applications has increased significantly over the years. The usage includes agriculture, search and rescue operations, object's inspection, and environmental mapping. The main goal of autonomous exploration is to produce a map of a priori unknown environment by minimising exploration time and map uncertainty. Different SLAM algorithms make a passive localisation and mapping by processing sensors' information while MAV is operating, whereas the GPS signal is most likely denied or very inaccurate in indoor environments. Navigation algorithms are employed for path planning with a goal to move MAV to the required position as optimally as possible with respect to its model physics and controls and avoiding obstacles. Having mapping, localisation and navigation, the exploration algorithms are used to find the best next view candidate to navigate MAV to in order to maximise the information gain of the currently chosen action. In this survey, current state-of-the-art methods and the main problems of exploration algorithms for indoor environments are reviewed.

Intrusion Detection Based on Keystroke Biometrics and Siamese Neural Networks

Arnoldas Budžys, Olga Kurasova, Viktor Medvedev

Institute of Data Science and Digital Technologies
Vilnius University

arnoldas.budzys@mif.stud.vu.lt

Cyber security is becoming one of the most important topics in today's critical infrastructure to ensure a secure connection between the administrator and the session. In recent years, the development of intrusion detection systems to protect against insiders has been a relevant and complex challenge in critical infrastructures. If the system administrator's password is compromised or otherwise misappropriated, it could cause significant damage to the critical infrastructure. An insider threat is a harmful activity against an organisation by users with legitimate access to the organisation's infrastructure, software, or databases. These users may be current or former employees with access to the organisation's data. A methodology for user authentication of critical infrastructure systems based on a deep learning network is proposed. Behavioural biometric data or user behavioural characteristics are converted into an image and further used in the proposed methodology for authentication. The keystroke data obtained from the login password is converted into a format (image in our case) that is more acceptable to convolutional neural networks. Siamese neural networks can be employed for image similarity detection to distinguish a legal user from an insider. In this study, similarity measures were defined to identify the user based on his biometric data converted into an image and composed of keyboard inputs. Experiments were carried out using public datasets. The results are promising, demonstrating that using a deep learning-based approach to analyse images obtained from user keystroke data can improve intrusion detection accuracy and perform user authentication more efficiently.

Training System for Elite Athletes

Eglė Butkevičiūtė, Liepa Bikulčienė,
Tomas Blažauskas, Andrius Paulauskas

Kaunas University of Technology

egle.butkeviciute@ktu.lt

Smart coaching systems rely on decision-making and machine-learning methods and could be applied in various fields, including professional sports. From practical experience, it is known that elite athletes can have deficiencies in certain cognitive factors that may influence their performance. The aim of this research is to create a training and testing WebVR platform adjusted to professional athletes' cognitive-mental abilities improvement. The designed system consists of two platforms: a testing environment that includes sensors, an oculus quest, two virtual reality glasses and a specialist client device; a training environment that is designed for daily use in front of a personal computer or tablet. Both platforms include three cognitive exercises: attention transfer task, anticipation task, and concentration task. In the initial phase, all athletes perform all three tasks in the testing environment and get an individualised training plan based on memory, concentration, reaction time, attention peculiarities, decision-making, focus, anticipation results and heart rate variability (HRV) results (obtained using Polar belt V10). The testing process is repeated in a month. The results have shown that, on average, participants made some improvement in all three tasks. For example, attention transfer abilities improved a few times, and for most participants, attention transfer results reached 100%. However, the analysis has shown that there was no significant difference in reaction time.

On Multi-Criteria Decision-Making Methods in Finance Using Explainable Artificial Intelligence

Jurgita Černevičienė, Audrius Kabašinskas

Kaunas University of Technology

jurmark@ktu.lt

The influence of Artificial Intelligence is growing, as is the need to make it as explainable as possible. Explainability is one of the main obstacles that AI faces today on the way to more practical implementation. In practice, companies need to use models that balance interpretability and accuracy to make more effective decisions, especially in the field of finance. The main advantages of the multi-criteria decision-making principle (MCDM) in financial decision-making are the ability to structure complex evaluation tasks that allow for well-founded financial decisions, the application of quantitative and qualitative criteria in the analysis process, the possibility of transparency of evaluation and the introduction of improved, universal and practical academic methods to the financial decision-making process. This article presents a review and classification of multi-criteria decision-making methods that help to achieve the goal of forthcoming research: to create artificial intelligence-based methods that are explainable, transparent, and interpretable for most investment decision-makers.

Acknowledgement: The research of A.K. was partially supported by COST (European Cooperation in Science and 541 Technology) Action 19130.

Improving Network Intrusion Detection Applying Hybrid Machine Learning Algorithms

Karina Čiurlienė, Denisas Stankevičius

Vilnius Gediminas Technical University

karina.ciurliene@vilniustech.lt

Network intrusion detection is a relevant cybersecurity research field. The growing number of intrusions requires more sophisticated methods to protect computer networks. Different machine learning algorithms are used to identify intrusions and anomalies by analysing network-level and host-level data, however, their accuracy is limited. In this research, we address the problem of improving network-level intrusion detection by applying hybrid machine-learning algorithms. Such kinds of algorithms combine several machine learning algorithms and allow increasing accuracy of the classification problem. Two publicly available datasets were used for the analysis, i.e. CSE-CIC-IDS 2018 and NSW-NB-15. Both datasets include benign network data as well as data of cyberattacks such as SQL injection, infiltration, Brute force attack, DoS, DDoS, and others. TCP data and event logs were collected in datasets and labelled based on cyberattack type. First of all, $\chi 2$ test was performed to determine significant attributes. Then three hybrid algorithms consisting of three different machine learning algorithms were proposed and analysed using both datasets. Analysis of the resulting accuracy of the hybrid algorithms showed that the highest accuracy of 99.24% was achieved. This result has a higher value of 5.41% compared to the best machine learning. The study also showed that the hybrid algorithms allow to achieve better performance. Finally, all investigated machine learning algorithms were ranked using three different ranking techniques and accuracy, and the most appropriate algorithms for intrusion detection were proposed.

Influence of Oxygen Consumption Rate Modulation on Bacterial Pattern Formation Models

Boleslovas Dapkūnas, Romas Baronas, Žilvinas Ledas

Institute of Computer Science
Vilnius University

boleslovas.dapkunas@mif.vu.lt

Mathematical bacterial pattern formation models have been studied since the 1970s. Most models are based on Keller-Segel equations. In this approach, the dynamics of the bacteria population is modelled using a system of nonlinear reaction-diffusion-chemotaxis partial differential equations.

Escherichia coli exhibits attraction to self-excreted chemoattractant. It was also shown that the activity of *E. coli* depends on available oxygen. The dynamics of oxygen consumption rate play an important part in the pattern formation. Multiple methods of oxygen consumption rate modulation have been used in different studies. The interactions between several active processes lead to very complex dynamic systems that are still poorly understood.

The aim of this work is to examine the effects of several different oxygen consumption rate modulation functions on the spatiotemporal pattern formation of luminous bacteria. The model involving chemoattractant and oxygen dynamics is used to simulate the 2D patterns in bacterial populations near the inner lateral surface of a cylindrical micro-container. The numerical simulation was conducted using the finite difference technique.

Natural Language Generation Problems

Konstantinos Diamantaras

International Hellenic University, Greece

kdiamant@ihu.gr

Natural Language Generation problems like question-answering, text summarisation, and machine translation are nowadays tackled using mainly transformer-based machine learning models such as GPT-3, T5, and BART. Although these sophisticated models achieve very good performance in most NLG tasks, they suffer from a fundamental lack of common sense. For this reason, they often generate implausible and "strange" sentences or sentences that are short and simple, avoiding the rich and natural structures generated by humans. Recently there is an increasing trend to incorporate common sense reasoning in text generation. The aim is to enhance/enrich the process of natural language generation using external knowledge, which exists in many data sources like Wikipedia, knowledge bases, and knowledge graphs. Common-sense knowledge is one example of knowledge that can be acquired from knowledge bases/graphs and can be used in the generation process by creating embeddings. The focus of this talk is to present methods that incorporate external knowledge available in various common-sense knowledge bases into state-of-the-art natural language generation models.

Dynamic Knowledge PRediction And eduCaTIonal Content rEcommendation system – DK PRACTICE

Marina Delianidi, Konstantinos Diamantaras, Ioannis Moras

International Hellenic University, Greece

d.marina@iee.ihu.gr

Learning is an integral part of human activity from the moment of birth. Knowledge is cultivated and evolves over time through learning. The way of learning, the ease of understanding, or the pace of learning contribute to the evolution of knowledge over time. In the educational process, knowledge is built by studying various knowledge components or concepts using appropriate educational content. E-learning is either done by choice or is a necessity in cases where in-person learning is difficult or impossible. In recent years it has become very popular, providing a large amount of online educational material. Digital learning materials can be found from various sources, for example, in organised courses in learning management platforms, in teacher websites, in educational social network platforms, in e-books, or scattered across the internet. There is a variety of digital educational material in terms of format, such as presentations, text, video, audio, images, etc. The instructors' goal is to achieve the best learning outcomes for their students. Using personalised learning is an important factor in this endeavor. If it is possible to estimate the student's knowledge state, then the personalised learning method and the appropriate educational material can significantly improve the student's performance. Educational data stored and provided through e-learning platforms and Learning Management Systems, after appropriate processing, can be used to model the students' knowledge of specific educational concepts. The added value offered by student knowledge modelling is the dynamic assessment of the learners' knowledge state, the recommendation of educational content, and the suggestion of learning methods aimed at the positive evolution of

knowledge. The representation of the student's knowledge state in relation to certain knowledge components or concepts is achieved through knowledge tracing. One of the fields of application and utilisation of this information is educational content recommendation systems. With the recommendation of appropriate educational content, the students are called to focus on improving their knowledge in the specific components where weaknesses are detected. In our work, we present an architecture of an educational recommendations system, where through a specific case study, we aim at recommending educational content in the knowledge area of computer architecture in the context of a course at the International Hellenic University. We use anonymised data from student responses to multiple-choice questions with only one correct answer and only one attempt to answer each question. The system, using the students' history of responses, is trained to dynamically estimate their current knowledge state and to recommend the specific learning materials in the concepts that, according to the predictions, the student has weaknesses and needs improvement. The purpose of the study is to improve the performance of students through the prediction of their knowledge state based on their past interactions and the positive evolution of knowledge over time by recommending appropriate, personalised educational content. In addition, the system can be used to check the student's knowledge before a test or as an auxiliary tool for the instructors to know the knowledge level of their students.

All for One and One for All: How to Assess Performance When There Are Different Dimensions and Different Stakeholders' Priorities?

Giovanna D'Inverno

Department of Economics and Management
University of Pisa, Italy

giovanna.dinverno@unipi.it

Composite indicators are often used to aggregate several dimensions in one single score, so to provide an overall performance measure. In the evaluation process, there are key elements that need to be considered. First, it is important to measure whether and to what extent set performance targets are met. Second, the aggregation must reflect and harmonise the different preferences of the involved stakeholders. In this talk, we discuss a new composite indicator that integrates the Goal Programming Synthetic Indicator methodology with the Analytic Hierarchy Process. We showcase its potential by evaluating to which extent European countries fulfil the European Union requirements in terms of municipal waste management while taking into account preferences as expressed by a panel of experts.

Transition From Proof-of-Work to Proof-of-Stake Blockchains: Why It Matters More Than Ever?

Ernestas Filatovas, Aleksandr Igumenov, Viktor Medvedev, Remigijus Paulavičius

Institute of Data Science and Digital Technologies
Vilnius University

ernestas.filatovas@mif.vu.lt

Blockchain and underlying distributed ledger technology attracted widespread attention recently due to its transparency, decentralisation, and security properties. However, it still faces many challenges that have to be solved. One of the most discussed ones is the enormous power consumption of Bitcoin, and other Proof-of-Work (PoW) based blockchain networks. Nowadays, when the world faces a global energy crisis, the matter of the sustainability of blockchains has become particularly sensitive. The blockchain community has already made progress in resolving the high energy consumption challenge. The utilisation of renewable energy used for mining is constantly increasing, reducing carbon footprint. Also, tremendous stress is done on the adoption of energy-efficient Proof-of-Stake (PoS) consensus protocols. Among recently developed blockchain networks, PoS-based ones dominate the market. Moreover, in September 2022, the world's most used blockchain, the Ethereum network, moved from PoW to PoS. This caused a transformation in the mining power distribution. In this study, we propose a new methodology and tool to accurately estimate the energy usage of Ethereum-like PoW-based blockchain networks. We also analyse the impact of the PoW-based Ethereum network transition to PoS on the energy consumption used for mining in the industry. With this work, we aim to contribute towards developing more sustainable blockchains. This research has received funding from the Research Council of Lithuania (LMTLT), agreement No. S-MIP-21-53.

Clustering Healthcare Data

Pasi Fränti

School of Computing, University of Eastern Finland

franti@cs.uef.fi

Clustering can a powerful tool in analysing healthcare data. We show how clustering based on k-means and its variants can be used to extract new insight from various data with the aim to better optimise the health care system. We first show that simple variants of k-means and random swap algorithms can provide highly accurate clustering results. We demonstrate how k-means can be applied to categorical data, sets, and graphs. We model health care records of individual patients as a set of diagnoses. These can be used to cluster patients, and also create co-occurence graph of diagnoses depending on how often the same pair of diseases are diagnosed in the record of the same patient. Taking into account the order of the diagnoses, we can construct a predictor for the likely forthcoming diseases. We also provide a clustering algorithm to optimise the location of health care systems based on patient locations. As a case study, we consider coronary heart disease patients and analyse in what way the optimisation of the locations can affect the expected time to reach the hospital within the given time. All the results can provide additional statistical information to healthcare planners and also medical doctors at the operational level to guide their efforts to provide better healthcare services.

A Deep Dive Into Industry-Inspired Research Questions: From Causal Inference and Machine Learning to Complexity Science

Jev Gamper

Vinted UAB

jevgenij.gamper@vinted.com

In this talk, I will motivate several open research directions inspired by real industry scenarios and constraints at Vinted. For each research direction, I will present a business case, its mathematical formalisation and derived open questions, prior literature, as well as avenues for developing answers to these scientific questions. The topics I will cover include but are not limited to: Causal discovery and experiment design; Machine learning and feedback loops; Complex system simulation; Machine learning, surrogacy and mediation analysis.

A Generalised Surrogate Model Selection Method for Surrogate Metrics

Jevgenij Gamper, Agnė Reklaitė, Giedrius Blažys

Vinted UAB

jevgenij.gamper@vinted.com

Estimating the long-term effects of online controlled experiments is one of the top challenges for industry practitioners (Gupta el., 2019). Athey et al. (2019) proposed the Surrogate Index approach, whereby one uses a machine learning model to predict the long-term treatment effects given the short-term proxies. Surrogate index methodology enabled industry practitioners to develop a range of surrogate metrics for estimating the long-term effects of online experiments (Weitao et al., 2021; Chandar et al., 2022). However, the predictive power of machine learning models used for surrogate metrics may come from association rather than causation. Treatment may cause an increase in the surrogate metric but have little to no impact on the long-term outcome. As a result, one may gain additional confidence in their surrogate metric and the underlying predictive model if the selection criteria are based on a testable causal model. Given that the surrogacy assumption is a causal mediation process, we draw inspiration from causal mediation analysis literature and contribute by 1) developing and comparing a set of causal models for surrogate model selection for surrogate metrics; 2) showing how these models allow to not only select the best model for a surrogate metric but also naturally incorporate the variance and other covariates of interest when performing long-term metric treatment-effect estimation using a surrogate metric.

System Call-Based Malware Detection Using Hybrid ML Methods Trained on the AWSCTD Dataset

Nana Kwame Gyamfi, Nikolaj Goranin, Dainius Ceponis, Antanas Čenys

Vilnius Gediminas Technical University

nana-kwame.gyamfi@vilniustech.lt

Automatic malware detection methods can be classified into two main classes: signature-based and anomaly-based. The signature-based approach is considered to be reliable and having a low false-positive rate for known malware types, but is not able to detect new and zero-day attacks. It also has other drawbacks, like a signature database increase, as well. Anomaly-based approach is considered as a perspective for detecting new and zero-day malware. Typically they are statistical analysis or machine learning based and require training on classified data. The majority of research on anomaly-based malware detection is currently concentrated on the use of system call sequences generated by the analysed suspicious programs as input data. Although, formally, the achieved results seem promising and the obtained accuracy rate is higher than 99%, it is still necessary to mention that it is possible to get such accuracy only in case when a long sequence of system calls (>600) is provided for analysis. Such data input prompts that in a real environment, malware would be able to perform all planned actions before being detected. Optimisation of this method can be seen in utilising adjacent (and metadata) data to minimise the length of system call sequence needed for reliable attack detection, thus minimising the reaction time, optimisation of data structures and parameters of currently utilised artificial intelligence methods as well as search and/or development of the new artificial intelligence methods/architectures, mostly suitable for the anomaly detection task. In this research, we present the experimental evaluation of several classical (KNN) and hybrid (Pso-naïve Bayes; Pso-KNN; Pso-SVM; Pso-multilayer perceptron; BestFirst-SVM) ML methods trained on our previously generated AWSCTD dataset. Training and

experiments were performed on an open-source training and testing platform Google Colaboratory. The best accuracy (97.35%) results were achieved with the BestFirst-SVM method, which outperformed earlier used ML methods on the AWSCTD dataset. Still, it is behind the earlier tested deep-learning methods, such as AWSCTD-CNN-S, by accuracy, although winning the speed competition.

Computational Models of Heat and Mass Transfer in the Textile Structures

Aušra Gadeikytė, Rimantas Barauskas

Department of Applied Informatics
Kaunas University of Technology

ausra.gadeikyte@ktu.lt

The aim of this study is to develop computational models of the heat and mass exchange processes that occur in modern composite textile structures of three-dimensional internal structure. The proposed finite element models enable the prediction of air permeability, water-vapour resistance, thermal resistance, and heat transfer coefficients of textile structures in the early design stage. Furthermore, the finite element models were verified by comparing the sample variants of the calculated objects with the experimental measurements presented in the scientific literature. These models might be used in the development of protective clothing, airbags, fireproof seat covers, passive and active cooling systems, and others. In order to facilitate the use of computational models for numerical analysis in manufacturing departments or test laboratories, the proposed finite element models were extended to the simulation apps using Comsol Multiphysics Application Builder.

Algal Bloom Monitoring Using Multi-Spectral Satellite Data

Dalia Grendaitė[1], Linas Petkevičius[2]

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

linas.petkevicius@mif.vu.lt

Algal and cyanobacterial blooms are a natural phenomenon that are caused by the sudden proliferation of algal and cyanobacterial biomass. These organisms take up carbon dioxide and release oxygen during photosynthesis. The blooms result in low water transparency and limit light penetration to deeper water layers. Thus, ecosystem structure changes and benthic plants may disappear. Currently, bloom monitoring is based on in situ data, that are collected infrequently. Remote sensing allows monitoring of the situation on a weekly basis. We collected and analysed in situ data of chlorophyll-a concentration and water transparency from national environmental agencies of Lithuania, Latvia, and Estonia for the years 2017–2021. The machine learning models that use optical Sentinel-2 Multispectral Imager (MSI) data for predicting the blooms and chlorophyll-a concentration were successfully created. Finally, the monitoring dashboard was created and deployed for weekly blooming and chlorophyll-a concentration predictions.

On Global Optimization and Machine Learning

Eligius M. T. Hendrix

University of Malaga, Spain

eligius@uma.es

Machine learning of predictive and classification models can be viewed from an optimization perspective. Lately, there has been an increased interest due to a seminar and European program called "Machine Learning Needs Optimization".

In the application of deep learning, we observe that many times stochastic gradient approaches implemented in learning algorithms may suffer from effective convergence. This pushes researchers to go for random approaches based on a randomly selected training set of data. The idea of design of experiments in statistics taught us that in fact, the training set may be selected in a way that more information becomes available. An interesting question in hyperspectral image analysis is that correlation between the selected points appear, due to an overlap of used observation windows to smooth signals. We will discuss a way to deal with this overlap in a conscious way using LP type models.

We focus on the characterization of the underlying optimization landscape of estimation and parameter estimation problems posing some questions on the effectiveness of algorithms. In parameter estimation, ill conditioning may appear easily. Moreover, parameter identification and symmetry may lead to infinitely many parametrizations providing similar performance. In parameter estimation, this is related to the so-called identifiability problem. In deep learning, the term over-parametrization is used. Attention is paid to use design of experiments to select the training data set. Specifically, we present an MILP model to maximize intra class signal variation and minimize overlap of observation windows in earth observations.

Several topics characterize the optimization problem of learning:

- Identifiability in parameter estimation
- Ill-conditioning

- Symmetry
- Optimization landscape
- Over-parametrization
- Local versus Global search
- Design of experiments
- Correlation due to using windows of observations

We use several small instances to showcase the underlying characteristics of the corresponding optimization problem.

Using Domain-Specific Word Embeddings to Boost Keyword-Based Commit Classification

Tjaša Heričko, Boštjan Šumak

Faculty of Electrical Engineering and Computer Science
University of Maribor, Slovenia

tjasa.hericko@um.si

During the lifetime of a software project, various modifications are performed to the source code to correct, adapt, and perfect software under maintenance. Using source code management tools that effectively facilitate the management of software changes, the committed changes are tracked and accompanied by short messages from committers communicating the code changes in natural text. Existing research has shown that keywords extracted from commit messages based on word frequency analysis are good indicators of commit intent. However, developers who compose commit messages often use jargon terms, acronyms, misspelt words, and synonyms or related words, making it challenging to classify commits based on a small set of fixed keywords alone. This paper proposes an enhancement to the keyword-based commit classification approach by extending the set of keywords by exploiting semantic similarities between the words of commit messages. Word embedding vectors were generated by training on a domain-specific corpus containing 2.5 million commit messages from 500 code repositories. Each word in the corpus vocabulary was assigned a single vector representation in a vector space of 100 dimensions. The N most similar words for each predefined keyword were extracted by computing cosine distances. Then, every keyword was extended by a set of M words with the highest cosine similarity values of N words in such a way that there is no overlap between M words of all keywords. This was ensured by keeping only the duplicated word with the highest cosine similarity value over the duplicated related words of all keywords prior to selecting the M most similar words for every keyword. To evaluate the proposed approach, different machine learning models were built on a labelled dataset using Random

Forest, Gradient Boosting Machine, and CART algorithms, considering different similarity thresholds. While the existing approach showed that only 65% of the sample commit messages examined contained at least one keyword, the applied approach demonstrated that up to 82% contained at least one of the extended keywords. Comparison with the existing approach showed that enriching keywords with the words semantically closest to them can benefit the predictive performance of the models while providing deeper insight into how developers use written communication through commit messages.

Evolution of Nucleotide Sequences Over Passing Time

Kamilija Jablonskaitė, Tomas Ruzgas

Department of Applied Mathematics
Faculty of Mathematics and Natural Sciences
Kaunas University of Technology

tomas.ruzgas@ktu.lt

All DNA sequences contain four types of nucleotides, which in turn hold all genetic information inherited by an organism. However, DNA can mutate while replicating itself, which means that it is possible to lose a number of nucleotides and/or to gain different fragments of the original sequence; in other words, the initial DNA sequence can differ from its duplicate. Knowing this, DNA sequence over passing time can be depicted as a discrete-time homogeneous Markov chain, while sequence evolution in space can be described as an action which depicts new element addition to the sequence. Theoretically, evolution in space simulates DNA sequence formation. In the stationary case, the distribution of a random sequence does not depend on the fixed time moment. It is hard to find any data regarding DNA sequence evolution in space – usually, only one sequence can be found. It is possible to reconstruct the transition matrix or the properties of that matrix from the stationary distribution of the Markov chain during the evolution over passing the time, yet this problem is ill-posed. In general, said the problem has a lot of solutions which could be found only by using some additional assumptions and regularisation methods. However, the solution could be found more easily using the local balance equation if the DNA sequence is reversed and the transition matrix only depends on a relatively small number of unknown parameters. The mathematical model of the described genetic sequence should not disagree with already known facts of genetic science and lean on these biological assumptions:

- Introns (non-coding DNA fragments) do not directly influence the possible survival of an individual or species in general. This means the regularities of inanimate nature have more of an influence than

natural selection. The fragments in question are not as important as fragments in the sequence that hold information, so it would be appropriate to search non- informative sequences in introns.

- Intron evolution over time has a simple structure, and evolution itself is only affected by local random factors. For example, it is assumed that the insertion or deletion of DNA fragment does not exist and nucleotides simply swap with each other with the probability that only depends on their nearest neighbours (simple local evolution). While this assumption is intuitive, it is not necessarily true as it has not been fully proved.
- It is assumed that the process of evolution over time of the DNA sequence in question is stable. If the said process is not stable, any part of the DNA sequence can become informative.

Stationary distribution of nucleotides residing in introns is assumed as non-informative in the case of simple local evolution.

How Much Do We Collect and How Much Do We Use for Policy and Research? Education Data. Case of Lithuania

Audronė Jakaitienė, Rimantas Želvys, Rita Dukynaitė

Institute of Data Science and Digital Technologies
Vilnius University

audrone.jakaitiene@mf.vu.lt

Education in all senses is one of the fundamental factors of development. Datafication refers to the transformation of different aspects of education (tests scores, school inspection reports, etc.,) into digital data. Making information about education into digital data allows it to be inserted into databases, where it can be measured, calculations can be performed on it, and through which it can be turned into charts, tables, and other forms of graphical presentation. Data and the ways of its presentation do not necessarily provide an unbiased view of education; it can also be misleading.

We review various sources of educational data (e.g., large-scale international studies and data registers) and their use for policy decisions and research in Lithuania. It has been shown that a lot of data has already been collected and that more is being accumulated. Up to 20 per cent of the information gathered in Lithuania is used for policy purposes and even less in research. It is noted that most of the data collected are useful for the economic paradigm. We will present a case study demonstrating that national population-based studies and international achievement studies can send different messages and cannot be considered in isolation.

Disease Trajectory Prediction Among Cancer Patients

Gabrielė Jenciūtė[1,2], Gabrielė Kasputytė[1,2], Adomas Bunevičius[3], Romas Bunevičius[3], Šarūnas Bagdonas[3], Jonas Venius[4,5], Tomas Krilavičius[1,2]

[1] Vytautas Magnus University
[2] CARD – Centre for Applied Research and Development
[3] ProIT
[4] Medical Physics Department, National Cancer Institute
[5] Biomedical Physics Laboratory, National Cancer Institute

gabriele.jenciute@vdu.lt

Cancer is the leading cause of morbidity and mortality in Lithuania and around the world. In order to monitor patients between clinical encounters, passively collected data taken from smartphones could revolutionise the way in which cancer patients are cared for by enabling real-time data analysis and individual patient monitoring. This information can also be used to predict how patients will behave in the future. Using ARIMA (autoregressive integrated moving average), Holt's Trend, TBATS, Simple Exponential Smoothing, and Naive models, this study aims to forecast the activity period per day estimated by passive data. The most suitable model is selected after evaluating the training error (MAPE), and forecasts are provided for the next 30 days. Root mean square error (RMSE), mean absolute percentage error (MAPE) and mean absolute error (MAE) are used to evaluate the accuracy of the models. According to the results, there is no single most accurate model that can be used for all 30 patients. There are, however, two models that are most appropriate: ARIMA and Holt.

Detecting Drivers of First Invasions Using Relational Event Models

Rūta Juozaitienė[1], Ernst C. Wit[2]

[1] Vytautas Magnus University
[2] Università della Svizzera italiana, Lugano, Switzerland

ruta.juozaitiene@vdu.lt

Spatio-temporal processes play a key role in ecology, from genes to large-scale macroecological and biogeographical processes. For analysing such spatiotemporally structured data, timing and time-ordering of events are particularly important. In this research, we present a generic method – relational event modelling – for studying spatio-temporal patterns of biological invasions at large spatial scales and by including variables that drive these dynamics. Relational event modelling (REM) relies on temporal interaction dynamics that encode sequences of relational events connecting a sender to a recipient at a specific point in time. We present a relational event model case study of the spread of alien species, which are species that are introduced accidentally or deliberately into geographical regions outside their native range. In this context, a relational event represents the new occurrence of an alien species, given its former distribution. By considering the bipartite network of species and regions, we embed the first records process into a relational event setting to detect drivers of invasions in the presence of complex confounding. Thus, taking the temporal sequence of occurrences into account, the relational event model identifies commonalities among species' spread and their relation to underlying variables. Using years of first records of 16,019 established alien species from major taxonomic groups, we have shown that the relational event model can be used constructively in environmental science to model time-stamped ecological interactions, such as the invasive species process, without having to resort to more traditional simplifications. Combining the first records data with other spatio-temporal information enables us to discover which factors have been driving the spread of species across the globe.

Measurement of Performance of Lithuanian II Pillar Pension Funds Using Benchmark and Rolling Window Technique

Audrius Kabašinskas, Miloš Kopa, Kristina Šutienė, Aušrinė Lakštutienė, Aidas Malakauskas

Kaunas University of Technology

audkaba@ktu.lt

In this research, we analyse how to measure the performance of Lithuanian II pillar pension funds using a benchmark. In a previous conference (Mathematical Models in Economics 2022), we presented how to present the dynamics of simple performance measures (correlation, VaR, CVaR, Average Recovery time, adjusted Sharpe ratio, etc.). However, it turned out that simple measures describe dynamical performance only partially. Moreover, pension funds typically have their benchmark that they track. However, straightforward measurement of tracking error also has a lot of weaknesses. Therefore, we analyze the performance of pension funds by using only two benchmarks "MSCI world index" and "BB EURO 1−5yr bond index". There are many ratios based on comparison of asset performance with performance of benchmark. In this research, we use nearly 40 different ratios and reveal the most important ones.

Deeper analysis of funds returns suggest to split data analyzed in to specific regime periods (no crisis, crisis in stock markets, crisis in bond market and global financial crisis) and to analyze them separately.

Our results reveal that in different regimes performance ratios correlate to each other quite differently. On the other hand, fund returns are stronger correlated during periods of crisis. It is interesting to note that some performance ratios begin to behave "strange" even before true financial crisis has started. Hence, they could be used as early warning indicators of local or global financial crisis.

Acknowledgement: This project has received funding from the Research Council of Lithuania (LMTLT), agreement No S-MIP-21-32.

Lithuanian Hate Speech Classification Using Deep Learning Methods

Eglė Kankevičiūtė, Milita Songailaitė, Justina Mandravickaitė, Danguolė Kalinauskaitė, Tomas Krilavičius

Vytautas Magnus University

egle.kankeviciute@vdu.lt

The ever-increasing amount of online content and the opportunity for everyone to express their opinions online leads to frequent encounters with social problems: bullying, insults, and hate speech. Some online portals are taking steps to stop this, such as no longer allowing comments to be made anonymously, removing the possibility to comment under the articles. Also, some portals employ moderators who identify and eliminate hate speech. However, given a large number of comments, an appropriately large number of people is required to do this work. The rapid development of artificial intelligence in the language technology area may be the solution to this problem. Automated hate speech detection would allow to manage the ever-increasing amount of online content.

In this work, we report the comparison of hate speech detection models for the Lithuanian language. We used three deep learning models for hate speech detection: *Multilingual* BERT, *LitLat* BERT, and *Electra*. The latter model we trained from scratch ourselves with Lithuanian texts that made ~70 million words. All three models were further re-trained to classify Lithuanian user-generated comments into three main classes: hate, offensive, and neutral speech. To adapt the models to the hate speech detection task, we prepared a pre-processed and annotated dataset. It has had 25 219 user-generated comments (hate speech – 2082, offensive – 7821, neutral – 15 316). The collected text corpus was also analyzed using topic modeling to reveal the most frequent topics of hate speech comments. The trained models were evaluated with accuracy, precision, recall, and F1-score metrics. *LitLat* BERT performed the best with a weighted average F1-score of 0.72. *Multilingual* BERT was in the second place with a weighted average F1-score of 0.63, and *Electra* took the third place with a weighted average F1-score of 0.55. As *Electra* models have the potential to perform better than BERT models, our future plans include retraining our base *Electra* model for Lithuanian with larger datasets.

Application of Machine Learning for Improving Contemporary ETA Forecasting

Gabrielė Kasputytė[1,2], Arnas Matusevičius[1,2], Tomas Krilavičius[1,2]

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

gabriele.kasputyte@vdu.lt

Cost-effective transportation of various goods has always been a necessary and challenging task for logistics companies to solve. With road transportation being the most popular mode for transporting goods, decreasing the time and mileage it takes to deliver freight can significantly increase a company's profitability. In this research, the possibility of improving the estimated time of arrival forecasts by ranking the drivers based on their behavioural data and estimating deviations from planned arrival time using different machine learning methods are analysed. TOPSIS and VIKOR methods are used for ranking the drivers, while forecasting of deviations is performed using five machine learning algorithms: Decision Tree, Random Forest, XGBoost, Support Vector Machine and k-nearest Neighbours. To ascertain the feasibility of the forecasting models, they are evaluated using the adjusted coefficient of determination, root square mean error and mean absolute error metrics. The research concludes that the ranking of drivers should be constructed using the VIKOR method. Moreover, the machine learning evaluation metrics reveal that the best forecasts are achieved using an ensemble model based on the Random Forest and Support Vector Machine models.

Handling Two-Dimensional Net Constraint: Its Formulation in MIP Terms and Empirical Studies of Solvers

Mindaugas Kepalas, Julius Žilinskas

Institute of Data Science and Digital Technologies
Vilnius University

mindaugas.kepalas@mif.stud.vu.lt

Suppose you want to place a number of points which are restricted to be on a two-dimensional net consisting of a set of straight-line segments (like a spider's net) in a way such that a certain objective is minimized. For example, the points could represent facilities, and the net constraint can represent a road net on the map, and we can interpret this in a way that a facility must be opened close to the road. With this placement, there is some cost associated, and we want to minimize it. We show that this net constraint can be formulated as a set of linear-mixed-integer equalities and inequalities, and we present our formulation here. We further perform an analysis of how well the currently available MIP and MIQP solvers manage this net constraint when the number of facilities increases or when the net becomes more complicated. This is done for two artificial problems, which in the end, are formulated as a mixed-integer-programming or mixed-integer-quadratic-programming problem.

Enrich Knowledge Graphs and Test Pre-trained Language Models in Graph2seq Tasks

Gražina Korvel[1], Alkiviadis Katsalis[2], Konstantinos Diamantaras[2], Elena Lloret[3]

[1] Institute of Data Science and Digital Technologies Vilnius University
[2] Department of Information and Electronic Engineering International Hellenic University, Thermi, Greece
[3] Department of Software and Computing Systems, University of Alicante, Alicante, Spain

grazina.korvel@mif.vu.lt

Modern natural language generation methods usually take the form of an encoder-decoder, which encodes the input sequences into latent space and predicts a collection of words based on the latent representation. Sequence-to-sequence (Seq2Seq) learning is one of the most widely used encoder-decoder based paradigms in this field. Recently, researchers find that structural knowledge is beneficial to addressing some troublesome challenges, e.g., long-dependency problems, and thus propose the graph neural network (GNN) techniques. For tackling particular tasks, graph-to-sequence (Graph2Seq) GNNs-based encoder-decoder models are increasingly discussed in the literature. Graph2Seq models have shown superior performance in comparison with Seq2Seq models in various tasks including neural machine translation, text summarization, and question generation. Disregarding success, Graph2Seq models also inherit challenges, for example encoding relations between distant nodes. This work explores the possibilities of enriching the graphs with external knowledge and testing pretrained language models in text generation tasks.

Speech Signal Enhancement for Audio Forensics: An Initial Study

Gražina Korvel[1], Gintautas Tamulevičius[1], Jelena Devenson[2]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Forensic Science Centre of Lithuania

gintautas.tamulevicius@mif.vu.lt

Audio Forensics is the field of Forensic Science relating to analyzing, processing, and evaluating audio recordings. The primary aspects of Audio Forensics are:

- determining the authenticity of audio evidence and integrity of audio recordings;
- reconstructing crime or accident scenes and timelines based on acoustic events;
- identification of speaking persons;
- transcription of speech content;
- performing enhancement of audio recordings to improve speech intelligibility and the audibility of noisy or low-level sounds, etc.

The main problems and challenges arise with the quality of audio recordings. Operational recordings are characterized by high additional speech and noise sources, echo and reverberation effects, lost segments of the target signal, limited bitrate, etc. Efficient enhancement techniques must be used to provide a reliable audio forensic analysis. Audio enhancement involves reducing unwanted sounds and enhancing the dialogue and other target sounds, improving speech intelligibility and the audibility of low-level voice within audio evidence recording. The fundamental challenge of forensic audio enhancement is to ensure that this process does not inadvertently degrade the speech inflections, nuances, and essential intelligibility needed for speech interpretation.

There is a need for research into the most reliable and effective enhancement techniques that can be explained and demonstrated to the Court. For this purpose, a systematic review and evaluation of enhance-

ment techniques are provided. Also, we examine the initial assumptions for applying these approaches and techniques to the Lithuanian language. The results of initial experiments with forensic audio are provided. The experience of the Forensic Science Centre of Lithuania in this field, methods, and tools used are also presented.

Evaluation of Measures Applied to Monitor the Lombard Speech Signal in the Presence of Noise

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

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

grazina.korvel@mif.vu.lt

The noisy environment decreases the quality and intelligibility of the speech signal. Contrarily, the so-called Lombard effect occurring in a person's talk while exposed to environmental noise enhances the speech signal. Even though the noise may contaminate the Lombard speech, it remains more intelligible. Our long-term goal is to build a machine learning-based system for generating speech with the Lombard effect that can automatically adapt to noise inference. In this research, a method enabling to monitor of the Lombard speech signal in the presence of noise is to be proposed. For this purpose, the variation of frequency characteristics of Lombard speech at different noise distortions is investigated. The signal frequency domain is considered in the form of frequency tracks containing the location of fundamental frequency and formats. To quantify the effect of noise on Lombard effect, an average formant track error is calculated. The effect of noise is investigated at varying levels of SNR, from -10 dB to 40 dB. For environmental noise, real-life noise recordings are employed. Experiments are carried out on the recordings made in the studio with a room with acoustic treatment. To obtain the Lombard effect while speaking, closed headphones playing back the interfering noise are used. The recording scenario includes both reading sentences and a conversation between two people. For quality assessment, the traditional objective image quality metrics and the authors' proposed method are used. Based on the experiment results, the recommendations for machine learning-based system for detecting the Lombard effect before enhancing speech automatically are made.

Regularisation Algorithms for Conic and Copositive Programming Problems

Olga Kostyukova[1], Tatiana Tchemisova[2]

[1] Institute of Mathematics, National Academy of Sciences of Belarus
[2] University of Aveiro, Portugal

tatiana@ua.pt

Regularisation of conic and copositive programming problems consists of transforming a problem to an equivalent form, where the Slater condition is satisfied and, therefore, the strong duality holds. In the talk, we will present a new regularisation algorithm Reg-LCoP based on a new concept of immobile indices suggested in our study of semi-infinite, semidefinite, and copositive problems and compare it with a Face Reduction Approach (FRA) suggested by J. M. Borwein and H. Wolkowicz for the abstract convex problem, and a similar regularisation approach of H. Waki and M. Muramatsu for conic problems. Based on the comparison of these approaches, we can conclude that being applied to linear copositive problems, the approach based on the immobile indices is more constructive since it permits to formulate the regularised problem explicitly and in a finite number of steps.

Challenges in Biomedical Signal and Image Analysis: Lessons Learned and Future Perspectives

Algimantas Kriščiukaitis[1,2], Robertas Petrolis[1,2], Renata Paukštaitienė[1]

[1] Dept. Physics, Mathematics and Biophysics
Lithuanian University of Health Sciences
[2] Neuroscience Institute
Lithuanian University of Health Sciences

algimantas.krisciukaitis@lsmuni.lt

Biomedical signal and image analysis has become a major part in many areas of biomedical research. Numerous papers are published on various analysis topics, including registration, segmentation, quantification or even artificial intelligence-based clinical decision support. The validation and evaluation of newly elaborated methods usually were based on the authors' personal data sets. The rapid development of computation and communication technologies during several past decades allowed the establishment of open-access databases of biomedical signals and/or images containing specialised data for algorithm development and direct comparison of the results achieved by several competing solutions. Around the same time, several initiatives to solve globally actual biomedical diagnostic problems related to biomedical signal or image analysis arose. The evocation of Worldwide interest in the solution of such problems and benchmarking of newly elaborated methods is realised in so-called "Challenges" – the events where volunteer participants are provided with annotated specialised data and their achieved results are compared by quantitative estimates. Usually, motivation of eventual participants is boosted offering some valuable awards for the best results. But, are such initiatives indeed revealing the best solution for the particular clinical diagnostic problem? - The question which often rises for many participants of such events. The importance of principles of "Challenge" design and choice of proposed scoring metric for results evaluation will be discussed based on published analysis and personal experience obtained participating in "Challenges" during past 15 years.

Conceptual Framework of Data Science for Good Squad

Dalia Kriksciuniene, Virgilijus Sakalauskas, Giedrius Romeika

Vilnius University

dalia.kriksciuniene@knf.vu.lt

The idea of the research is based on the EU initiative «Data science for good» (DS4G), which aims to effectively use data collected by public organisations due to digital transformation of society. These data are related to health, education, legal environment and security, development of labour resources, energy, transport, sustainability, climate change problems, and other solutions.

The DS4G concept tackles the problem of how to help public organisations not only to collect but also to extract useful information from data, allowing them to make decisions useful to the society. The use of data is complicated not only due to the abundance of collected data but also due to their different sources, collection formats, insufficient data «cleanliness», and the lack of specialists in IT, data science and analytics.

The efforts to bring together specialist communities and solve the data analytics problems raised by the DS4G initiative are already being made. Leading companies in the field of technology, such as Facebook, Google, and Amazon are becoming more and more actively involved in solving these issues.

The objective of the research is to propose the conceptual framework to create an innovative data science hub model (Data for Good squad), which would ensure the creation of effective workflows of operational processes, preparation of training materials, methodologies for identifying company problems and needs, management of risk and sustainability factors, and a building a relevant network of data science specialists. To implement this idea, volunteer teams of specialists in the field of data science and analytics are being assembled.

The research is inspired by the international EU project EPSILON «European Platform for Data Science: Incubation, Learning, Operations and Network». Project coordinator – Harz University (Germany), Nova SBE Institute of Science, Business and Economics (Portugal), University of Cyprus (Cyprus), Vilnius University (Lithuania).

Dimensionality Reduction for Financial Distress Detection

Dovilė Kuizinienė, Tomas Krilavičius

Vytautas Magnus University

dovile.kuiziniene@vdu.lt

Identification of a business's financial distress is a topic that became relevant together with the birth of money and business. Different methods and tools were used to identify it, and over the last several hundred years different quantitative and statistical methods were getting standard, e.g. for credit scoring. However, with the omnipresence of digital technologies, big data, and Artificial Intelligence, new methods and approaches are being investigated and applied for credit scoring, insolvency, financial distress, and other business indicators analysis and detection.

In this study, many data sources related to the subject are analyzed, including tax office, social security (e.g. SODRA), courts, Central Bank of Lithuania, Statistics department, etc. It leads of determination 977 different features, including not only financial indicators but also other equally important features like changes in employees, judicial events, managers, etc. The purpose of this study is to compare different dimensionality reduction techniques for financial distress recognition and important variable extraction. In this research financial distress is understood as a 'bad' situation of the company, which was detected at least in one Lithuanian government register. The analysis period is from 2016 to 2022. The research sample is 274105, which contains 2.8 % of financial distress. The data was normalized and split to train and test samples. Information from last year is included in the testing sample, which makes up 26.3% of the whole sample. For the training sample used the SMOTE technique to ensure the balance of classes. The research compares different feature selection techniques with machine learning algorithms for financial distress detection. The evaluation of models is used different metrics (accuracy, AUC, specificity, sensitivity, F-score).

Threat Modeling in RPA-Based Systems

Anastasiya Kurylets, Nikolaj Goranin

Vilnius Gediminas Technical University

anastasiya.kurylets@vilniustech.lt

Robotic process automation (RPA) is a family of business process automation approaches and technologies based on the use of software robots and artificial intelligence. Nowadays, RPA is being actively developed and used in IT industry market. As any new technology, RPA technology has a number of potential cybersecurity weaknesses, caused either by fundamental logical mistakes in the approach or by cyber-human mistakes made during the implementation, configuration and operation phases. Taking into consideration the widespread of RPA in many fields, especially in the banking sector, fighting RPA cyber threats becomes a critical task. The classical threat management approach suggests that before finding the countermeasures, it is necessary to identify the existing threats. One of the methods for doing so is threat modelling. Threat modelling is a description of a global security problem, an approach to identifying and categorising possible risks, such as vulnerability or lack of protection from protection, as well as setting priorities for eliminating threats or satisfying consequences, considering a set of problems, such as security analysis of conventional security measures based on nearby information systems and surveys, the most likely attacks, their methods, targets and search engines. Still, despite the RPA security topic importance and notoriety of the threat modelling approach in cybersecurity, not too much research is being done. In this research, we present threat modelling for RPA case scenarios in the financial sector. Microsoft Threat Modeling Tool was used as a threat modelling tool that allows the identification of threats at the software design phase, thus minimising further risks. The construction of case models made it possible to identify risks, the categories to which they belong, their description with the possibility of raising or lowering the priority of the threat, and analyse trust boundaries for software robots on a real sample.

A Multi-Objective Optimization Algorithm for EO Data Processing Based on Dask Library

Arthur Lalayan[1,2], Hrachya Astsatryan[1], Gregory Giuliani[3,4]

[1] Institute for Informatics and Automation Problems of NAS RA, Armenia
[2] National Polytechnic University of Armenia
[3] Institute for Environmental Sciences, University of Geneva, Switzerland,
[4] UNEP/GRID Geneva, Switzerland

arthurlalayan97@gmail.com

Earth observation (EO) data are widely used for environmental monitoring, urban occupation analysis, or risk detection. The parallel processing of EO data using High-Performance Computing (HPC) or cloud resources improves the processing performance of the growing amounts of geospatial data. Parallel computing frameworks divide the workflow into chunks for further parallel processing on several independent computational worker nodes to speed up the simulation time.

The distributed computing libraries scale the workflow using mainly a master-slave architecture, such as the open-source parallel python Dask library. Multiple computational and data platforms and environments can be configured and customized for the Dask, including virtual, physical, cloud-based, and on-premises solutions.

For efficient and optimal processing of EO data, it is necessary to manage and consider several aspects, such as the cost and the utilization of a set of running computational nodes. Therefore, finding a trade-off between cost and performance is an actual challenge, as it depends on the input data size, the number and the complexity of processing instructions, and other factors. The article aims to propose a Pareto-based collaborative multi-objective optimization algorithm, which is implemented on the suggested EO data scalable processing platform. A multi-objective optimization non-dominated sorting genetic algorithm is applied to the simulation dataset to find the optimal point considering cost and performance.

The evaluation results of the algorithm for several EO data processing workflow is presented, which offers users the optimal amount of resources. The normalized difference vegetation index (NDVI) index has been used for the experiments using Dask clusters with a different number of worker nodes. For each case, the processing time is found, and based on it, the price is calculated.

Facility Location with Ranking of Location Candidates Using High-Performance Computing Systems

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

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

algirdas.lancinskas@mif.vu.lt

Facility location problems deal with finding optimal locations for facilities providing goods or services to customers in a given geographical area. There are various models of facility location problems which vary in properties such as customer behaviour rules, location space, which can be continuous or discrete, constraints for the location of facilities, etc. Depending on the properties of the problem, appropriate solution methods should be used to effectively solve the problem. This paper focuses on the discrete competitive facility location problem for an entering firm, which is important for firms entering the market by choosing locations for the new facilities from a given set of location candidates, taking into account competition for the market share with preexisting facilities owned by other firms. The heuristic algorithm based on the ranking of location candidates is proposed and experimentally investigated by solving different instances of the facility location problem with the Pareto-Huff customers behaviour rule. The parallel version of the algorithm is developed to enable the solution of complex problems using high-performance computing systems. The results of the experimental investigation demonstrate the advantage of the proposed strategy of ranking location candidates over the previously proposed strategy. The parallel version of the algorithm demonstrates an efficient solution to the problem using more than one hundred computing nodes with almost linear speed-up.

Big Data Processing System for Lithuania Economic Activity Nowcasting

Mantas Lukauskas[1,2], Vaida Pilinkienė[2], Jurgita Bruneckienė[2], Alina Stundžienė[2], Andrius Grybauskas[2], Tomas Ruzgas[1]

[1] Department of Applied Mathematics, Kaunas University of Technology
[2] School of Economics and Business, Kaunas University of Technology

antas.lukauskas@ktu.lt

The assessment of economic activity is an important assessment of the state's economy, which allows assessing the current situation, as well as predicting future prospects. The increasing amount of data every year allows this data to be used in the forecasting of economic processes. However, due to the large amount of data, its rapid renewal, and its diversity, it is difficult to evaluate it in traditional ways. Traditional methods of assessing economic activity use monthly or quarterly data, which are no longer appropriate in the face of various economic shocks. A good example of this is the COVID-19 pandemic or the war in Ukraine, which affects the state's economy quite quickly. For this reason, it becomes important to evaluate not only traditional economic indicators, but also various alternative ones collected from various openly available sources. The purpose of this work is to present a possible economic activity assessment system that collects, processes, transforms and visualizes Lithuanian economic activity. The developed economic activity forecasting system allows you to automatically collect text information, prices of products and services, real estate and others. And using machine learning methods, this data is turned into valuable insights, which can be used in state, business and other decision-making.

Data Clustering Based on the Modified Inversion Formula Density Estimation

Mantas Lukauskas, Tomas Ruzgas

Department of Applied Mathematics
Kaunas University of Technology

mantas.lukauskas@ktu.lt

Data research is widely used in various fields such as business, production, online trade, consumer services, and other fields. Due to such a large application of data mining, the field is receiving a lot of attention. Data clustering is an unsupervised type of machine learning that is also widely used in data mining. In data clustering, the main goal is to divide objects into separate, unknown groups to have as many similar objects as possible in one group. Making such groups allows you to find hidden relationships between data. Data clustering is applied in such areas as bioinformatics, feature selection, pattern recognition and others. Although there are many different methods in data clustering, data clustering itself is a complex task. Due to different data structures, different clustering methods work well only under certain conditions, so the need for these methods remains high. One of the most used data clustering methods is the k-means method, which is relatively simple, but can work effectively in good conditions. Most clustering methods perform poorly in the presence of outliers in the data, and the previously mentioned k-means method suffers from this drawback, as do GMM, BGMM, and some other methods. Recently, various researchers have been paying a lot of attention to different density estimation methods, as well as robust modifications of these methods, such as soft constrained neural networks, and others. Due to such a demand for density estimation, this paper aims to evaluate the accuracy of a new clustering method based on modified inversion formula density estimation. The obtained results show that this developed method is competitive compared to the currently most popular methods (K-means, GMM, BGMM). Based on the clustering results, it can be observed that the MIDEv2 method works the best with generated data with noise in all datasets (0.5%, 1%, 2%, 4%

noise). The interesting point is that a new method can cluster the data even if the data do not have noise/outliers, for example, one of the most popular Iris dataset. However, there are also possible shortcomings of the method, since in case of large dimensions (d>15) the method is difficult to apply, but this problem will be solved in further iterations of the method.

Towards Generation of Phishing Texts

Justina Mandravickaitė, Eglė Kankevičiūtė, Milita Songailaitė, Bohdan Zhyhun, Danguolė Kalinauskaitė, Tomas Krilavičius

Vytautas Magnus University

bohdan.zhyhun@vdu.lt

We present work in progress for thematic email generation for phishing prevention. One of the ways to prevent data theft is to develop solutions that enable the automatic identification of fraudulent emails, thereby reducing the likelihood of individuals, companies, or public institutions becoming victims of fraud and preventing the possible leakage of sensitive, confidential, or other information. The development of our solution includes a natural language generation methodology based on machine learning, natural language processing (NLP) as well as rules. NLP involves semantic analysis as well as language and vocabulary control. Document plans for email generation are constructed by employing a rule-based engine to control the content of the generated text by defining the logical structure of the text to be generated, expressed communication goals, and data usage within the text. This combination allows variation of text structure to avoid repetitiveness and rigidity in the generated emails. For the development of our tool, we use anonymised emails (fraudulent and regular) obtained from publicly available datasets, as well as those provided (in anonymised form) specifically for research purposes by the project partner. The solution focuses on simulations of the types of data theft that require users to click on web links to occur. The target audience for the use of the tool is universities and other research institutions, R&D companies, business enterprises, the military (e.g. for wargaming simulations), and institutions ensuring state security (for simulating various cyber actions and training).

Latvia's Regional Disparities: Comparison of Interwar Period with Modern Latvia

Jurgita Markevičiūtė[1], Zenonas Norkus[2], Adomas Klimantas[2]

[1] Institute of Applied Mathematics, Vilnius University
[2] Institute of Sociology and Social Work, Vilnius University

jurgita.markeviciute@mif.vu.lt

There is a lack of information, based on quantitative data, about societal and economic developments in the Baltic states during the last 100 years. The data we have usually is fragmented and often not comparable. Thus the cross-time comparison is challenging to perform. This work aims to provide quantitative data about economic transformations in the regions of Latvia over 100 years. We evaluate the benchmark GDP of Latvia in 1935 and the time series of GDP per capita for the interwar period. Further, applying a methodology tested in recent research on the economic development of Europe's regions since 1900, we decompose the GDP series to a regional level to explore trends in the economic disparities between regions inside Latvia. Next, we apply a cross-time adjustment for the regions, since they changed during 100 years and compare the regional inequality during the interwar period and modern Latvia.

Computer Vision for Used Car Parts Recognition

Jonas Matuzas, Mindaugas Šipelis

Institute of Computer Science
Vilnius University
UAB Ovoko

jonas.matuzas@mif.vu.lt

A circular economy is now gaining popularity. It is important not only for reusing used car parts but also for ecology (a lot of used car scrapyards exist, and some of them pollute the environment). Uploading as much and as fast as possible used car parts are important for selling and for utilisation (if it is not sellable). In the company, Ovoko developed a Computer Vision (CV) system which detects all necessary part numbers, category, an automaker this information helps prefill all necessary fields for the selling platform. CV systems are faced with challenges: used parts part numbers are printed in very different ways: different fonts and surfaces (plastic, metallic, glass, etc.). Because it is not new parts - they are rusty, part numbers can be damaged. Photo quality is with various rotation angles and zooms. CV pipeline consists of 6 deep-learning models. Information later goes to the LSH (not exact match) algorithm to search the existing database. All models deployed in the AWS cloud using sagmaker multi-container endpoints. The system simplifies used car part upload, and now, you do not have to be an automotive expert to do that.

Extraction of Microservices from Monolithic Software Based on the Database Model

Dalius Mažeika, Edgaras Kazimieras Kazlauskas

Vilnius Gediminas Technical University

dalius.mazeika@vilniustech.lt

Most of the modern enterprise-level applications are built based on microservice architecture. It allows to improve software development agility and flexibility and makes this process faster. Microservices can be deployed independently, so it opens up opportunities to release application updates in a shorter time. Decomposed monolithic systems into microservices can increase the performance of the application by scaling those microservices that actually need resources. In this research, we analysed the software reengineering problem related to microservice identification in monolithic software systems through relational database decomposition. A physical database model is designed based on business processes and workflows; therefore, database entities and types of relationships can be used to identify microservice candidates. The database schema was used to extract the entities and relationships, build a graph, and calculate the weight coefficient of the relationships. Weight coefficients were obtained by transforming the extracted data into a graph, where vertices represent database tables, and the edges are the relations between them. The weights were calculated based on the number of edges, their direction, degrees of ingress and egress at vertices, type of entity connection, and additional information such as business domain knowledge. Candidate microservices were identified based on the clusters of database entities defined by weight coefficients. The community fast unfolding algorithm was used for clustering. Five open-source ERP systems were tested to validate the proposed method. Extracted data from the database schema was used to build a graph applying the ForceAtlas2 layout algorithm. Investigation of microservice extraction was performed by changing resolution while class modularity was used for comparison of the results. It was found that the number of

clusters depends on the resolution, while the best results were achieved when resolution values of 1 or 2 were used. Analysis of modularity revealed that higher values of modularity lead to better cluster identification accuracy.

Comparison of Fuzzy Sets Based on the Concept of Imprecision

Jolanta Miliauskaitė[1], Diana Kalibatienė[2]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Vilnius Gediminas Technical University

diana.kalibatiene@vilniustech.lt

Since 1971 introduced different types of fuzzy sets, the question arose as to how these fuzzy sets differ and which of these fuzzy sets are more suitable for certain application domains. Some of them aimed at solving the problem of developing the membership degrees of the elements. Others focused on representing the uncertainty linked to the considered problem. Nevertheless, from the application viewpoint of different types of fuzzy sets, there should be a method to implement the imprecise concept by those fuzzy sets in a reasoning engine and use it for automated inference. Consequently, we are dealing with the computational complexity, complexity of fuzzy rules, complexity of developing membership functions, and data complexity. Therefore, a more comprehensive study is needed to compare different types of fuzzy sets. In this research, we present an initial study of different types of fuzzy sets based on the concept of imprecision and their historical occurrence. The results of this research allows us to supplement existing knowledge on fuzzy sets by systematising them.

ICPUTRD: Image Cloud Platform for Use in Tagging and Research on Decomposition

Audris Mockus[1,2], Anna-Maria Nau[2]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] University of Tennessee, USA

audris@utk.edu

Big data systems have the potential to transform how forensic anthropology research is conducted if the typical research workflow of experimenting with physical artefacts (e.g., bones) could be replaced or augmented by similar activities based on digitised archive data. The aim of this study was to develop and evaluate the feasibility of Image Cloud Platform for Use in Tagging and Research (ICPUTRD) that enables forensic research and casework on a large collection of digital photos documenting longitudinal human decomposition. The paper describes key requirements, design considerations, aspects of implementation, and a user study evaluating the feasibility of conducting several forensic tasks involving forensic experts and non-experts., The tasks included search for and tagging of photos with features based on a forensic nomenclature (i.e., anatomic, stages of decomposition, scavenging, etc.). The user study confirmed the feasibility of using ICPUTRD in forensic research and suggests its potential as a practical tool for law enforcement.

On Recognising Emotion of Sadness in Images of a General Nature Using CNN

Modestas Motiejauskas, Gintautas Dzemyda

Institute of Data Science and Digital Technologies
Vilnius University

modestas.motiejauskas@mif.stud.vu.lt

An increasing number of users are expressing their emotions through social media. Recognising emotions in images is becoming very important. The recognition of emotions in general images is gaining more and more attention from researchers. Such emotion recognition is more complex and different from conventional computer tasks. Due to people's subjectivity, ambiguous judgements, and cultural and personal differences, there is no single model for emotion assessment. Current models cannot directly train visual emotion features for a given emotion. We decided to analyse emotions using a categorical model, as this is a natural way of human expression in psychological research. Studies in psychology propose the presence of six distinct, basic and universal emotion categories: happiness, anger, sadness, surprise, disgust and fear. We chose to analyse the emotion recognition of sadness as a starting point to evaluate the performance of the models. The main points of reference for comparison were set by the Xception and EfficientNetV2 convolutional neural network (CNN) models. Preliminary experiments show that the EfficientNetV2 family of convolutional neural network models perform better compared to the older Xception model. The EfficientNetV2 models achieve up to 75.57% accuracy, while the Xception model achieves up to 72.53% accuracy in predicting whether an emotion image expresses a sadness class. Finally, we examined the choice of network parameters and their importance for the model results. Our findings show that different optimisers produce quite different results, with studies on appropriate learning strategies and indicator selection showing the importance of model-tuning strategies. We also tested and measured accuracy and F1-scores for predicting sadness recognition in the presence of different types of emotion representations. Preliminary experimental results show good perspectives for the recognition of a broader range of emotions in images of general nature.

Application of Convolutional Deep Neural Network for Human Detection in Through the Wall Radar Signals

Dalius Navakauskas, Julius Skirelis, Eldar Šabanovič, Mantas Kazlauskas, Borisas Levitas, Irina Naidionova, Michail Drozdov, Aleksandr Prisiažnyj, Matvei Kazharov

Vilnius Gediminas Technical University

dalius.navakauskas@vilniustech.lt

Detecting people through walls or ruins is considered essential for tactical and rescue operations. Ultra-wideband radars provide just this capability to detect human movement, breathing, or heartbeat through one or more walls. Frequency-modulated continuous-wave radars with at least one transmitting and two receiving antennas are particularly useful because they provide information about the distance, angle, and speed of objects relative to the radar. Typically, the signals are pre-processed using Fast Fourier Transform to obtain 3D spectrograms. The position of a human can be very difficult to discern in the 3D spectrograms because humans have multiple moving body parts that can move in different directions and at different speeds. Therefore, conventional maximum search gives results that are not accurate enough. Convolutional neural networks (CNNs) provide state-of-the-art performance for recognising people in visual images by analysing complex patterns. Radar signal spectrograms can be considered images in which people's patterns can be detected using CNN. It is clear that deep neural networks need large datasets for supervised learning, and for each specific radar, the specific dataset should be provided. Therefore, we developed an automated method to annotate radar signals based on synchronised video images. Human coordinates in space were estimated in calibrated camera videos by popular in robotics area visual tags. This process included camera calibration, ArUco tag detection, estimation of human position from the tags, 3D transformation from camera coordinate system to radar coordinate system and synchronisation with interpolation of coordinates. The new dataset with Range-Azimuth-Doppler and human coordinates was

created with more than 1100 pairs of input and output data. This dataset was used with a modified version of the RADDet neural network for real-time processing of radar signals on an embedded computer wirelessly connected to the radar. This network was based on ResNet backbone and YOLO detection heads. We obtained promising results for human detection with the possibility of changing the detection threshold. The processing speed in the embedded system is 16,7 ± 0,43 ms for each radar data frame. Currently, it takes 176 to 352 ms to acquire a radar data frame; therefore, in theory, more than four radar data frames can be processed during that time. We demonstrate our dataset creation pipeline and CNN training process, along with initial real-time processing results. This work presents research that is a part of the technological development project "Prototype of gateway for Artificial Intelligence-based Through-Wall Imaging Radar data" funded by the Innovation Agency of Lithuania (TPP-04-032). The project is a collaboration between researchers at Vilnius Gediminas Technical University and JSC "Geozondas".

Evaluating Synthesized Speech: The Cognitive Approach

Gediminas Navickas, Gerda Ana Melnik-Leroy

Institute of Data Science and Digital Technologies
Vilnius University
gediminas.navickas@mif.vu.lt

Modern speech technologies allow the synthesizing of almost perfect synthetic speech. However, despite the increase in computational power, the size of the datasets and the elaboration of the algorithms, this "almost" has still never been surpassed. We propose that one of the major reasons for the impossibility to come up with a 100% naturally sounding synthetic speech is the lack of appropriate quality evaluation methods. When it comes to the assessment of synthesized speech quality, two techniques, employing so called "objective" and "subjective" measures, are usually used. While the objective ones are suitable to evaluate the resulting physical speech signal and its properties, they do not reflect how this signal is perceived by the human listener. Yet, a vast body of literature in psychoacoustics and psycholinguistics show that even tiny distortions of the speech signal can have detrimental effects on speech processing quality and speed. On the other hand, the subjective measures traditionally used (Mean opinion score; intelligibility and comprehension tests (Soares et al., 2018; Viswanathan & Viswanathan, 2005)) are not sufficiently informative, lack methodological precision and statistical analyses to assess the small, but essential distortions in modern synthetic speech.

Since speech perception has been widely studied in psycholinguistics (Lieberman et al., 1965; Studdert-Kennedy & Hadding, 1973; Niebuhr, 2007) and cognitive science in general, we propose that the measures of synthesized speech quality should be created with regards to the cognitive mechanisms underlying the processing of linguistic phenomena. In other words, we argue that a genuinely efficient quality measure should be sensitive to the peculiarities of human speech perception. This can

be achieved by applying behavioral methods, that have been already developed and shown to be effective in cognitive science (Malisz et al., 2019; Wagner et al., 2019; Winters & Pisoni, 2004).

We will provide an overview of the most suitable options available and discuss their applicability to assess modern TTS.

Agile Application Development Management Using Causal Knowledge

Karolis Noreika, Saulius Gudas

Institute of Data Science and Digital Technologies
Vilnius University

karolis.noreika@mif.vu.lt

Agile management methods and tools are being widely used to improve Enterprise Application Software (EAS) development. The paper deals with the problem of misalignment between business strategy and application software development. Our experience using Agile management tools like Atlassian «Jira» shows the lack of coordination between software development management and business management content. This paper discusses an approach to enhance Agile management tools with Artificial Intelligence capabilities to deal with software and business alignment. The causal modelling paradigm is used to construct an internal model of an AI-based system. The paper aims to base the development of Agile applications on causal modelling and define the architecture of an intelligent project management tool. The management transaction (MT) is the causal model used to redefine the Agile management hierarchy. The content of feedback between two adjacent levels in the Agile hierarchy, such as theme – initiative – epic – user story, is revealed using the MT framework. Enterprise architecture framework MODAF is used to specify the content of MT and establish the links between Agile management hierarchy levels. By defining the internal model of the Agile management hierarchy using MT and MODAF constructs, the obtained causal knowledge is expressed as a new attribute in the Agile management tool. This ensures checking the integrity of project content with enterprise strategy objectives and reduces the misalignment between business strategy execution and software development solutions.

The 1D Wada Index for the Classification of Digital Images of Concrete Cracks

Ugnė Orinaitė, Juratė Ragulskienė

Kaunas University of Technology

jurate.ragulskiene@ktu.lt

The Wada index has been recently introduced for the detection if a given basin boundary is a Wada boundary. The Wada index is based on the weighted and truncated Shannon entropy and does represent the number and the distribution of different colours (attractors) in the two-dimensional phase space of initial conditions. The Wada index is based on the standard box counting algorithm. That makes the algorithm for the computation of the Wada index conveniently applicable for different basins of attraction represented as color digital images.

With the recent advances in machine learning, the development of ANN- and CNN-based algorithms has become a popular approach for the automated detection and identification of concrete cracks. However, most of the proposed models are trained on images taken in ideal conditions and are only capable of achieving high accuracy when applied to the images of concrete cracks devoid of irregular illumination conditions, shadows, shading, blemishes, etc..

A 1D modification of the Wada index is presented in this paper. It is demonstrated that the 1D Wada index algorithm can be used as an efficient pre-processing tool for digital images contaminated by the additive and/or optical noise. The 1D Wada index algorithm helps to reduce the additive noise in images of concrete cracks what enables better classification based on deep learning algorithms.

Alexnet convolutional neural network (CNN) is used to train the classification model on Mendeley Concrete Crack Images for Classification dataset. It is demonstrated that the application of the 1D Wada index algorithm helps to improve the classification accuracy of concrete crack images contaminated by noise up to 98% what corresponds to the industrial standard in the field of autiomatic crack identification.

Improving the Implementation of Quantum Comparators

Francisco Orts[1], Gloria Ortega[1],
Ester M. Garzón[1], Ernestas Filatovas[2]

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

francisco.orts@ual.es

Quantum computers allow it to speed up the resolution of specific problems such as number factorisation or searches in disordered data sets. However, it is still at an early stage of development, and current quantum computers have few resources available, in addition to a high error rate. Common operations, such as addition, subtraction, or the one that concerns us in this work, comparison, are widespread in quantum algorithms. For the reasons above, having optimised versions of these basic operations in terms of resources and error tolerance is particularly valuable. It can make the difference between a quantum algorithm being executable or not in current quantum devices. In this work, we present a comparator that optimises the number of qubits and error tolerance with respect to the rest of the comparator circuits available in state-of-the-art quantum computing. The main contribution of this research is that the circuit is completely built using only Clifford+T gates.

Digital Twins in Manufacturing

Vytautas Ostasevicius

Kaunas University of Technology

vytautas.ostasevicius@ktu.lt

Digital twins, the Internet of Things (IoT) and the electronic, physical system are key concepts in the „Industry 4.0“ Digital twins use cloud-connected machine sensors to load real-time operational data, creating state-of-the-art virtual simulations of real-world machines, where IoT is the key to modelling large-scale digital twins. The term „digital twin“ covers virtual and physical replicas of a product, a machine, or an entire manufacturing process that are used as a specific test-bed for the process or product, where the changes made can be simulated before being implemented in real life. The desired result is the quality of the product, often related to the dynamics of the cutting tools, which could be assessed using virtual or physical twins and predicted by artificial intelligence (AI) methods. The paper presents detailed methodologies developed using artificial neural networks and machine learning, supported by process modelling in digital twins, based on technological solutions and data validation.

Extrinsic Evaluation of Word Embedding Models Using Semisupervised NLP Tasks: The Case of Sentiment Analysis

Mindaugas Petkevičius, Daiva Vitkutė-Adžgauskienė

Vytautas Magnus University

daiva.vitkute@vdu.lt

Two main approaches are used for evaluating the performance of a word embedding model - intrinsic and extrinsic evaluation. Intrinsic evaluation metrics measure the quality of a model independent of a specific application. However, the best way to evaluate a model is by applying it to an Natural Language Processing (NLP) application and measuring how much this application improves. Thus, different NLP tasks, such as sentiment analysis, are employed for model evaluation.

Semisupervised learning approaches can be used as a testbed for evaluating word embedding and transformer-based models in NLP tasks. Training a classifier in a supervised learning approach requires large amounts of labelled data in order to achieve acceptable quality measures. Meanwhile, a semisupervised model is capable of achieving comparable quality with small training sets by additionally employing pre-trained models.

In this research, models are tested and compared by applying them to a semisupervised social text sentiment analysis task for Lithuanian language social texts. A small dictionary, initially derived from a relatively small training dataset of labelled 500 reviews, is further expanded using several word embedding models.

A dataset containing 10,000 online Lithuanian language reviews from the e-commerce domain was used for training neural network classifiers and word embedding models in sentiment analysis experiments. For comparison, two instances of pre-trained word embedding models were also used. One of these pre-trained models uses Common Crawl and Wikipedia texts, while the second one was trained, additionally using crawled social media texts.

As a result, a semisupervised learning approach with a fastText word embedding model trained on a domain-specific dataset was able to reach the F1 score of 81.6%, which is comparable to the F1 score of 83% for a CNN classifier in our sentiment analysis experiment.

It is shown that the accuracy of sentiment analysis using a semisupervised learning approach with the expanded dictionary is considerably higher, compared to the accuracy of the supervised learning approach with a small training dataset, and is close to the accuracy of a supervised learning approach with large training datasets. It is also shown that word embedding models for morphologically rich Lithuanian language, trained on a domain-specific dataset, outperform corresponding pretrained word embedding models in semisupervised learning tasks.

Decision Support for Many-Objective Optimisation

Dmitry Podkopaev

Systems Research Institute
Polish Academy of Sciences, Poland

dmitry.podkopaev@gmail.com

The abundance of data and the development of information technologies enable solving of large-scale decision-making problems at high levels of detail. In some applications, it is reasonable to consider many individual objectives to accurately address the balance of interests. Moreover, if multiple scenarios are introduced to model uncertainty, the number of objectives increases many-fold. However, dealing with many objectives using traditional decision-support tools is problematic due to the limitations of human cognitive capacities. We present recent developments of techniques that provide interactive decision support for solving many-objective optimisation problems.

Risks and Solutions of Bitcoin Instant Payments With Zero-Confirmations at Physical Points of Sale and Services

Eimantas Rebždys, Saulius Masteika, Kęstutis Driaunys

Vilnius University

saulius.masteika@knf.vu.lt

The aim of this study is to identify the risks associated with double spending fraudulent activity when accepting bitcoin payments with zero-confirmations at physical points of sale and services without the use of an additional layer of lightning network or creating an inner user's sub-network with the crypto exchange service provider. The study details and compare potential cyber-attacks like Sybil attack, Eclipse attack, Replay attack, 51% attack, Vector-76 and Replace by Fee mechanism vulnerability on the bitcoin network when processing instant payments. The objectives of the research is to examine the mechanism of possible double spending attacks, to evaluate the risks and relevance and to present possible solutions and preventive measures that can be applied to address loopholes in Bitcoin instant payments with zero confirmations. The results can be interesting for businesses with non-e-commerce activities and be applied to the development of an instant bitcoin payment prototype and use case solutions.

Acknowledgement: This project has received funding from European Regional Development Fund (project No 13.1.1-LMT-K-718-05-0006) under grant agreement with the Research Council of Lithuania (LMTLT). Funded as European Union's measure in response to Cov-19 pandemic.

EFECTAS: R Tool for International Large-Scale Assessment Data Analysis

Laura Ringienė, Audronė Jakaitienė

Institute of Data Science and Digital Technologies
Vilnius University

laura.ringiene@mif.vu.lt

The quality of education is an issue around the world. Researchers analyse not only national education data but also cross-country comparisons. High volume and special structure International Large-Scale Assessment data (ILSA) such as PISA (Programme for International Student Assessment), TIMSS (Trends in International Mathematics and Science Study), PIRLS (Progress in International Reading Literacy Study), and others, are used to compare students' achievement in mathematics, science, and reading within and between countries. Such data can be analysed using commercial software such as SPSS, SAS, Mplus, etc., or the open source software R.

We reviewed five open-source R software packages for statistical analysis of ILSA data: BIFIEsurvey, EdSurvey, intsvy, RALSA, and svyPVpack. After analysing the advantages and disadvantages of each package, we have developed a new tool, EFECTAS, based on the BIFIEsurvey and EdSurvey packages. The tool is designed to analyse PISA, TIMSS and PIRLS data from 2015 onwards. With EFECTAS, the user can calculate descriptive statistics, correlation, linear and logistic regression, and create multilevel analysis models. The results of the functions are informative and presented in tables. The results of most functions are presented on screen and in an MS Excel document.

The Relation of Motivational Constructs to Reading Achievement in EU Countries From PISA2018

Laura Ringienė[1], Saulė Raižienė[1], Inga Laukaitytė[2], Audronė Jakaitienė[1]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Department of Applied Educational Science,
Umeå University, Sweden

laura.ringiene@mif.vu.lt

Researchers usually focus on student achievement when analysing the quality of education. The Programme for International Student Assessment (PISA), is a large-scale assessment that provides data on mathematical, science, and reading literacy for 15-year-old students. PISA also collects information about students' motivational, family, and institutional factors, which can explain differences in student achievement at the individual, school, and country levels. The effect of motivational constructs on students' achievement is very rare in literature. Starting to look at the impact of motivational constructs first requires sufficient evidence of the measurement invariance (MI). Three motivational constructs the work mastery, competitiveness, and fear of failure, were chosen to evaluate MI in PISA 2018. Every motivational construct was evaluated by three questions. Reading performance was chosen as an educational outcome. Because there is evidence that economic, social, and cultural status and gender affect student achievement (e.g., Thomas & Stockton, 2003; OECD2013), we controlled for these variables. In this study, we tried to answer the following questions: Do we have MI of motivational constructs across EU countries in PISA 2018? Do the relationships between motivational constructs and reading performance differ in EU countries?

The sample included 183824 15-years-old students from 26 EU countries in PISA 2018. The study consists of three stages. The overall sample and single-group CFA models for each EU country were constructed in

the first stage, and MG-CFA model for all EU countries was constructed in the second stage to test the MI of motivational constructs. The results of these models confirmed the MI of the motivational constructs. In the third stage, a two-level MG-SEM model was constructed to examine the effect of three motivational constructs on students' reading achievement. The obtained model fit indices indicate an acceptable model fit. The results of the MGSEM model estimates for the 26 EU countries show that at least one motivational construct is statistically significant for reading achievement in all analysed EU countries.

Automated Propaganda Detection Using Deep Learning Methods: A Pilot Systematic Review Research

Ieva Rizgelienė, Darius Plikynas

Institute of Data Science and Digital Technologies
Vilnius University

paulaviciute.ieva@gmail.com

In the context of the global information war, propaganda is disseminated massively and systematically through mass media, news portals and social networks to manipulate public mood and behaviour. This is aimed at demoralising, having a psychological effect and indoctrinating. Social groups, caught in the bubbles of propagandistic information flows, become radicalised and polarised not only in terms of political but also social behaviour, which leads to the loss of (i) mutual agreement, (ii) trust in government institutions, (iii) the cohesion of civil society, (iv) the ability to resist external influences, and (v) the overall social capital of society. The objective of this study is to review the latest five-year deep learning applications for automated propaganda detection. We apply PRISMA systemic review research method and corresponding meta-analysis. Obtained insights and observations serve to develop further a decision-support tool to detect propaganda in news articles and social media. For that matter, we explore the best methods to detect propaganda signs, classify them according to types of propaganda and rank them in English and Lithuanian language. In the presented research, we pay particular attention to the best language models and transformers capable of fine-tuning the textual data to the specific language context.

To Merge or Not to Merge Datasets? What Do the Experiments Show?

Paulius Savickas, Dovilė Kuizinienė, Tomas Krilavičius

Vytautas Magnus University

paulius.savickas@card-ai.eu

«To be or not to be» is a legendary question. We face a similar question when we have limited and hard-to-access data. In this case, using synthetically generated data, we do not know whether the assumptions used in generating it are correct and similar in reality. This study uses three different synthetically generated, publicly available datasets of money laundering cases. By applying Random Forest, Generalized Linear Regression, XgBoost, Isolation Forest and Ensemble machine learning methods, it was tested whether the merged datasets were better or not. After training the models on one dataset, the methods performed well when tested on the rest of the same set, but when tested on other sets, the models classified most of the values into one class, and the AUC score was around 50%. This showed that the datasets are generated based on different assumptions and cannot be verified. When the datasets were merged, and the models were tested on other datasets as well as the remaining test sample from the merged dataset, they performed more accurately. XgBoost correctly identified 84.4% of all money laundering cases and 96.3% of legitimate payments.

Security Risk Assessment for Blockchain-Based Identification and Authentication Method of Internet of Things Objects

Raimundas Savukynas

Faculty of Mathematics and Informatics
Vilnius University
raimundas.savukynas@mif.vu.lt

The Internet of Things (IoT) takes promises the possibility of connecting billions of objects into networks providing a wide variety of information.

The exponent growing number of connected heterogeneous objects brings security issues for the identification and authentication method of IoT. The major security risks for the identification and authentication method of IoT objects are identity spoofing, network disruption, signal jamming, software modification, unauthorised access, and information disclosure. A security risk assessment is an essential part of the information security process that identifies critical assets, threats, vulnerabilities in the IoT and then removes found hazards or minimizes the level of their risk by adding control measures and taking precautions with respect to security issues. Therefore, the ultimate goal of the security risk assessment is to measure the risk using a testing process of the identification and authentication method of IoT objects in order to obtain numerical values that could be used to compare the obtained security risk for each asset and process. Existing identification and authentication methods of IoT objects are based on a centralized model which has risk with a single point of failure.

Meanwhile, a blockchain-based identification and authentication method of IoT objects is considered an alternative that eliminates a single point of failure and uses a cryptographic hash function as the core of security.

A blockchain-based identification and authentication can help prevent the loss of identities, effectively detect frauds and mitigate critical risk issues, provide transparent and secure authentication of different

IoT objects. The aim of this work is to propose a structured methodological approach to identifying security threats and assessing related security risks for blockchain-based identification and authentication method of IoT objects. Resulting of this assessment, countermeasures are proposed to improve blockchain-based identification and authentication method of IoT objects.

Exploring the Limits of Early Predictive Maintenance Applying Anomaly Detection Technique

Artūras Serackis, Mindaugas Jankauskas

Vilnius Gediminas Technical University

arturas.serackis@vilniustech.lt

The aim of the presented investigation is to explore the time gap between anomaly appearance in continuously measured parameters of the device and a failure related to the end of the remaining resource of the device-critical component. In this investigation, we propose to use a recurrent neural network to model the time series of the healthy device parameters in order to detect anomalies by comparing predicted values with actually measured ones. An experimental investigation was performed on SCADA estimates received from different wind turbines with failures. A recurrent neural network was used to predict the oil temperature of the gearbox. The comparison of the predicted oil temperature values and actual measured ones showed that anomaly in the gearbox oil temperature could be detected up to 17 days in advance before the failure of the device-critical component. Performed investigation compares different models that can be used for temperature time-series modelling and the influence of selected input features to the performance of temperature anomaly detection.

Recommending Music Using Music Information Retrieval Methods and Deep Learning

Milita Songailaitė, Tomas Krilavičius

Vytautas Magnus University, CARD

milita.songailaite@vdu.lt

Various technological changes and the ever-increasing digitisation of information have forced the music industry to move gradually from physical recordings and live performances to online space. However, this also made the amount of available online content increase rapidly. Therefore, various recommendation systems were developed to deal with the vast amounts of available music. In this work, we propose a music information retrieval and deep learning-based music recommendation method, which combines various musical features to adapt to the human music similarity perception. We combined seven musical feature extraction methods to create this recommendation model: Mel Frequency Cepstral Coefficients, Chromagram, Tempogram, Zero-Crossing rate, Autoencoder, Variational Autoencoder, and OpenL3 embeddings models. First, the model was trained to identify the audio similarities using a database of 4039 songs from 11 popular genres. Then, the recommendation model was evaluated by comparing the methods' recommendation results with the experts' music similarity perception and music ranking ability results. The evaluation results showed that the Chromagram features gave the results which were the closest to the human music similarity perception.

Fabulator: A Synthetic Social Media Data Generator

Milita Songailaitė, Justina Mandravickaitė, Veronika Gvozdovaitė, Danguolė Kalinauskaitė, Tomas Krilavičius

Vytautas Magnus University

justina.mandravickaite@vdu.lt

There is a need for vast amounts of training and testing data for research and the development of artificial intelligence technologies. However, the usage of real-world data (especially social media data) is often protected by data protection regulations, such as GDPR. This makes a significant proportion of data unavailable to use in research. The problem can be overcome by generating synthetic data that imitates real-world data but still abides by data protection regulations. *Fabulator* is a synthetic generator of social media data where synthetic graph structures and synthetic text are combined. For text generation, we took pretrained dialogue response generation model (*DialoGPT-medium*) (Zhang et al., 2019) and retrained it with two different datasets, so the generation of conversations following different topics and perspectives would be possible. We used relevant *Reddit* data for this retraining, which resulted in 2 dialogue response generation models ("political" and "conspiratorial") which interact with each other. Our retrained models reached 0,58 mean *ROUGE* value, and 11,63 mean perplexity score, indicating good quality models. We got a low mean *BLEU* score which indicates that generated texts differ in their structure, though they are grammatically correct and meaningful. For the graph generation part, we chose to apply numbers of different events and links to be generated randomly by *fake social* (https://github.com/berfr/fakesocial) social network generator. It generates a simple social network consisting of fake users who have connections, makes posts, comment, and like these posts as well. User profiles utilise generated images as profile pictures, while generated text serves as posts and comments. The whole social network is packaged as a website. Generated graph met the requirements in terms of user profile and

network structure. We believe that the combination of graph and text provides data that is more relevant and that can be used for further research of social media interactions. Our generator is initially oriented towards *Reddit* data, though further research will focus on other platforms as well in order to find a more comprehensive, generalised solution.

A New Genetic Tourist Trip Design Algorithm for a Highly Personalised Globe-Trot Traveling Experience

Linas Stripinis, Remigijus Paulavičius, Ernestas Filatovas

Institute of Data Science and Digital Technologies
Vilnius University

linas.stripinis@mif.vu.lt

In this work, we consider the tourist trip design problem. To maximise tourists' satisfaction during their visit, choosing the most appropriate attractions from an extensive set is necessary. In addition to standard constraints such as the maximum tour duration, the constructed tourist trip design problem also considers practical conditions such as mandatory visits, time limits for different types of locations, the limits of similar attractions, etc. As this work is under active development, new constraints may be needed as the project progresses.

Due to the NP-hard nature, a new genetic-type metaheuristic algorithm is proposed to find the most optimal or near-optimal tourist trips for complex cases. We used an actual dataset from the city of London (UK) to demonstrate the proposed algorithm's effectiveness and potential. Results indicate that the algorithm solved the personalised tourist trip design problems efficiently and achieved high accuracy in a reasonable time.

This research has received funding from the Research Council of Lithuania (LMTLT), agreement No. S-MIP-21-53.

Signal Relationship Analysis of Prostate mpMRI T2w, DCE, DWI Sequences for Cancer Localization

Roman Surkant[1], Jolita Bernatavičienė[1], Jurgita Markevičiūtė[2], Ieva Naruševičiūtė[3], Mantas Trakymas[3], Povilas Treigys[1]

[1] Institute of Data Science and Digital Technologies, Vilnius University
[2] Institute of Applied Mathematics, Vilnius University,
[3] National Cancer Institute

roman.surkant@mif.stud.vu.lt

Prostate cancer is one of the leading causes of cancer death worldwide. Among males, prostate cancer has the second highest incidence rate after lung cancer (2018). Although death rates have been decreasing in some countries, it remains a considerable disease affecting many patients and early diagnosis and treatment are critical. Preliminary identification of cancer involves biopsy PSA protein screening, elevated levels of which indicate an increased likelihood of prostate cancer. PSA screening is then followed by a biopsy. Unfortunately, such testing is invasive and prone to false-negative and false-positive results, so a less invasive and more reliable procedure is needed. Currently, evaluation is based on multi-parametric MRI, which contains several imaging sequences, each having its own acquisition methods and purpose, and the final diagnosis is formulated based on all of them in conjunction. According to Prostate Imaging-Reporting and Data System (PI-RADS), a structured reporting scheme for MRI-based evaluation of prostate cancer, the main imaging sequences are T2-weighted (T2w), Diffusion Weighted Imaging (DWI), and Dynamic Contrast Enhancement (DCE). The decision-making of cancer assessment follows a rule-based system which differs depending on lesion location – either peripheral or transition zones of the prostate. For the peripheral zone, the primary sequence is DWI while the secondary is DCE; for the transition zone, primary and secondary imaging sequences are T2w and DWI, respectively. T2w imaging helps with characterising the size, form, and texture of the lesion. DWI characterisation is based

on proton diffusion-determined signal intensity, and clinically significant cancer appears brighter due to lower diffusion. DCE determines the rate of enhancement – rapid and early signal increase is associated with malignant tissue. This work is dedicated to investigating the quantitative relationship between the three imaging sequences. Such evaluation is accomplished by doing a pairwise comparison between T2w, DWI, and DCE signal intensities of benign and malignant prostate regions verified by MRI-targeted biopsy.

Discrete-Time Risk Models for Non-Life Insurance Business

Jonas Šiaulys

Institute of Mathematics
Vilnius University

jonas.siaulys@mif.vu.lt

The so-called risk renewal model with certain supplements describing the investment environment is commonly used to describe the non-life insurance business. The main part of such models is described by the flow of claims. If claims are considered to be integer-valued, then we get the so-called discrete time risk renewal model. If we additionally know that claims are identically distributed, then a homogeneous discrete-time risk model is obtained. The critical characteristics of such a model can be calculated using recursive formulas. If the homogeneity of the model is abandoned, it is also possible to apply analogous recursive formulas to calculate the critical characteristics of the model. However, additional problems arise related to finding initial values. The presentation will discuss two fundamentally different ways of finding the necessary starting values for a seasonal discrete-time risk model. The results presented in articles [1] and [2] will be discussed during the presentation.

[1] A. Grigutis, J. Šiaulys, Recurrent sequences play for survival probability of discrete time risk model, Symmetry, 12(12), 2111–2131, 2020.

[2] A. Grigutis, J. Jankauskas, J. Šiaulys, Multi seasonal discrete time risk model revisited, arXiv:2207.03196v1.

Company Recommendation Model: Empowering the Accounting System and Publicly Available Data

Rokas Štrimaitis, Pavel Stefanovič,
Simona Ramanauskaitė, Asta Slotkienė

Department of Information Technology
Vilnius Gediminas Technical University

simona.ramanauskaite@vilniustech.lt

Business management requires constant decision-making. Usually, the selection of suitable partners for collaboration is very intuitive, experience-based skill or requires the usage of data analytics. The most important aspect to get an explainable recommendation for possible collaboration between companies is data. In this research, a company recommendation model is created to incorporate both the accounting system and publicly available data. The accounting system data is used to estimate the collaboration effectiveness of existing collaboration cases. The publicly available data include company registration data and news portal article texts. The news article data include the detection of the company mentioning, estimation of its sentiment and context category. Separate models were developed for different data extraction and analysis. A combined company recommendation system was designed. Model validation with gathered test cases demonstrated 70% recommendation accuracy.

Acoustic Analysis of Pathologic Voice: What Is Done and What Is Next?

Gintautas Tamulevičius[1], Nora Šiupšinskienė[2], Monika Danilovaitė[1]

[1] Institute of Data Science and Digital Technologies, Vilnius University
[2] Department of Otolaryngology, Hospital of Lithuanian University of Health Sciences Kauno Klinikos

gintautas.tamulevicius@mif.vu.lt

Acoustic analysis-based pathologic voice assessment is not a new task and research activity. Sixty years ago, researchers focused on the changes in the voice generation process, the physiological causes of these changes, and the acoustic features of pathologies. Nowadays, artificial intelligence-based approaches dominate studies: end-to-end pathologic voice analysis, intelligent feature selection techniques, and evaluation of pathology degrees.

Despite the apparent progress in the field of pathologic voice analysis, some signs of research stagnation or cyclicality can be seen. Acoustic features proposed and studied 60 years ago make a comeback today with a deep networks-based classification paradigm. This paradigm also revived the attempts to deal with the problem by applying artificial neural networks, the idea that had vanished during the last century ending decades. Even today, we cannot define the relationship between the results of the objective approaches we apply and the subjective techniques widely used in clinical practice.

In this study, we present the results of the acoustic analysis of the pathologic voice review. We have analyzed studies starting from the first results until recent ones. We have summarized and grouped these results by decades. We have tried highlighting the critical achievements and trending directions in pathological voice analysis: features and feature sets, feature selection, and decision-making. Based on this grouping and group attributes, we have tried to detect problematic and weak points in research and their impact on results. At the end of this study, we have attempted to summarize the analysis results and tried to identify further directions in this domain: what more should be done to make an acoustic analysis one of the standard screening testing techniques acceptable to clinicians.

Convolutional Neural Network Approach for Anomaly-Based Intrusion Detection on IoT-Enabled Smart Space Orchestration System

Vikas Upman, Nikolaj Goranin, Antanas Čenys

Vilnius Gediminas Technical University

antanas.cenys@vilniustech.lt

Artificial intelligence approaches have been used in a variety of industries, including business, engineering, management, science, the military, and finance, thanks to the rapid development of AI in information processing applications. Technologies were competing to offer the finest service to people. The concept of smart infrastructures, such as the smart grid, smart factories, or smart hospitals, was created by developing Internet of Things (IoT) technology. The Internet of Things is a new technology that is rapidly growing in popularity nowadays. IoT devices, on the other hand, are a soft target and open to attack. Security analysts and attackers are engaged in an ongoing conflict. Because attacks are always changing, researchers and security analysts are under pressure to adapt to current threats by strengthening their defences. Finding anomalies in IoT device-based data is extremely difficult to maintain due to the complexity and diversity of today's malicious activities. In the age of artificial intelligence, this study focuses primarily on IoT-based anomaly detection applications and development trends in the area of IoT security. This study offers a deep learning-based intelligent system for detecting anomalies in the distributed smart space orchestration systems DS2OS dataset to prevent security breaches. The proposed convolutional neural network approaches achieved, i.e., ReLu_CNN, test accuracy of 99.1% with a false-positive rate of 0.9% and ReLu_CNN-LSTM, test accuracy of 98.8% with a false positive rate of 0.8%. These cunning methods investigate anomalies and attacks in IoT-enabled systems. Also, this paper successfully classifies the eight classes of attacks; Normal (NL); Scan (SC); Malicious Operations (MO); Denial of Service (DoS); Spying (SP); Data Probing (DP); Wrong Setup (WS); Malicious Control (MC) for which both techniques; ReLu_CNN and ReLu_CNN-LSTM, perform and optimise results differently.

High-Frequency Cryptocurrency Trading in the Face of Covid-19

Mantas Vaitonis, Konstantinas Korovkinas

Vilnius University

mantas.vaitonis@knf.vu.lt

The number of financial institutions that do include cryptocurrencies in their portfolios has increased in recent years. These are the first purely digital assets that are included in hedge funds and asset managers portfolio. Cryptocurrencies price prediction is attracting growing attention from researchers and investors. Although they have some similarities with most traditional assets, they do have their own behaviour as an asset is still in the process of being understood. It is, therefore important to test this highly noisy and risky financial instrument, especially during critical financial periods. Market fear is a critical macroeconomic construct which demands in-depth and thorough monitoring due to its strong nexus with critical financial assets, even in normal circumstances. A lot of works has been done from the beginning of the pandemic, which focused on the impact of COVID – 19 pandemic to financial instruments. However, there is little information of the effect to cryptocurrencies, especially in high-frequency environments. In this research, authors implement their created testing method for the automated high-frequency trading strategy, which was previously not tested using cryptocurrency data and was not implemented during a critical finance environment. HFT algorithm testing method allows to parallelise data normalisation, trading pair selection, position opening/closing, deletion of unnecessary trade pairs, and closing of long-held transactions that are performed simultaneously at the same time.

Impact of Timestamp and Segmentation Map Selection for Cancerous Prostate Regions in DCE MRI Classification

Aleksas Vaitulevičius[1], Jolita Bernatavičienė[1], Jurgita Markevičiūtė[2], Ieva Naruševičiūtė[3], Mantas Trakymas[3], Povilas Treigys[1]

[1] Institute of Data Science and Digital Technologies, Vilnius University
[2] Institute of Applied Mathematics, Vilnius University
[3] National Cancer Institute

aleksas.vaitulevicius@mif.stud.vu.lt

Dynamic contrast enhancement image sequence is one of the multiparametric magnetic resonance imaging modalities used to detect cancerous regions in the prostate. This image sequence is acquired by capturing the prostate region several times, resulting in prostate region images acquired in different timestamps with an interval of several seconds. The previous investigation covered the following steps. Firstly, a single timestamp image is selected, and the prostate region is segmented. Secondly, obtained segmentation map is projected on all images at timestamps, allowing to monitor prostate segment's intensity change at different timestamps. Then, functional data analysis methods are applied to each segment's data. Lastly, multivariate data obtained were used for machine learning modelling resulting in the classifier of the prostate segments into cancerous and benign classes. By investigating the previous research more thoroughly, this study aims to inspect whether the timestamp selection influences classification performance. The tests are accomplished using the XGBoost classification algorithm and the obtained features from functional data analysis. The experiment results indicate that choice of timestamp for segmentation is statistically insignificant. The segmentation map selection is tested on functional data using the nearest centroid classifier. Map selection investigates whether the proportionate number of Simple Linear Iterative Clustering regions to the prostate outperforms the fixed region number map. The comparison indicates that a fixed number of regions yields statistically more accurate results. Finally, the experiments show that non-registered functional data produces more accurate results statistically.

Evaluation of Pjezoelectric Bending Actuator-Based Ultrasonic Action to Human Blood Circulatory System Functional Status

Vincentas Veikutis[1], Algimantas Bubulis[2], Vytautas Jurenas[2], Joris Vezys[2], Augustas Skaudickas [1], Rokas Janavicius[1]

[1] Lithuanian University of Health Sciences, Institute of Cardiology, Kaunas
[2] Kaunas University of Technology

vincentas.veikutis@lsmu.lt

It is known that ultrasound-based (Usb) therapy can activate both central and peripheral blood flow, thus improving tissue macro/macro perfusion, especially in cases of pathology. Piezoelectric transducers can also be used for that purpose, the high-frequency oscillations of which create stimulate blood flow in tissues, improve metabolism, activate lymph circulation, stimulate microcirculation of the skin and subcutaneous tissue, increase the proliferation of leukocytes and other immunocompetent cells, increase the mobility and number of erythrocytes in the blood. It is also known that neurosensory structures are the most sensitive to ultrasound, and all blood vessels, without exception, have integrated nerve fibres in their wall, through which both contraction and relaxation of the blood vessel can be successfully regulated and modulated.

We developed an ultrasonic blood flow stimulation device which enables a relatively localized increase in micro/macro blood flow efficiency by creating a resonant vibration generation system combining a piezoelectric buzzer transducer, an ultrasonic generator, and a controller. We evaluated general blood tests and specialized thromboelastometry tests using 30W, 60W and buzzer-type Usb action. Computer modelling was performed using the Comsol Multiphysics software package.

No structural changes on general blood tests were founded by using 30W and the buzzer Usb application. Using of 60W Usb we found a decrease on RBC, HGB, HCT and an increase on MCHC, RDW, PLT expression. No significant changes were founded on EXTEM, INTEM, but

increased CT and CT(A5) on FIBTEM by using 30-60W Usb. Also, we observed a gradually increasing of CFT and ML expression with no changes in EST, which could be clinically important. Looking at specific platelet functional status, we found increased APTT, FbC, SPA/INR and D-dimer expression.

In conclusion: disorders of hemostasis mainly manifest as bleeding or thrombosis. Using extracorporeal Usb in therapeutic parameters can successfully adjust disorders of the blood plasma coagulation system that occur due to insufficient activity of coagulation factors or their deficiency, identify other hemostasis disorders and evaluate the success of treatment.

Development of Recommendation System for Pupil's Informal Education Based NLP and LSTM Network

Julius Venskus[1,3], Agnė Brandišauskienė[2]

[1] Institute of Applied Mathematics, Vilnius University
[2] Nacionalinis švietimo centras, JSC
[3] Vytautas Magnus University

julius.venskus@mif.vu.lt

Data from international pupil achievement surveys show that the general learning outcomes of Lithuanian pupils in the international context remain quite average (EBPO PISA, 2018), while the national achievements of pupils are also not high. It is important to look for a variety of tools that can help students achieve higher learning outcomes. Non-formal education is one of the directions that help to solve this problem. This research is looking for recommendation systems based on machine learning models that can recommend to the pupil the non-formal education field to improve learning outcomes.

A generator of recommendations for informal education services has been developed, forming personal recommendations for users of formal education based on the newly developed algorithms of machine learning. The recommendation generator consists of parts such as a data preparation aggregator that collects and transforms data. One of the main parts of the model is the classification of text-based feedback messages from teachers into established categories using the NLP (Natural language processing) technique. A multi-layered LSTM network is trained to prepare the data, with the help of which a recommendation for the non-formal learning direction is subsequently provided.

The developed non-formal science recommendation generator is available as a tool that can help students achieve higher learning outcomes.

Interpolation Methods Impact on Eye Fundus Optic Disc and Optic Cup Segmentation

Sandra Virbukaitė, Jolita Bernatavičienė

Institute of Data Science and Digital Technologies
Vilnius University

sandra.virbukaite@mif.vu.lt

The Optic Disc and Optic Cup are the key parameters in eye health assessment which is manual and time-consuming. With the help of computer-aided systems, this assessment can be automated and work as an advisory opinion for doctors. Here, accurate segmentation of the Optic Disc and Optic Cup is essential as the ratio of these parameters is used in various eye diseases such as glaucoma diagnosis. One of the computer-aided systems is a Convolutional Neural Network that uses eye fundus images of the same size. The variety of images size caused by different eye fundus cameras is aligned using an image resizing technique where the image interpolation occurs. In this research, we applied the three most common interpolation methods, such as bilinear, nearest neighbour, and bicubic, and evaluated the impact on Optic Disc and Optic Cup segmentation caused by these interpolation methods. The experiments on the Drishti dataset demonstrate that Optic Disc segmentation by Dice increased from 0.92 to 0.96 and Optic Cup – from 0.83 to 0.85, resizing eye fundus images to a size of 256x256 by bicubic interpolation than bilinear.

Asymmetric Univariate and Multivariate Financial Time Series Models and Their Applications

Diana Vereškaitė, Remigijus Leipus

Institute of Applied Mathematics
Vilnius University

veresk.diana@gmail.com

The aim is to go through the process of evaluating the risk of a stock in view of its own past time information and inter-relations between stocks. By modelling risk, attention is drawn to asymmetry, heavy-tailed distribution and external regressors. The importance of these features is proven by comparing the diagnostic of models with and without them. Models used for risk analysis are univariate and multivariate GARCH (Generalised AutoRegressive Conditional Heteroskedasticity), together with change point analysis which are further applied to budget European airline companies' stocks. The findings are that asymmetry, leptokurtosis and some exogenous regressors are important in risk analysis as they increase accuracy. Furthermore, in the means of own and other stock's past time information, companies can be ranked from the riskiest to the least risky.

Optimisation of Electric Vehicles Charging Station

Monika Zdanavičiūtė, Tomas Krilavičius

Vytautas Magnus University

monika.zdanaviciute@vdu.lt

As the number of electric vehicles grows, the need for charging stations and the relevance of the limited resources problem increases. Many charging stations are powered only by a limited amount of electricity. Therefore, not all arriving cars can charge to their maximum potential. The distribution of electricity power to consumers must therefore be optimised. To ensure a positive charging experience for the customer, it is important to consider their needs when designing the charging station: while some prioritise fast charging, others aim for the lowest cost possible. Our research focused on developing a methodology for optimising electric car charging, which divides users into three categories based on their priorities (time and money costs). Whenever there are several electric cars of the same category at the charging station, a linear programming problem is created in order to distribute electric power according to arrival time and battery capacity. Experimental studies were conducted to assess whether the arrival time of an electric car or its battery capacity is more important in a linear programming problem. In order to conduct the research, we used a set of generated data, which includes the technical characteristics of the vehicles, the user category selected, and the time of arrival of the vehicles. Kruskal-Wallis test and Dunn's statistical test were applied in order to determine whether the difference in the obtained models results is statistically significant. It was found during the research that the best-performing model is based on equal weights assigned to two parameters: the arrival time of the electric vehicle and its battery capacity.

Classification of Industrial Conveyor Load Status Using Rubber Belt Tension and Deep Learning Models

Tadas Žvirblis[1], Linas Petkevičius[2], Damian Bzinkowski[3], Mirosław Rucki[3], Artūras Kilikevičius[1]

[1] Institute of Mechanical Science
Vilnius Gediminas Technical University
[2] Institute of Computer Science
Vilnius University
[3] Faculty of Mechanical Engineering
Kazimierz Pułaski University of Technology and Humanities, Poland

tadas.zvirblis@vilniustech.lt

Industrial conveyors are used in production process to ensure its efficiency in terms of timely transportation of loose materials and assembled units. In industrial applications, new conveyor belt solutions may substantially reduce overall production costs. The investigated object was the model of belt conveyor (CB) with strain gauges placed on the roller to measure belt strain in real-time work conditions. Objectives of the study were to develop ML models for classification load-ed and unloaded conditions of CB and to identify optimal signal length of tensile pressure which enables achieving the best classification accuracy. Test campaign included measuring static tension under 2 kg load in different points of the CB and measurements in dynamic conditions. Ten-sile signals were divided into 0.2 s, 0.4 s, 0.8 s, 1.6 s, 3.2 s and 5.0 s length. There were developed 5 models (LR, SVM, RF, LSTM, and Transformer) for distinguishing loaded and unloaded conditions of CB. In shallow machine learning models (LR, SVM, and RF) the accuracy of the model increased by 4% on the average each time when the signal length was doubled. RF was the most accurate among the three models and was able to classify 3.2 s and 5.0 s-length signals with an accuracy of 79% and 78%, which was by 3% higher than that of LR or SVM. The accuracy of deep learning models (LSTM and Transformer) increased very rapidly with increasing signal length and an accuracy of 100% was achieved when using both

models with the longest signals. The ac-curacy of Transformer increased on average by 8% when doubling the signal length and after training with the longest signal of 5.0 s, its accuracy reached 100%. The accuracy of LSTM model grew even faster and after training with the 1.6 s signal, the accuracy reached 100%.

Notes

13th 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

Printed by UAB „Baltijos kopija"
Kareivių St. 13b, LT-09109 Vilnius
Print run 130 copies

PARTNER

GENERAL SPONSORS

NOVIAN

Vinted

SPONSORS

asseco

baltic amadeus

VITP
Visorių informacinių technologijų parkas