Book of Abstracts

Workshop on AI for Sustainable Energy Systems and Green AI

Brdo pri Kranju

11th – 12th March 2025

Editor(s) Klemen Žbontar, Bernhard C. Geiger, Amadej Pavšič
Layout Miloš Babić, Bernhard C. Geiger
Cover Miloš Babić, Barbara Gstöttenmayr
Cover picture(s) Brdo Estate

2025 Verlag der Technischen Universität Graz
www.tugraz-verlag.at

ISBN 978-3-99161-045-8
DOI 10.3217/978-3-99161-045-8

Book of Abstracts

Workshop on AI for Sustainable Energy Systems and Green AI

Klemen Žbontar, Bernhard C. Geiger, Amadej Pavšič (eds.)

Brdo pri Kranju, 11th-12th March 2025

Contents

Preface **5**

List of Organizers **7**

Overview of the ENFIELD Project **9**

Agenda **11**

Keynotes **13**

Abstracts **15**

Grid State-Estimation with Least Square Methods and Minimum PMU Infrastructure,
Rene H. Jilg and Wilfried Elmenreich 17

Synthetic Data Generation for Electric Vehicle Charging,
Iroshani Jayawardene, Magdalena Skrzycki, Xiang Ma, Ahmet Soylu and Dumitru Roman 21

Forecasting Solar Power Production by Using Satellite Images,
Dimitar Stefanov and Jure Demšar 23

A Brief Conceptual Comparative Analysis of ASP Reasoners and Multi-Agent Reasoners for Predictive Maintenance and Failure Prediction in Battery Systems,
Mohamed El Bahnasawi, Martin Gebser and Kyandoghere Kyamakya 25

Serverless Large Language Models: Edge vs. Cloud Deployment Trade-offs,
Reza Farahani and Radu Prodan 29

Swarm Intelligence for Sustainable Logistics: Enabling Privacy for the Green Freight Delivery from the Bottom-Up,
Marija Gojković, Melanie Schranz and Wilfried Elmenreich 33

Estimating the Energy Consumption of AI in Smart Meter Data Analysis,
Carolina Fortuna, Vid Hanzel and Dumitru Roman 35

Eco-RETINA: a green flexible algorithm for model building,
Javier Capilla, Alba Alcaraz, Ángel Valarezo, Alfredo García-Hiernaux and Teodosio Pérez-Amaral 37

Neuralizing Cloned-Structured Causal Graphs for World Model Learning,
Tristan M. Stöber, Ali Dasmeh, Erik J. Husom and Sagar Sen 41

Hybridization of data and expert knowledge: Towards inherently interpretable AI models,
Ricardo J. Bessa, Francisco S. Fernandes, Lucas Bost and João Peças Lopes 43

LexAlign: Towards a Multiagent AI System for Regulatory Compliance of Data/AI Pipelines,
Sagar Sen, Carl-Henrik Lien, Nikolay Nikolov, Adela Nedisan Videsjorden, Erik Johannes Husom, Shukun Tokas, Arda Goknil and Dumitru Roman 45

Towards a Vision-Language Foundation Model for Critical Infrastructure Integrity Interpretation,
Sagar Sen, Simeon Tverdal and Erik Johannes Husom 47

Geometric and Information-Theoretic Compression in Neural Classifier Training,
Linara Adilova and Bernhard.C. Geiger 49

Preface

Artificial Intelligence (AI) is increasingly used for modelling, both for the purpose of understanding certain phenomena and as well as for design and development in engineering. Among other things, AI has been successfully used for creating, improving, and controlling energy systems, and for understanding their roles in socio-technical systems. With approaches such as AI-based design, surrogate modeling for digital twins, and data-driven modeling of large-scale interactions between energy systems and individuals, AI has contributed to more sustainable energy conversion and transmission. At the same time, the computational cost of AI – most notably of large foundational models for generative AI – has risen substantially. There is thus a growing need and wider recognition that AI systems must, like all systems, be designed considering their environmental costs during development (training costs) and deployment (inference costs).

This workshop, hosted by the Horizon Europe project ENFIELD at Brdo Estate, Brdo pri Kranju, Slovenia, was dedicated to bringing together experts to discuss 1) the role of AI in the development and control of energy systems and 2) approaches that reduce the energy demand of AI applications. Our program featured presentations on how AI can help estimating the state of the electrical grid and on how AI can be used for energy management tasks, e.g., in battery systems or ship energy systems. Also large language models received substantial attention, both in terms of quantifying and reducing their energy consumption, an as tools for assessing the integrity of critical infrastructure and regulatory compliance of AI pipelines. Presentations on the topic of Green AI featured both novel methods that are inherently resource-efficient as well as fundamental research investigating the interplay between compression and generalization performance.

In addition to interesting presentations and engaging discussions (that, together with a strike of the German air personnel, had overthrown our initial agenda), the workshop attendants also enjoyed two excellently delivered keynotes as well as being spoiled culinarily. We hope that the connections built during these two days in Brdo will be long-lasting, and contribute to a greener future.

The Organizers
Klemen Žbontar, Bernhard C. Geiger, Amadej Pavšič

Acknowledgement

The ENFIELD: European Lighthouse to Manifest Trustworthy and Green AI project has received funding from the European Union's HORIZON Research and Innovation Programme under the grant agreement No. 101120657.

Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

List of Organizers

Conference and Scientific Chairs

- Klemen Žbontar
  Iskraemeco
- Bernhard C. Geiger
  Signal Processing and Speech Communication Laboratory, Graz University of Technology and
  Know Center Research GmbH
- Amadej Pavšič
  Iskraemeco

Organization Committee (alphabetical order)

- Mohammed Salah Al-Radhi (Social Media Channels)
  BME
- Miloš Babić (Book of Abstracts)
  Know Center Research GmbH
- Mátyás Bartalis (Website)
  BME
- Barbara Gstöttenmayr (Agenda and Preparations)
  Know Center Research GmbH
- Irena Justin (Marketing Material)
  Iskraemeco
- Anja Primožič (Graphics Design)
  Iskraemeco

Overview of the ENFIELD Project

ENFIELD is a European Center of Excellence and aims to drive research in Adaptive, Green, Human-Centric, and Trustworthy AI. Focusing on applying research results to sectors such as healthcare, energy, manufacturing, and space, it directly provides valuable results for critical parts of the industry in the EU.

The main focus areas of the **Green AI** research pillar are: (a) Green AI metrics monitoring, which aims at standardization and adoption of Green AI metrics for assessing AI models regarding their carbon footprint; (b) Green AI on the edge, which aims to develop environmentally sustainable and efficient AI technologies for deployment on the resource-constrained edge; and (c) Green AI on the edge-cloud continuum, which aims to advance sustainability in AI by developing energy-efficient strategies and methods to schedule and/or jointly execute AI tasks across edge and cloud systems.

The research pillar on **Adaptive AI** focuses on: (a) Adaptive AI on the edge, which is focused on creating lightweight, adaptive AI models for real-time processing on resource-constrained edge devices; (b) Adaptive robustness and trustworthiness, which proposes adaptive AI approaches that are more robust (e.g., to uncertainty, degraded operating conditions, or out-of-distribution data) and trustworthy (in the sense of reliability); and (c) Brain-inspired adaptive AI, which aims to identify those elements of the learning machinery of the brain that enables it to adapt to its ever-changing environment, and incorporate them into AI algorithms.

Human-Centric AI focuses on: (a) Novel explainable AI methods for decision-making, which aims to explain how AI algorithms based on black box models, such as deep neural networks, arrive at their decisions; (b) Methodologies for evaluating human-AI shared decision making, which aim to develop methodologies for supporting and evaluating trust, reliance, and robustness in human-AI shared decision making, complementing research on the design, development, and deployment of such systems; (c) Interpretable data-driven decision support systems, which aims to explore and create integrated supporting environments in which humans (with different expertise) and AI tools can collaborate to make decisions; and (d) Societal impacts of AI-aided decision making, which explores methods to detect and mitigate bias in AI systems.

Finally, **Trustworthy AI** focuses on (a) Security of AI systems, which explores robust security frameworks for AI systems to ensure their trustworthiness and robustness; (b) Privacy and compliance of AI systems, which establishes privacy-by-design approaches for training and

inference stages of AI systems and develops methodologies that ensure compliance to standards for AI systems; and (c) AI in distributed systems, which investigates and analyzes the security and trust implications of distributed AI architectures and develops corresponding secure protocols, methods, and architectures for software engineering of distributed AI solutions.

ENFIELD prioritizes the acquisition of the best talents and resources toward those goals to eventually strengthen the EU's AI competitiveness. The initiative unites 30 consortium members from 18 countries, including academia, industry, and public sector leaders. Through Open Calls, ENFIELD supports more than 70 researchers and 18 small-scale projects. It will also organize a series of training activities, such as summer schools and hackathons. Outreach efforts will foster engagement, growth, and long-term impact.

Georgios Spathoulas
Research Project Advisor for ENFIELD

Agenda

Time	Activity
	Day 1 - 11.3.2025
8:30-9:00	Arrival
9:00-9:40	Welcome Address, Overview of ENFIELD and the Workshop
9:40-9:55	Presentation on Open Calls for Innovation and Exchange, *Mariana Malta* (remote)
10:00-10:45	**Keynote: Digitalising the Energy system, *Uroš Salobir***
10:45-11:00	Coffee Break
11:00-11:20	Grid State-Estimation with Least Square Methods and Minimum PMU Infrastructure, *Rene H. Jilg, Wilfried Elmenreich*
11:20-11:40	Synthetic Data Generation for Electric Vehicle Charging, *Iroshani Jayawardene, Magdalena Skrzycki, Xiang Ma, Ahmet Soylu, Dumitru Roman* (remote)
11:40-12:00	Forecasting Solar Power Production by Using Satellite Images, *Dimitar Stefanov, Jure Demšar*
12:30-13:30	Lunch Break
12:00-12:20	Comparative Analysis of ASP Reasoners and Agent-Based AI for Predictive Maintenance and Failure Prediction in Battery Systems, *Mohamed El Bahnasawi, Martin Gebser, Kyandoghere Kyamakya* (remote)
13:30-14:15	**Keynote: Efficient Implementation of AI on Edge Devices, *Veljko Pejović***
15:00-15:15	Coffee Break
14:15-14:35	Serverless Large Language Models: Edge vs. Cloud Deployment Trade-offs, *Reza Farahani, Radu Prodan*
15:20-15:40	Swarm Intelligence for Sustainable Logistics: Enabling Privacy for the Green Freight Delivery from the Bottom-Up, *Marija Gojković, Melanie Schranz, Wilfried Elmenreich*
16:00-16:20	Estimating the Energy Consumption of AI in Smart Meter Data Analysis, *Carolina Fortuna, Vid Hanzel, Dumitru Roman*
16:20-16:40	Open Discussion and Closing

Time	Activity
Day 2 - 12.3.2025	
8:30-9:00	Arrival
9:00-9:15	Welcome Address
9:15-9:35	Eco-RETINA: a green flexible algorithm for model building, *Javier Capilla, Alba Alcaraz, Ángel Valarezo, Alfredo García-Hiernaux, Teodosio Pérez-Amaral* (remote)
9:35-9:55	Neuralizing Cloned-Structured Causal Graphs for World Model Learning, *Tristan M. Stöber, Ali Dasmeh, Erik J. Husom, Sagar Sen* (remote)
9:55-10:15	LLMs at the Edge - A Study of Energy Efficiency and Performance, *Erik J. Husom* (presentation-only)
10:15-10:35	Developing a Framework for Green AI practices, *Jeriek Paul Van den Abeele* (presentation-only)
10:45-11:00	Coffee Break
11:00-11:20	Hybridization of data and expert knowledge: Towards inherently interpretable AI models, *Ricardo J. Bessa, Francisco S. Fernandes, Lucas Bost, João Peças Lopes*
11:20-11:40	SEASHINE - Safe Intelligent Agent to optimize SHIp eNErgy management, *Udayanto Dwi Atmojo* (presentation-only)
11:40-12:00	Energy Management using AI-Based Digital Twins, *Astik Samal* (remote, presentation-only)
12:00-12:20	LexAlign: Towards a Multiagent AI System for Regulatory Compliance of Data/AI Pipelines, *Sagar Sen, Carl-Henrik Lien, Nikolay Nikolov, Adela Nedisan Videsjorden, Erik Johannes Husom, Shukun Tokas, Arda Goknil, Dumitru Roman*
12:20-12:40	Towards a Vision-Language Foundation Model for Critical Infrastructure Integrity Interpretation, *Sagar Sen, Simeon Tverdal, Erik Johannes Husom*
14:35-14:55	Geometric and Information-Theoretic Compression in Neural Classifier Training, *Linara Adilova, Bernhard C. Geiger*
12:45-13:45	Lunch Break
13:45-16:30	Closed Session for ENFIELD Participants
16:30-17:00	Workshop Wrap-Up

Keynotes

Uroš Salobir – Digitalising the Energy System

Keynote Speaker Biography

Uroš Salobir has been working at the Slovenian transmission system operator ELES for over 25 years. He has worked on management and coordination in power system operations, electricity markets, infrastructure projects, and technical information systems during this period. He pioneered ELES's strategic innovation activities and was responsible for implementing some of the most challenging innovation projects in ELES's history. In 2017, he helped establish the Strategic Innovation Department he is currently leading. In addition, he and his team recently launched the first cross-sector partnerships and projects with the mobility, clean gases, and heating and cooling sectors. He is the Chair of the ENTSO-E Research and Development Committee (RDIC) and holds prominent positions in CIGRE and EGVIAfor2Zero.

Keynote Abstract

The keynote explored the transformative potential of digitizing the energy system, with a focus on integrating AI and digital twins into the core business processes of large utility companies. Emphasis was placed on the challenges and opportunities of applying large language models (LLMs) and reinforcement learning in the energy sector, highlighting the necessity of understanding AI agents to optimize digital twins effectively. The evolution of digital twins from simulators to sophisticated tools that enhance system operations and planning was discussed.

The European Union's action plan for the digitalisation of the energy system, known as DESAP, was detailed. This initiative aims to define future projects and calls for the European Commission, involving collaboration between Distribution System Operators (DSOs) and Transmission System Operators (TSOs) to address existing challenges using modern digital tools. The importance of customer-oriented business models, advanced system planning, and the development of future control rooms to improve observability and operational efficiency was underscored.

The significance of cross-sector collaboration, particularly between the power system and the mobility sector, was addressed to optimize energy use and reduce costs. Solution concepts *E8* and *Pentlja* on how to enhance the interaction between electric vehicles and the power grid, were introduced. Another concept for a community-based heating solution, called *KODO*, was presented. These concepts aim to help in achieving green energy goals and underline the necessity of integrating digitalisation efforts across various sectors to create a resilient and efficient energy system.

Veljko Pejović – Efficient Implementation of AI on Edge Devices

Keynote Speaker Biography

Veljko Pejović received his PhD in computer science from the University of California Santa Barbara, and worked as a Research Fellow at the Computer Science Department, University of Birmingham, UK. From 2015 he is with the Faculty of Computer and Information Science, University of Ljubljana, where he is an Associate Professor and is leading research on mobile computing, focusing on resource-efficient mobile systems, human-computer interaction, and cybersecurity in ubiquitous systems.

Keynote Abstract

The keynote addressed the growing environmental impact of Information and Communication Technology (ICT) and the urgent need for sustainable practices. It was highlighted that ICT's greenhouse gas emissions are projected to increase significantly, potentially reaching 35% of global emissions if current trends continue. This underscores the necessity for policies and holistic approaches to mitigate ICT's environmental footprint.

Research on resource-efficient mobile systems was presented, focusing on approximate mobile computing. This approach aims to reduce computational demands by adjusting the precision of operations based on user needs and context. The Mobiprox framework was introduced, allowing dynamic approximation of neural network operations on Android devices, enabling significant energy savings without compromising user experience. The framework supports various approximation techniques, such as row and column perforation and filter sampling, and includes a cross-platform profiler to ensure optimal performance on mobile devices.

Additionally, the application of approximate computing in real-world scenarios was discussed, such as human activity recognition and spoken keyword recognition on mobile phones. These applications demonstrated substantial energy savings while maintaining acceptable accuracy levels.

Another research was also presented that showed the potential of on-device AI for drones in agriculture. The Squeeze Slim Unit, a neural network architecture optimized for low-cost drones to perform real-time weed detection, was introduced. This innovation promises to enhance agricultural practices by enabling immediate action based on real-time data, despite the constraints of limited computational resources and network connectivity.

The keynote concluded with an emphasis on the importance of efficiency and sustainability in mobile computing. The presented research offers practical solutions to reduce ICT's environmental impact while maintaining high performance and user satisfaction.

Abstracts

Grid State-Estimation with Least Square Methods and Minimum PMU Infrastructure

Rene H. Jilg[1] **W. Elmenreich**[1]

[1]Alpen-Adria University, Klagenfurt, Austria

1 Introduction

The transition toward a sustainable energy future relies on the widespread adoption of renewable energy sources (RES) such as wind and solar power (Zhou et al., 2019). However, their inherent volatility presents challenges for grid integration, particularly in Distribution Networks (DN). At the same time, rising electricity demand from electrification—such as electric vehicles and heat pumps—further strains the grid, especially when local energy generation and consumption are not properly balanced. Without effective grid management and control (GMC), fluctuations in key electrical parameters like voltage, current, and power can lead to inefficiencies, energy losses, and potential safety risks when operational limits are exceeded. A key for effective GMC is accurate grid state estimation (SE). It directly influences grid stability but also supports a more sustainable energy system by reducing unnecessary energy losses, optimizing resource use, and minimizing the environmental impact of monitoring infrastructure. In this paper, we present investigation results of grid observability analysis with minimal Phasor Measurement Unit (PMU) infrastructure.

1.1 The Necessity of Grid Power Management and Grid State-Estimation

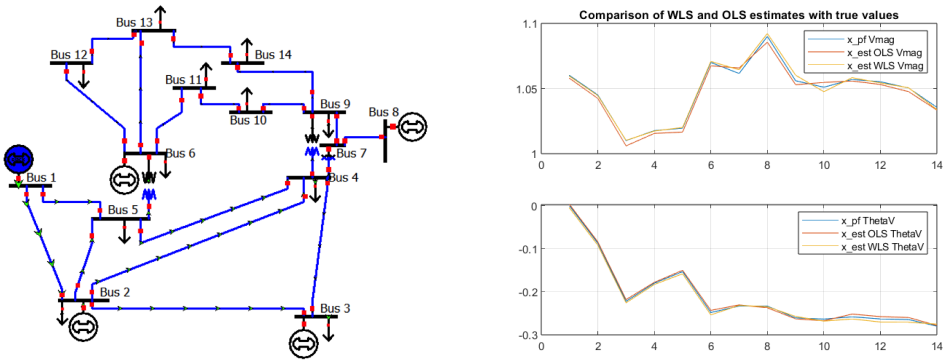
Power management as a control task is the procedure to actively manage controllable grid devices (CGDs) following a control strategy. Examples of CGDs are power generators, transformer tap positions, battery energy systems (BESS), STATCOM, etc. A medium-term (e.g., over the next 24 hours) "base control trajectory" for the CGDs can be obtained by performing an Optimal Power Flow Calculation (OPF). The OPF calculation can be done offline and incorporates optimization models, including minimizable cost functions (e.g., minimization of power losses) and constraints (e.g., considering voltage and current safety limits). Inputs to the OPF are, besides the network model, estimated power productions and consumptions in the network. For the volatile RES, the productions can be estimated by means of weather forecasting values and production models.
In any case, during operation, deviations from the predicted grid quantities will occur. Reasons include stochastic production due to clouds and wind, deviations in predicted consumptions (e.g., EV charging times), inaccurate models, or outages and failures of grid-relevant components. Depending on the impact on the critical grid quantities (voltages, currents), grid control needs to react in real time. Hence, it is necessary to know the physical state of the grid in real time. The process of obtaining these grid state values (state vectors (SV)) is called Grid State Estimation (SE).

1.2 Challenges in Grid State Estimation

Key challenges in power system state estimation for distribution networks include:

- Measurement Inaccuracies (Gómez-Expósito and Abur, 2004)
- Data Synchronization Issues (Monticelli, 2000)
- Limited Observability in Distribution Networks (Abur and Exposito, 2004)
- Unbalanced Loads and Network Configurations (Karimi et al., 2019)
- Low X/R-ratio especially in DG (Zhao et al., 2017)
- Cybersecurity Concerns (Chung et al., 2018)

This paper is focusing on the limited observability of a network.



(a) IEEE-14 transmission network (University of Illinois at Urbana-Champaign)

(b) OLS and WLS estimates with true values obtained from Power Flow Calculation

Figure 1: IEEE-14 transmission network and a comparison of OLS and WLS estimates.

2 Method and Experiments

2.1 Measurement Configuration

In this simulation setup only PMUs are used, while each PMU is able to measure the voltage phasor (voltage magnitude $V_{mag,i}$ and voltage phase angles $\theta_{V,i}$) at the installation bus i and the current phasors (current magnitudes $I_{mag,ij}$ and current phase angles $\theta_{I,ij}$) of the connected branches between bus i and bus j. For all the calculations the reference bus is defined at bus 1 ($\theta_{V,1} = 0$). To perform simuluations, the IEEE-14 bus network is used, which comprises $N = 14$ buses and $M = 20$ branches (figure 1a). Althought the IEEE-14 bus is a Transmission Network (TN), conclusions can also be taken for DN. The goal is to determine the minimum PMUs needed for network state estimation using OLS and WLS, ensuring cost-effective deployment.

2.2 Applied State-Estimation Methods

When performing SE a mathematical model is necessary, which relates the vector of electrical states of the system $\mathbf{x} = [V_{mag,1}, .., V_{mag,N}, \theta_{V,1}, .., \theta_{V,N}]^T$ to the vector of available measurements $\mathbf{z} = [\mathbf{V}_{mag}, \theta_V, \mathbf{I}_{mag}, \theta_I]^T$ by means of the measurement equation $\mathbf{z} = \mathbf{h}(\mathbf{x}) + \mathbf{e}$. In that equation $\mathbf{h}(\mathbf{x})$ is a nonlinear vector function containing the measurement functions, $\mathbf{e}$ is the vector of disturbances (e.g. measurement noise) and $\mathbf{V}_{mag}$, θ_V, $\mathbf{I}_{mag}$, θ_I are block-vectors which length depends on the numer of measurements. (Remark: while the relationsship between State-Variables and measurements of voltage phasor quantities is linear, the relationship of State-Variables and current phasor quantities is non-linear (Abur and Exposito, 2004)). When using OLS or WLS algorithms the nonlinear relationsship between SV and measurements needs to be linearized, so that the measurement function becomes $\mathbf{z} = \mathbf{H}(\mathbf{x}) \cdot \mathbf{x} + \mathbf{e}$, where $\mathbf{H}(\mathbf{x})$ is the Jacobian matrix with elements $H_{i,j} = \frac{\partial h_i}{\partial x_j}$. The SE calculates iteratively a new state estimate at every step according $\Delta\hat{\mathbf{x}}_k = (\mathbf{H}_k{}^T \mathbf{W}_k \mathbf{H}_k)^{-1} \mathbf{H}_k{}^T \mathbf{W}_k \Delta \mathbf{z}_k$, where $\mathbf{H}_k$ again is the measurement Jacobian at iteration step k and $\mathbf{W}_k$ is the weighting matrix at iteration step k. The latter is equal to the identity matrix in case of OLS. The states are finally iteratively estimated by $\hat{\mathbf{x}}_{k+1} = \hat{\mathbf{x}}_k + \Delta\hat{\mathbf{x}}_k$. According to observation theory, a necessary condition that such a system is fully observable is, that the rank of the Jacobian $\mathbf{H}$ is equal to the number of the state-variables (Monticelli, 1999). Observability in that context means that the SV $\mathbf{x}$ can be reconstructed from the measurement $\mathbf{z}$. In the following, two case are considered:

1. PMU placement at the buses [2,6,7,9] which leads to a theoretically fully observable system (Hussain et al., 2021)
2. PMU placement just at buses [2,6] which leads to a theoretically not fully observable system

2.3 Results

Both methods (OLS and WLS) have been used in simulation with different measurement noise variances and initial conditions over multiple runs. The rank of matrix $\mathbf{H}$ and the condition $\kappa(\mathbf{H^T H})$ as well as $\kappa(\mathbf{H^T W H})$ were analyzed. Simulations reveal an interesting result: even in a setup at which theoretical observability is not given (violation of the necessary condition from above) OLS and WLS methods are capable of consistently estimating the system state. As an example, in figure 1b compares estimated states from both methods and the true values which were obtained from a Power Flow Calculation. The simulation was done with PMUs installed at buses [2,6] which results in $rank(\mathbf{H}) = 18$ which is lower than the number of State-Variables, which is 28.

3 Discussion and Outlook

The simulation results have shown, that also for the case 2 (theoretically not fully observable), both LS estimation methods are able to estimate the state of the system accurately. This fact can be summarized by the following effects:

1. **Redundancy in Current Phasor Measurements**: A single PMU measures voltage at its bus and currents in connected branches. Since currents depend on neighboring voltages via the admittance matrix, this redundancy enhances numerical observability, even without full theoretical observability (Zhao et al., 2020).
2. **Impact of Network Topology and Admittance Matrix**: The power grid's topology determines how measurement information propagates. The Jacobian matrix structure often retains full rank numerically, allowing OLS and WLS to perform well (Zhao et al., 2020; Monticelli, 2000).
3. **Small Measurement Noise and Well-Conditioned Problem Structure**: PMUs offer highly accurate voltage and current phasor measurements. The low noise levels in these measurements reduce numerical instability, allowing state estimators to converge despite limited data (Abur and Exposito, 2004).

It has been shown that under some circumstances it is possible to obtain consistent state-estimations even in the case of theoretical non fully observability with simple LS algorithms based on PMU measurements being sufficient. If the sparsity of the measurement matrix increases, alternative methods such as Sparse Bayesian Learning (SBL) become more beneficial.

References

Abur, A. and Exposito, A. G. (2004). *Power System State Estimation: Theory and Implementation*. CRC Press.

Chung, C. Y., Lee, Y. O., and Huang, J. Q. (2018). Cybersecurity challenges in power system state estimation. *IEEE Transactions on Smart Grid*, 9(2):1451–1461.

Gómez-Expósito, A. and Abur, A. (2004). *Power System State Estimation: Theory and Implementation*. CRC Press.

Hussain, A., Baloch, M. H., Uqaili, M. A., Albogamy, F. R., and Baig, M. A. (2021). Optimal placement of phasor measurement unit considering system observability and redundancy index. *Heliyon*, 7(8):e07892.

Karimi, H., Javadi, M. S., and Van Cutsem, T. (2019). Power system state estimation considering unbalanced conditions and measurement errors. *IEEE Trans. on Power Systems*, 34(1):325–334.

Monticelli, A. (1999). *State Estimation in Electric Power Systems: A Generalized Approach*. Springer.

Monticelli, A. (2000). Electric power system state estimation. *Proceedings of the IEEE*, 88(2):262–282.

Zhao, J., Kar, S., and Moura, J. M. (2020). Pmu-based state estimation for power transmission networks: Theory and implementation. *IEEE Transactions on Power Systems*, 35(5):3755–3766.

Zhao, J., Netto, M., and Mili, L. (2017). A robust iterated extended kalman filter for power system dynamic state estimation. *IEEE Transactions on Power Systems*, 32(4):3205–3216.

Zhou, X., Zhang, Y., and Wang, J. (2019). Challenges and opportunities of renewable energy integration in future power systems. *IEEE Transactions on Sustainable Energy*, 10(3):1445–1458.

Synthetic Data Generation for Electric Vehicle Charging

Iroshani Jayawardene[1] **Magdalena Skrzycki**[2] **Xiang Ma**[1] **Ahmet Soylu**[3] **Dumitru Roman**[1,4]

[1]SINTEF AS, Norway
[2]Leiden University, Netherlands
[3]Kristiania University of Applied Sciences, Norway
[3]OsloMet - Oslo Metropolitan University, Norway

1 Introduction

The increasing adoption of electric vehicles (EVs) significantly contributes to reducing carbon emissions but also complicates energy systems due to increased demands for residential charging. A fundamental challenge is the lack of detailed, high-quality data on EV charging patterns.

High-quality synthetic datasets are crucial to effectively manage and predict future needs of integrated energy systems, as traditional data sources often lack the necessary granularity to capture the complex inter-dependencies between energy components and user behaviors. This abstract proposes a generative model for EV charging data that incorporates temporal, spatial, and environmental conditions, leveraging real-world open-source datasets and employing advanced techniques such as data augmentation, and imputation. This approach supports customized scenario analyses for energy management research, enhancing the EMS's optimization for improved grid stability and efficiency. Additionally, a framework to evaluate synthetic data quality and a graphical interface for researchers, utility companies, and policymakers is proposed. The anticipated results include a versatile tool for data synthesis, providing new insights to enhance grid stability and energy management practices amid the increasing prevalence of EV charging. This work is done in the context of a prototype energy management system (EMS) that includes a 10 kW photovoltaic (PV) system, 193 kWh of battery storage (split into four 48.4 kWh units), an EV charging station, and a grid connection (Jayawardene et al. (2023)).

2 Method and Case Study

This work focuses on developing a generative diffusion-based model (Cao et al. (2023)) to create realistic synthetic data for EV charging, accommodating temporal, spatial, weather conditions, and energy pricing factors. The model will incorporate advanced machine learning techniques, including Long short-term memory (LSTMs) and cross-attention mechanisms (Zheng et al. (2021)), to enhance the realism of time-series data.

2.1 Data Description

The primary datasets include the dataset of EV charging records (Sørensen (2024)) and historical weather data from Open-meteo [1], complemented by Ember energy dataset [2] for energy pricing in Norway. These datasets will provide the foundational real-world data necessary for training and validation.

2.2 Data Pre-processing

Data pre-processing targets normalization and imputation to manage gaps and ensure robust modeling inputs. The focus will be on ensuring data quality, especially regarding the temporal and spatial consistency needed for precise synthetic data generation.

[1]`https://open-meteo.com`
[2]`https://ember-energy.org/data`

2.3 Model Development

The modeling approach will utilize a conditional diffusion framework integrated with time-series specific layers to accurately simulate individual charging sessions. A Graphical User Interface (GUI) is under development[3], enabling users to customize data generation based on specific scenarios and conditions.

2.4 Evaluation

Synthetic data will be validated against real-world data using statistical tests like the Wasserstein distance SciPy (2023) and Kullback-Leibler divergence Unknown (2023) to ensure data quality and realism. The synthetic outputs will also be compared against benchmarks from established synthetic data generation models to gauge performance enhancements.

3 Discussion and Outlook

Based on the initial investigations and literature review, the integration of daily and session-specific weather variables alongside hourly electricity prices through synthetic data generation is poised to significantly enhance the model's precision in predicting EV charging behaviors and cost estimations. As we advance, continuous refinement of data pre-processing and model integration of dynamic variables like real-time weather and fluctuating energy prices is crucial. These advancements in synthetic data generation are key to accurately mimicking real-world complexities in our datasets. Ultimately, the methodologies developed are expected to greatly improve the granularity and accuracy of EV charging data models, supporting the evolution of energy management systems that adeptly adjust to user behavior and environmental shifts, thereby driving forward the development of sustainable, efficient urban energy frameworks.

Acknowledgements. The work is funded partially through the projects SEP SynDaGen (SINTEF internal funding), enRichMyData (HE 101070284), Graph-Massivizer (HE 101093202), and UPCAST (HE 101093216).

References

Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., Heng, P.-A., and Li, S. Z. (2023). A survey on generative diffusion models. *Journal of LaTeX Class Files*, 1. Senior Member, IEEE, and Fellow, IEEE.

Jayawardene, I., Roman, D., Zhao, Y., Ulyashin, A. G., Soylu, A., and Ma, X. (2023). Towards an open energy management system for integrated energy storage and electric vehicle fast charging station. SINTEF Digital, OsloMet - Oslo Metropolitan University, High North Quality AS, SINTEF Industry, Kristiania University College. Accessed: 2023-02-27.

SciPy (2023). Scipy.stats.wasserstein_distance. Available online: `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wasserstein_distance.html` (Accessed on February 27, 2023).

Sørensen, K. (2024). Electric vehicle charging dataset with 35,000 charging sessions from 12 residential locations in norway.

Unknown, A. (2023). Understanding kl divergence. `https://medium.com/towards-data-science/understanding-kl-divergence-f3ddc8dff254`. Accessed: 2023-02-27.

Zheng, S., Chen, F., and Wang, X. (2021). Semantic matching for short texts: A cross attention mechanism. *Journal of Physics: Conference Series*, 1757(1):012087.

[3] `https://github.com/IroshaniJ/SynDaGen.git`

Forecasting Solar Power Production by Using Satellite Images

Dimitar Stefanov[1,2] **Jure Demšar**[1]

[1]Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia
[2]Sunesis, Ljubljana, Slovenia

1 Introduction

In recent years, there has been a significant increase in the integration of photo-voltaic capacities across the world, which in turn has made the electrical grid more difficult to maintain. The reason for this is the high variability of solar power generation. Hence, to be able to operate the grid successfully, grid operators require accurate forecasts of solar power generation.

For this purpose, our work focused on predicting 2-hour ahead solar power generation at 15-minute intervals – a typical resolution requirement for solar power plants and system operation measurements. We made solar power generation forecasts for 233 different locations across Slovenia – more extensive research regarding the number of photo-voltaic locations than what can be found in the literature on this topic. We showed that the state-of-the-art deep learning architecture called *Temporal Fusion Transformer* (TFT) (Lim et al., 2021) outperforms well-established benchmarks in solar forecasting by significant margins across all metrics and training settings considered.

2 Method and Experiments

TFT, as visualized in Figure 1, is among the first transformer-based architectures designed for time series forecasting. It was designed to build separate feature representations for different input types (static, known, and observed input types) passed to the model. In addition, the model has a modular structure, meaning that each part of the model tackles a specific issue related to the dataset or forecasting problem at hand. These 2 model features have led to strong model performance on a wide range of problems (Lim et al., 2021).

We compared the performance of TFT against several well-established benchmarks in the field of solar forecasting: climatology (Yang, 2019), naive persistence and smart persistence models (Pedro and Coimbra, 2012), as well as a multi-layer perceptron (MLP) (Rumelhart et al., 1986). The results of the experiments are presented in Table 1. The table lists MSE, MAE, and the skill score metric (in parentheses), along with their corresponding standard errors for all models considered.

3 Discussion and Outlook

As visible in Table 1, TFT is the best-performing model across all metrics and training settings considered. However, we believe that there are additional neural network architectures that could even better utilize the information present in satellite images. A fitting example for this task is the series of weather models developed by Google DeepMind and Google Research: MetNet-1 (Sønderby et al., 2020), MetNet-2 (Espeholt et al., 2021) and MetNet-3 (Andrychowicz et al., 2023). Adapting the MetNet model architectures for the task of solar forecasting is a path worth exploring.

Acknowledgements. The numerical weather predictions were provided as part of a 1-year direct access to the European Centre for Medium-Range Weather Forecasts (ECMWF) Archive Catalogue, available at http://apps.ecmwf.int/archive-catalogue/.

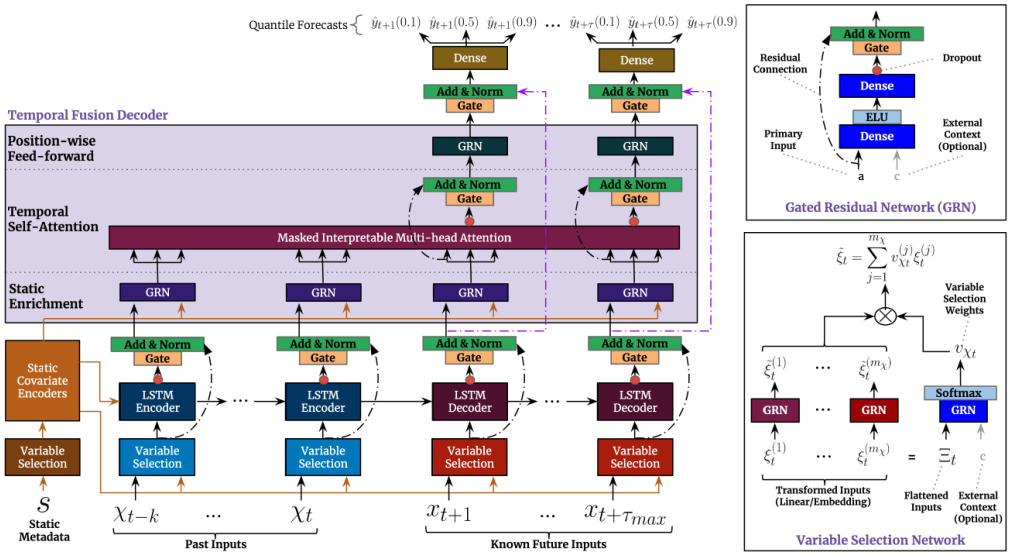


Figure 1: **A visualization of the TFT architecture**. We can see the gated residual networks (GRN) and variable selection networks in more detail on the right-hand side. The figure is borrowed from the work by Lim et al. (2021).

Table 1: **Model evaluation results**. The results of the best-performing model are in **bold**.

Model	MSE	MAE
Climatology	10169.68 ± 0.00 (-3.62 ± 0.00)	65.66 ± 0.00 (-1.93 ± 0.00)
Naive persistence	3390.90 ± 0.00 (-0.46 ± 0.00)	30.31 ± 0.00 (-0.35 ± 0.00)
Smart persistence	2199.73 ± 0.00 (0.00 ± 0.00)	22.42 ± 0.00 (0.00 ± 0.00)
MLP	7310.65 ± 0.62 (-2.32 ± 0.00)	61.60 ± 0.00 (-1.75 ± 0.00)
TFT	**1736.49 ± 0.39 (0.21 ± 0.00)**	**21.69 ± 0.00 (0.03 ± 0.00)**

References

Andrychowicz, M. et al. (2023). Deep learning for day forecasts from sparse observations. *arXiv preprint arXiv:2306.06079*.

Espeholt, L. et al. (2021). Skillful twelve hour precipitation forecasts using large context neural networks. *CoRR*, abs/2111.07470.

Lim, B. et al. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764.

Pedro, H. T. and Coimbra, C. F. (2012). Assessment of forecasting techniques for solar power production with no exogenous inputs. *Solar Energy*, 86(7):2017–2028.

Rumelhart, D. E. et al. (1986). *Learning representations by back-propagating errors*, volume 323. Springer.

Sønderby, C. K. et al. (2020). Metnet: A neural weather model for precipitation forecasting. *arXiv preprint arXiv:2003.12140*.

Yang, D. (2019). Making reference solar forecasts with climatology, persistence, and their optimal convex combination. *Solar Energy*, 193:981–985.

A Brief Conceptual Comparative Analysis of ASP Reasoners and Multi-Agent Reasoners for Predictive Maintenance and Failure Prediction in Battery Systems

Mohamed El Bahnasawi[1,2,*] **Martin Gebser**[2,*] **Kyandoghere Kyamakya**[1,*]

*Equal contribution .
[1]Institute of Smart Systems Technologies, University of Klagenfurt
[2]Institute of AI und Cybersecurity, University of Klagenfurt

1 Introduction

Accurate prediction of battery failure is crucial for ensuring reliability, safety, and longevity in energy storage systems across diverse applications, from electric vehicles to renewable energy grids. Effective battery management relies on analyzing multivariate time-series data to capture the dynamic and evolving characteristics of battery degradation Lipu et al. (2022). Traditional predictive maintenance methods, including reactive (run-to-failure) and scheduled (time-based) approaches, often entail high costs, significant risks, and either premature or delayed maintenance interventions, Fioravanti et al. (2020). Consequently, there is a critical need for more sophisticated methods that can deliver early failure detection and accurate degradation forecasting, thus enhancing operational safety, system reliability, and significantly reducing maintenance costs and downtime.

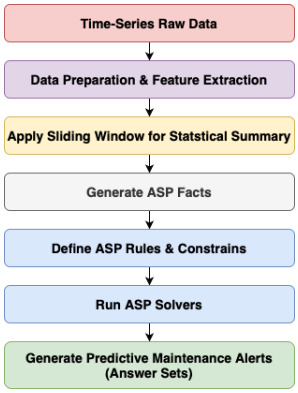
2 Method and Experiments

Recent advancements in predictive maintenance have introduced a range of methodologies, such as in Chen (2020); Cavus et al. (2025) from a rule-based threshold checks to complex data-driven models. In this work, we propose a brief conceptual comparative analysis of two symbolic AI reasoning approaches: Answer Set Programming (ASP) as in Lifschitz (2019) and Multi-Agent Systems (MAS) as in Šarunas Raudys and Zliobaite (2006); Salvador Palau et al. (2019), utilizing the NASA Prognostics Batteries dataset introduced by Saha and Goebel (2007). This dataset comprises comprehensive cycle-by-cycle measurements, including voltage, current, temperature, capacity, and impedance, from Li-ion batteries subjected to varied operational profiles (charge, discharge, and impedance)

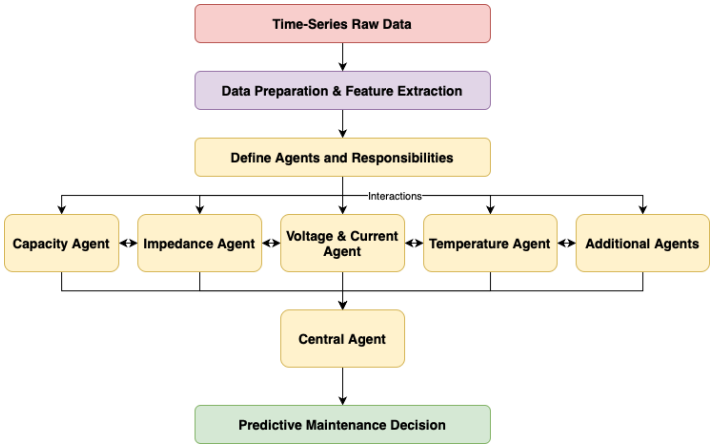
In the ASP-based framework as illustrated in Figure 1a, raw battery cycle data is initially subjected to preprocessing and feature extraction. A sliding window method is applied to the processed time-series data to mitigate the effects of transient anomalies and measurement errors. Subsequently, the data within each sliding window is converted into a structured set of logical facts suitable for reasoning in ASP. Expert domain knowledge is then encoded into declarative rules, defining specific conditions for battery failure. or example, a rule may state that if the measured capacity during a discharge cycle falls below 1.4 Ahr (a 30% fade from the rated capacity of 2 Ahr), then the battery is considered to have reached its end-of-life. ASP solvers then compute answer sets that explicitly reveal the battery's state by enforcing these constraints. This method offers high transparency and a clear traceability of the decision-making process, which is especially valuable for diagnosing specific degradation events

In the proposed MAS framework shown in Figure 1b, we aim to deploy multiple specialized agents, each monitoring a specific aspect of battery data such as capacity, impedance, voltage,

and temperature. Each agent processes incoming real-time data streams, assesses them against predefined thresholds, and communicates potential anomalies through established message-passing protocols. For instance, a dedicated 'Capacity Agent' identifies and flags cycles where the battery capacity falls below critical thresholds, while an 'Impedance Agent' monitors abnormal impedance trends. A central coordinator, the 'Diagnosis Agent,' aggregates alerts from these diverse agents, synthesizes the findings, and generates dynamic, real-time predictive maintenance decisions. This decentralized structure ensures adaptability and real-time responsiveness, essential for promptly adjusting assessments as new data becomes available.



(a) Detailed Workflow of the ASP-based approach



(b) Workflow of the ASP-based approach

Figure 1: Overall caption describing both subfigures.

3 Discussion and Outlook

At this stage, we anticipate distinct strengths and limitations from each method. We expect that ASP would excel in transparency and explainability because it explicitly uses human-readable logical rules to interpret battery states. However, we also foresee potential limitations regarding computational complexity and scalability, particularly for large datasets or real-time scenarios, due to the resource-intensive nature of grounding and rule-solving processes. Conversely, we anticipate that MAS would demonstrate advantages in real-time responsiveness, adaptability, and scalability because of its distributed and modular architecture. However, MAS may offer only moderate transparency due to the complexity arising from multiple interacting agents. This conceptual analysis provides an initial framework to guide future practical evaluations and helps in identifying which criteria will be crucial for eventual benchmarking.

In summary, a conceptual framework comparing two symbolic AI reasoning methods: ASP and MAS for predictive battery maintenance has been presented. In future research practical validation and quantitative comparison will be provided. We aim to implement and experimentally validate both approaches, assessing key performance metrics including accuracy, computational efficiency, transparency, scalability, and real-time responsiveness. Additionally, we plan comprehensive benchmarking against established predictive methods such as classical machine learning and deep learning to confirm their practical applicability.

Acknowledgements. Funding is acknowledged from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101137725 (BatCAT)

References

Cavus, M., Dissanayake, D., and Bell, M. (2025). Next generation of electric vehicles: Ai-driven approaches for predictive maintenance and battery management. *Energies*, 18(5).

Chen, W. (2020). A rule-based expert system for predictive maintenance of a hybrid bus.

Fioravanti, R., Kumar, K., Nakata, S., Chalamala, B., and Preger, Y. (2020). Predictive-maintenance practices: For operational safety of battery energy storage systems. *IEEE Power and Energy Magazine*, 18(6):86–97.

Lifschitz, V. (2019). Answer set programming. *Answer Set Programming*, pages 1–190.

Lipu, M. S. H., Ansari, S., Miah, M. S., Meraj, S. T., Hasan, K., Shihavuddin, A. S., Hannan, M. A., Muttaqi, K. M., and Hussain, A. (2022). Deep learning enabled state of charge, state of health and remaining useful life estimation for smart battery management system: Methods, implementations, issues and prospects. *Journal of Energy Storage*, 55:105752.

Saha, B. and Goebel, K. (2007). Battery data set. `https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/#battery`.

Salvador Palau, A., Dhada, M. H., Bakliwal, K., and Parlikad, A. K. (2019). An industrial multi agent system for real-time distributed collaborative prognostics. *Engineering Applications of Artificial Intelligence*, 85:590–606.

Šarunas Raudys and Zliobaite, I. (2006). The multi-agent system for prediction of financial time series. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4029 LNAI:653–662.

Serverless Large Language Models: Edge vs. Cloud Deployment Trade-offs

Reza Farahani[1] **Radu Prodan**[1,2]

[1]Institute of Information Technology, University of Klagenfurt, Austria
[2]Department of Computer Science, University of Innsbruck, Austria

1 Introduction

Recently, *Large Language Models* (LLMs) are becoming popular, powering applications such as chatbots, content generation, and intelligent assistants [1]. Cloud computing has long been the default infrastructure for hosting LLMs, providing on-demand scalability, access to high-performance GPUs, and global availability. Yet, cloud instances for LLMs can be inefficient, as bursty workloads lead to idle resources and unnecessary costs, especially for sporadic inference requests, e.g., chatbots experiencing peak traffic during business hours but remaining underutilized at night.

Serverless computing has emerged as a complementary paradigm for cloud computing, offering a cost-efficient, flexible alternative that dynamically scales resources while eliminating the need for manual infrastructure management [3, 4]. Platforms like *Hugging Face's Inference Endpoints*[1] leverage this paradigm to provide serverless access to thousands of LLM models. However, *cold start* latency, i.e., container initialization before processing requests, plus network overhead, delays LLM inference for token generation and limits cloud suitability for real-time or restricted-data applications, driving interest in edge computing. *Edge-based LLM inference* on servers or low-power CPUs, GPUs, or TPUs can reduce network latency, cloud costs, and energy overhead while enabling offline operation, allowing LLM models to function without Internet connectivity and efficient processing through task-specific models. However, hardware constraints and limited scalability remain key challenges.

This short paper compares serverless LLM deployed on Google Cloud Functions with an edge-based deployment on an NVIDIA Jetson Nano (NJN) running open-source serverless platform. We evaluate *latency* and *scalability* to analyze the trade-offs between serverless edge and cloud inference, contributing to *"Systems for AI"* [2] by assessing infrastructure impact on LLM efficiency.

Keywords

Serverless Computing, Edge-Cloud Computing, Large Language Models (LLM), Systems for AI.

2 Evaluation Setup

We conducted all experiments on Google Cloud Run[2], a cloud-based serverless platform that dynamically allocates resources with an NVIDIA A100 GPU for cloud deployment, and NVIDIA Jetson Nano (NJN) edge device, featuring a Quad-core ARM Cortex-A57 CPU@1.43 GHz with 4 GB RAM, running `OpenFaaS`[3], an open-source serverless platform deployed atop `Kubernetes`[4]. Both platforms hosted `LLaMA-2` (7B parameters, 4-bit Q4_K_M quantization) in `.gguf` format, downloaded from Hugging Face as a dockerized serverless function. To ensure a diverse workload, we selected 100 unique queries from `MMLU` (Massive Multitask Language Understanding) [6], covering general

[1]`https://huggingface.co/inference-endpoints`
[2]`https://cloud.google.com/run`
[3]`https://www.openfaas.com/`
[4]`https://kubernetes.io/`

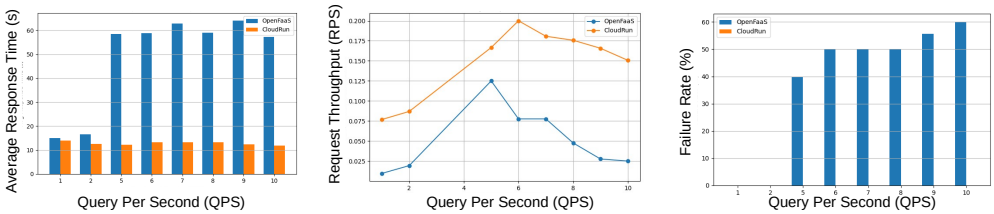


Figure 1: Impact of query per second (QPS) on (a) average response time, (b) request throughput (RPS), and (c) failure rate across deployment environments.

knowledge, reasoning, and code generation tasks. For load testing, we employed `Locust`[5], where queries were formatted as `JSON` payloads, dynamically sampled, and sent at varying concurrency levels (1 to 10 queries per second (QPS)), returning sequence of tokens as outputs for further analysis. We monitored resource utilization via `Prometheus`[6] at the edge and `Google Cloud Monitoring`[7] on the cloud.

3 Experimental Results

For each query, we measured the total response time, capturing the full sequence generation latency, including request transmission, processing, and response return. Cold start latency was evaluated by deliberately forcing function inactivity before execution, triggering container reinitialization in Google Cloud Run and function reloading in OpenFaaS. As shown in Fig. 1 (a), increasing QPS leads to higher response times due to increased processing overhead. While edge deployments benefit from reduced cold start latency, their limited computational resources exacerbate performance degradation under load. Moreover, larger container sizes and increasing loads on resource-constrained edge devices further contribute to failures (Fig. refresult (c)), highlighting the scalability advantage of cloud deployments for handling high query loads.

To evaluate the scalability of cloud and edge deployments, we measured requests per second (RPS) as $RPS = \frac{\#Queries}{Total\_Time}$, where a higher RPS indicates better scalability and throughput. As QPS increased, RPS initially scaled but eventually plateaued or degraded due to resource saturation (Fig. 1 (b)). This confirms the earlier discussion, where cloud platforms sustain higher throughput under load while the edge experiences bottlenecks from computational capacity and queuing delays.

4 Discussion and Outlook

This work explores LLM deployment in serverless edge and cloud environments, emphasizing the trade-offs between efficiency and scalability. It advances research on LLM deployment within "Systems for AI" context through the following insights: *(i) Edge serverless LLM deployment* enables offline usage, reduces costs during off-peak periods, and enhances privacy by keeping data local. It also lowers energy consumption and CO_2 emissions, exemplified by 16 ChatGPT queries emitting as much as boiling a kettle[8]; *(ii) Cloud serverless LLM deployment* remains superior in handling high query loads due to its on-demand scalability and computational power.

Future research directions include enhancing *fine-tuning* of LLMs at the edge for personalized adaptation, optimizing *quantized* models with *Retrieval-Augmented Generation* (RAG) for efficiency, and implementing *intelligent query routing* for cost-aware execution. Advancements in *adaptive resource allocation* and *bi-objective scheduling* solutions like [5] will further improve the scalability and sustainability of LLM inferences.

[5] `https://locust.io/`
[6] `https://prometheus.io/`
[7] `https://cloud.google.com/monitoring`
[8] `https://piktochart.com/blog/carbon-footprint-of-chatgpt/`

Acknowledgements. This work received funding from the Horizon Europe research and innovation program (project 101093202 “Graph-Massivizer”) and the Austrian Research Promotion Agency (FFG project 909989 “AIM AT Stiftungsprofessur für Edge AI”).

References

[1] Zoha Azimi, Reza Farahani, Christian Timmerer, and Radu Prodan. Towards an Energy-Efficient Video Processing Tool with LLMs. In *Proceedings of the 4th Mile-High Video Conference*, 2025.

[2] Reza Farahani, Zoha Azimi, Christian Timmerer, and Radu Prodan. Towards Ai-assisted Sustainable Adaptive Video Streaming Systems: Tutorial and Survey. *arXiv preprint arXiv:2406.02302*, 2024.

[3] Reza Farahani, Dragi Kimovski, Sashko Ristov, Alexandru Iosup, and Radu Prodan. Towards sustainable serverless processing of massive graphs on the computing continuum. In *Companion of the 2023 ACM/SPEC International Conference on Performance Engineering*, 2023.

[4] Reza Farahani, Frank Loh, Dumitru Roman, and Radu Prodan. Serverless Workflow Management on the Computing Continuum: A Mini-Survey. In *Companion of the 15th ACM/SPEC International Conference on Performance Engineering*, 2024.

[5] Reza Farahani, Narges Mehran, Sashko Ristov, and Radu Prodan. HEFTLess: A Bi-Objective Serverless Workflow Batch Orchestration on the Computing Continuum. In *2024 IEEE International Conference on Cluster Computing (CLUSTER)*. IEEE, 2024.

[6] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Swarm Intelligence for Sustainable Logistics: Enabling Privacy for the Green Freight Delivery from the Bottom-Up

Marija Gojković[1,2] **Melanie Schranz**[1] **Wilfried Elmenreich**[2]

[1]Lakeside Labs GmbH, Klagenfurt, Austria, email: {gojkovic, schranz}@lakeside-labs.at
[2]Alpen-Adria University, Klagenfurt, Austria, email: wilfried.elmenreich@aau.at

1 Motivation

The logistics industry faces a critical challenge with the significantly growing number of empty truck drives. Eurostat data from Austria reveals an increase in empty truck runs from 40% in 2010 to 45% in 2019 in Austria, alongside a rise in empty kilometers from 31% to 34% (VCÖ Mobilität mit Zukunft, 2019). These empty runs not only contribute to increased operational costs for logistics providers, but also have a detrimental impact on environmental sustainability.

Traditionally, logistics providers have relied on intermediaries, such as brokers, to facilitate order transfers while maintaining confidentiality. These intermediaries act as trusted third parties, ensuring that sensitive data remains protected from competitors (Fig. 1). While these intermediaries ensure data confidentiality, they can limit the scope of optimization, reduce transparency, and hinder adaptability in dynamic markets. True collaborative optimization, without a central intermediary, has the potential to significantly reduce empty runs, improve resource utilization, and lower overall costs. However, the fear of data leakage, particularly the disclosure of sensitive information like cost structures, significantly hinders the widespread adoption of such collaborative models. This challenge calls for innovative solutions that enable secure and efficient collaboration while ensuring data confidentiality.

To reduce empty runs and associated CO_2 emissions, we propose a decentralized framework utilizing swarm intelligence. Inspired by nature, this approach allows logistics providers to collaboratively optimize order allocation following simple rules (Schranz et al., 2021) while sharing anonymized data. Unlike resource-intensive methods, swarm intelligence is computationally lightweight, minimizing environmental impact. This leads to improved resource utilization, reduced operational costs, and significant environmental and economic benefits, making it an environmentally conscious AI system.

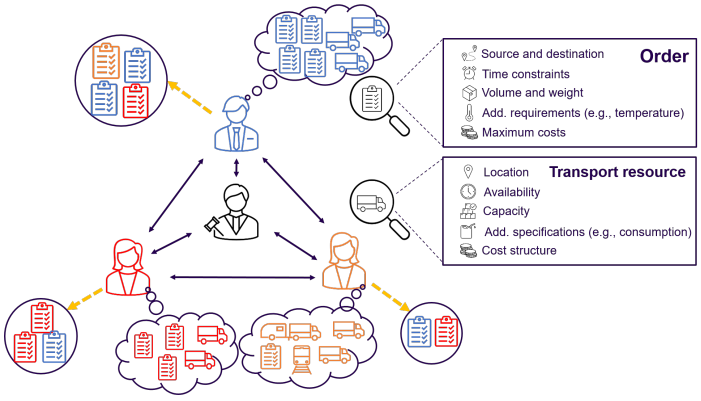


Figure 1: Traditional system architecture involving three parties and one broker that is considered to be trustworthy (Gojković and Schranz, 2024).

2 Method and Experiments

To enable secure collaboration, we model the task scheduling problem using agent-based modeling (ABM). ABM approach eliminates the need for a central broker, enhancing privacy by restricting data exchange between agents: Freighters cannot access the origin of orders placed by other freighters. Orders can only access limited details about other orders that share the same truck. The freighter who placed an order (order origin) is always hidden. However, a freighter has access to a pool of available trucks to place its order without accessing information on ownership of a particular truck (e.g., other freighter id). All trucks available in a pool aren't able to share any information with one another (e.g., ownership, destination). Lastly, trucks do not have access to the origin of orders. This decentralized framework, inspired by swarm intelligence, leverages the ABC algorithm to optimize order allocation while preserving privacy. By modeling trucks as food sources and orders as bees, the algorithm optimizes resource utilization, minimizes empty runs, and ensures timely deliveries while adhering to privacy constraints. More details on this implemented method are given in (Gojković and Schranz, 2024).

Table 1: Performance of swarm-based system compared to benchmark (Gojković and Schranz, 2024).

	Ins.11	SI	Imp %	Ins.139	SI	Imp %	Ins.180	SI	Imp %
Average empty	9.72	9.72	0	8	8.92	-11.5	3.78	3.01	20.37
Total drives	25	25	0	25	26	-4	294	231	21.43

This study evaluates the swarm intelligence approach inspired by the ABC algorithm in three distinct logistics scenarios, each with varying configurations of trucks, freighters, and orders, leading to different swarm sizes. The goal is to design a privacy-preserving system that avoids the exchange of sensitive information while achieving performance comparable to a benchmark. The results show that the swarm-based system exhibits varying performance across the scenarios. The largest swarm size (scenario based on instance 180) yields the most promising results, even in the worst-performing scenario (instance 139). The system maintains privacy by ensuring no sensitive information is exchanged. In the small scenario, the system achieves performance close to the benchmark, in the middle scenario there is a degradation of 11.5% and in the large scenario there is an improvement of 21.43%.

3 Conclusion

This work[1] demonstrates the feasibility of a privacy-preserving logistics system using a distributed, swarm-based approach. By enabling anonymized collaboration, it reduces empty runs, optimizes resource utilization, and lowers CO_2 emissions, contributing to a more sustainable and efficient freight sector. The results highlight the significant influence of swarm size on system performance, underscoring the need for further research to refine configurations that balance privacy, efficiency, and environmental benefits. The lightweight nature of swarm intelligence makes it a promising solution for sustainable AI-driven logistics.

References

Gojković, M. and Schranz, M. (2024). Preserving privacy in logistics by using swarm intelligence from the bottom-up. In *2024 IEEE 12th International Conference on Intelligent Systems*, pages 1–7.

Schranz, M., Di Caro, G. A., Schmickl, T., Elmenreich, W., Arvin, F., Şekercioğlu, A., and Sende, M. (2021). Swarm intelligence and cyber-physical systems: concepts, challenges and future trends. *Swarm and Evolutionary Computation*, 60:100762.

VCÖ Mobilität mit Zukunft (2019). Anzahl Lkw-Leerfahrten in Österreich stark gestiegen – jeden 3. Kilometer fahren Lkw leer. https://vcoe.at/presse/presseaussendungen/detail/vcoe-anzahl-lkw-leerfahrten-in-oesterreich-stark-gestiegen-jeden-3-kilometer-fahren-lkw-leer. Accessed 25-June-2024.

[1]This work was performed in the course of the project MUPOL (MUlti Party Optimization for Logistics) supported by FFG under contract number FO999902669.

Estimating the Energy Consumption of AI in Smart Meter Data Analysis

Carolina Fortuna[1] **Vid Hanzel**[1] **Dumitru Roman**[2]

[1]Jozef Stefan Institute, Ljubljana, Slovenia
[2]SINTEF AS, Norway

1 Introduction

In the energy domain, Large Language Models (LLMs) have been applied, trained, or fine tuned to automate the synthesis of complex regulatory texts or forecast energy trends, supporting decision-making by generating precise, context-aware insights from unstructured data with the help of retrieval augmented generation (RAG) (Fortuna et al. (2024)). Furthermore, AI Assistants have become pivotal in processing and generating domain-specific text in the energy sector (Majumder et al. (2024)).

Estimating the energy consumption of AI models in such scenarios is a challenging task. A new metric, eCAL, that measures the end-to-end energy cost of AI lifecycle including data collection, pre-processing, training and inference has been proposed (Chou et al. (2025)). The metric has been introduced for AIoT systems and has been also proposed as suitable for measuring the energy consumption of the emerging open and softwarized cellular systems (Chou et al. (2024)). In this abstract we elaborate on the architecture of the open source calculator that enables computing the end-to-end energy cost of AI lifecycle for a household smart meter scenario, based on the eCAL metric.

2 Method and and Case Study

Figure 1 depicts the computation blocks and the flow of the eCAL calculator on an example with a household smart meter. The horizontal blue arrow shows the smart meter data flow from collection to abstraction in an AI model. The data collected by the smart meter is assumed to be transmitted via wireless connectivity to some computing device where the measurements are pre-processed and then used to train and evaluate an AI system. The resulting model is then used for inference.

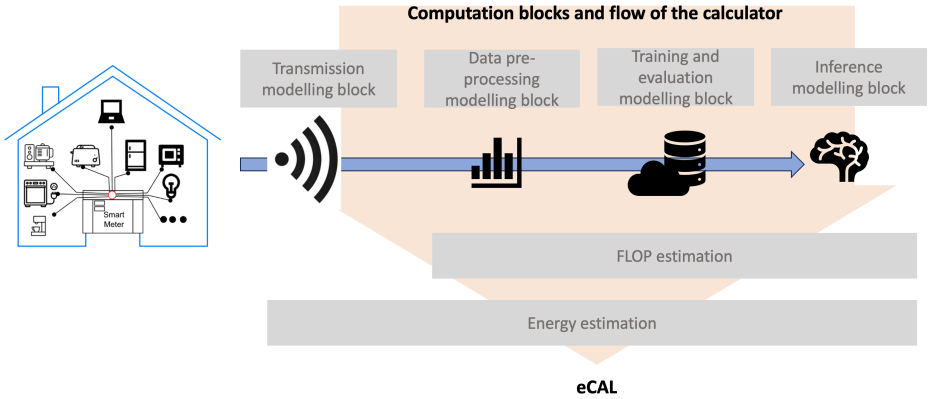


Figure 1: Computational blocks and flow of the eCAL calculator on an example with a household smart meter.

The calculator, available on GitHub[1], contains a transmission modeling block that can be configured to use generic value for data and control plane overheads as well as per OSI layer overheads. Additionally, it can also be configured to include specific protocols such as the ones from the TCP/IP stack. It also includes a data processing block currently implementing three pre-processing models: normalization, min-max scaling and Grammian angular field transformation. The training and evaluation block contains methods for modeling multilayer perceptrons, convolutional neural networks, Kolmogorov Arnold Networks and transformer decoder. However custom models or exiting models can also be added. The inference block is able to model all the architectures supported by the training and evaluation block. For the computational blocks, the calculator first estimated floating points operations (FLOPs) that are required for performing the respective data manipulation operation corresponding to that block. Finally, for all blocks, the calculator computed per block energy estimation based on configurable processor and transmitter power. Finally, eCAL is computed based on configurable number of inferences the model is expected to make during its lifetime.

3 Discussion and Outlook

In this abstract, we showcased a modular and open source calculator that can be used as is or extended to estimate the energy consumption of the entire lifecycle of AI models in a household smart meter scenario, from data collection to their last inference. It can be used for a large variety of AI tasks including to estimate the energy of developing models for energy forecasting, disaggregation, data analysis through conversational agents with configurable communication technologies for smart grids and neural architectures including MLPs, CNNs, KANs and transformers. In our existing and future work on AI applied to smart infrastructures, we plan to add eCAL results to quantify the energy footprint of the developed technology and we invite the community to use and extend the calculator.

Acknowledgements. The work is funded partially through the Slovenian Research Agency under the grant P2-0016SEP, and the projects SynDaGen (SINTEF internal funding), enRichMyData (HE 101070284), and UPCAST (HE 101093216).

References

Chou, S.-K., Hribar, J., Hanžel, V., Mohorčič, M., and Fortuna, C. (2025). The energy cost of artificial intelligence of things lifecycle. *arXiv*, pages 1–12.

Chou, S.-K., Hribar, J., Mohorčič, M., and Fortuna, C. (2024). Towards the standardization of energy efficiency metrics of the ai lifecycle in 6g and beyond. In *2024 IEEE Conference on Standards for Communications and Networking (CSCN)*, pages 187–190.

Fortuna, C., Hanžel, V., and Bertalanič, B. (2024). Natural language interaction with a household electricity knowledge-based digital twin. In *2024 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pages 8–14.

Majumder, S., Dong, L., Doudi, F., Cai, Y., Tian, C., Kalathil, D., Ding, K., Thatte, A. A., Li, N., and Xie, L. (2024). Exploring the capabilities and limitations of large language models in the electric energy sector. *Joule*, 8(6):1544–1549.

[1] https://github.com/sensorlab/eCAL

Eco-RETINA: A Green, Flexible Algorithm for Model Building

Javier Capilla[1] **Alba Alcaraz**[1] **Ángel Valarezo**[1,2] **Alfredo García-Hiernaux**[1,2] **Teodosio Pérez-Amaral**[1,2]

[1]Universidad Complutense de Madrid
[2]Instituto Complutense Análisis Económico

Eco-RETINA is an innovative and eco-friendly algorithm explicitly designed for out-of-sample prediction. Functioning as a regression-based flexible approximator, it is linear in parameters but nonlinear in inputs, employing a selective model search to optimize performance. The algorithm adeptly manages multicollinearity while emphasizing speed, accuracy, and environmental sustainability. Its modular and transparent structure facilitates easy interpretation and modification, making it an invaluable tool for researchers in developing explicit models for out-of-sample forecasting. The algorithm generates outputs such as a list of relevant transformed inputs, coefficients, standard deviations, and confidence intervals, enhancing its interpretability.

Now implemented in Python and soon available on GitHub, Eco-RETINA introduces several new features, including measuring CO_2 emissions and energy consumption. These enhancements, alongside improved data transformations, bottleneck elimination, and a user-friendly interface, significantly boost its performance. The algorithm achieves remarkable reductions in carbon footprint and power consumption (ranging from 50% to 90%) while significantly reducing computational time. Empirical results indicate that Eco-RETINA is not only a sustainable alternative to conventional neural networks but also surpasses them in certain aspects, offering a competitive edge in accuracy, interpretability, and environmental impact.

1 Motivation and Characteristics

The power consumption and carbon footprint associated with AI-related algorithms have become pressing concerns for researchers, industry, and policymakers. Kaack et al. (2022), OECD (2022) OECD (2022), Barbierato and Gatti (2024), among others, have highlighted the critical importance of addressing these issues, with some researchers predicting the potential depletion of available energy resources by 2030. To tackle these challenges, major technological firms and initiatives such as the European Commission's Enfield project are investing heavily in sustainable AI solutions.

The energy costs of developing, training, and deploying deep learning models have surged in recent years, driven by the increasing complexity of neural networks and their reliance on computationally intensive resources like GPUs and TPUs. As a result, the development of efficient and sustainable algorithms has become imperative. However, many of these innovations come at the cost of reduced performance compared to their original counterparts.

Eco-RETINA addresses this gap by providing a green algorithm specifically designed for out-of-sample prediction. Building on the foundational work of Perez-Amaral et al. (2003), Eco-RETINA (Capilla, 2024) integrates substantial improvements over previous versions. It operates as a regression-based flexible approximator, linear in parameters for improved convergence and nonlinear in inputs for enhanced flexibility. The algorithm's selective search process allows faster execution without compromising accuracy. It effectively handles multicollinearity through a threshold-based approach and incorporates robust mechanisms for outlier detection and management.

Eco-RETINA is modular and transparent, enabling users to replace individual components with alternative modules tailored to specific problems. Unlike black-box models, its transparent nature

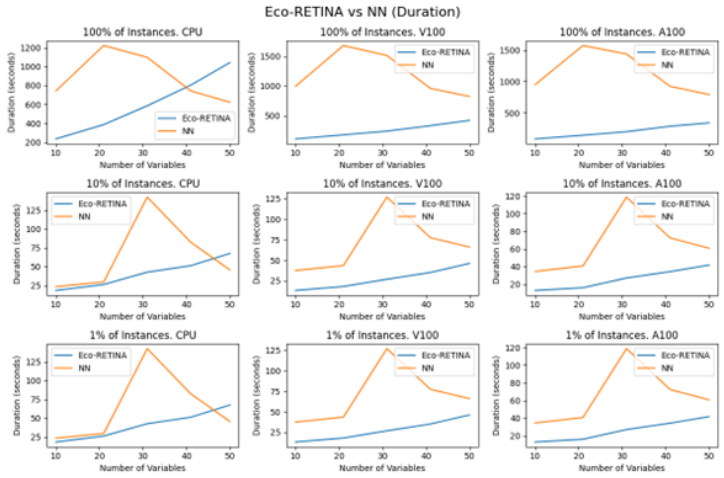


Figure 1: The duration of Eco-RETINA vs. a relevant NN

allows for modification and understanding by the user. Its outputs, akin to those of traditional econometric packages, include explicit forecasting models, point estimates, parameter standard deviations, and confidence intervals, all of which are interpretable.

Written in Python and soon available on GitHub, Eco-RETINA incorporates functionalities for measuring CO_2 emissions and energy expenditure. The algorithm also allows for a range of data transformations, including ratios, cross-products, inverses, and logarithmic transformations of inputs. Bottleneck elimination during programming ensures optimal performance by avoiding repetitive, time- and energy-intensive operations. A newly developed user interface enhances accessibility and usability.

Eco-RETINA is faster over much of the relevant range of the number of variables (see Figure 1). In an example modeling the prices of diamonds as a function of their characteristics (e.g., carat weight, measured lenghts and widths, quality of polish, etc.), the algorithm gives us a parsimonious model using 9 (possibly transformed) inputs plus the constant (see Figure 2). Eco-RETINA also provides the usual statistics for regression.

2 Conclusions

Eco-RETINA represents a significant advancement over its predecessors by integrating additional capabilities such as outlier management, enhanced multicollinearity handling, expanded input transformations, bottleneck elimination, and the ability to measure power consumption, carbon footprint, and execution time. Experiments demonstrate that Eco-RETINA achieves prediction errors comparable to earlier versions while significantly reducing power consumption and computational time (up to 4.25 times faster).

When compared to neural networks, Eco-RETINA shows several advantages. While neural networks may achieve slightly lower prediction errors, Eco-RETINA's training emissions are substantially smaller in most experiments. Additionally, it outperforms neural networks in accuracy when used with a moderate number of inputs. Its openness, interpretability, and speed position it as a strong contender among Green AI algorithms.

Eco-RETINA provides an explicit forecasting model based on transformed inputs, along with comprehensive diagnostics such as estimated coefficients, t and F statistics, and coefficient of determination. As a transparent and efficient tool, it complements neural networks and serves as a cost-effective exploratory data tool for identifying promising subsets of transformed inputs.

```
eco_retina.model.summary()
```
✓ 0.0s

OLS Regression Results

Dep. Variable:	obj_variable	R-squared:	0.839
Model:	OLS	Adj. R-squared:	0.839
Method:	Least Squares	F-statistic:	5.631e+04
Date:	Sun, 14 Apr 2024	Prob (F-statistic):	0.00
Time:	18:39:16	Log-Likelihood:	-7.0650e+05
No. Observations:	86553	AIC:	1.413e+06
Df Residuals:	86544	BIC:	1.413e+06
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
constant	2406.9761	1093.325	2.202	0.028	264.068	4549.884
carat_weight^1*meas_width^1	1012.5979	1.530	662.004	0.000	1009.600	1015.596
meas_width^-1*meas_length^1	265.2951	18.546	14.304	0.000	228.944	301.646
polish_Excellent*depth_percent	110.3873	25.818	4.276	0.000	59.784	160.991
polish_Excellent*table_percent	-75.9719	27.756	-2.737	0.006	-130.373	-21.571
polish_Very Good	2149.2133	603.394	3.562	0.000	966.566	3331.860
polish_Excellent	257.7628	584.608	0.441	0.659	-888.064	1403.590
polish_Very Good*depth_percent	91.2128	26.122	3.492	0.000	40.013	142.412
polish_Very Good*table_percent	-91.0710	27.773	-3.279	0.001	-145.506	-36.636
table_percent^-1*depth_percent^1	-5306.3593	1510.614	-3.513	0.000	-8267.151	-2345.568

Omnibus:	13326.421	Durbin-Watson:	2.006
Prob(Omnibus):	0.000	Jarque-Bera (JB):	166250.786
Skew:	0.336	Prob(JB):	0.00
Kurtosis:	9.756	Cond. No.	2.34e+16

Figure 2: Eco-RETINA output summary

In an era where sustainable AI is increasingly essential, Eco-RETINA's emphasis on speed, accuracy, and environmental sustainability underscores its critical role in the future of algorithm development. By offering a green, interpretable, and high-performing alternative, Eco-RETINA sets a new standard for sustainable machine learning and forecasting tools.

Acknowledgements. This work was supported by the European Union's HORIZON Research and Innovation Programme under grant agreement No 101120657, project ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI).

References

Barbierato, E. and Gatti, A. (2024). Towards green ai. a methodological survey of the scientific literature. *IEEE Access*.

Capilla, J. (2024). *Modelado automático con RETINA desde un enfoque de Green AI*. PhD thesis, Facultad de Económicas y Empresariales, Universidad Complutense de Madrid.

Kaack, L. H., Donti, P. L., Strubell, E., Kamiya, G., Creutzig, F., and Rolnick, D. (2022). Aligning artificial intelligence with climate change mitigation. *Nature Climate Change*, 12(6):518–527.

OECD (2022). Measuring the environmental impacts of artificial intelligence compute and applications: The ai footprint. *OECD Digital Economy Papers*.

Perez-Amaral, T., Gallo, G. M., and White, H. (2003). A flexible tool for model building: the relevant transformation of the inputs network approach (retina). *Oxford Bulletin of Economics and Statistics*, 65(s1):821–838.

Neuralizing Cloned-Structured Causal Graphs for World Model Learning

Tristan M. Stöber[1, 2] **Ali Dasmeh**[3] **Erik J. Husom**[4] **Sagar Sen**[4]

[1]Institute for Neural Computation, Faculty of Computer Science, Ruhr University Bochum, Germany
[2]Goethe University Frankfurt, Epilepsy Center Frankfurt Rhine-Main, Department of Neurology, University Medical Center Frankfurt, Frankfurt, Germany
[3]University of Europe for Applied Sciences, Germany
[4]SINTEF Digital, Oslo, Norway

1 Introduction

Current AI technology lacks true understanding. Even in advanced transformer architectures, learning remains statistical. Unlike mammalian brains, they lack an internal model of the world which would allow them to genuinely extract hidden causes behind observations by counterfactual reasoning. The lack of world models in AI systems has dire consequences: Despite the vast amounts of training data, they still generalize poorly to unseen situations.

Cloned-structured causal graphs (CSCGs) are an innovative approach for world model learning (George et al., 2021). CSCGs are based on Hidden Markov Models, with multiple hidden states for a given observation - so-called clones - and a sparse emission matrix (Dedieu et al., 2019). Trained on sequences of observation and action pairs, clones learn to distinguish the context of similar observations. The transitions between the clones define an interpretable graph, which captures the structure of the world that generated the observations.

While CSCGs explain learning dynamics in the brain (Sun et al., 2025) and enable desirable computations, such as schema learning and hierarchical planning, their current matrix-based formulation is rigid and limited to well-defined scenarios (Raju et al., 2024).

2 Method and Experiments

In this work, we demonstrate, as a proof-of-principle, that CSCGs can be implemented using deep neural networks. For this purpose, we replace both transition and emission matrices with multilayer feed-forward neural networks and optimize them by reformulating the training process as gradient optimization (Rabiner, 1989; Tran et al., 2016).

The transformation of emission and transmission matrices in CSCGs into neural architectures enables greater adaptability and generalization. Both emission and transition matrices, defining the relations between observation and hidden states, can be replaced with neural networks. In its current form, we focus on replacing the transition matrix and train it as in (Tran et al., 2016).

We compare our approach with the classical CSCG formulation, focusing on learning performance and data efficiency. As input data we use sequences of observation and action pairs generated by an agent randomly exploring a grid world.

3 Discussion and Outlook

Neuralizing CSCGs opens multiple avenues for future research. A neural network implementation of CSCG is not only relevant from a neuroscience perspective, but also allows us to subsequently integrate this approach into existing model-based reinforcement learning architectures, such as Dreamer (Hafner et al., 2023). We hypothesize that a CSCG-enhanced version of Dreamer will excel

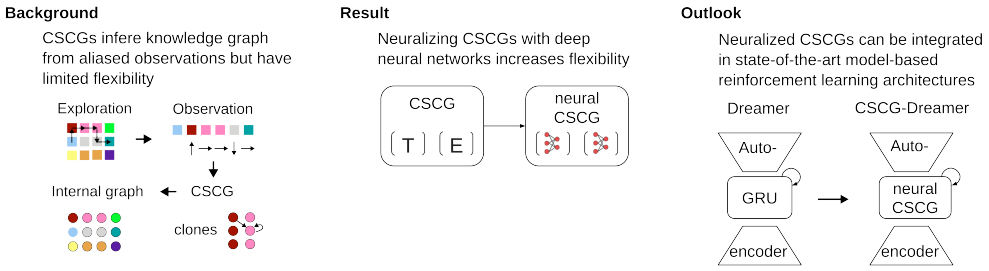


Figure 1: **Neuralizing Cloned-Structured Causal Graphs for World Model Learning**
Background CSCGs are trained on sequences of observation and action pairs of an agent exploring a room. A graph representation can be extracted from the transition probabilities between hidden clones (George et al., 2021).
Result We replace emission and transition matrices by neural networks and compare their performance.
Outlook Neuralizing CSCGs offers various advantages, such as integration with existing reinforcement architectures like Dreamer (Hafner et al., 2023).

in challenging navigation tasks, such as the Memory Maze (Pasukonis et al., 2022). Further, we plan to address how to dynamically allocate clones, how to efficiently reuse learned schemas, and how to implement this algorithm on energy-efficient neuromorphic hardware.

References

Dedieu, A., Gothoskar, N., Swingle, S., Lehrach, W., Lázaro-Gredilla, M., and George, D. (2019). Learning higher-order sequential structure with cloned hmms. *arXiv preprint arXiv:1905.00507*.

George, D., Rikhye, R. V., Gothoskar, N., Guntupalli, J. S., Dedieu, A., and Lázaro-Gredilla, M. (2021). Clone-structured graph representations enable flexible learning and vicarious evaluation of cognitive maps. *Nature Communications*, 12(1):1–17.

Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. (2023). Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*.

Pasukonis, J., Lillicrap, T., and Hafner, D. (2022). Evaluating long-term memory in 3d mazes. *arXiv preprint arXiv:2210.13383*.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Raju, R. V., Guntupalli, J. S., Zhou, G., Wendelken, C., Lázaro-Gredilla, M., and George, D. (2024). Space is a latent sequence: A theory of the hippocampus. *Science Advances*, 10(31):eadm8470.

Sun, W., Winnubst, J., Natrajan, M., Lai, C., Kajikawa, K., Bast, A., Michaelos, M., Gattoni, R., Stringer, C., Flickinger, D., Fitzgerald, J. E., and Spruson, N. (2025). Learning produces an orthogonalized state machine in the hippocampus. *Nature*.

Tran, K., Bisk, Y., Vaswani, A., Marcu, D., and Knight, K. (2016). Unsupervised neural hidden markov models. *arXiv preprint arXiv:1609.09007*.

Hybridization of data and expert knowledge: Towards inherently interpretable AI models

Ricardo J. Bessa[1,*] **Francisco S. Fernandes**[1,2,*] **Lucas Bost**[1,*] **João Peças Lopes**[1,2,*]

[*]Equal contribution.
[1]INESC TEC – The Institute for Systems and Computer Engineering, Technology and Science
[2]Faculty of Engineering of the University of Porto

1 Introduction

The widespread integration of renewable energy sources is driving a shift toward operating conditions characterized by low system inertia, significant temporal variations in generation, and an increasing penetration of distributed energy resources (DER). This transition significantly enhances human operators' supervisory and control responsibilities in control rooms while also requiring innovative power system control strategies (Marot et al., 2022). Examples include frequency-sensitive modes, fast frequency response, and hidden inertia. Among these, grid-forming (GFM) control has garnered considerable attention due to its promising ability to enhance system stability and resilience (Musca et al., 2022).

In this evolving context, ensuring system security becomes increasingly challenging. The growing prevalence of DER introduces greater susceptibility to instability, as their behaviors are influenced by both their primary energy sources and control mechanisms. This complexity leads to large and intricate state-space models, rendering traditional model-based techniques impractical for online dynamic security assessment (DSA). Moreover, while GFM converters have shown great potential to outperform conventional synchronous machines, their capabilities remain underutilized. This highlights the urgent need to redesign and optimize GFM control strategies to harness their advantages fully in modern power systems.

2 Method and Experiments

For these problems, data-driven approaches are promising alternatives (Wehenkel, 1997). However, the benefits of these approaches are often undermined by the "black-box" nature of many artificial intelligence methods. Thus, interpretability is essential for effectively deploying data-driven decision-support systems. This work introduces a new framework, the Evolving Symbolic Model (ESM) – depicted in Figure 1 – designed to create highly interpretable data-driven models for DSA and design of controllers, namely for system security classification and the real-time definition of preventive measures, as well as search for the best control design of the GFM primary control loop.

This ESM concept has different modes of application, namely: (a) learning a symbolic model from scratch using historical data and/or synthetic data generated in a digital environment, (b) imitating a "black-box" model (e.g., deep neural network) and distilling its knowledge into a more compact and interpretable model, (c) enhancing an existing expert system by integrating insights from data and/or a digital environment, and (d) redesigning an expert system to learn a new task or objective using data and/or a digital environment.

The ESM framework uses a meta-heuristic as the core data-driven optimizer of a symbolic and/or graph template that a human expert defines. For instance, the results for the Madeira Island case study indicate that ESM achieves classification accuracy by using pruned decision trees while offering greater global interpretability. Additionally, it outperforms an artificial neural network (Salimans et al., 2017) in identifying preventive actions. It also shows the method's

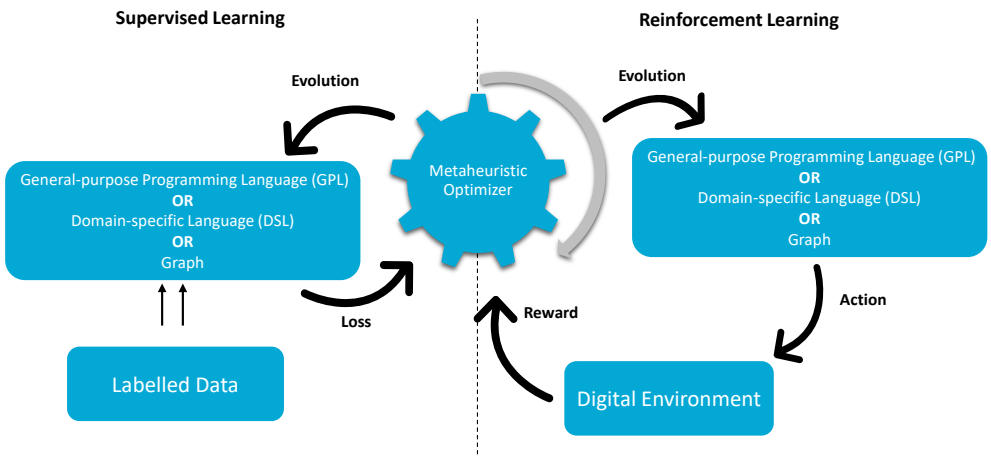


Figure 1: Evolving symbolic model framework for inherently interpretable AI models

capability to learn efficient controllers without prior knowledge while ensuring interpretable solutions, representing a step towards automating the design of control systems.

3 Discussion and Outlook

Using model-driven or mathematical formulations has long been a standard practice in the energy sector and is well-accepted by human decision-makers and operators. However, AI can still play a valuable role in complementing these traditional approaches by extracting additional knowledge from data and augmenting this pre-existing knowledge. The ESM concept can applied to other use cases, such as improving a rule-based expert system to control hybrid energy storage systems (Bessa et al., 2024), or adaptive protection systems in distribution grids.

Acknowledgements. The research leading to this work is being carried out as a part of the ENFIELD (*European Lighthouse to Manifest Trustworthy and Green AI*) project, European Union's Horizon Research and Innovation Programme, Grant Agreement No. 101120657. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

References

Bessa, R., Lobo, F. S., Fernandes, F. S., and Silva, B. (2024). Data augmented rule-based expert system to control a hybrid storage system. In *IEEE MELECON 2024*, Porto, Portugal.

Marot, A., Kelly, A., Naglic, M., Barbesant, V., Cremer, J., Stefanov, A., and Viebahn, J. (2022). Perspectives on future power system control centers for energy transition. *Journal of Modern Power Systems and Clean Energy*, 10(2):328–344.

Musca, R., Vasile, A., and Zizzo, G. (2022). Grid-forming converters. a critical review of pilot projects and demonstrators. *Renewable and Sustainable Energy Reviews*, 165:112551.

Salimans, T., Ho, J., Chen, X., and Sutskever, I. (2017). Evolution strategies as a scalable alternative to reinforcement learning. *arXiv:1703.03864*, pages 1–13.

Wehenkel, L. (1997). Machine learning approaches to power-system security assessment. *IEEE Expert*, 12(5):60–72.

LexAlign: Towards a Multiagent AI System for Regulatory Compliance of Data/AI Pipelines

Sagar Sen[1] **Carl-Henrik Lien**[2] **Nikolay Nikolov**[1] **Adela Nedisan Videsjorden**[1]
Erik Johannes Husom[1] **Shukun Tokas**[1] **Arda Goknil**[1] **Dumitru Roman**[1]

[1]SINTEF Digital, Oslo, Norway
[2]University of Oslo, Oslo, Norway

1 Introduction

Data and AI pipelines drive modern technology, seamlessly operating across the Internet of Things and the computing continuum. They process diverse data streams—time-series sensor data, structured/unstructured text, source code, images, video, sound, 3D point clouds, chemical data types, and biological omics data—transforming raw inputs into actionable insights. These intelligent systems drive automation and decision-making across numerous domains, including agriculture, life sciences, manufacturing, process industries, robotics, smart communities, transportation, energy, and even space exploration. Embedded across edge devices and the cloud, they process vast amounts of data, shaping industries and daily life. Despite their rapid deployment, concerns persist about their harmful effects on society including privacy violations, energy cannibalization, fair use abuse and algorithmic exploitation. Another major issue is *perverse instantiations*, where AI misinterprets intended goals, a risk that Bostrom warns could lead to violations of human intentions (Nick (2014)). To mitigate these risks, regulatory frameworks such as the AI Act, Data Act, GDPR have been introduced (Pathak (2024)) by the EU. Generally, enforcement has lagged behind the rapid pace of digitalization, though in some cases, these regulations have already resulted in fines. Meanwhile, jurisdictions prioritizing market leadership over strict regulation face ongoing copyright infringement cases such as against Stability AI and Microsoft's GitHub Copilot, fueling debates on fair data use (Samuelson (2023)). Therefore, we ask, *can we proactively embed compliance verification into Data/AI pipelines, ensuring alignment with evolving regulations to avoid stepping into a minefield of regulatory violations while ultimately fostering genuine societal benefit?*

2 Method and Case Study

We present LexAlign, a proof-of-principle multi-agent AI system using foundation models and software tools to detect regulatory violations and enforce compliance within Data/AI pipelines (Figure 1). By ensuring *compliance by design*, it prevents legal risks. LexAlign employs a divide-and-conquer approach, deploying specialized agents that retrieve knowledge from legal documents, contracts, responsible AI tools, and legal precedents. Built on AutoGen[1], agents invoke (multimodal) foundation models via APIs to verify violations in pipeline artifacts, including data, source code that is compiled, containerized and deployed, performance and resource usage metrics. Various observability (e.g. SIM-PIPE [2]) and repository mining tools (e.g. PyDriller) can support pipeline artifacts extraction. Furthermore, context filtering of artifacts (e.g., CodeQL) improves relevance and reduces context length for agents.

LexAlign agents perform parallel compliance verification, analyzing datasets, code, metrics from software analytics and execution traces for violations. Findings are processed by guardrails (Dong et al. (2024)) (e.g., GuardRails-AI[3]), generating compliance statements at a policy decision point.

[1]https://github.com/microsoft/autogen
[2]https://github.com/DataCloud-project/SIM-PIPE
[3]https://www.guardrailsai.com/

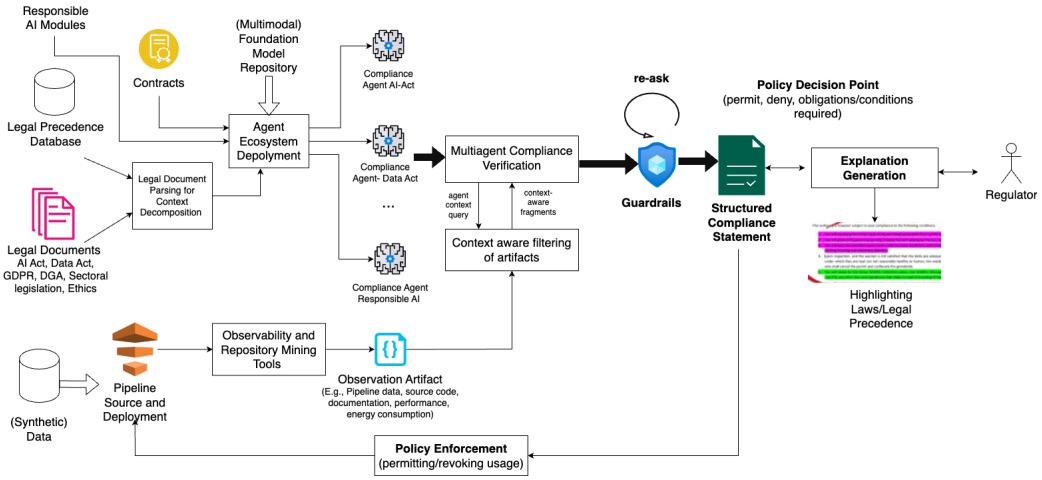


Figure 1: The LexAlign Framework for Compliance of AI and Data Pipelines

LexAlign then enforces policy by permitting, denying, or conditioning artifact flow while offering regulatory explanations via an interface. This automated process helps ensures *accountability*.

We showcase LexAlign in a remote patient monitoring (RPM) compliance scenario, observed via SIM-PIPE. The workflow includes Retrieve (data collection), Process (FHIR structuring, storage), and Notify (monitoring, alerts). An AI act lawyer agent verified articles 10, 40, and recital 29 on sensor data transmitted to RPM finding *no violation* in data-quality for high-risk, harmonized EU standards, and manipulative AI practices.

3 Discussion and Outlook

LexAlign is an experimental system using foundation model-powered agents to assess compliance in Data/AI pipelines. Due to the lack of benchmarks, we propose mutation analysis—artificially injecting compliance violations across privacy, fairness, transparency, and oversight. We injected an Extreme Outlier mutant into input data for RPM resulting in highly skewed values for weight, heart rate, and blood pressure. LexAlign detected Article 10: Non-compliant – Data contains extreme, erroneous, and unrepresentative values. Similarly, we can develop mutation operators to modify privacy settings, inject bias, obscure changes, bypass audits, manipulate data, distort AI decisions, remove safeguards, extend retention, or alter logs. LexAlign can detect these violations, with performance measured as $100 \times \frac{hits}{hits+misses}$.

Acknowledgements. Funded by HEU ENFIELD(Project number: 101120657) and DataPact (Project number: 101189771)

References

Dong, Y., Mu, R., Jin, G., Qi, Y., Hu, J., Zhao, X., Meng, J., Ruan, W., and Huang, X. (2024). Building guardrails for large language models. *arXiv preprint arXiv:2402.01822*.

Nick, B. (2014). Superintelligence: Paths, dangers, strategies.

Pathak, M. (2024). Data governance redefined: The evolution of eu data regulations from the gdpr to the dma, dsa, dga, data act and ai act. *DSA, DGA, Data Act and AI Act.(February 6, 2024)*.

Samuelson, P. (2023). Generative ai meets copyright. *Science*, 381(6654):158–161.

Towards a Vision-Language Foundation Model for Critical Infrastructure Integrity Interpretation

Sagar Sen[1] **Simeon Tverdal**[1] **Erik Johannes Husom**[1]

[1]SINTEF Digital, Oslo, Norway

1 Introduction

Critical infrastructure, such as *electric power lines*, plays a crucial role in ensuring economic stability and social resilience at both national and continental levels. Recent advancements in UAV technology, equipped with multimodal sensors like LiDAR and image sensors, now enable rapid and efficient monitoring of these essential systems. These technologies generate vast amounts of 3D point cloud and image data, which can be analyzed to assess risks, enhance maintenance strategies, and improve preparedness against potential threats. Domain experts (e.g. civil/structural engineers, field inspectors, remote sensing specialists) can further enrich this multi-modal data with textual reports enhancing the interpretability and usability of the data for informed decision-making. In an era of increasing infrastructure vulnerabilities and climate-related challenges, can we automatically interpret critical infrastructure integrity in natural language by learning from available sensor data and textual reports? We address this question by introducing *CriticalCLIP*, a vision-language foundation model that establishes connections between UAV sensor data and textual reports, enabling a semantic understanding of events affecting critical infrastructure. CriticalCLIP fine-tunes OpenAI's Contrastive Language-Image Pretraining (CLIP) (Radford et al. (2021)) with datasets containing images and reports obtained from monitoring critical infrastructure. CLIP is vision-language foundation model that has been pre-trained on millions of images and text to learn a shared representation space between images and textual description, enabling the model to understand and associate visual data with an explanation. Fine-tuning CLIP to specific datasets (e.g. STN PLAD (Vieira-e Silva et al. (2021))) enables us to build a highly speciailized vision-language model for interpreting critical infrastructure integrity.

2 Method and Experiments

CriticalCLIP is a proof-of-principle foundation model designed for critical infrastructure analysis, trained on multimodal datasets that include aerial imagery and textual reports accumulated over time. It incorporates an image encoder for processing visual infrastructure data and a text encoder for interpreting expert assessments, as illustrated in Figure 1 (a). This enables advanced semantic understanding and text-based queries for infrastructure monitoring and risk assessment, as demonstrated in Figure 1 (b). To ensure robustness, CriticalCLIP will be trained using diverse critical infrastructure datasets, including STN PLAD(Vieira-e Silva et al. (2021)), nuScenes(Caesar et al. (2020)), and Kitti-360 (Liao et al. (2022)). In cases where textual reports are unavailable, state-of-the-art large language models, such as OpenAI's GPT-4o, will be leveraged to generate synthetic reports along with labeled annotations for the training data, enhancing the model's ability to learn meaningful infrastructure representations. Additionally, we explore strategies to enhance CriticalCLIP's energy efficiency through knowledge distillation, enabling deployment on edge devices such as UAVs. This is complemented by event-triggered AI, where edge devices perform primary filtering, transmitting only high-priority data to cloud models, thereby reducing computational overhead and optimizing resource usage.

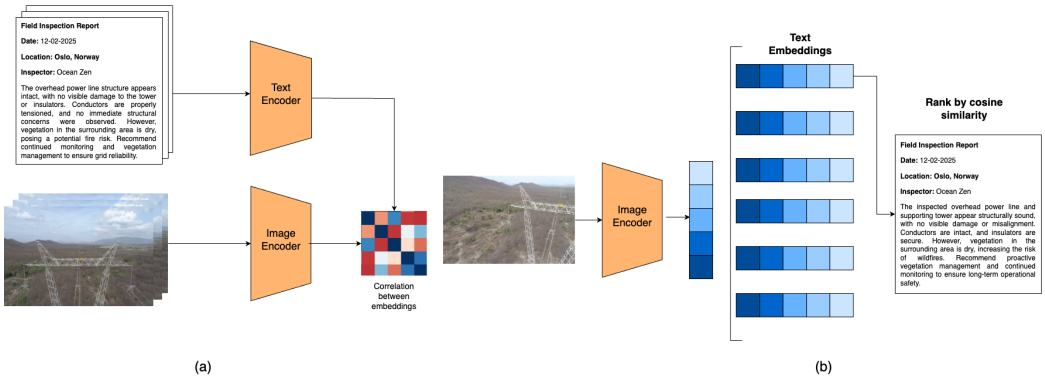


Figure 1: (a) CriticalCLIP training for connecting images and textual descriptions (b) Retreival of textual descriptions for new images

3 Discussion and Outlook

This abstract presents a conceptual solution for addressing energy sector challenges using multimodal AI for critical infrastructure monitoring. CriticalCLIP links UAV sensor data, LiDAR, images, and textual reports for risk assessment and decision-making. While it proposes methods for semantic understanding and edge AI deployment, no experimental results are provided. Future enhancements include integrating LiDARCLIP (Hess et al. (2024)), which bridges LiDAR point clouds and text by leveraging image-LiDAR data as an intermediate representation, enabling improved infrastructure analysis.

Acknowledgements. Funded by Horizon Europe ENFIELD: European Lighthouse to Manifest Trustworthy and Green AI (Project number: 101120657)

References

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. (2020). nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631.

Hess, G., Tonderski, A., Petersson, C., Åström, K., and Svensson, L. (2024). Lidarclip or: How i learned to talk to point clouds. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7438–7447.

Liao, Y., Xie, J., and Geiger, A. (2022). Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Vieira-e Silva, A. L. B., de Castro Felix, H., de Menezes Chaves, T., Simões, F. P. M., Teichrieb, V., dos Santos, M. M., da Cunha Santiago, H., Sgotti, V. A. C., and Neto, H. B. D. T. L. (2021). Stn plad: A dataset for multi-size power line assets detection in high-resolution uav images. In *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 215–222. IEEE.

Geometric and Information-Theoretic Compression in Neural Classifier Training

Linara Adilova[*1] **Bernhard C. Geiger**[*2,3]

[*]Equal contribution.
[1]Faculty of Computer Science, Ruhr University Bochum
[2]Signal Processing and Speech Commmunication Laboratory, Graz University of Technology
[3]Know Center Research GmbH

Occam's Razor – the principle that among all explanations, the simplest shall be preferred – has recently received renewed interest in deep learning. Measuring the simplicity of an explanation (e.g., a model or a latent representation) via information-theoretic quantities has led to the discovery of novel generalization bounds, e.g., (Hellström et al., 2025), and to training objectives for neural networks, e.g., (Alemi et al., 2017). The information bottleneck principle, a mathematical formulation of Occam's Razor, has furthermore been proposed for understanding the success of deep learning. Indeed, it has been claimed that stochastic gradient descent inherently removes irrelevant information from latent representations, which in turn makes overfitting impossible (Shwartz-Ziv and Tishby, 2017). Subsequent works questioned this claim, e.g., (Saxe et al., 2018), triggering a discourse in the scientific community that extends to this day. In particular, recently it has been suggested that the compression observed in classical information plane analyses is geometric, either inherently due to the stochasticity of the considered networks (Goldfeld, 2019) or via the properties of common mutual information estimators (Geiger, 2022). Thus, it is currently hypothesized that information-theoretic and geometric compression are strongly connected.

1 Experimental Setup

To shed some light on this hypothesis, we investigate the behavior of information-theoretic and geometric compression measures during network training. Specifically, for this extended abstract, we used the conditional entropy bottleneck framework of Fischer (2020) that trains a stochastic encoder $p_\theta(z|x)$ and a deterministic decoder $\hat{y} = f_\psi(z)$ via minimizing the following functional:

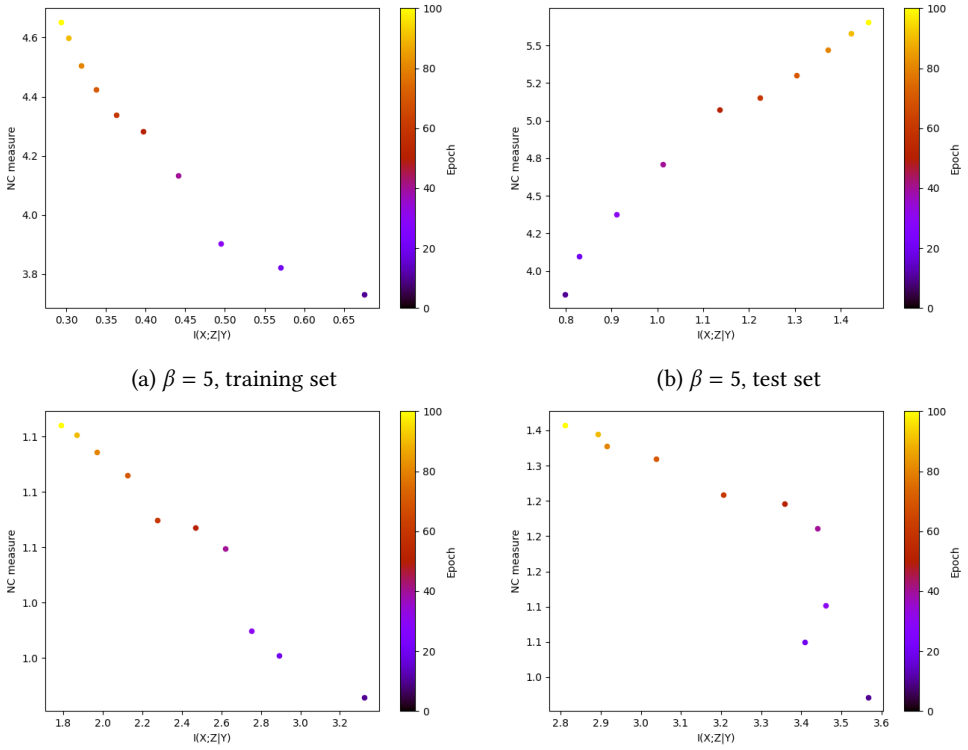
$$I(x;z|y) + \beta L_{\mathrm{CE}}(y; f_\psi(z))$$

where $I(\cdot)$ measures information-theoretic compression via mutual information, $L_{\mathrm{CE}}(\cdot)$ is the cross-entropy loss, and β trades between these objectives. We used a variational approximation of this functional to train a LeNet5 (LeCun et al., 1998) encoder and an MLP decoder with one hidden layer of size 1024 on FashionMNIST. We used ReLU activation functions and selected the dimensionality of the bottleneck variable z to 64.

For this network and for various values of β, we plot how information-theoretic and geometric compression change throughout training. We quantify information-theoretic compression via the loss component $I(x;z|y)$, which is readily available for each epoch. To quantify geometric compression, we compute, for selected epochs, the unified neural collapse characteristic measure proposed by Galanti et al. (2021):

$$NC = \frac{1}{(c-1)c} \sum_{i,j=1}^{c} \frac{\mathrm{Var}_i + \mathrm{Var}_j}{2\|\mu_i - \mu_j\|^2}$$

where c is the number of classes and μ_i and Var_i are the mean vectors and traces of covariance matrices of representations z generated for samples x belonging to class $y = i$. A strong class-specific clustering structure of z leads to small values of NC.



(a) $\beta = 5$, training set

(b) $\beta = 5$, test set

(c) $\beta = 100$, training set

(d) $\beta = 100$, test set

Figure 1: Information-theoretic compression $I(x;z|y)$ vs. geometric compression NC at different epochs of training a stochastic LeNet5 encoder and an MLP decoder on FashionMNIST. Small β (top) indicates a strong relative weight on regularizing $I(x;z|y)$. Compression measures were evaluated on the training (left) and test sets (right).

2 Results

The results in Figure 1 show that when the networks train successfully (approx. 99% accuracy on the training and 90% accuracy on the test set), information-theoretic and geometric compression pursue different paths: Representations z appear to compress information-theoretically, but expand geometrically during training, as can be seen by increasing values of NC. We observed such a negative correlation between NC and $I(x;z|y)$ also for end-of-training values for different datasets and neural architectures. This indicates that the purported connection between information-theoretic and geometric compression is not as universal as initially believed and appears even reversed for stochastic networks regularized with mutual information. For small values of β, leading to underfitting networks due to strong information compression, the behavior of NC and $I(x;z|y)$ during training is less consistent. Indeed, when evaluated on the test set, NC and $I(x;z|y)$ both increase throughout training (Figure 1b). This indicates that the network seems to overfit on the dominant regularization term $I(x;z|y)$, which is minimized on the training set, but increases on the test set. This phenomenon of overfitting on a loss component that is aimed at *avoiding* overfitting seems to be quite interesting and deserves future study.

Acknowledgements. This work was supported by the European Union's HORIZON Research and Innovation Programme under grant agreement No 101120657, project ENFIELD (European

Lighthouse to Manifest Trustworthy and Green AI), and by the subgrant “Information-theoretic analysis of generalization in deep learning” (oc1-2024-TES-01).

References

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. (2017). Deep variational information bottleneck. In *Proc. Int. Conf. on Learning Representations (ICLR)*.

Fischer, I. (2020). The conditional entropy bottleneck. *Entropy*, 22(9):999.

Galanti, T., György, A., and Hutter, M. (2021). On the role of neural collapse in transfer learning. In *Proc. Int. Conf. on Learning Representations (ICLR)*.

Geiger, B. C. (2022). On information plane analyses of neural network classifiers – a review. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):7039–7051. arXiv:2003.09671 [cs.LG].

Goldfeld, Z. (2019). Estimating information flow in deep neural networks. In *Proc. Int. Conf. on Machine Learning (ICML)*.

Hellström, F., Durisi, G., Guedj, B., and Raginsky, M. (2025). *Generalization Bounds: Perspectives from Information Theory and PAC-Bayes*. Foundations and Trends® in Machine Learning.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Saxe, A. M., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A., Tracey, B. D., and Cox, D. D. (2018). On the information bottleneck theory of deep learning. In *Proc. Int. Conf. on Learning Representations (ICLR)*.

Shwartz-Ziv, R. and Tishby, N. (2017). Opening the black box of deep neural networks via information. arXiv:1703.00810v3 [cs.LG].