ML4ProM
NireHApS
EKAmBa
FH
Bielefeld
UNIVERSITÄT
Lemgo
TH OWL
ModularDRL
RoSe
Paderborn
UNIVERSITÄT
TU
Dortmund
X-FI
Bochum
RUHR-UNIVERSITÄT
Hagen
FERNUNIVERSITÄT
(RL)³
GAIA
TH
Köln
RWTH
Aachen

dataninja.nrw

sAIOnARA

2024 Conference

Shaping Trustworthy AI: Opportunities, Innovation & Achievements for Reliable Approaches

DATES

25-27th
JUNE
2024

or visit www.dataninja.nrw

Illustration: Christoph J Kellner // studio animanova

Funded by
Ministry of Culture and Science
of the State of
North Rhine-Westphalia

Research training group *DataNinja*: Trustworthy AI for Seamless Problem
Solving: Next Generation Intelligence Joins Robust Data Analysis
Research Institute for Cognition and Robotics
Bielefeld University
CITEC
Inspiration 1
33619 Bielefeld

BieColl –Bielefeld eCollections

Editor
Ulrike Kuhl
Bielefeld University
Bielefeld, Germany

Email: contact@dataninja.nrw

August 2024

Contents

Shaping Trustworthy AI: An Introduction to This Issue

Ulrike Kuhl UKUHL@TECHFAK.UNI-BIELEFELD.DE
Machine Learning Group, Faculty of Technology, Bielefeld University, Bielefeld, Germany

1. Introduction

The rapid advancement of artificial intelligence (AI) has brought about a paradigm shift in various domains, from healthcare to finance, and from autonomous systems to natural language processing. As AI systems become increasingly integrated into our daily lives, ensuring their trustworthiness is paramount. The DataNinja sAIOnARA 2024 Conference, centered around the theme "Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches", brought together cutting-edge research aimed at addressing the multifaceted challenges of creating reliable and ethical AI systems. This collection of scientific abstracts represents a broad spectrum of innovative work that contributes to the overarching goal of trustworthy AI.

The contributions are grouped under three prevailing hot topics in XAI: fairness and ethics, interpretability and transparency, and reliability and robustness, highlighting the multifaceted approaches to developing AI systems that are both innovative and trustworthy.

2. Fairness and Ethics

A central theme related to trustworthy AI is in how far systems can be considered fair and ethical. As automated decision-making increasingly impacts individuals and communities directly, concerns about bias, equity, and transparency become critical.

Balestra [1] delves into the fundamental question of fairness in algorithmic rankings. Rankings are a ubiquitous feature in modern life, from search engines to personalized recommendations. They draw attention to the fact that-while fairness may not seem essential when ranking depersonalized items-it becomes deeply relevant when individuals are being ranked. In such cases, disparities in how people are represented or treated can have significant consequences. Highlighting the often conflicting relationships between group fairness, individual fairness, and diversity in rankings, Balestra [1] draws attention to the inherent trade-offs and complexities involved in attempting to optimize all these aspects simultaneously. Moreover, their exploration combines Shapley values [2], known for promoting individual fairness, with a diversity measure to ensure group fairness in rankings. This approach introduces a framework that balances individual contributions with the need for diverse representation, offering a potential pathway for future research to operationalize fairness in rankings.

The work presented in Hellwig and Maier [3] extends the conversation on fairness into the domain of workplace leadership, where AI may increasingly take over functions like rewarding employee performance or allocating resources. Drawing on the "Resource Theory of Social Exchange" [4], they present a study plan to explore whether certain resources, such as affiliation or emotional support, hold the same value when allocated by AI compared to human leaders. This raises profound ethical questions about the "humanness" of resource allocation and whether AI can truly fulfill the nuanced role of a leader in fostering relationships and providing more than just material rewards. This perspective introduces an important dimension to the discussion of trustworthy AI: while AI systems may be technically proficient at optimizing resources, their ability to consider the social and emotional impacts of their decisions remains questionable.

Shifting the focus from fairness in decision-making processes to the social dynamics of human-AI interactions, Arlinghaus and Maier [5] outline a research framework to explore these interactions in workplaces where humans and robots collaborate. They seek to explore how individuals experience social exclusion when working with robots compared to exclusion by human colleagues, highlighting a unique challenge in the development of AI systems: the human perception of social interaction with machines. The research will focus on addressing the psychological needs of belonging, control, and self-esteem, examining how these needs are impacted differently depending on whether the source of exclusion is a human or a robot. By questioning the validity of the "Computers Are Social Actors" theory [6], their work promises to open the door to a deeper understanding of how humans attribute social qualities to AI and robots and how these attributions influence the effectiveness and fairness of human-robot collaborations.

Taking a more technological perspective to the fairness and ethics discussion, Sanaullah et al. [7] focus on the issue of data privacy. They tackle the challenge of preserving privacy in machine learning models without sacrificing performance in the context of different encryption techniques, highlighting the trade-offs between model accuracy, memory usage, training time, and security. By demonstrating how different encryption methods impact both the robustness of the models and their interpretability, this work addresses a core concern in the development of trustworthy AI systems. Their findings provide crucial insights into how we can develop ML models that are both effective and respectful of individual privacy, ensuring that fairness is maintained even when data is securely encrypted.

In the domain of process segmentation in operational systems, the work presented by Norouzifar and van der Aalst [8] highlights six potential research questions related to leveraging information not just from desirable, but also from undesirable events in process mining tasks. This research has significant implications for fairness in operational decision-making, as the method provides a more nuanced view of process data to help organizations ensure that cases are treated equitably based on their complexity and risk. Thus, the work presented by Norouzifar and van der Aalst [8] highlights the pathway to fairer outcomes in process management that will eventually support organizations in delivering more balanced and transparent decisions.

The contributions in this section collectively highlight the complexity and multifaceted nature of fairness and ethics in AI. Taken together, they illustrate the breadth of ethical challenges we face as AI systems take on increasingly important roles in society: from ensuring fair representation in rankings to maintaining ethical considerations in resource allocation and leadership; from fostering inclusive social dynamics in AI-human collaboration to protecting privacy while maintaining performance. They call for ongoing research, thoughtful design, and ethical foresight to ensure that AI systems not only perform their tasks efficiently but also align with broader human values of fairness, equity, and trust.

3. Interpretability and Transparency

While the relationship between interpretability and trust is more nuanced than often assumed, a prevailing notion remains that systems that are more interpretable and transparent are generally easier to trust [9]. The contributions in this section explore various approaches to making AI systems more interpretable and transparent, ensuring that these systems can be trusted not just for their performance but also for their ability to provide insight into their inner workings. The research presented here investigates how feature importance, logical constraints, and inherently interpretable models may contribute to this goal.

The work by Kolpaczki [10] addresses one of the most pressing issues in interpretability: understanding which features contribute most to a model's prediction. In many machine learning models, particularly those involving high-dimensional data, it is critical to assign importance scores to features to understand the model's behavior. These importance scores are not only crucial for interpretability but also serve practical purposes, such as feature selection, which can reduce the complexity of data and models. Kolpaczki [10] addresses the challenge of computational complexity and performance. Via empirical evaluation, they demonstrate an advantage of stratification methods, which may be attributed to the influence of feature subset size on overall correlation. The significance of this work lies in its practical application: by improving how we quantify feature importance efficiently, Kolpaczki [10] helps to make machine learning models more transparent.

A key challenge to interpretability is the "black-box" nature of deep neural networks. Despite their impressive performance across numerous domains, the complexity of neural networks often renders them opaque, making it difficult to interpret or trust their predictions [11]. This opacity poses a barrier when deploying these models in high-stakes environments such as healthcare, finance, or autonomous systems, where understanding the rationale behind a decision is just as important as the decision itself. Lutz and Neider [12] propose a framework that uses Linear Temporal Logic to create inherently interpretable machine learning models. By its very nature, Linear Temporal Logic provides a transparent and structured way of expressing temporal relationships and dependencies within data. Corresponding formulas are intuitive for human experts and can be easily validated against known rules or domain knowledge, ensuring that the models appear trustworthy in practical settings. At the same time, the approach presented by Lutz and Neider [12] offers remarkable flexibility, as it can be applied in both supervised and unsupervised learning settings, and it can be adapted to handle noisy data and incorporate expert knowledge. Thus, their work demonstrates how inherently interpretable models can bridge the gap between high-performing AI and trustworthiness.

In a similar vein, the research plan proposed by Katzke et al. [13] explores how deep learning models can benefit from the incorporation of logical constraints. They outline a strategy for extending existing frameworks to integrate constraints grounded in foundational prior knowledge. Ultimately, this approach may enable the model to operate with formal guarantees regarding its behavior. Thus, by embedding logical rules within the model, their approach aims to not only enhance model performance but also ensure that essential properties are preserved throughout the inference process. The significance of this work lies in its ability to provide formal guarantees, which is an essential feature in applications where the reliability and trustworthiness of AI systems are critical.

While logical constraints provide a structured way to enhance the reliability of deep learning models, another promising approach to interpretability lies in concept extraction methods. Traditionally applied to image models, Holzapfel et al. [14][1] extend these techniques to time series models and demonstrate the practical insights these methods offer. By analyzing the extracted concepts, their work reveals which features the model relies on for its predictions, as well as those contributing to errors. This level of analysis provides users with valuable insights into the biases within the model or dataset. The preliminary results presented by Holzapfel et al. [14] confirm that the adapted algorithm can successfully identify meaningful features in time series data, making these concepts critical for enhancing the interpretability of the model. Thus, their work promises to directly advance trustworthy AI by allowing users to identify and address potential biases or misleading features in critical fields like healthcare, finance, or any domain that relies on complex time series data.

Collectively, the presented contributions highlight the growing importance of interpretability and transparency in AI systems, particularly as these systems become more complex and integral to decision-making in high-stakes environments. They underscore the necessity of building AI systems that can explain their decisions, adhere to logical principles, and be scrutinized by human users. Interpretability and transparency are not just desirable features—they are essential for building trust in AI systems, especially as AI continues to evolve and take on more significant roles in society.

4. Reliability and Robustness

It is essential for a system to maintain consistent performance across varying conditions. AI systems specifically must be resilient enough to manage unexpected inputs, noisy data, and novel scenarios without failure. The contributions in this section are dedicated to designing AI models and systems that not only deliver high performance reliably, but are also robust enough to handle the complexities and uncertainties of real-world environments. The presented research approaches reliability and robustness from different perspectives, including active learning, physical sensor integration, decentralized robotic control, and bias correction in neural networks.

Zelba et al. [15] introduce a system designed to monitor technical processes and detect anomalies with minimal training data. Their COMETH system leverages active learning, a technique where the system efficiently queries the most informative data points to improve model performance, significantly reducing the amount of data needed for training. This approach is particularly important in real-world industrial applications, such as heating, ventilation, or air conditioning systems, and industrial machinery, where collecting large volumes of labeled training data is often impractical. A key strength of COMETH lies in its capacity to integrate feedback into its learning process, thereby enhancing the reliability and robustness of anomaly detection. Moreover, Zelba et al. [15] introduce an intriguing extension by integrating large language models (LLMs), adding another layer of robustness by incorporating context-aware insights that allow the system to provide more specific and actionable recommendations to users. This fusion of machine learning techniques

[1] Holzapfel et al. [14] were awarded with the 1st Place Best Poster Award at the DataNinja sAIOnARA 2024 Conference, Bielefeld.

underscores the importance of adaptability and feedback in creating AI systems that are both reliable and trustworthy.

While anomaly detection as presented by Zelba et al. [15] tackles issues of system reliability in real-time monitoring, Krebs et al. [16] shift the focus to long-term process improvement in battery manufacturing, using virtual measurements to balance efficiency and quality assurance. Virtual measurements reduce testing efforts while maintaining high standards of quality. In the presented contribution, Krebs et al. [16] identify the main requirements to facilitate trustworthy virtual measurements even for complex process chains found in battery manufacturing, with the goal of making virtual data measurements as robust as physical tests.

Shifting from industrial applications to healthcare, Grimmelsmann et al. [17] and Vieth [18] both focus on improving predictive accuracy and sensor-based systems for enhancing human physical abilities. Grimmelsmann et al. [17] expand the use of AI in biomechanics through the development of exoskeletons and the prediction of limb movements. Relying on surface electromyography signals, their study trains virtual sensors to predict the movement of deep muscles, providing an intuitive method for controlling exoskeletons in rehabilitation and physical enhancement settings. Thus, this approach paves the way for exoskeletons that can function effectively across a range of scenarios, making them more resilient to variations in muscle activity and improving the overall stability and reliability of AI-driven biomechanical systems. Vieth [18] focuses on improving the placement of pressure sensors in a smart shoe insole by exploring nonlinear modeling techniques. In previous work [19], they used a linear model to predict weight distribution on the foot and leg based on data from pressure sensors. While the linear model was effective, the current study demonstrates that the number of sensors can be reduced with nonlinear modeling techniques while maintaining robustness. Thus, their work reflects core aspects of robustness: maintaining reliable performance with fewer resources and demonstrating resilience in the face of reduced sensor input. In applications like smart insoles for healthcare, where consistent and accurate monitoring of weight distribution is crucial for diagnosing and treating mobility issues, this enhancement is critical.

Building on the theme of physical movement, we turn from human biomechanics to robotic locomotion. Hermes et al. [20] address the challenges of legged locomotion in robots, which is inherently more complex than wheeled or tracked movement due to the intricate coordination required between legs. Inspired by the biological coordination seen in insects, Hermes et al. [20] propose a decentralized control system for hexapod robots, using a Graph Neural Network to model inter-leg coordination. The decentralization of control mimics biological systems, where different parts of the body communicate and adapt locally, ensuring robustness in navigating difficult terrains. This decentralized approach enhances the robustness of the robot by allowing it to make local adjustments to its leg movements based on the specific challenges of the terrain. Preliminary results demonstrate how this gives rise to a stable tripod gait, highlighting how decentralized and flexible coordination, inspired by biology, produces robust and reliable solutions, particularly in autonomous robotics.

Just as Hermes et al. [20] seek to optimize control and coordination in robots, Posada-Moreno and Trimpe [21] focus on optimizing model performance by correcting biases and aligning predictions with expert knowledge. Concept extraction is a useful approach in explainable AI to identify model biases that may affect transparency and fairness (see also [14], this issue, for an extension to time-series data). Posada-Moreno and Trimpe [21] extend this idea by introducing Concept Regularization, a method that goes beyond simply identifying biases by embedding a regularization term during the retraining process, adjusting the model's sensitivities based on feedback from human experts. Importantly, the proposed Concept Regularization method addresses the robustness of AI systems by ensuring that identified biases do not compromise model performance. Thus, Posada-Moreno and Trimpe [21] demonstrate the critical role of feedback loops in enhancing the robustness and reliability of AI systems, particularly when addressing complex issues like bias.

From this extension of concept extraction, we move to an extension of the Multi-Armed Bandit framework in reinforcement learning, i.e., the Dueling Bandit problem. While traditional Dueling Bandit algorithms assume immediate feedback on which option performs better after the learner chooses two options, Brandt et al. [22] introduce a strategy that can start a new duel even if the feedback from the previous duel has not yet been observed. They demonstrate that this approach significantly improves the time efficiency of the algorithm by balancing the expected information gain and feedback delay. Considering that feedback delays are common on many dynamic environments, the contribution by Brandt et al. [22] offers a reliable solution for real-time applications, such as online recommendation systems and adaptive learning environments.

Finally, Moriz et al. [23] address the robustness of deep learning models trained on synthetic data and seeks to identify and mitigate factors contributing to performance gaps in real-world applications. With the increasing use of synthetic data for model training, the sim-to-real gap remains a significant challenge for broader adoption. Focusing on the impact of texture variation in synthetic validation sets for object detection, the work presented by Moriz et al. [23] concludes that texture properties alone do not fully account for the observed performance gap between synthetic and real datasets. This finding suggests that other factors, such as object size or illumination, may play a more critical role and warrant further investigation. By improving the reliability of synthetic training data, this research enhances the robustness of AI systems when deployed in real environments.

Overall, the contributions in this section illustrate the importance of building AI systems that are both reliable and robust in the face of real-world challenges. They offer solutions on how AI systems may be made more reliable, robust, and thus adaptable even in the face of real-world challenges

5. Bridging Themes

The development of trustworthy AI encompasses various domains, from performance optimization to the ethical considerations of human-AI interactions. However, achieving truly trustworthy AI systems requires integrating multiple perspectives and methodologies, building connections between diverse fields such as machine learning, human-computer interaction, and real-world applications. The research in this section highlight these interdisciplinary connections, offering innovative approaches that bridge the gap between technical advancements, user-centric designs, and transparency.

The contribution presented by Fischer and Bunse [24] lays the foundation for assessing the reliability and accountability of AI systems by presenting a "Sustainable and Trustworthy Reporting" (STREP) framework. The STREP framework presents a structured method for reporting performance indicators of AI systems, combining empirical data, theoretical algorithmic properties, and evaluation context. While AI systems are often evaluated in controlled environments, real-world applications introduce variables that can influence performance and trustworthiness. STREP contributed to understanding and communicating these nuances, thus enabling stakeholders to assess AI systems in a more comprehensive and transparent manner. This connection between data, knowledge, and context reflects the need for multidisciplinary approaches to create AI systems that are both technically robust and socially responsible.

Improving the performance and explainability of reinforcement learning by incorporating cognitive-inspired methods, Lange et al. [25] introduce techniques inspired by human cognition, such as enhanced state representations and causal reasoning. While traditional reinforcement learning systems rely heavily on trial-and-error learning, this extension provides an intriguing approach that may allow agents to reason about the past and future, explore hypothetical options, or learn from mistakes. Thus, the work presented by Lange et al. [25] demonstrates how AI systems can be designed to align better with human cognitive processes, creating models that are not only efficient but also more understandable and trustworthy.

The nuanced discussions in Belosevic and Buschmeier [26] and Schmidt and Cimiano [27] underscore the role of AI in sensitive settings, highlighting the need for reliable and understandable AI applications. Belosevic and Buschmeier [26] take a linguistic-based approach to understanding trust in interactions between humans and LLM-based chatbots, focusing on trust calibration, i.e., the process by which users adjust their trust in AI systems based on their interactions. Addressing the need for enhancing AI literacy and preventing overtrust or misuse of chatbot-provided information, particularly in educational contexts, their research offers critical insights into how users can be better supported in managing their trust levels. As a first step, Belosevic and Buschmeier [26] present a case study on the route to conversational agents capable of delivering proactive dialogue actions that assist students in controlling their trust in chatbot responses. This study begins by identifying the linguistic units perceived as trust cues in human-chatbot interactions, providing a foundation for recommendations on designing effective communication strategies that promote trustworthy AI. Schmidt and Cimiano [27] demonstrate how AI systems can leverage real-world data to impact healthcare outcomes.

Their study focuses on extracting information from online healthcare forums to automate the process of answering quality-of-life questionnaires for cancer patients. The results of their ongoing work could significantly reduce the burden on both patients and healthcare professionals, making it easier to assess patient well-being in real-time.

Bridging the gap between mathematical modeling and real-world applications, the contribution by Besginow et al. [28][2] leverages Gaussian processes for equation discovery in dynamical systems. Their method aims to uncover the differential equations that govern the physical processes observed in time series data. In contrast to traditional time series analysis with Gaussian processes, their approach seeks to identify the most frequently occurring differential equations within the data, offering a more refined understanding of the underlying system dynamics. Thus, their work provides a deeper understanding of complex, multi-component systems, contributing to the development of robust AI systems that can more accurately model and predict the behaviors of dynamical systems.

Finally, Sanaullah et al. [29][3] offer a comprehensive review of the evolution and optimization of artificial neural networks. Their paper examines improvements in network architecture, training algorithms, optimization techniques, and hardware acceleration, all of which have significantly enhanced the capabilities of neural networks. By analyzing the progression of deep learning models, such as convolutional neural networks and spiking neural networks, the review highlights their impact on critical areas like natural language processing and computer vision. Their contribution categorizes neural networks into distinct generations, emphasizing key milestones that have improved performance, scalability, and transparency—essential factors for building AI systems that are not only powerful but also reliable, interpretable, and aligned with human values, thus contributing to the broader goal of trustworthy AI.

6. Conclusions

The development of trustworthy AI is a multidimensional challenge, requiring attention to fairness, interpretability, reliability, and the integration of interdisciplinary approaches. Throughout this collection, various contributions highlight these essential facets, showcasing the importance of creating AI systems that are not only high-performing but also aligned with human values.

In the realm of fairness and ethics, research that explores fairness in rankings [1] or AI-led resource allocation in leadership roles [3] emphasizes the need for ethical decision-making frameworks in AI. These contributions underscore the importance of ensuring equitable outcomes and maintaining user trust in AI-driven processes. Interpretability and transparency are considered to be central to fostering trust [9]. The analysis of feature importance using Shapley values [10] and the development of inherently interpretable models [12] pave the way to empower users to make educated and well-informed evaluations of AI systems, particularly in high-stakes scenarios. Without transparency, even the most technically advanced AI systems risk losing credibility. Reliability and robustness are critical for AI systems operating in dynamic and unpredictable real-world environments. Advances in anomaly detection [15] or sensor optimization [18] highlight the necessity of building systems that maintain consistent performance under varying conditions, ensuring that AI systems can handle complexity without sacrificing reliability. Finally, it is important to recognize that achieving trustworthy AI requires more than just technical innovation: It is equally critical to integrating diverse perspectives and interdisciplinary approaches throughout the development, evaluation, and deployment of AI systems.

To conclude, the collective contributions to the DataNinja 2024 sAIOnARA Conference illustrate the multifaceted nature of building trustworthy AI. Technical performance must be complemented by ethical considerations, user transparency, and robust system design to ensure AI systems meet the needs of human users and society.

Acknowledgments

This research was supported by the research training group “Dataninja” (Trustworthy AI for Seamless Problem Solving: Next Generation Intelligence Joins

[2] Besginow et al. [28] were awarded with the 3rd Place Best Poster Award at the DataNinja sAIOnARA 2024 Conference, Bielefeld.

[3] Sanaullah et al. [29] were awarded with the 2nd Place Best Poster Award at the DataNinja sAIOnARA 2024 Conference, Bielefeld.

Robust Data Analysis) funded by the German federal state of North Rhine-Westphalia.

References

[1] Chiara Balestra. Is it possible to characterize group fairness in rankings in terms of individual fairness and diversity? In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 10–12, 2024. doi: 10.11576/dataninja-1157.

[2] Lloyd S Shapley. A value for n-person games. *Contribution to the Theory of Games*, 2, 1953.

[3] Paul Hellwig and Günter W. Maier. Distributive justice of resource allocation through artificial intelligence. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 75–77, 2024. doi: 10.11576/dataninja-1177.

[4] Edna B Foa and Uriel G Foa. Resource theory of social exchange. *Handbook of social resource theory: Theoretical extensions, empirical insights, and social applications*, pages 15–32, 2012.

[5] Clarissa Sabrina Arlinghaus and Günter W. Maier. Feeling socially excluded when working with robots. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 36–38, 2024. doi: 10.11576/dataninja-1165.

[6] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78, 1994.

[7] Sanaullah, Hasina Attaullah, and Thorsten Jungeblut. Trade-offs between privacy and performance in encrypted dataset using machine learning models. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 39–42, 2024. doi: 10.11576/dataninja-1166.

[8] Ali Norouzifar and Wil van der Aalst. Leveraging desirable and undesirable event logs in process mining tasks. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 32–35, 2024. doi: 10.11576/dataninja-1164.

[9] Umang Bhatt, Pradeep Ravikumar, et al. Building human-machine trust via interpretability. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9919–9920, 2019.

[10] Patrick Kolpaczki. Comparing shapley value approximation methods for unsupervised feature importance. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 13–15, 2024. doi: 10.11576/dataninja-1158.

[11] Feng-Lei Fan, Jinjun Xiong, Mengzhou Li, and Ge Wang. On interpretability of artificial neural networks: A survey. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 5(6):741–760, 2021.

[12] Simon Lutz and Daniel Neider. Interpretable machine learning via linear temporal logic. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 72–74, 2024. doi: 10.11576/dataninja-1176.

[13] Tim Katzke, Simon Lutz, Emmanuel Müller, and Daniel Neider. Provable guarantees for deep learning-based anomaly detection through logical constraints. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 65–68, 2024. doi: 10.11576/dataninja-1174.

[14] Antonia Holzapfel, Andres Felipe Posada-Moreno, and Sebastian Trimpe. Concept extraction for time series with eclad. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 78–80, 2024. doi: 10.11576/dataninja-1178.

[15] Franziska Zelba, Stefanie Hittmeyer, and Gesa Benndorf. Cometh—an active learning approach enhanced with large language models. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and

Achievements for Reliable Approaches, pages 23–25, 2024. doi: 10.11576/dataninja-1161.

[16] Lukas Krebs, Tobias Müller, and Robert H. Schmitt. Trustworthy virtual measurements in battery manufacturing. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 81–83, 2024. doi: 10.11576/dataninja-1179.

[17] Nils Grimmelsmann, Malte Mechtenberg, Markus Vieth, Barbara Hammer, and Axel Schneider. Prediction of intermuscular co-contraction based on the semg of only one muscle with the same biomechanical direction of action. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 47–49, 2024. doi: 10.11576/dataninja-1168.

[18] Markus Vieth. Nonlinear prediction in a smart shoe insole. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 50–52, 2024. doi: 10.11576/dataninja-1169.

[19] Markus Vieth, Nils Grimmelsmann, Axel Schneider, and Barbara Hammer. Efficient sensor selection for individualized prediction based on biosignals. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 326–337. Springer, 2022.

[20] Luca Hermes, Barbara Hammer, and Malte Schilling. Bioinspired decentralized hexapod control with a graph neural network. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 53–55, 2024. doi: 10.11576/dataninja-1170.

[21] Andres Felipe Posada-Moreno and Sebastian Trimpe. Closing the loop with concept regularization. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 62–64, 2024. doi: 10.11576/dataninja-1173.

[22] Jasmin Brandt, Björn Haddenhorst, Viktor Bengs, and Eyke Hüllermeier. Dueling bandits with delayed feedback. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 29–31, 2024. doi: 10.11576/dataninja-1163.

[23] Alexander Moriz, Dominik Wolfschläger, and Robert H. Schmitt. Study on the influence of texture variation on the validation performance of a synthetically trained object detector. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 69–71, 2024. doi: 10.11576/dataninja-1175.

[24] Raphael Fischer and Mirko Bunse. Improving trust in ai through sustainable and trustworthy reporting. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 56–57, 2024. doi: 10.11576/dataninja-1171.

[25] Moritz Lange, Raphael C. Engelhardt, Wolfgang Konen, and Laurenz Wiskott. Beyond trial and error in reinforcement learning. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 58–61, 2024. doi: 10.11576/dataninja-1172.

[26] Milena Belosevic and Hendrik Buschmeier. Linguistic-based reflection on trust calibration in conversations with llm-based chatbots. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 19–22, 2024. doi: 10.11576/dataninja-1160.

[27] David M. Schmidt and Philipp Cimiano. Question answering from healthcare fora. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 16–18, 2024. doi: 10.11576/dataninja-1159.

[28] Andreas Besginow, Jan David Hüwel, Markus Lange-Hegermann, and Christian Beecks. Finding commonalities in dynamical systems with gaussian processes. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI:*

Opportunities, Innovation, and Achievements for Reliable Approaches, pages 26–28, 2024. doi: 10.11576/dataninja-1162.

[29] Sanaullah, Shamini Koravuna, Ulrich Rückert, and Thorsten Jungeblut. Advancements in neural network generations. In *DataNinja sAIOnARA 2024 Conference on Shaping Trustworthy AI: Opportunities, Innovation, and Achievements for Reliable Approaches*, pages 43–46, 2024. doi: 10.11576/dataninja-1167.

DOI: 10.11576/dataninja-1157

Is it Possible to Characterize Group Fairness in Rankings in Terms of Individual Fairness and Diversity?

Chiara Balestra CHIARA.BALESTRA@CS.TU-DORTMUND.DE
TU Dortmund, Germany

Abstract

Rankings are ever-present in everyday life. Examples are the results of personalized recommendations and web search queries. Rankings can result from an algorithm, importance scores and human-based rankings of items. Till we are not concerned with societal applications, the "fairness" of the ranking is often irrelevant; however, problems appear when switching from depersonalized items to individuals. Then, suddenly, fairness becomes an issue.

We investigate the relationships among group fairness, individual fairness, diversity, and Shapley values. Far from being a comprehensive survey of fairness-related papers or proposing a new method, we want to raise awareness of the chaos we are trying to navigate and propose some new research direction we are trying to follow.

Keywords: Shapley values, individual fairness, group fairness, diversity.

1. Fairness and Shapley values

We start from where it "all" started. Fairness in machine learning is a relatively new branch; the necessity to study algorithms from a fairness perspective derives from the unpleasant discovery that some algorithms, implemented in critical contexts, were being racist. Implementing machine learning algorithms for risk assessment [1] by financial institutions, training image classification models on biased data, and selecting candidates for job positions represent a few of the critical societal applications where fairness is essential; different groups of individuals need to be fairly and "equally" treated.

But what is "fairness"? How can we say that a prediction, a ranking, or an algorithm is fair? We start by introducing "individual fairness", and particularly by introducing Shapley values, said "fair" scores in Cooperative Game Theory [2]. A *cooperative game* is a pair $(\mathcal{N}, f)$ where $\mathcal{N}$ is a finite set of players $\mathcal{N} = \{1, \dots, N\}$ and f is a function over the power set of players $\mathcal{P}(\mathcal{N})$, i.e., $f : \mathcal{P}(\mathcal{N}) \mapsto \mathbb{R}$. f is the *value function* of the game; the role of the value function is to assign to sets of players a real number, and it is usually assumed to satisfy some mathematical properties, i.e., $f(\emptyset) = 0$, it is "non-negative" and it is "monotone". Under the monotonicity assumption, the *grand coalition* $\mathcal{N}$ is the set assuming the maximum of the value function f.

Shapley values assign to each player his worth in the game $(\mathcal{N}, f)$, their values sum up to $f(\mathcal{N})$ and they are "fair" concerning the value brought by each player to the coalitions. The Shapley value of player i is formally defined as

$$\phi_f(i) = \sum_{\mathcal{A} \subseteq \mathcal{N} \setminus i} \frac{1}{N\binom{N-1}{|\mathcal{A}|}} \left[f(\mathcal{A} \cup i) - f(\mathcal{A}) \right]. \qquad (1)$$

Shapley values derive their popularity from their "nice to have" properties, i.e., the *Pareto optimality*, the *dummy*, the *linearity*, and the *symmetry property* [2]. Particularly interesting for us are the *dummy property*, stating that given $i \in \mathcal{N}$ such that $f(\mathcal{A} \cup \{i\}) = f(\mathcal{A})$ for each $\mathcal{A} \subseteq \mathcal{N}$ it holds $\phi_f(i) = 0$, and the *symmetry property*, claiming that given two players $i, j \in \mathcal{N}$ such that $f(\mathcal{A} \cup \{i\}) = f(\mathcal{A} \cup \{j\})$ for each $\mathcal{A} \subseteq \mathcal{N}$ it holds $\phi_f(i) = \phi_f(j)$.

The definition of fairness in the Oxford Dictionary reads, "Fairness is the quality of treating people equally or in a way that is reasonable". The definition is quite fuzzy; nevertheless, it was needed to formalize it mathematically. This resulted in several proposals of mathematical definitions; the first distinction is between "individual" and "group fairness". Individual fairness refers to the similar treatment of "similar" individuals. Group fairness refers to the treatment of "different" groups and usually includes the ethical concerns of gender parity, race, and sexual orientation; the so-called "protected attributes" are usually defined by law and morality concerns. Individual fairness, instead, does "not necessarily" care of morality issues: the similarity among individuals can be defined with respect to "any" attribute, either protected or not by law.

2. Contradictions within fairness

The intrinsic fairness of Shapley values derives from two of their properties, i.e., the dummy and symmetry properties [2], which guarantee that two players with similar characteristics obtain similar Shapley values. On the other side, two recent works [3, 4] show how the result of these two properties is essentially a "redundancy unawareness" of the importance scores obtained through Shapley values. The concepts of "redundancy unawareness" and "individual fairness" are eventually the same. So why do we claim in some contexts that they represent an advantage and in others that they represent a disadvantage? To understand this, we need to relate it with the *(group) fairness* in rankings.

2.1. Fairness in rankings

Rankings are spread in any field, from everyday life to complex machine learning algorithms. Rankings are nothing more than ordered lists of elements, items, or individuals. Rankings often go hand in hand with importance scores; however, if rankings are trivial to obtain from the corresponding importance scores, the opposite does not hold.

Given the several applications in society [5–7], issues relative to the fairness of the rankings and their evaluation play an essential role. The need to explore the rankings fully, which is not fulfilled in most real-world contexts, implies that elements not ranked in the top positions suffer from low visibility; this fact is usually referred to as *position bias*, and it is particularly relevant when it affects the items belonging to different various groups in a dissimilar manner. Position biases can affect differently protected and unprotected communities and potentially propagate gender, sex, and sexual orientation biases against marginalized groups. To address the issue, one could define *group fairness* constraints, with the aim of guaranteeing the same treatments in the various groups; Singh and Joachims [8] define constraints for "fairness of exposure" in ranking outputs, and the work by Zehlike et al. [9] deals with the fair top-k ranking problem. We stated that Shapley values satisfy the "individual fairness", but this might still be in contrast with the definition of "group fairness".

2.2. Diversity and individual fairness achieve group fairness?

So far, we have introduced group fairness, individual fairness, and Shapley values. The individual fair ranking derived from Shapley values does not respect any group fairness condition; this can be easily concluded by observing that elements in groups, e.g., highly correlated groups, are similarly ranked. Furthermore, the pruning criteria proposed in [3, 4] avoid that highly correlated elements are similarly ranked; in other words, the pruning criteria include "diversity" in the rankings. Diversity [10] was introduced in Recommender Systems and refers to the property of the recommendations to propose items that are new to the user and not "too similar" to the already seen elements in various parts of the ranking.

We claim that a combination of "diversity" and "individual fairness" in importance scores can induce a ranking that respects the "group fairness" property. The claim can be potentially generalized to any importance scores, independently of their derivation and it is not limited to Shapley values importance scores. Our claim is supported by the works by Balestra et al. [3, 4]; furthermore, the concern is becoming relevant to the community [11], where the authors study the connection between fairness and novelty in RS. Although some preliminary experiments showed the connection between group fairness and the simultaneous satisfaction of diversity and individual fairness in Shapley values, additional analysis must be performed to prove it more generally. Therefore, our claim is still far from being generally proven.

3. Conclusions

We introduced the relationship between Shapley values, well known for providing individual fair rankings, and the lack of diversity in the provided rankings. We propose a new theory, claiming that the pruning criteria proposed in [3, 4] can be interpreted as adding "diversity" in the rankings; more generally, we claim that under some specific (and still under study) conditions, the equation "diversity" plus "individual fairness" equals "group fairness" holds.

Acknowledgments

This research was supported by the research training group Dataninja funded by the German federal state of North Rhine-Westphalia.

References

[1] Jonathan N Crook, David B Edelman, and Lyn C Thomas. Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183, 2007.

[2] L. S. Shapley. A value for n-person games. In *Contributions to the Theory of Games*, volume II. 1953.

[3] Chiara Balestra, Florian Huber, Andreas Mayr, and Emmanuel Müller. Unsupervised features ranking via coalitional game theory for categorical data. In *DaWaK*, 2022.

[4] Chiara Balestra, Carlo Maj, Emmanuel Müller, and Andreas Mayr. Redundancy-aware unsupervised ranking based on game theory: Ranking pathways in collections of gene sets. *Plos one*, 18, 2023.

[5] Peter Emerson. The original borda count and partial voting. *Social Choice and Welfare*, 40, 2013.

[6] Christian List. Social choice theory. 2013.

[7] Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. Rank aggregation methods for the web. In *WWW*, 2001.

[8] Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *KDD*, 2018.

[9] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. Fa*ir: A fair top-k ranking algorithm. In *CIKM*, 2017.

[10] Pablo Castells, Neil Hurley, and Saul Vargas. Novelty and diversity in recommender systems. In *Recommender systems handbook*. 2021.

[11] Yuying Zhao, Yu Wang, Yunchao Liu, Xueqi Cheng, Charu Aggarwal, and Tyler Derr. Fairness and diversity in recommender systems: a survey. *arXiv preprint arXiv:2307.04644*, 2023.

DOI: 10.11576/dataninja-1158

Comparing Shapley Value Approximation Methods for Unsupervised Feature Importance

Patrick Kolpaczki PATRICK.KOLPACZKI@UPB.DE
Paderborn University, Germany

Abstract

Assigning importance scores to features is a common approach to gain insights about a prediction model's behavior or even the data itself. Beyond explainability, such scores can also be of utility to conduct feature selection and make unlabeled high-dimensional data manageable. One way to derive scores is by adopting a game-theoretical view in which features are understood as agents that can form groups and cooperate for which they obtain a reward. Splitting the reward among the features appropriately yields the desired scores. The Shapley value is the most popular reward sharing solution. However, its exponential complexity renders it inapplicable for high-dimensional data unless an efficient approximation is available. We empirically compare selected approximation algorithms for quantifying feature importance on unlabeled data.

Keywords: Shapley values, feature importance scores, unsupervised learning

1. Unsupervised Feature Importance

The increasing complexity of machine learning models as well as dimensionality of collected data is calling for a method to make both interpretable to the human user. A universally applicable approach are additive feature explanations which divide an observed numerical effect among the available features. Choosing this effect to be explained appropriately allows to interpret each feature's share as its contribution to the behavior of interest. In particular, the Shapley value [1] has emerged as the most frequently applied scoring rule. Popular examples include the features' contributions to a model's generalization performance [2, 3] and prediction value for a selected instance [4]. In the realm of unlabeled data and absence of a prediction model, Shapley-based feature importance scores have been utilized to perform dimensionality reduction [2]. Balestra et al. [5] refined this approach by proposing a feature ranking based on Shapley values that reduces redundancy among the selected features. Aiming at preserving the information contained in the data while minimizing correlation between the selected feature subset Balestra et al. employ the total correlation of shared by all all available features of the dataset as the numerical effect to be divided. For any subset S it is given by

$$C(S) = \sum_{X \in \mathcal{S}} H(X) - H(S) \qquad (1)$$

where $H(X)$ and $H(S)$ denote the Shannon entropy of a single feature X and a set of features S respectively. This is made feasible by viewing the set of all feature values as observed realizations of a random variable.

2. Cooperative Games

A *cooperative game* is formally given by a pair $(\mathcal{N}, \nu)$ containing a finite set of *players* $\mathcal{N} = \{1, \dots, n\}$ and a *value function* $\nu : \mathcal{P}(\mathcal{N}) \to \mathbb{R}$ that assigns a real-valued *worth* to each *coalition* $S \subseteq \mathcal{N}$. This simple formalism is expressive enough to model feature subsets as coalitions that share some total correlation. The most popular solution to the question of how to divide the achieved worth $\nu(\mathcal{N})$ among all players is the *Shapley value* [1] as it is provably the only solution to fulfill certain axioms [1] that plausibly capture a notion of fairness. It assigns to each $i \in \mathcal{N}$ the share

$$\phi_i = \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \frac{1}{n \cdot \binom{n-1}{|S|}} \cdot [\nu(S \cup \{i\}) - \nu(S)] \qquad (2)$$

and can be interpreted as a weighted average of marginal contributions $\Delta_i(S) := \nu(S \cup \{i\}) - \nu(S)$. Given the context of high-dimensional data yielding large player numbers, the computational complexity caused by the exponential number of coalitions renders any attempt to exactly calculate ϕ_i futile.

3. Shapley Value Approximation

The rapid increase of the Shapley value's popularity in recent years, spanning over various machine learning fields [6] and beyond, incentivized the research on how to approximate it, facilitating its practical usage. The approximation problem consists of the task of computing precise estimates $\hat{\phi}_1, \dots, \hat{\phi}_n$ of all Shapley values with minimal resource consumption.

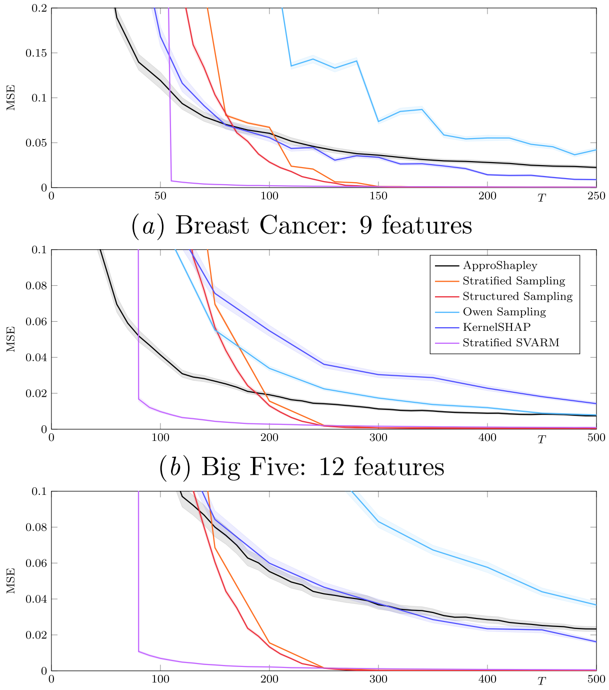
We consider the *fixed-budget setting* in which the number of times an approximation algorithm is allowed to access ν is limited by a *budget* $T \in \mathbb{N}$. This is motivated by the observation that the evaluation of large models or data poses a bottleneck, possibly even causing monetary costs when the access is provided remotely by another party. The quality of the estimates is measured by the mean squared error (MSE) averaged over all players which is to be minimized:

$$\text{MSE} := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\left(\hat{\phi}_i - \phi_i \right)^2 \right] .$$

We shortly describe selected algorithms that we use for our experiments in Section 4. The first and simplest class of approximation methods leverages the fact that ϕ_i can be interpreted as player i's expected marginal contribution. This allows to obtain a mean estimate by randomly sampling marginal contributions. Castro et al. [7] propose with *ApproShapley* an algorithm that draws random permutations of $\mathcal{N}$. It extracts a marginal contribution of each player by iterating through a permutation. Following the spirit, *Stratified Sampling* [8] partitions the population of a player's marginal contributions into strata, each containing marginal contributions to coalitions S of the same size. This technique can increase estimation quality if $|S|$ has an influence on $\Delta_i(S)$. Closely related, *Structured Sampling* [9] modifies sampled permutations such that the marginal contributions to coalitions of different sizes appear in the same frequency. Departing from the discrete sum, *Owen Sampling* [10] updates an integral representation of the Shapley value [11]. Introducing another representation, Kolpaczki et al. [12] sample with *Stratified SVARM* single coalitions instead of marginal contributions. In combination with stratification it reaches higher sample efficiency as all players' estimates are updated with each coalition. Adopting a different view, *KernelSHAP* [4] solves a weighted least squares problem, filled by randomly drawn coalitions, of which the Shapley values are the solution.

4. Empirical Evaluation

We compare the approximation quality of selected algorithms depending on the available budget T for unsupervised feature importance. In particular we use three real-world datasets: Breast Cancer, Big Five Personality Test, and FIFA 21 prepared as in [5]. A cooperative game is built from each dataset by interpreting the features as players and applying the total correlation as the corresponding coalition's worth. The approximation algorithms are run for a range of different budget values for multiple repetitions. In order to track the MSE, we calculate the Shapley values exhaustively beforehand. From Figure 1, *Stratified SVARM* emerges as significantly superior once it completes its warmup. *Stratified Sampling* and *Structured Sampling* perform on par or marginally better for higher budget ranges. The advantage of stratifying methods is likely to be caused by the impact of the feature subset size on the total correlation. In contrast, other methods including *KernelSHAP* perform clearly worse, except for *ApproShapley* displaying the lowest MSE given exteremely small budget.



(a) Breast Cancer: 9 features

(b) Big Five: 12 features

(c) FIFA 21: 12 features

Figure 1: Averaged MSE and std. error over 50 repetitions depending on available budget T.

Acknowledgments

This research was supported by the research training group Dataninja, funded by the German federal state of North Rhine-Westphalia.

References

[1] L. S. Shapley. A value for n-person games. In *Contributions to the Theory of Games (AM-28), Volume II*, pages 307–318. Princeton University Press, 1953.

[2] Shay B. Cohen, Eytan Ruppin, and Gideon Dror. Feature selection based on the shapley value. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pages 665–670, 2005.

[3] Ian Covert, Scott M. Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. In *Proceeedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[4] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceeedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 4768–4777, 2017.

[5] Chiara Balestra, Florian Huber, Andreas Mayr, and Emmanuel Müller. Unsupervised features ranking via coalitional game theory for categorical data. In *Proceedings of Big Data Analytics and Knowledge Discovery (DaWaK)*, pages 97–111, 2022.

[6] Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Oliver Kiss, Sebastian Nilsson, and Rik Sarkar. The shapley value in machine learning. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5572–5579, 2022.

[7] Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.

[8] Sasan Maleki, Long Tran-Thanh, Greg Hines, Talal Rahwan, and Alex Rogers. Bounding the estimation error of sampling-based shapley value approximation with/without stratifying. *CoRR*, abs/1306.4265, 2013.

[9] Tjeerd van Campen, Herbert Hamers, Bart Husslage, and Roy Lindelauf. A new approximation method for the shapley value applied to the wtc 9/11 terrorist attack. *Social Network Analysis and Mining*, 8(3):1–12, 2018.

[10] Ramin Okhrati and Aldo Lipani. A multilinear sampling algorithm to estimate shapley values. In *Proceedings of International Conference on Pattern Recognition (ICPR)*, pages 7992–7999, 2020.

[11] Guillermo Owen. Multilinear extensions of games. *Management Science*, 18(5):64–79, 1972. ISSN 00251909, 15265501.

[12] Patrick Kolpaczki, Viktor Bengs, Maximilian Muschalik, and Eyke Hüllermeier. Approximating the shapley value without marginal contributions. In *Proceeedings of AAAI Conference on Artificial Intelligence (AAAI)*, pages 13246–13255, 2024.

DOI: 10.11576/dataninja-1159

Question Answering from Healthcare Fora

David M. Schmidt DAVID.SCHMIDT@UNI-BIELEFELD.DE
Bielefeld University, Germany

Philipp Cimiano CIMIANO@CIT-EC.UNI-BIELEFELD.DE
Bielefeld University, Germany

Abstract

Assessing the quality of life of cancer patients is an important aspect of patient-focused drug development and real-world evidence generation. Specialized quality of life questionnaires exist for this purpose, and different types of cancer, such as breast cancer or lung cancer, can be assessed. However, conducting these surveys is a time-consuming process for both patients and clinical staff. At the same time, many patients discuss their experiences with and symptoms of their specific diseases in online healthcare fora. These forum posts may contain information that could be used to answer quality of life questions. Our objective is to determine whether forum posts can be used to answer quality of life questionnaires and, if so, whether this process can be automated successfully.

Keywords: Question Answering, Quality of Life, Healthcare Fora

1. Introduction

In cancer treatment, survival time is not the only important factor for evaluating the success of different therapies. The quality of life of a cancer patient is also an important part of patient-focused drug development in order to ensure patients do not only live as long as possible but also as well as possible considering their respective disease. For evaluating the various aspects that affect the quality of life of cancer patients, specialized questionnaires such as those provided by the *European Organisation for Research and Treatment of Cancer (EORTC)*[1] exist.

However, those filling those questionnaires and evaluating the results is tedious and time-consuming for both patients and clinical experts, although they are of key importance for both patient-focused drug development as well as real-world evidence generation.

At the same time, many cancer patients also frequently use social media and in particular specialized healthcare fora like *Inspire*[2], *Breast Cancer Now*[3] or *Macmillan Cancer Support*[4]. In these fora, patients and relatives of patients discuss how their disease affects their lives, which symptoms they are experiencing, or they support each other getting through those tough times. Many of those topics include aspects of their lives which are relevant to quality of life questions.

2. Research Questions

Considering the mayor investment of time and resources which is necessary to evaluate the quality of life of patients using standard tools like EORTC questionnaires, this observation raises the question whether at least an approximation of the results could be (automatically) extracted from those forum posts. This leads us to the following research questions:

1. *Feasibility:* Is it possible to extract answers to quality of life questions from those forum posts?
2. *Automatization:* Can we extract answers to the questionnaires automatically using AI methods? With which level of accuracy?

3. Research Plan

In order to answer those research questions, first some ground-truth data needs to be collected, consisting of a number of forum posts on the one hand as well as filled *EORTC QLQ-C30* and *EORTC QLQ-BR23* questionnaires on the other hand. We plan to conduct this data by posting a survey in one such healthcare forum. More precisely, we currently cooperate with

[1] See https://www.eortc.org/

[2] See https://www.inspire.com/

[3] See http://breastcancernow.org/

[4] See https://www.macmillan.org.uk/

Inspire to realize that data collection. Taking into account the sensitivity of that (partially medical) data, multiple steps were necessary in advance to ensure the compliance of the planned survey with ethical and data privacy standards as well as the consent of the respective patients. These steps include:

1. Plan survey structure and quality of life questionnaires to use (EORTC QLQ-C30 and EORTC QLQ-BR23) ✓
2. Ensure compliance with *General Data Protection Regulation (GDPR)* ✓
3. Get approval by the Ethics Review Board of Bielefeld University ✓
4. Contact healthcare fora about cooperation in posting and conducting the survey ✓
5. Conducting the study (ongoing)

We plan to conduct data from 100 Inspire community members as an initial dataset. In order to answer the first research question, we already started developing annotation guidelines for the data. After the data collection has been completed, the whole dataset will be annotated by three different people, followed by an evaluation whether the data in the forum posts is rich enough to approximate answers to quality of life questions from them.

If this is the case, the annotated data will be used to develop different approaches for extracting the quality of life information from those posts. As the title of this work already suggests, we plan to frame this task as a question answering problem, pointing to different parts of the input context, i.e. the forum posts, which are relevant to the asked question, i.e. the respective quality of life question.

4. Methods

In parallel to preparing the data collection, we already worked on multiple lines of research which are relevant to the second research question. The current state and relevance to the addressed problem are briefly described in the following.

4.1. Baseline Method - Information Extraction from Clinical Trials

The extraction of *PICO (Patient, Intervention, Comparison, Outcomes)* [1, 2] information from *Randomized Controlled Trials (RCTs)* is an important part of creating systematic reviews, which form the foundation of the evidence-based medicine paradigm. Our work [3–5] aims to automatically extract PICO elements in form of nested templates from RCT abstracts. These information extraction approaches could be especially well-suited to act as a baseline for the extraction of quality of life information from forum posts due to the shared medical domain and similar structure of the task.

4.2. Question Answering using *Dependency-based Underspecified Discourse REpresentation Structures (DUDES)*

DUDES [6, 7] are a compositional approach to representing meaning of words, phrases and sentences which can be used to generate SPARQL queries for textual questions given as an input. This DUDES-based approach differs a lot from recent purely machine learning or LLM-based approaches as it is an white box approach by construction. By relying both on fixed composition rules as well as neural dependency parsers etc., DUDES offer an explainable alternative combining the best of both worlds in contrast to the various black box machine learning approaches presented in recent times. This way, it also might be useful for solving the presented question answering task. This work will soon be published in a paper about the DUDES question answering approach.

5. Conclusion

All in all, automated prediction of the quality of life of cancer patients bears the potential to automate parts of the time-consuming quality of life evaluation process and this way potentially allows both clinical experts and cancer patients to focus more on the success of the corresponding therapy.

The necessary data is currently being collected and will soon allow interesting insights into the potential use of social media data for clinical purposes and automation of tedious processes.

References

[1] Connie Schardt, Martha B Adams, Thomas Owens, Sheri Keitz, and Paul Fontelo. Utilization of the pico framework to improve searching pubmed for clinical questions. *BMC medical informatics and decision making*, 7:1–6, 2007.

[2] W Scott Richardson, Mark C Wilson, Jim Nishikawa, and Robert S Hayward. The well-built clinical question: a key to evidence-based decisions. *ACP journal club*, 123(3):A12–A13, 1995.

[3] Christian Witte, David M Schmidt, and Philipp Cimiano. Comparing generative and extractive approaches to information extraction from abstracts describing randomized clinical trials. *Journal of Biomedical Semantics*, 15, 2024.

[4] David M Schmidt and Philipp Cimiano. Grammar-constrained decoding for structured information extraction with fine-tuned generative models applied to clinical trial abstracts. *Frontiers in Artificial Intelligence*, 2024. (in review).

[5] Christian Witte and Philipp Cimiano. Intra-template entity compatibility based slot-filling for clinical trial information extraction. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 178–192, 2022.

[6] Philipp Cimiano. Flexible semantic composition with dudes (short paper). In *Proceedings of the Eight International Conference on Computational Semantics*, pages 272–276, 2009.

[7] Philipp Cimiano, Christina Unger, and John McCrae. *Ontology-based interpretation of natural language*. Springer Nature, 2022.

Proceedings of the DataNinja sAIOnARA 2024 Conference 19-22 DOI: 10.11576/dataninja-1160

Linguistic-Based Reflection on Trust Calibration in Conversations with LLM-Based Chatbots

Milena Belosevic MILENA.BELOSEVIC@UNI-BIELEFELD.DE
German Linguistics, Faculty of Linguistics and Literary Studies, Bielefeld University, Bielefeld, Germany

Hendrik Buschmeier HBUSCHME@UNI-BIELEFELD.DE
Digital Linguistics Lab, Faculty of Linguistics and Literary Studies, Bielefeld University, Bielefeld, Germany

Abstract

This paper presents a linguistic approach to trust in human conversations with LLM-based chatbots. Using the concept of trust calibration [1] as a starting point, we aim to address the question of how to increase user AI literacy and prevent misuse of as well as overtrust in the information provided by LLM-based chatbots in educational contexts. We propose a linguistic-based model of trust calibration that supports users in adopting a critical perspective on trust calibration and controlling their trust level. The method combines previous studies on trust in human interaction, specifically linguistic trust cues displayed by human trustors to indicate their level of trustworthiness in naturally occurring contexts [see 2] with studies on proactive human-computer interaction [3] and the social influence of conversational agent's embodiment in educational contexts [4].

Keywords: trust calibration; linguistic trust cues; LLM-based chatbots

1. Background

Trusting information provided by large language models (LLMs) has received growing attention with the advent of LLM-based chatbots. Currently, users have high expectations regarding the desired capabilities of LLM-based chatbots. As Wang et al. [5] note, "users expect LLMs to be multifaceted, capable of accurately solving complex professional tasks, and rich in providing personalized or novel responses". Such expectations can have serious consequences, especially in educational contexts where students typically lack literacy in artificial intelligence (AI). In this paper, we propose and test a linguistic method for trust calibration in educational settings, which supports students in reflecting on and controlling their trust in a chatbot's responses, thus increasing their AI literacy, and agency [6]. Previous studies have focused on adjusting the level of trust in LLMs by training models to display confidence levels [7]. Another group of studies is concerned with developing specific prompting strategies [8], or training the models to elicit the appropriate level of trustworthiness [9] or conducting user studies to elicit users' experience and expectations regarding the desired design of LLM-based chatbots [10]. However, although "the choice of words is a vehicle for establishing trust in interpersonal online communication – regardless of whether it is written or spoken and whether the interaction is with another human or an artificial interlocutor" [11], to our knowledge, trust calibration has not been approached from linguistic perspectives [but see 12].

Starting from the hypothesis that linguistic trust cues that trustees use to indicate their trustworthiness in human interaction (e.g., markers of (un)certainty, referring to experts and numbers, lexical alignment, etc.) can be accidentally generated by LLMs as next words in particular contexts, our model complements previous research on trust in LLMs by focusing on how these cues can help students to adopt a critical stance towards the information provided by LLM-based chatbots and support them to engage in critical reflection on how to control their trust in LLM-based chatbots. To this end, we introduce a new phase of trust calibration, the 'reflection phase', which complements the existing aspects of trust calibration (e.g., overtrust, undertrust, etc. [13]) by including conversational agents in students' interaction with LLM-based chatbots.

2. Methodology

Our model is designed for the context of higher education [14, 15] and is based on the following idea: A conversational agent designed to act as a learning assistant helps students remain aware that LLMs are merely next-word predictors that should not generally be trusted in the same manner as we trust humans. By providing optional assistance in the form of (non)verbal dialogue actions, the conversational agent supports students' critical reflection and control over their trust in the chatbot's responses [4]. Suppose, for example, that the chatbot's response contains the verb 'understand', such as in 'I *understand* what you are trying to say.' This should be regarded as problematic because this linguistic unit can be associated with a set of trust cues denoting the orientation toward the trustor's needs and goals in human interaction [16, 17].

We argue that such linguistic units should not be perceived as trust cues but should instead serve to motivate and support users' critical reflection on the reliability of the response. This can be achieved by designing conversational agents to display verbal, nonverbal, and prosodic-acoustic metacommunicative expressions (e.g., distance markers/quotation marks) embedded in the agent's proactive dialogue actions [3], such as notifications, suggestions, or interventions. Using specific conversational acts (e.g., acknowledgment, see [18]), conversational agents can assist students in assessing the trustworthiness of the chatbot's response based on the verbal trustworthiness cues displayed in the response. Importantly, the agent first offers help, but students can decide whether they want the agent's assistance.

3. Case study

To design a conversational agent capable of providing proactive dialogue actions that help students control their trust in the chatbot's responses, it is first necessary to identify which linguistic units are perceived as linguistic trust cues in human interactions with LLM-based chatbots. To this end, we conducted a rating study to test the influence of one type of linguistic units that potentially influence users' trust, namely grounding acts [18]. As noted by Chiesurin et al. [19], current LLM-based dialogue systems usually guess what the user intended instead of leveraging grounding acts, which may lead to miscalibrated trust and overconfidence. We tested the following two hypotheses: (H1) The responses in the 'baseline' condition will receive lower trustworthiness ratings than in two alternative conditions (see also [20]). (H2) The perceived trustworthiness is higher in the 'anthropomorphic' than in the 'grounding act' condition (contrary to [21]).

In a within-subject design study, students ($N = 32$; 17 female, 15 male; 14 German native speakers, 12 bilingual, 6 non-native speakers of German; age: $\mathrm{M} = 25.96, \mathrm{SD} = 4, \mathrm{Mdn} = 26$) were exposed to items from these three conditions[1]. The acknowledgment and the anthropomorphic condition comprised the *Other-Acknowledgment* speech-act pattern, specifically the *inform → ackn+mrequest* pattern [22], where a student (other) presents a math task (inform) selected from the MathDial dataset [23] and a virtual tutor (ChatGPT 3.5) was prompted to respond by acknowledging the information (*ackn*). In the acknowledgment condition, the acknowledgment is verbalized by 'In Ordnung', 'Alles klar', etc. and followed by a request for clarification of some part of the information to verify understanding (*mrequest*). In the anthropomorphic condition, the math task presented by the student was followed by the tutor's acknowledgment and a follow-up question verbalized by anthropomorphic verbs (e.g., 'understand'). The baseline condition comprised the tutor's direct response to the student's task. The perceived trustworthiness of the tutor's response was measured indirectly by having participants rate the following statements: "The virtual chatbot tutor can help the student." The participants responded by selecting a value on a four-point Likert scale. Both hypotheses were confirmed by statistical analyses: The responses in the baseline condition received lower mean trustworthiness ratings (M = 1.93) than in the anthropomorphic (M = 3.23) and the acknowledgment condition (M = 2.98). As indicated by these values, the perceived trustworthiness received higher ratings in the anthropomorphic than in the acknowledgment condition (see H2). The differences between the three conditions are statistically significant (Friedman Test, $\chi = 28.79, p < 0.001$).

4. Conclusions and future work

The results obtained in the questionnaire study can be used as a starting point for providing recommendations for designing communication strategies for trustworthy AI [8, 10]. In the next step, we will use these results to test whether enriching grounding acts with nonverbal aspects affects students' perceived trust in the chatbot's responses. To this end, we plan to include a social robot in the conversation with LLM-based chatbots.

[1] See the supplementary material for study design, dataset, and results: https://doi.org/10.17605/OSF.IO/FYQ3P.

References

[1] Bonnie M. Muir. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27:527–539, 1987. doi: 10.1016/S0020-7373(87)80013-5.

[2] Mie Femø Nielsen and Ann Merrit Rikke Nielsen. *Revisiting Trustworthiness in Social Interaction*. Routledge, New York, NY, USA, 2022. doi: 10.4324/9781003280903.

[3] Matthias Kraus, Nicolas Wagner, Zoraida Callejas, and Wolfgang Minker. The role of trust in proactive conversational assistants. *IEEE Access*, 9:112821–112836, 2021. doi: 10.1109/ACCESS.2021.3103893.

[4] Daniel Leyzberg, Samuel Spaulding, Mariya Toneva, and Brian Scassellati. The physical presence of a robot tutor increases cognitive learning gains. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34, pages 1882–1887, Sapporo, Japan, 2012.

[5] Jiayin Wang, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. Understanding user experience in large language model interactions. *arXiv:2401.08329*, 2024. doi: 10.48550/arXiv.2401.08329.

[6] Don Passey, Miri Shonfeld, Lon Appleby, Miriam Judge, Toshinori Saito, and Anneke Smits. Digital agency: Empowering equity in and through education. *Technology, Knowledge and Learning*, 23:425–439, 2018. doi: 10.1007/s10758-018-9384-x.

[7] Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022. doi: 10.48550/arXiv.2012.14983.

[8] Joel Wester, Tim Schrills, Henning Pohl, and Niels van Berkel. "as an ai language model, I cannot": Investigating LLM denials of user requests. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Honululu, HI, USA, 2024. doi: 10.1145/3613904.3642135.

[9] Jingyu Zhang, Marc Marone, Tianjian Li, Benjamin Van Durme, and Daniel Khashabi. Verifiable by design: Aligning language models to quote from pre-training data. *arXiv:2404.03862*, 2024. doi: 10.48550/arXiv.2404.03862.

[10] Mateusz Dubiel, Sylvain Daronnat, and Luis A Leiva. Conversational agents trust calibration: A user-centred perspective to design. In *Proceedings of the 4th Conference on Conversational User Interfaces*, pages 1–6, Glasgow, UK, 2022. doi: 10.1145/3543829.3544518.

[11] Regina Jucks, Gesa A. Linnemann, Franziska M. Thon, and Maria Zimmermann. Trust the words: Insights into the role of language in trust building in a digitalized world. In Bernd Blöbaum, editor, *Trust and Communication in a Digitized World: Models and Concepts of Trust Research*, pages 225–237. Springer, Cham, Switzerland, 2016. doi: 10.1007/978-3-319-28059-2_13.

[12] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. "I'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, Rio de Janeiro, Brazil, 2024. doi: 10.48550/arXiv.2405.00623.

[13] Ewart J. De Visser, Marieke M. M. Peeters, Malte F. Jung, Spencer Kohn, Tyler H. Shaw, Richard Pak, and Mark A. Neerincx. Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*, 12:459–478, 2020. doi: 10.1007/s12369-019-00596-x.

[14] Melissa Donnermann, Philipp Schaper, and Birgit Lugrin. Social robots in applied settings: A long-term study on adaptive robotic tutors in higher education. *Frontiers in Robotics and AI*, 9:831633, 2022. doi: 10.3389/frobt.2022.831633.

[15] Thomas Beelen, Khiet Truong, Roeland Ordelman, Ella Velner, Vanessa Evers, and Theo Huibers. A child-friendly approach to spoken conversational search. In *Proceedings of the 2nd Workshop on Mixed-Initiative Conversational Systems (MICROS)*, Atlanta, GA, USA, 2022.

[16] Pavla Schäfer. *Linguistische Vertrauensforschung: Eine Einführung.* de Gruyter, Berlin, Germany, 2016. doi: 10.1515/9783110451863.

[17] Martha Kuhnhenn. *Glaubwürdigkeit in der politischen Kommunikation Gesprächsstile und ihre Rezeption.* UVK, Konstanz, Germany, 2014.

[18] David R. Traum and Elizabeth A. Hinkelmann. Conversation acts in task-oriented spoken dialogue. *Computational Intelligence*, 8:575–599, 1992. doi: 10.1111/j.1467-8640.1992.tb00380.x.

[19] Sabrina Chiesurin, Dimitris Dimakopoulos, Marco Antonio Sobrevilla Cabezudo, Arash Eshghi, Ioannis Papaioannou, Verena Rieser, and Ioannis Konstas. The dangers of trusting stochastic parrots: Faithfulness and trust in open-domain conversational question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 947–959, Toronto, Canada, 2023. doi: 10.18653/v1/2023.findings-acl.60.

[20] Gesa Alena Linnemann and Regina Jucks. 'Can I trust the spoken dialogue system because it uses the same words as I do?' – Influence of lexically aligned spoken dialogue systems on trustworthiness and user satisfaction. *Interacting with Computers*, 30:173–186, 2018. doi: 10.1093/iwc/iwy005.

[21] Nanna Inie, Stefania Druga, Peter Zukerman, and Emily M. Bender. From "AI" to probabilistic automation: How does anthropomorphization of technical systems descriptions influence trust? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, Rio de Janeiro, Brazil, 2024. doi: 10.48550/arXiv.2404.16047.

[22] David G. Novick and Stephen Sutton. An empirical model of acknowledgement for spoken-language systems. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 96–101, Las Cruces, NM, USA, 1994. doi: 10.3115/981732.981746.

[23] Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore, 2023. doi: 10.18653/v1/2023.findings-emnlp.372.

DOI: 10.11576/dataninja-1161

COMETH - An Active Learning Approach Enhanced With Large Language Models

Franziska Zelba FRANZISKA.ZELBA@IOSB-INA.FRAUNHOFER.DE
Fraunhofer IOSB-INA, Lemgo, Germany

Stefanie Hittmeyer STEFANIE.HITTMEYER@IOSB-INA.FRAUNHOFER.DE
Fraunhofer IOSB-INA, Lemgo, Germany

Gesa Benndorf GESA.BENNDORF@IOSB-INA.FRAUNHOFER.DE
Fraunhofer IOSB-INA, Lemgo, Germany

Abstract

We present a system for supervision of technical processes, called COMETH, which involves an active learning approach. The system is able to identify anomalies with very little training data, through an efficient feedback process. COMETH has been successfully applied in the context of heating ventilation and air conditioning systems and in industrial machinery. Here, we describe the idea of combining the time series analysis COMETH with large language models to integrate further context information and thus provide the user with specific recommendations.

Keywords: anomaly detection, active learning, large language models

1. Introduction

Appropriate maintenance of machines and industrial processes has been thoroughly discussed in literature and practice [1, 2]. An ideal maintenance scheme avoids down-time of machines and ensures high throughput while keeping maintenance efforts at a low level. With the ongoing shortage of skilled labour this aspect gains increasing attention. Moreover, significant economical benefits of adequate maintenance strategies can be observed [3].

Although today more and more sensor data from machines become available, it often remains a challenge to make use of the acquired data in automized condition-based maintenance systems [4]. This is mainly due to a lack of labelled data for the training of machine learning models [2] and eventual changes in the set up of the machines or the environment, which drive models out of their prediction range. Therefore, robust methods which require few training data and quickly adapt to changes are needed in order to establish automated and flexible maintenance strategies. Promising approaches to mitigate the labelling problem are provided by active learning methods [5, 6] and human-in-the-loop strategies [7, 8].

Here, we consider an active learning method, called COMETH which at the same time provides supplementary information about detected anomalies to the user (technician). Thus, the method can be seamlessly integrated into existing maintenance procedures and support the technician by identifying anomalous behaviour, while continuously collecting labelled data to further improve the prediction accuracy.

COMETH was proposed in [9] and first applied to HVAC (heating ventilation and air-conditioning) systems. Later, its application in an industrial context was demonstrated [10]. Here, we present the idea to extend the approach further, using large language models for the generation of useful recommendations, based on the detected anomalies and additional textual machine documentation.

2. Description of the method

COMETH is based on two machine learning methods which are applied in parallel to the same data. If both methods yield the same result, the classification is likely to be correct, i.e. the user gets a notification in case of a fault or nothing happens in case of no fault. If the two methods yield distinct results a warning is generated and the user has the possibility to give a feedback telling whether the detected anomaly is a true fault or not. This information is then used to retrain the corresponding method with the labelled data. A schematic overview of this procedure is shown in Fig. 2.

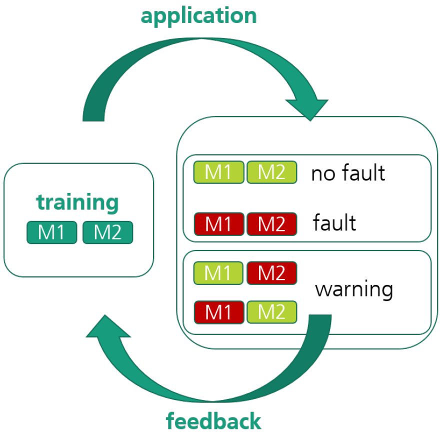


Figure 1: Schematic overview of the procedure of COMETH [10].

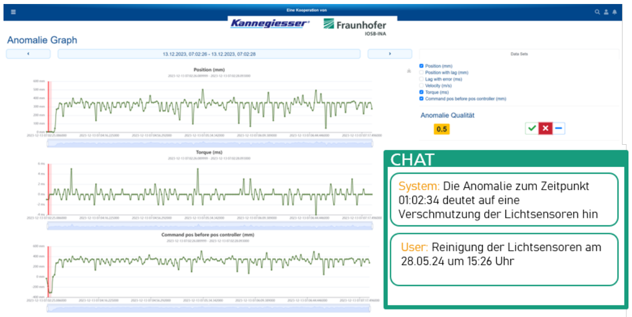
In order to ensure high classification accuracy and low feedback rates it is essential for M1 and M2 to be complementary, which can be realized by choosing an outlier detection method for M1 and a classification method for M2. This implies that M1 is trained only on fault-free data and tends to have a high sensitivity. M2, on the contrary, is trained on data from at least two distinct classes (fault-free and faulty data) and tends to have a high specificity.

The exact choice of M1 and M2 is arbitrary as long as the above-mentioned criteria are fulfilled. Typically, we choose a density-based clustering (DBSCAN) method for M1 and a decision tree method for M2, as this combination has proven to be successful in other applications [9, 11, 12].

To further support the feedback process and to enhance the trustworthyness of the results the user can, in case of a warning, be provided with additional information about the detected anomaly. More precisely, the most responsible variables for the detected anomaly are evaluated. For DBSCAN a fault identification mechanism was proposed in [11]. The resulting responsibility pattern can guide the user to reject the fault or take an adequate countermeasure.

3. Coupling to LLMs

As a further direction of current research we are planning to couple the results of the fault detection and identification with COMETH to large language models (LLMs), which provide specific information about the considered machine. This can be done via Retrieval Augmented Generation (RAG) [13] approaches where context in terms of handbooks, log book information or service documentation is integrated. As a result the user can interact with the system via a custom chat window and directly receives proposed countermeasures in case of a detected fault, based on the user manual of the machine. The idea is to further reduce the barriers to give feedback by providing an intuitive and conversational interface. Furthermore, the validity of the results can be double-checked by analysing time series data and additional textual documentation via the LLM.



System: Die Anomalie zum Zeitpunkt 01:02:34 deutet auf eine Verschmutzung der Lichtsensoren hin

User: Reinigung der Lichtsensoren am 28.05.24 um 15:26 Uhr

Figure 2: Illustration of the envisioned interactive GUI with chat functionality.

4. Conclusion

For the integration of an active learning approach into existing maintenance processes the user experience and user acceptance have to be carefully taken into account. Along these lines, further improvements have to be achieved concerning the user interface and enhanced decision support. The latter includes for example an automated generation of recommendations in case of faults or warnings.

To keep the required feedback from the user at a minimum level indirect feedback from maintenance log books or further external information sources, like handbooks can be used. Recent progress in natural language processing and the rapid advancements of large language models in many fields provide new opportunities to enhance the interaction of technicians with data analytics solutions and thus to foster the employment of active learning methods in machine services.

References

[1] John H Williams, Alan Davies, and Paul R Drake. *Condition-based maintenance and machine diagnostics*. Springer Science & Business Media, 1994.

[2] Giuseppe Ciaburro. Machine fault detection methods based on machine learning algorithms: A review. *Mathematical Biosciences and Engineering*, 19(11):11453–11490, 2022.

[3] Imad Alsyouf. The role of maintenance in improving companies' productivity and profitability. *International Journal of production economics*, 105(1):70–78, 2007.

[4] Gen Li and Jason J Jung. Deep learning for anomaly detection in multivariate time series: Approaches, applications, and challenges. *Information Fusion*, 91:93–102, 2023.

[5] Burr Settles. Active learning literature survey. 2009.

[6] Burr Settles. From theories to queries: Active learning in practice. In *Active learning and experimental design workshop in conjunction with AISTATS 2010*, pages 1–18. JMLR Workshop and Conference Proceedings, 2011.

[7] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *AI magazine*, 35 (4):105–120, 2014.

[8] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, 56(4):3005–3054, 2023.

[9] Gesa A Benndorf, Dominik Wystrcil, and Nicolas Réhault. A fault detection system based on two complementary methods and continuous updates. *IFAC-PapersOnLine*, 51(24):353–358, 2018.

[10] Franziska Zelba and Gesa Benndorf. Active learning for condition-based maintenance of industrial machinery using COMETH. *to be published*, 2024.

[11] Gesa Angelika Benndorf and Nicolas Réhault. Density-based clustering algorithm for fault detection and identification in HVAC systems. *Proceedings of BauSIM*, pages 243–250, 2016.

[12] K Chavan, N Réhault, and T Rist. Transfer learning methodology for machine learning based fault detection and diagnostics applied to building services. In *Journal of Physics: Conference Series*, volume 2600, page 082038. IOP Publishing, 2023.

[13] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wentau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

DOI: 10.11576/dataninja-1162

Finding Commonalities in Dynamical Systems with Gaussian Processes

Andreas Besginow[*] ANDREAS.BESGINOW@TH-OWL.DE
Technische Hochschule Ostwestfalen-Lippe, Germany

Jan David Hüwel[*] JAN.HUEWEL@FERNUNI-HAGEN.DE
University in Hagen, Germany

Markus Lange-Hegermann MARKUS.LANGE-HEGERMANN@TH-OWL.DE
Technische Hochschule Ostwestfalen-Lippe, Germany

Christian Beecks CHRISTIAN.BEECKS@FERNUNI-HAGEN.DE
University in Hagen, Germany

Abstract

Gaussian processes can be utilized in the area of equation discovery to identify differential equations describing the physical processes present in time series data. Furthermore, automatically constructed models can be split into components that facilitate comparisons between time series on a structural level. We consider the potential combination of these two methods and describe how they could be used to detect shared physical properties in multiple recordings of dynamical systems as time series. This approach provides insights into the underlying dynamics of the observed systems, facilitating a deeper understanding of complex processes.

Keywords: Gaussian Process, Dynamical Systems, Frequent Itemset Mining, Equation Discovery

1. Gaussian Processes (GPs)

Formally, a GP $g(x) = \mathcal{GP}(\mu(x), k(x, x'))$ defines a probability distribution over the space of functions $\mathbb{R}^d \to \mathbb{R}^\ell$, such that the outputs $g(x_i)$ at any set of inputs $x_i \in \mathbb{R}^d$ are jointly Gaussian [1]. Such a (multi-input multi-output) GP is defined by its mean function (often set to zero for the prior)

$$\mu : \mathbb{R}^d \to \mathbb{R}^\ell : x \mapsto \mathbb{E}(g(x))$$

and its (multi-input multi-output) positive semidefinite covariance function (also called kernel)

$$\begin{aligned} k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^\ell_{\succeq 0} : \\ (x, x') \mapsto \mathbb{E}((g(x) - \mu(x))(g(x') - \mu(x'))^T). \end{aligned}$$

where (X, y) with $X \in \mathbb{R}^{n \times d}$ and $y \in \mathbb{R}^{n \times \ell}$ are a dataset with n observations. The kernel defines the general form of the GP and usually contains additional hyperparameters, like a lengthscale or the length of a periodic pattern [2]. These hyperparameters directly impact the calculation of the covariance matrix K of observation locations X.

2. CATGP

A GP's kernel encodes prior assumptions about the data, like smoothness or periodicity. The choice of kernel greatly impacts the model's performance, as a strong correlation between these prior assumptions and the actual data allows the model to generate more accurate predictions and fit the data with more precision. Conversely, automatic kernel searches can find a descriptive kernel for a given dataset by evaluating the model performance of different kernels [2, 3]. For complex structures in the data, the best kernel is often a sum or product of simple kernels. In such cases, sums can be interpreted as modelling independent subprocesses which make up the dataset.

Recently, the Component Analysis in Time Series with Gaussian Processes (CATGP) has emerged as a way to use this principle for further analysis of GP models [4]. This algorithm finds commonly appearing kernel components (subkernels) in a collection of GPs by interpreting kernels as sets of such components and applying frequent itemset mining. In the previous publications about this method, the Apriori algorithm [5] was used as a basis, but the same prin-

[*] These authors contributed equally

ciple can be applied to any frequent itemset mining algorithm.

In this work we outline a potential combination of this method with kernels with inductive bias on systems of differential equations to infer knowledge from data with potential dynamical systems behaviour.

3. LODE-GPs

Consider a system of linear homogenous ordinary differential equations with constant coefficients

$$A \cdot \mathbf{f}(t) = 0 \tag{1}$$

with operator matrix $A \in \mathbb{R}[\partial_t]^{m \times n}$ determining the relationship between the smooth functions $f_i(t) \in C^\infty(\mathbb{R}, \mathbb{R})$ of $\mathbf{f}(t) = \begin{pmatrix} f_1(t) & \dots & f_n(t) \end{pmatrix}^T$. For such systems the main result of [6] holds:

Theorem 1 *(LODE-GPs) For every system as in Equation* (1) *there exists a GP g, such that the set of realizations of g is dense in the set of solutions of $A \cdot \mathbf{f}(t) = 0$.*

As the authors of [6] demonstrate, these LODE-GPs can be constructed algorithmically and are guaranteed to satisfy the original system of linear homogenous system of ordinary differential equations with constant coefficients given by $A \cdot \mathbf{f}(t) = 0$. This, combined with the previously described kernel search approach, enables users to find a fitting system of differential equations for a given dataset.

4. Proposed Method

We propose a combination of these two methods. The resulting process is depicted in Figure 1. While the system in the real world can not be directly examined to identify process components, it's reasonable to assume that these components correspond to time series components in a suitable decomposition. We achieve this decomposition by employing GP models, which adapt to structures present in the data [2], categorizing subkernels by the ordinary differential equation that they are based on, and create analyzable sets of kernels representing present structures in each dataset.

The intent of our method falls under the class of equation discovery methods, which try to discover a fitting dynamic system description for a given dataset [7–9]. Where other works make use of learning this behaviour through direct GP regression for a whole dataset, we propose to find the *most frequently occurring* differential equations as follows.

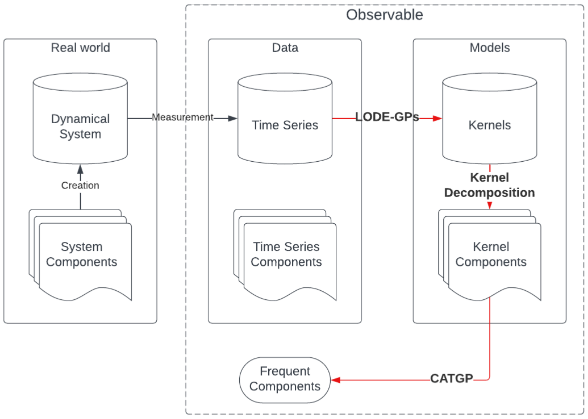


Figure 1: The correlation between different aspects in our proposed method. The red connections symbolize our contribution. Overall, process components correspond to time series components in the recorded data, which in turn correspond to kernel components of descriptive GP kernels. The figure is based on a figure in [4].

The intended objects of analysis for this method are systems, that accumulate multiple subprocesses, that follow physical equations, where the exact correlation between these subprocesses is unknown to the user. To generate insights into such physical systems, we first select descriptive LODE-GPs for time series recordings of those systems. The selected kernels are additive combinations of kernels, that correspond to systems of ordinary differential equations. Thus, each kernel can be equated to a set of kernel components, which can in turn be analysed via CATGP. This analysis identifies the most frequent types of differential equations that appear as subprocesses in the data.

Acknowledgments

This research was supported by the research training group "Dataninja" (Trustworthy AI for Seamless Problem Solving: Next Generation Intelligence Joins Robust Data Analysis) funded by the German federal state of North Rhine-Westphalia.

References

[1] Carl Edward Rasmussen and Christopher K Williams. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.

[2] David Duvenaud, James Lloyd, Roger Grosse, Joshua Tenenbaum, and Ghahramani Zoubin. Structure discovery in nonparametric regression through compositional kernel search. In *International Conference on Machine Learning*, pages 1166–1174. PMLR, 2013.

[3] Fabian Berns, Jan David Hüwel, and Christian Beecks. Automated model inference for gaussian processes: An overview of state-of-the-art methods and algorithms. *SN Comput. Sci.*, 3(4):300, 2022.

[4] Jan David Hüwel and Christian Beecks. Frequent component analysis for large time series databases with gaussian processes. In *EDBT*, pages 617–622. OpenProceedings.org, 2024.

[5] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499. Santiago, Chile, 1994.

[6] Andreas Besginow and Markus Lange-Hegermann. Constraining gaussian processes to systems of linear ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

[7] Kevin Egan, Weizhen Li, and Rui Carvalho. Automatically discovering ordinary differential equations from data with sparse regression. *Communications Physics*, 7(1), 2024.

[8] Yi-Chen Zhu, Paul Gardner, Ramon Fuentes, David J Wagg, Elizabeth J Cross, and Robert J Barthorpe. On robust equation discovery: a sparse bayesian and gaussian process approach. *ISMA Conference on Noise and Vibration Engineering and International Conference on Uncertainty in Structural Dynamics*, 2020.

[9] Yi-Chen Zhu, Paul Gardner, David J Wagg, Robert J Barthorpe, Elizabeth J Cross, and Ramon Fuentes. Robust equation discovery considering model discrepancy: A sparse bayesian and gaussian process approach. *Mechanical Systems and Signal Processing*, 168:108717, 2022.

Proceedings of the DataNinja sAIOnARA 2024 Conference 29-31 DOI: 10.11576/dataninja-1163

Dueling Bandits with Delayed Feedback

Jasmin Brandt JASMIN.BRANDT@UNI-PADERBORN.DE
University of Paderborn, Germany

Björn Haddenhorst
University of Paderborn, Germany

Viktor Bengs
LMU Munich, Germany,
Munich Center for Machine Learning, Germany

Eyke Hüllermeier
LMU Munich, Germany,
Munich Center for Machine Learning, Germany

Abstract

Dueling Bandits is a well-studied extension of the Multi-Armed Bandits problem, in which the learner must select two arms in each time step and receives a binary feedback as an outcome of the chosen duel. However, all of the existing best arm identification algorithms for the Dueling Bandits setting assume that the feedback can be observed immediately after selecting the two arms. If this is not the case, the algorithms simply do nothing and wait until the feedback of the recent duel can be observed, which is a waste of runtime. We propose an algorithm that can already start a new duel even if the previous one is not finished and thus is much more time efficient. Our arm selection strategy balances the expected information gain of the chosen duel and the expected delay until we observe the feedback. By theoretically grounded confidence bounds we can ensure that the arms we discard are not the best arms with high probability.

Keywords: Multi-Armed Bandits, Dueling Bandits, Online Learning

1. Introduction and Related Work

The Dueling Bandits or also called Preference-based Multi-Armed Bandits is a variant of the standard Multi-Armed Bandits (MAB) problem in which a learner can compare a pair of choice alternatives that are called arms in the following in a sequential manner (see [1] or [2]). Instead of a numerical reward the learner can only observe a winner information in form of a binary feedback for each selected duel. As in the standard MAB problem, we assume this feedback to be stochastic. The goal of the learner is to identify the ”best” arm as fast as possible.

Dueling Bandits algorithms were sucessfully applied to other settings in which the goal is to find the best among different choice alternatives by sequential comparisons. A classical example is the Algorithm Configuration (AC) setting in which we want to find the best parameter configuration for a given target algorithm by repeatedly racing two of them against each other. However, these target algorithms are often complex and need a long runtime until we can observe which parameter configuration performs best. All of the existing Dueling Bandits algorithms wait paitiently until the winner feedback is observed and do nothing in between. By this, a huge part of the overall runtime of the algorithm is wasted by just waiting for the observation. While there are some existing MAB methods that can deal with such a delayed feedback like [3], this setting is not studied until now for the Dueling Bandits case.

2. Problem Formulation

Assume that we have given a finite set of m choice alternatives that we denote by their indices $[m] = \{1, \dots, m\}$ and a finite number of allowed parallel duels $K \in \mathbb{N}_{>0}$. In each time step $t \in \mathbb{N}$ the learner has to choose a new duel $S_t = (i, j)$ for $i, j \in [m]$ if the number of active duels has not reached the number of allowed duels K yet. For each duel, the environment creates a pair $(\tau_t, \omega_t) \sim \mathbb{P}_{\tau,\omega}(\cdot|S_t)$ of a delay τ_t and winner ω_t of the selected duel. The learner can observe the winner $\omega_t = 1_{\{i \succ j\}}$ at time step $s = t + \tau_t$.

The goal is to identify the "best" arm as fast as possible, where fast as possible means the least wall-clock time here. The best arm ist defined as follows.

Definition 1 (Condorcet Winner) *We call arm $i^* \in [m]$ the Condorcet Winner (CW) iff $\mathbb{P}(i^* \succ j) > \mathbb{P}(j \succ i^*)$ for all $j \in [m]$ with $j \neq i^*$.*

3. Algorithm

To identify the CW in the Dueling Bandits problem with delayed feedback, we propose an algorithm that chooses the duels in each time step according to a trade-off of the expected information gain and the expected delay to observe the feedback. This is done by choosing the duel with the minimal ratio between the average observed delay for this duel and the gap between the winning probabilities of the arms. If this gap is huge, the chance is high that we can eliminate the worse arm soon. Let $\hat{\mathbb{P}}$ denote in the following the empirical probability, then we can solve the problem as given in the pseudo-code in algorithm 1.

For the theoretical derivation of the confidence bounds used in algorithm 2, we need the following assumptions.

Algorithm 1: Best arm identification in Dueling Bandits with Delay

```
Input: confidence $1-\gamma$, set of arms $[m]$,
       number of parallel slots $K$
Output: the CW of $[m]$
Initialize remaining winner candidates $R_1 = [m]$,
  epoch $e = 1$
while $|R_e| > 1$ do
    Divide $R_e$ in duels $\{S_1, \dots, S_{|R_e|/2}\}$
    $E_e \leftarrow$ ELIMINATE($\{S_1, \dots, S_{|R_e|/2}\}$)
    $R_{e+1} \leftarrow R_e \backslash E_e$
    $e \leftarrow e + 1$
end
Return remaining arm in $R_e$
```

Assumption 2 *(1) A CW exists.*
(2) No ties in duels are allowed and we have for each duel set $\{i, j\} \subset [m]$ that $|\mathbb{P}(i \succ j) - \mathbb{P}(j \succ i)| \geq h$.
(3) The delays are upper bounded by $\tau \leq b$.

With this, we can derive the following confidence bounds which proofs are beyond the scope of this extended abstract.

Algorithm 2: Eliminate

```
Input: set of duels $\mathcal{S} = \{S_1, \dots, S_n\}$
Output: set $E$ of arms to eliminate
Initialize arms to eliminate $E = \emptyset$, active duels
  $A = \emptyset$, time step $t = 0$
while $E = \emptyset$ do
    for each set $S = \{i, j\} \in \mathcal{S}$ do
        estimate winning probability gap
          $\hat{\Delta}_t(S) = |\hat{\mathbb{P}}_t(i \succ j) - \hat{\mathbb{P}}_t(j \succ i)|$
        estimate average observed delay $\hat{\tau}_t(S)$
        estimate confidence bound $c_t(S)$
    end
    if number of active duels $|A| < K$ then
        Play duel $S_t = \text{argmin}_{S \in \mathcal{S}} \frac{\hat{\tau}_t(S)}{\hat{\Delta}_t^2(S)} + c_t(S)$
        add $S_t$ to set of active duels $A = A \cup S_t$
    end
    for each set in active duels $S \in A$ do
        Possibly observe feedback $(\tau_S, w_S)$
        if feedback for S is observed then
            remove from active duels $A = A \backslash S$
            Update arms to eliminate
              $E = E \cup \{i \in S | \hat{\Delta}_t(S) \geq c_t^\delta(S)\}$
        end
    end
end
```

Theorem 3 (Confidence bound) *For the confidence bound $c_t(S) = \sqrt{\frac{2b^2}{h^2 t} ln\left(\frac{t}{4}\right)}$, we have $\mathbb{P}\left(\left|\frac{\tau(S)}{\Delta^2(S)} - \frac{\hat{\tau}_t(S)}{\hat{\Delta}_t^2(S)}\right| \geq c_t(S)\right) \leq t^{-1}$.*

Theorem 4 (Eliminated arms) *For $c_t^\delta(S) = \max\left\{3h, \sqrt{-\frac{9t}{2} ln\left(\frac{\delta}{4}\right)}\right\}$, we eliminate a wrong arm only with probability δ in algorithm 2.*

4. Conclusion

We introduced the dueling bandits with delayed feedback problem and to the best of our knowledge are the first ones who study this problem. Our proposed algorithm for the best arm identification is guaranteed to only discard good arms with low probability.

References

[1] Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits

problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

[2] Róbert Busa-Fekete, Eyke Hüllermeier, and Adil El Mesaoudi-Paul. Preference-based online learning with dueling bandits: A survey. *CoRR*, abs/1807.11398, 2018.

[3] Tal Lancewicki, Shahar Segal, Tomer Koren, and Y. Mansour. Stochastic multi-armed bandits with unrestricted delay distributions. In *International Conference on Machine Learning*, 2021.

DOI: 10.11576/dataninja-1164

Leveraging Desirable and Undesirable Event Logs in Process Mining Tasks

Ali Norouzifar ALI.NOROUZIFAR@PADS.RWTH-AACHEN.DE
RWTH Aachen University, Germany

Wil van der Aalst WVDAALST@PADS.RWTH-AACHEN.DE
RWTH Aachen University, Germany

Abstract

Traditional process mining techniques utilize one event log as input to offer organizational insights. In many applications, information regarding undesirable process aspects may exist. However, the literature lacks a comprehensive overview of their integration into process mining tasks. In our paper, we explore leveraging data from both desirable and undesirable event logs to augment existing process mining tasks and develop innovative applications. Our aim is to systematically outline the potential for enhancements in this realm.

Keywords: Process Mining, Desirable and Undesirable Behavior, Process Discovery

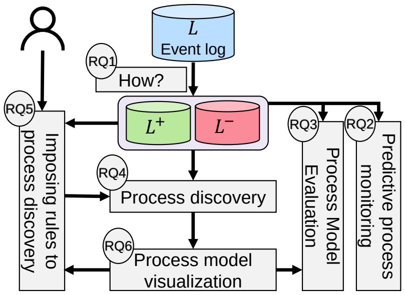

L
Event log
RQ1
How?
L^+
L^-
RQ5
Imposing rules to process discovery
RQ4
Process discovery
RQ6
Process model visualization
RQ3
Process Model Evaluation
RQ2
Predictive process monitoring

Figure 1: Overview of the research questions leveraging the desirable and undesirable event logs.

1. Introduction

Data extracted from information systems constitutes a rich and valuable resource, providing possibilities for diverse analyses. Process mining, as a broad discipline, encompasses a variety of applications applied to event data to extract meaningful insights [1]. In our research, we delved into leveraging information related to desirable and undesirable event logs to extract valuable insights and assist organizations in improving their processes by addressing performance and compliance problems. In Fig. 1, an overview of the research questions targeting desirable and undesirable cases is illustrated.

2. Potential Research Questions

Identification of Desirable and Undesirable Event Logs (RQ1) Some potential approaches to derive the desirability of the cases include domain-specific labeling [2], assessing the adherence to the normative behavior [3], rule-checking techniques [3], identification of outliers or strange cases [4], and automated detection of control flow variability [5]. Given that automatic labeling of cases as *desirable* or *undesirable* relies on interpretations and specific scenarios, user input may be necessary to choose an appropriate method. In [5], a framework for the identification of control flow variations across continuous dimensions like duration of cases is proposed. This framework takes an event log and a continuous dimension. Considering the cases are sorted based on their assigned value, a sliding-window-based algorithm utilizing the earth mover's distance is employed to find the change points in the control flow. Further analysis of the identified segments helps to categorize the cases into desirable or undesirable.

Predictive Process Monitoring (RQ2) Assuming that the desirability of the cases is known, the event logs can be encoded as suitable features for the machine learning techniques and the state-of-the-art architectures can be used to obtain predictive models or recommendations [6]. The process mining field can contribute to the improvements by providing meaningful features [7]. Extracting explainability from such predictive models helps to establish trustworthy predictions [8]. In addition to predictive models, many process variant analysis techniques from

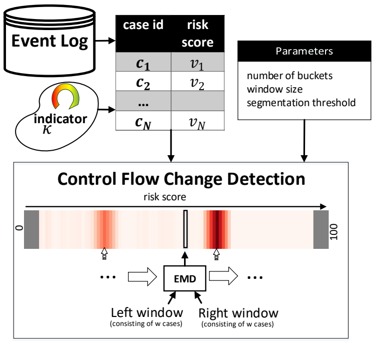


Figure 2: Process variant identification framework across continuous dimensions [5]

the literature could be used to effectively compare the event logs [9].

Process Model Evaluation (RQ3) Assessing a process model's capability to differentiate between desirable and undesirable behavior is challenging. Fitness is an evaluation metric used to assess the replayability of the observed behavior on the discovered model [3]. In [10], evaluation metrics are suggested, utilizing fitness criteria to ascertain model fit with the desirable event log while not aligning with the undesirable event log. Another approach proposed in [11] checks if the generalization of behavior allowed by the model conflicts with the undesirable behavior generated artificially from the desirable event log. Adaptations are required to make them applicable in scenarios with desirable and undesirable event logs.

Process Discovery (RQ4) The goal is to discover process models that support a desirable event log while avoiding an undesirable event log. Limited research has been conducted to involve desirable and undesirable event logs in process discovery. Declarative process discovery approaches like [2] and [12] discover constraint-based models from desirable and undesirable event logs. In [13] and [14], the discovery of procedural process models using these event logs is investigated. The IMbi algorithm is another relevant discovery technique introduced in [10]. In each recursion, the algorithm finds a process structure that has a low cost based on the desirable event log and a high cost based on the undesirable event log. In Fig. 3, one recursion of the proposed algorithm is illustrated [10]. The *ratio* parameter controls the involvement of the undesirable event log in the process discovery.

Imposing Rules to Process Discovery (RQ5) User-defined rules or discovered rules from event logs can help to enhance the quality of the discovered process models. Existing process discovery frameworks often disregard other valuable sources such as domain knowledge and documentation within business processes that provide insights into how the process functions. In [15], the IMr algorithm is proposed as a generalization of the IMbi framework that is capable of considering some declarative rules as input. One recursion of the IMr framework is illustrated in Fig. 4. This approach can be extended to scenarios with two event logs as input. In addition, a set of rules that can discriminate between the desirable and undesirable event logs can help to discover better process models.

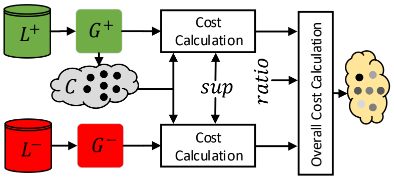


Figure 3: One recursion of IMbi framework [10], discovering a process model to support the desirable event log while avoiding the undesirable event log.

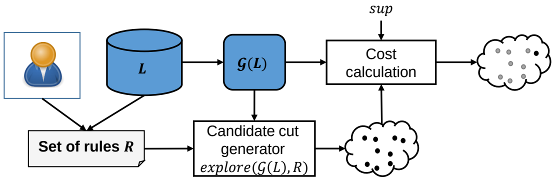


Figure 4: One recursion of IMr framework [15], allowing for a set of rules R as input.

Process Model Visualization (RQ6) The obtained insights from the previous research questions could be accompanied by visualization techniques. Conformance checking techniques from the literature or user feedback can be used to further improve the process models and align them better with reality.

3. Conclusion

Six potential research questions regarding the utilization of desirable and undesirable event logs in the process mining field are introduced in this extended abstract. Extending existing process mining frameworks with the integration of this information could yield enhanced insights.

Acknowledgments

This research was supported by the research training group "Dataninja" (Trustworthy AI for Seamless Problem Solving: Next Generation Intelligence Joins Robust Data Analysis) funded by the German federal state of North Rhine-Westphalia.

References

[1] Wil M. P. van der Aalst. *Process Mining - Data Science in Action, Second Edition*. Springer, 2016.

[2] Tijs Slaats, Søren Debois, and Christoffer Olling Back. Weighing the pros and cons: Process discovery with negative examples. In *International Conference on Business Process Management*, pages 47–64. Springer, 2021.

[3] Josep Carmona, Boudewijn F. van Dongen, Andreas Solti, and Matthias Weidlich. *Conformance Checking: Relating Processes and Models*. Springer, Cham, 1st edition, 2018. ISBN 3319994131.

[4] Jochen De Weerdt, Seppe K. L. M. vanden Broucke, Jan Vanthienen, and Bart Baesens. Active trace clustering for improved process discovery. *IEEE Trans. Knowl. Data Eng.*, 25(12): 2708–2720, 2013.

[5] Ali Norouzifar, Majid Rafiei, Marcus Dees, and Wil M. P. van der Aalst. Process variant analysis across continuous features: A novel framework. In *Enterprise, Business-Process and Information Systems Modeling - 25th International Conference, BPMDS 2024, and 29th International Conference, EMMSAD 2024, Proceedings*, volume 511 of *Lecture Notes in Business Information Processing*, pages 129–142. Springer, 2024.

[6] Irene Teinemaa, Marlon Dumas, Marcello La Rosa, and Fabrizio Maria Maggi. Outcome-oriented predictive process monitoring: review and benchmark. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(2):1–57, 2019.

[7] Mozhgan Vazifehdoostirani, Laura Genga, and Remco M. Dijkman. Encoding high-level control-flow construct information for process outcome prediction. In *4th International Conference on Process Mining, ICPM 2022*, pages 48–55. IEEE, 2022.

[8] Riza Velioglu, Jan Philip Göpfert, André Artelt, and Barbara Hammer. Explainable artificial intelligence for improved modeling of processes. In *Intelligent Data Engineering and Automated Learning - IDEAL 2022 , Manchester, UK, Proceedings*, volume 13756 of *Lecture Notes in Computer Science*, pages 313–325. Springer, 2022.

[9] Farbod Taymouri, Marcello La Rosa, Marlon Dumas, and Fabrizio Maria Maggi. Business process variant analysis: Survey and classification. *Knowledge-Based Systems*, 211:106557, 2021.

[10] Ali Norouzifar and Wil M. P. van der Aalst. Discovering process models that support desired behavior and avoid undesired behavior. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing, SAC 2023*, pages 365–368. ACM, 2023.

[11] Seppe KLM vanden Broucke, Jochen De Weerdt, Jan Vanthienen, and Bart Baesens. Determining process model precision and generalization with weighted artificial negative events. *IEEE Transactions on Knowledge and Data Engineering*, 26 (8):1877–1889, 2013.

[12] Federico Chesani, Chiara Di Francescomarino, Chiara Ghidini, Giulia Grundler, Daniela Loreti, Fabrizio Maria Maggi, Paola Mello, Marco Montali, and Sergio Tessaris. Shape your process: Discovering declarative business processes from positive and negative traces taking into account user preferences. In *International Conference on Enterprise Design, Operations, and Computing*, pages 217–234. Springer, 2022.

[13] Hernán Ponce De León, Lucio Nardelli, Josep Carmona, and Seppe KLM vanden Broucke. Incorporating negative information to process discovery of complex systems. *Information Sciences*, 422:480–496, 2018.

[14] Evelina Lamma, Paola Mello, Marco Montali, Fabrizio Riguzzi, and Sergio Storari. Inducing declarative logic-based models from labeled traces. In *Business Process Management: 5th International Conference, BPM 2007. Proceedings*, pages 344–359. Springer, 2007.

[15] Ali Norouzifar, Marcus Dees, and Wil M. P. van der Aalst. Imposing rules in process discovery: An inductive mining approach. In *Research Challenges in Information Science - 18th International Conference, RCIS 2024, Proceedings, Part I*, volume 513 of *Lecture Notes in Business Information Processing*, pages 220–236. Springer, 2024.

DOI: 10.11576/dataninja-1165

Feeling Socially Excluded When Working With Robots

Clarissa Sabrina Arlinghaus CLARISSA_SABRINA.ARLINGHAUS@UNI-BIELEFELD.DE
Bielefeld University, Germany

Günter W. Maier AO-PSYCHOLOGIE@UNI-BIELEFELD.DE
Bielefeld University, Germany

Abstract

Work is not just about money, but also about satisfying social needs. We examine processes of social inclusion and exclusion among human employees and robot employees. For our current study, we chose the restaurant industry as a contemporary use case where humans and robots work together as waiters. We assume that social exclusion from either human or robot colleagues will threaten peoples needs (i.e., belonging, control, meaningful existence, self-esteem) but will be interpreted differently depending on the excluding agent (i.e., human colleague or robot colleague). Assuming different attribution processes challenges the “Computers Are Social Actors” theory and could lead the rethinking human-robot interactions or even humans interacting with technology in general.

Keywords: social exclusion, robot coworker, work in progress

1. Introduction

Work goes beyond being a source of income; it also acts as a framework for time management, a platform for building social connections, a catalyst for pursuing collective objectives, a symbol of societal status, and a source of activity [1, 2]. Humans inherently seek social interaction [3]. Interactions with coworkers can either fulfill or compromise basic social needs [4].

As robots become more integrated into the workforce [5], this innate aspect of human behavior could face disruption, possibly leading to feelings of social exclusion. Social exclusion threatens fundamental human needs (i.e., a sense of belonging, meaningful existence, control, and self-esteem) and can lead to severe consequences (i.e., depression, alienation, helplessness) in the long run [6].

2. SAIL

The effects of human-robot team work are explored as one aspect of SAIL [7]. SAIL is an interdisciplinary research network in Germany, in which Bielefeld University, Paderborn University, University of Applied Sciences Bielefeld and the University of Applied Sciences Ostwestfalen-Lippe are involved [8]. The abbreviation SAIL stands for ”SustAInable Life-cycle of Intelligent Socio-Technical Systems” and illustrates the goal of developing the entire product life cycle of AI systems in a sustainable manner [8, 9]. The design of AI systems should ensure transparency, security and human self-determination [9].

3. Work in Progress

Within SAIL, we investigate processes of social inclusion and exclusion in human-robot interactions within work settings [7]. For our current study, we chose the restaurant industry as a contemporary use case since plenty of skilled workers have left the hospitality industry because of poor work-life balance, low job security, and little employee compensation [10]. As a consequence, many restaurants turned to employing robots as waitstaff to address the scarcity of skilled labor. However, these robots lack crucial communication skills, potentially causing human employees to experience feelings of exclusion.

To explore this, we planned a pre-registered [11] online study where people should imagine working in a restaurant with either human colleagues (see Figure 1) or robot colleagues (see Figure 2). During the study, participants read texts describing different behaviors of their colleagues in a randomized order. The behaviors depict various forms of social inclusion and exclusion. After each behavior, participants are asked to generate potential explanations for the previous behavior and indicate their need fulfillment through questionnaire items excerpted from [12].

Building on the temporal need-threat model [6], we anticipate that social exclusion poses a threat to essential human needs (e.g., belonging, self-esteem, meaningful existence) regardless of whether social exclusion originates from a human or a robot. Nevertheless, we speculate that individuals interpret their experienced exclusion differently depending on the excluding agent (human vs. robot). This assumption challenges the ”Computers Are Social Actors” theory [13], which suggests that individuals unconsciously apply social norms to computers based on interpersonal interactions. Such insights would represent a significant advancement in the field of human-robot interaction research.

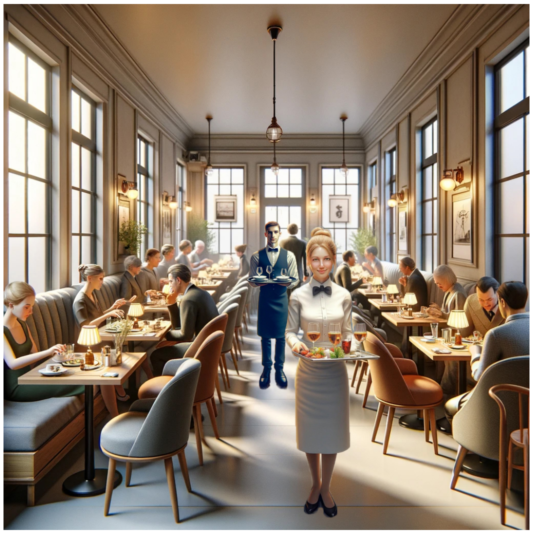


Figure 1: Restaurant Picture With Human Waiters

4. Outlook

Our next steps will be analyzing the data collected in this study and preparing a manuscript for publication. We comply with Open Science standards. Therefore, we pre-registered [11] our study and plan making the data available in the Open Science Framework (OSF) under the following link: https://osf.io/mnybf/

Depending on when you read this work-in-progress paper, the data may already be available there or will be added at a later date. We also plan to link to the final paper on the pre-registration [11] or in OSF project as soon as the final paper is published. However, this will take some time.

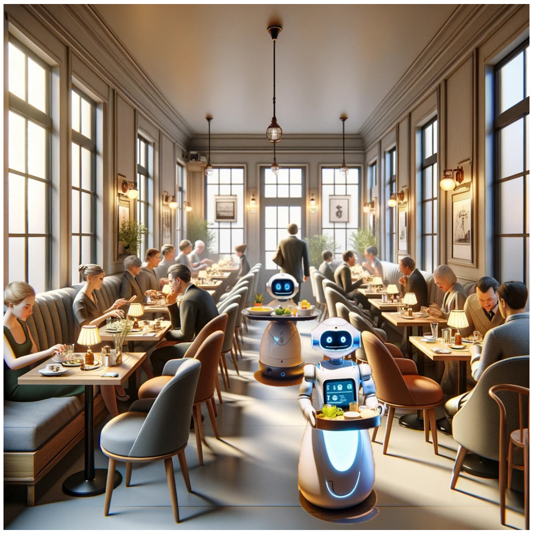


Figure 2: Restaurant Picture With Robot Waiters

Acknowledgments

SAIL is funded by the Ministry of Culture and Science of the State of North Rhine-Westphalia under the grant no NW21-05A.

References

[1] Karsten Ingmar Paul, Hannah Scholl, Klaus Moser, Andrea Zechmann, and Bernad Batinic. Employment status, psychological needs, and mental health: Meta-analytic findings concerning the latent deprivation model. *Frontiers in Psychology*, 14:1017358, 2023. doi: 10.3389/fpsyg.2023.1017358.

[2] Kerstin Isaksson. Unemployment, mental health and the psychological functions of work in male welfare clients in stockholm. *Scandinavian Journal of Social Medicine*, 17(2):165–169, 1989. doi: 10.1177/140349488901700207.

[3] Roy F. Baumeister and Mark R. Leary. The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117(3):497–529, 1995. doi: 10.1037/0033-2909.117.3.497.

[4] Jane O’Reilly and Sara Banki. Research in work and organizational psychology: Social exclusion in the workplace. In Paolo Riva and Jennifer

Eck, editors, *Social Exclusion*, pages 133–155. Springer International Publishing, Cham, 2016. doi: 10.1007/978-3-319-33033-4_7.

[5] Sonja K. Ötting, Lisa Masjutin, Jochen J. Steil, and Günter W. Maier. Let's work together: A meta-analysis on robot design features that enable successful human-robot interaction at work. *Human Factors*, 64(6):1027–1050, 2022. doi: 10.1177/0018720820966433.

[6] Kipling D. Williams. Ostracism: A temporal need-threat model. In M. P. Zanna, editor, *Advances in experimental social psychology*, volume 41, pages 275–314. Elsevier Academic Press, 2009. doi: 10.1016/S0065-2601(08)00406-1.

[7] SAIL. Projects, 2024. URL https://www.sail.nrw/research/projects/.

[8] SAIL. Sustainable life-cycle of intelligent socio-technical systems, 2024. URL https://www.sail.nrw/.

[9] SAIL. Research, 2024. URL https://www.sail.nrw/research/.

[10] Karine Grigoryan. Labor shortages in the hospitality industry: The effects of work-life balance, employee compensation, government issued unemployment benefits and job insecurity on employees' turnover intentions. *Westcliff International Journal of Applied Research*, 8(1):59–73, 2024. doi: 10.47670/wuwijar202481kg.

[11] Clarissa Sabrina Arlinghaus and Günter W. Maier. Different forms of social exclusion in a robo-restaurant, 2024. URL https://doi.org/10.17605/OSF.IO/ZAM24.

[12] Selma C. Rudert and Rainer Greifeneder. When it's okay that i don't play: Social norms and the situated construal of social exclusion. *Personality and Social Psychology Bulletin*, 42(7):955–969, 2016. doi: 10.1177/0146167216649606.

[13] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. Computers are social actors. In Beth Adelson, editor, *CHI 94: Human Factors in Computing Systems Conference*, pages 72–78. Association for Computing Machinery, 1994. doi: 10.1145/191666.191703.

Proceedings of the DataNinja sAIOnARA 2024 Conference 39-42 DOI: 10.11576/dataninja-1166

Trade-offs Between Privacy and Performance in Encrypted Datasets using Machine Learning Models

Sanaullah Sanaullah sanaullah@hsbi.de
Bielefeld University of Applied Sciences and Arts, Bielefeld, Germany

Hasina Attaullah hasina.attaullah@hsbi.de
Bielefeld University of Applied Sciences and Arts, Bielefeld, Germany

Thorsten Jungeblut thorsten.jungeblut@hsbi.de
Bielefeld University of Applied Sciences and Arts, Bielefeld, Germany

1. Introduction

In recent years, with the increasing importance of dataset privacy in machine learning (ML) applications, there has been an increased demand for secure and privacy-preserving solutions [1, 2]. Consequently, encryption techniques have become known as a critical tool for protecting data privacy in an era of massive data use, exchange, and analysis [3, 4]. Encryption protects data against illegal access and disclosure by changing it into unreadable ciphertext that can only be decrypted by authorized parties [5–7]. In the field of ML, where sensitive data is often utilized, in such a process the use of encryption techniques has significant potential for providing privacy-preserving model training and inference [7, 8].

Therefore, this article analyzes, investigates, and compares three widely used encryption techniques. Each encryption method offers unique advantages and trade-offs [9–11]. Thus, we evaluate the performance of Convolutional Neural Network (CNN) models trained on encrypted datasets using these encryption techniques to provide detailed information on the effectiveness, practical concerns, and applicability of various methods for real-world applications by completely analyzing them within the context of computer vision. We test the performance of CNN models trained on encrypted data with several encryption approaches using neural models based-architecture [12]. Parameters such as training time, memory usage, and classification accuracy are analyzed and compared between encryption methods. We also look into the effect of encryption on model interpretability and robustness against adversarial attacks. Furthermore, to support our study we demonstrate our approach by using practical implementation—to showcase the performance and efficiency of each encryption strategy in protecting data privacy while keeping model accuracy and testing in a real-time recognition application using an edge device such as NVIDIA Jetson. Through this comparative analysis, researchers and developers can achieve a more in-depth understanding of the importance and issues involved with the integration of encryption techniques into ML especially in computer vision application workflows.

2. Analysis Methodology

A CNN architecture is utilized for the classification task due to its effectiveness in handling image data [13–15]. The architecture comprises two convolutional layers followed by max-pooling layers that enable hierarchical feature extraction from the input images. Each convolutional layer is activated using the ReLU activation function to introduce nonlinearity. The resulting feature maps are downsampled using max-pooling layers to reduce spatial dimensions and extract dominant features. Subsequently, a flattened layer transforms the 2D feature maps into a 1D vector which helps the compatibility with densely connected layers. Two fully connected (dense) layers with ReLU activation functions further process the extracted features, promoting non-linear transformations and capturing intricate patterns in the data. Finally, a dense layer with a softmax activation function is employed for multi-class classification that generates probability distributions over the output classes. This CNN architecture uses the hierarchical nature of neural networks to effectively classify the handwritten digit images present in the MNIST dataset [16–18].

Table 1: Accuracy with Resource Consumption

Encryption Model	**Test Accuracy**	**CPU Time (sec.)**	**Memory (KB)**
Original Data	99.25%	85.50	306,140
XOR	96.70%	76.61	156
S.Cipher	11.35%	75.92	30,744
Homomorphic	98.02%	75.92	30744

2.1. Encryption Techniques

Three encryption techniques are utilized to protect the privacy of the MNIST dataset during model training and evaluation. First, XOR encryption involves applying the bitwise XOR operation between the image pixel values and a randomly chosen integer encryption key. This process effectively rearranges the pixel values based on the binary representation of both the image and the key. Second, substitution cipher encryption shifts pixel values by a fixed integer value (encryption key) using modulo addition. In this process, each pixel value in the image is shifted by the chosen encryption key, wrapping around if the resulting value exceeds the maximum pixel intensity. Last, homomorphic encryption enables computations on encrypted data without decryption, preserving data privacy during processing. An encryption homomorphic scheme is applied to the pixel values using a predefined encryption key for arithmetic operations on encrypted data while maintaining confidentiality.

3. Experimental Results

The experimental results showcase the performance and resource consumption of the CNN model trained on different encryption techniques applied to the MNIST dataset. Therefore, XOR and homomorphic encryption techniques showed reasonable performance with minimal computational overhead, and substitution cipher encryption significantly degraded the model's accuracy, indicating its limitations in preserving data integrity for image classification tasks, details of each model results can be seen in Table 1. These results highlight the trade-offs between data privacy and model performance. Furthermore, to support our study exploration, we analyze statistical properties. The statistical properties provide information into the distribution characteristics of pixel values in the encrypted datasets obtained using different encryption techniques. Therefore, XOR-Encrypted Data - Mean: 0.4887 and Variance: 0.0088, demonstrate a mean value close to 0.5, indicating a balanced distribution of pixel values around the midpoint. The variance is relatively small, suggesting moderate dispersion of pixel values around the mean. However, Substitution-Encrypted Data - Mean: 0.1966 and Variance: 1.4825e-06, similarly Homomorphic-Encrypted Data - Mean: 0.1236 and Variance: 0.0882, shows a significantly lower mean value compared to XOR-encrypted, indicating a shift in the pixel value distribution. The variance suggested a minimal dispersion of pixel values around the mean.

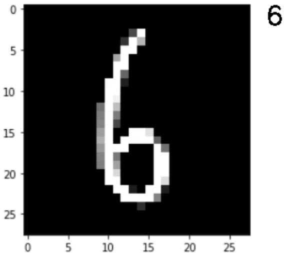


Figure 1: A handwritten digit on the left and on the right the model predicted number in real-time using XOR-encrypted dataset.

3.1. Real-Time Implementation

In the real-time implementation, we demonstrate the system's capability to predict handwritten digits from uploaded images seamlessly and efficiently while preserving data privacy using encryption techniques. To provide a concrete example, Figure 1 showcases an uploaded image of a handwritten digit on the left and on the right the model predicted number in real-time testing. Therefore, through this demonstration, we illustrate the system's ability to preprocess incoming images, feed them into the XOR encrypted model, and seamlessly provide accurate predictions without compromising data privacy. Additionally, all test results are available on our GitHub channel.

Acknowledgments

This research was supported by the research training group ”Dataninja” (Trustworthy AI for Seamless Problem Solving: Next Generation Intelligence Joins Robust Data Analysis) funded by the German federal state of North Rhine-Westphalia. This project is also supported by SAIL (grant no NW21-059B) and the "Transformation in Care and Technology" (funding program "Profiling 2020", and the grand number is PROFILNRW-2020-095), project is funded by the Ministry of Culture and Science of the State of North Rhine-Westphalia.

References

[1] Petr Velan, Milan Čermák, Pavel Čeleda, and Martin Drašar. A survey of methods for encrypted traffic classification and analysis. *International Journal of Network Management*, 25 (5):355–374, 2015.

[2] Sanaullah, Shamini Koravuna, Ulrich Rückert, and Thorsten Jungeblut. Snns model analyzing and visualizing experimentation using ravsim. In *International conference on engineering applications of neural networks*, pages 40–51. Springer, 2022.

[3] Rongxing Lu, Hui Zhu, Ximeng Liu, Joseph K Liu, and Jun Shao. Toward efficient and privacy-preserving computing in big data era. *IEEE Network*, 2014.

[4] Sanaullah, Hasan Baig, Jan Madsen, and Jeong-A Lee. A parallel approach to perform threshold value and propagation delay analyses of genetic logic circuit models. *ACS Synthetic Biology*, 9 (12):3422–3428, 2020.

[5] Karthik Nandakumar, Nalini Ratha, Sharath Pankanti, and Shai Halevi. Towards deep neural network training on encrypted data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[6] Sanaullah, S Koravuna, Ulrich Rückert, and Thorsten Jungeblut. Transforming event-based into spike-rate datasets for enhancing neuronal behavior simulation to bridging the gap for snns. In *IEEE Conference on Computer Vision (ICCV)*, 2023.

[7] Eva Papadogiannaki and Sotiris Ioannidis. A survey on encrypted network traffic analysis applications, techniques, and countermeasures. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[8] Shamini Koravuna, Ulrich Rückert, Thorsten Jungeblut, et al. Evaluation of spiking neural nets-based image classification using the runtime simulator ravsim. *International Journal of Neural Systems*, pages 2350044–2350044, 2023.

[9] Mohamed Elhoseny, Xiaohui Yuan, Hamdy K El-Minir, and Alaa Mohamed Riad. An energy efficient encryption method for secure dynamic wsn. *Security and Communication Networks*, 2016.

[10] Mohammed Abbas Fadhil Al-Husainy. A novel encryption method for image security. *International Journal of Security and Its Applications*, 6(1):1–8, 2012.

[11] Sana Ullah and Thorsten Jungeblut. Analysis of mr images for early and accurate detection of brain tumor using resource efficient simulator brain analysis. In *19th International Conference on Machine Learning and Data Mining MLDM*, 2023.

[12] Sanaullah, Shamini Koravuna, Ulrich Rückert, and Thorsten Jungeblut. Exploring spiking neural networks: a comprehensive analysis of mathematical models and applications. *Frontiers in Computational Neuroscience*, 17:1215824, 2023.

[13] Samir S Yadav and Shivajirao M Jadhav. Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big data*, 2019.

[14] Neha Sharma, Vibhor Jain, and Anju Mishra. An analysis of convolutional neural networks for image classification. *Procedia computer science*.

[15] Sanaullah Sanaullah. A hybrid spiking-convolutional neural network approach for advancing machine learning models. In *Northern Lights Deep Learning Conference*, pages 220–227. PMLR, 2024.

[16] Savita Ahlawat, Amit Choudhary, Anand Nayyar, Saurabh Singh, and Byungun Yoon. Improved handwritten digit recognition using convolutional neural networks (cnn).

[17] Sanaullah, Shamini Koravuna, Ulrich Rückert, and Thorsten Jungeblut. Streamlined training of gcn for node classification with automatic loss function and optimizer selection. In *International Conference on Engineering Applications*

of Neural Networks, pages 191–202. Springer, 2023.

[18] Feiyang Chen, Nan Chen, Hanyang Mao, and Hanlin Hu. Assessing four neural networks on handwritten digit recognition dataset (mnist). *arXiv preprint arXiv:1811.08278*, 2018.

Proceedings of the DataNinja sAIOnARA 2024 Conference 43-46 DOI: 10.11576/dataninja-1167

Advancements in Neural Network Generations

Sanaullah [*] SANAULLAH@HSBI.DE
Bielefeld University of Applied Sciences and Arts, Bielefeld, Germany

Shamini Koravuna SKORAVUNA@TECHFAK.UNI-BIELEFELD.DE
Bielefeld University, Germany

Ulrich Rückert RUECKERT@TECHFAK.UNI-BIELEFELD.DE
Bielefeld University, Germany

Thorsten Jungeblut THORSTEN.JUNGEBLUT@HSBI.DE
Bielefeld University of Applied Sciences and Arts, Bielefeld, Germany

1. Introduction

Innovations in Neural Network Generation demonstrate the continual evolution, optimization, and development of artificial neural networks (ANNs) over periods [1]. These improvements include a combination of methodologies, approaches, and technical breakthroughs aimed at increasing the efficiency and abilities of neural network models [1]. Researchers and engineers have repeatedly attempted to push the boundaries of neural network performance, scalability, and applicability across multiple fields. These improvements usually involve changes to network designs, training algorithms, optimization methodologies, and hardware acceleration methods. Moreover, the neural network generations are closely related to key achievements in the machine learning (ML) research domain, such as the development of deep learning (DL) designs like convolutional neural network (CNN) or spiking neural network (SNN) and using both neural generations to introduce natural language processing and advances in computer vision applications [2–4]. Thus, in the field of neural network study, researchers have categorized ANN models into generations based on their computational design and capabilities. Maass' classification approach [5] categorizes ANN evolution into three generations. Therefore, this research study explores the continual evolution and optimization of ANNs, highlighting advancements in methodologies and technical innovation. We discuss the different generations of ANN, based on computational design and capabilities, emphasizing their role in shaping achievements in ML research. The study underscores the significance of these generational milestones in enhancing the adaptability and efficacy of neural network models for computational tasks, such as image classification. Figure 1 demonstrates the visual representation of these generations.

2. First Generation of Neural Network

ML began with the perception neural network, a fundamental component of neural theory. Designed by Frank Rosenblatt in the late 1950s [6], the perceptron represented a unique technique for pattern recognition and classification. It symbolizes the first attempts to recreate the functioning of real neurons called a representation of biological neurons and create a human-like intelligence machine. The architecture was the first attempt to model biological brain network computers and it utilized simple threshold units. At its most basic explanation, the Perceptron is a single-layer neural network designed for binary classification tasks. Its primary element emphasizes its significance as an introduction to more complicated neural network topologies [7, 8]. Although relatively straightforward, first-generation neural networks encountered significant computing and conceptual challenges. The computational capability at the time was not sufficient for training large-scale networks or managing complicated learning algorithms. Therefore, the perceptron's linear decision limitations significantly restrict its ability to address nonlinear issues. Despite its limited extent and functionality, but it was an important step in the development of ANN [9].

[*] All authors contributed equally

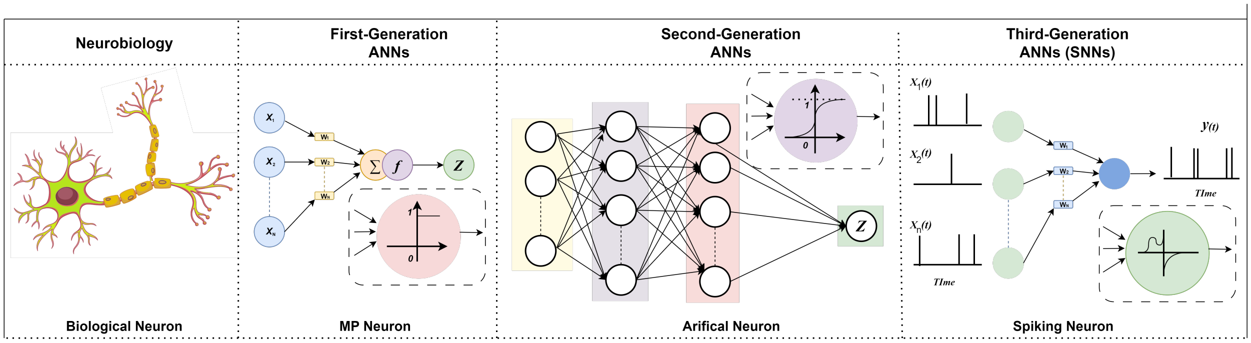


Figure 1: The basic overview of three generations of ANNs.

3. Second Generation of Neural Network

Second-generation neural networks consist of multilayer perceptrons (MLP), a class of neural networks containing one or more hidden layers [10, 11]. Unlike single-layer or perceptions neural networks, which can only be learned from linearly defined hyperparameters, MLPs can learn nonlinear mappings from the input and output data by incorporating one or more hidden layers with nonlinear activation functions. This enabled MLPs to reach complex functions and solve a wide range of ML challenges [12, 13]. However, compared to MLP, the most significant achievement in the second generation of ANNs was the development of CNN. CNN appeared as a marked improvement in the neural network history. In the mid-1980s, Kunihiko et al. [14] designed architectures for processing structured grid-like data, such as image-based datasets. They used the ideas of local connectivity and hyper-parameter sharing to effectively process hierarchical representations of graphical data. Additionally, Rumelhart and Williams et al. [15] presented new learning methods, including backpropagation, which transformed computer vision and showed a breakthrough in image recognition and object detection research domain.

4. Third Generation of Neural Network

SNNs are a class of ANN that draws inspiration from the human nervous system, such as the spiking mechanism of neurons in the brain [16, 17]. SNNs neural architecture-based sharing information using discrete spikes rather than continuous-valued signals, as compared to other generations of ANNs process. This spiking neuron function is the fundamental unit of an SNN, stimulating the activity of biological neurons by producing discrete spikes in response to input current. These spikes are frequently described as binary events that emerge at predetermined times and reflect both the timing and stability of neural activity. As a result, the temporal dynamics of the spiking process and propagation are important for information simulation in SNNs for allowing them to encode and interpret temporal patterns in input [18, 19]. Furthermore, SNNs have demonstrated promising performance in different applications, including event-driven processing, pattern recognition, and neuromorphic computing [20–22]. They are especially well-suited to applications that require processing spatiotemporal data, such as sensory processing, robotics, and object identification prediction [23, 24]. Unlike traditional neural network architectures that depend on the rate-based firing of neurons, SNN more closely mimics the behavior of biological neuron manners by communicating between neurons via discrete functions, commonly known as action potentials [25, 26]. Lastly, in terms of parallel processing and implementation on hardware or edge devices, SNNs perform incredibly well, due to their discrete spike trains. Therefore, this feature enables energy-efficient implementations on edge computing and is a particularly useful tool for low-power applications.

Acknowledgments

This research was supported by the research training group ”Dataninja” (Trustworthy AI for Seamless Problem Solving: Next Generation Intelligence Joins Robust Data Analysis) funded by the German federal state of North Rhine-Westphalia and the project SAIL. SAIL is funded by the Ministry of Culture and Science of the State of North Rhine-Westphalia under grant no NW21-059B.

References

[1] Oludare Isaac Abiodun, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Abubakar Malah Umar, Okafor Uchenwa Linus, Humaira Arshad, Abdullahi Aminu Kazaure, Usman Gana, and Muhammad Ubale Kiru. Comprehensive review of artificial neural network applications to pattern recognition. *IEEE access*, 7:158820–158846, 2019.

[2] Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.

[3] Sanaullah Sanaullah. A hybrid spiking-convolutional neural network approach for advancing machine learning models. In *Northern Lights Deep Learning Conference*, pages 220–227. PMLR, 2024.

[4] Yoav Goldberg. *Neural network methods for natural language processing*. Springer Nature, 2022.

[5] Wolfgang Maass. Networks of spiking neurons: the third generation of neural network models. *Neural networks*, 10(9):1659–1671, 1997.

[6] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6): 386, 1958.

[7] Silvia Curteanu and Hugh Cartwright. Neural networks applied in chemistry. i. determination of the optimal topology of multilayer perceptron neural networks. *Journal of Chemometrics*, 25 (10):527–549, 2011.

[8] Richard P Lippmann. Pattern classification using neural networks. *IEEE communications magazine*, 27(11):47–50, 1989.

[9] Imad A Basheer and Maha Hajmeer. Artificial neural networks: fundamentals, computing, design, and application. *Journal of microbiological methods*, 43(1):3–31, 2000.

[10] Simone Marinai, Marco Gori, and Giovanni Soda. Artificial neural networks for document analysis and recognition. *IEEE Transactions on pattern analysis and machine intelligence*, 27(1): 23–35, 2005.

[11] Sanaullah, Hasan Baig, Jan Madsen, and Jeong-A Lee. A parallel approach to perform threshold value and propagation delay analyses of genetic logic circuit models. *ACS Synthetic Biology*, 9 (12):3422–3428, 2020.

[12] Hind Taud and Jean-Franccois Mas. Multilayer perceptron (mlp). *Geomatic approaches for modeling land change scenarios*, pages 451–455, 2018.

[13] Sanaullah, Shamini Koravuna, Ulrich Rückert, and Thorsten Jungeblut. Streamlined training of gcn for node classification with automatic loss function and optimizer selection. In *International Conference on Engineering Applications of Neural Networks*, pages 191–202. Springer, 2023.

[14] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

[15] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[16] Samanwoy Ghosh-Dastidar and Hojjat Adeli. Spiking neural networks. *International journal of neural systems*, 19(04):295–308, 2009.

[17] Sanaullah, Shamini Koravuna, Ulrich Rückert, and Thorsten Jungeblut. Exploring spiking neural networks: a comprehensive analysis of mathematical models and applications. *Frontiers in Computational Neuroscience*, 17:1215824, 2023.

[18] Qiang Yu, Huajin Tang, Kay Chen Tan, and Haoyong Yu. A brain-inspired spiking neural network model with temporal encoding and learning. *Neurocomputing*, 138:3–13, 2014.

[19] Sana Ullah, Amanullah Amanullah, Kaushik Roy, Jeong-A Lee, Son Chul-Jun, and Thorsten Jungeblut. A hybrid spiking-convolutional neural network approach for advancing high-quality

image inpainting. In *International Conference on Computer Vision (ICCV) 2023*, 2023.

[20] Sanaullah, Shamini Koravuna, Ulrich Rückert, and Thorsten Jungeblut. Snns model analyzing and visualizing experimentation using ravsim. In *International conference on engineering applications of neural networks*, pages 40–51. Springer, 2022.

[21] Mike Davies, Andreas Wild, Garrick Orchard, Yulia Sandamirskaya, Gabriel A Fonseca Guerra, Prasad Joshi, Philipp Plank, and Sumedh R Risbud. Advancing neuromorphic computing with loihi: A survey of results and outlook. *Proceedings of the IEEE*, 109(5):911–934, 2021.

[22] Shamini Koravuna, Ulrich Rückert, Thorsten Jungeblut, et al. Evaluation of spiking neural nets-based image classification using the runtime simulator ravsim. *International Journal of Neural Systems*, pages 2350044–2350044, 2023.

[23] Sana Ullah, Shamini Koravuna, Ulrich Rückert, and Thorsten Jungeblut. A novel spike vision approach for robust multi-object detection using snns. 2023.

[24] Seamus Cawley, Fearghal Morgan, Brian McGinley, Sandeep Pande, Liam McDaid, Snaider Carrillo, and Jim Harkin. Hardware spiking neural network prototyping and application. *Genetic Programming and Evolvable Machines*, 12:257–280, 2011.

[25] Kashu Yamazaki, Viet-Khoa Vo-Ho, Darshan Bulsara, and Ngan Le. Spiking neural networks and their applications: A review. *Brain Sciences*, 12(7):863, 2022.

[26] Sana Ullah, Shamini Koravuna, Ulrich Rückert, and Thorsten Jungeblut. Design-space exploration of snn models using application-specific multi-core architectures. 2023.

DOI: 10.11576/dataninja-1168

Prediction of Intermuscular Co-Contraction Based on the sEMG of Only one Muscle With the Same Biomechanical Direction of Action

Nils Grimmelsmann NILS.GRIMMELSMANN@HSBI.DE
Malte Mechtenberg MALTE.MECHTENBERG@HSBI.DE
Biomechatronics and Embedded Systems Group, Institute of System Dynamics and Mechatronics,
Bielefeld University of Applied Sciences and Arts, Bielefeld, Germany

Markus Vieth MVIETH@TECHFAK.UNI-BIELEFELD.DE
Barbara Hammer BHAMMER@TECHFAK.UNI-BIELEFELD.DE
Machine Learning Group, Bielefeld University, Bielefeld, Germany

Axel Schneider AXEL.SCHNEIDER@HSBI.DE
Biomechatronics and Embedded Systems Group, Institute of System Dynamics and Mechatronics,
Bielefeld University of Applied Sciences and Arts, Bielefeld, Germany

Abstract

Research aims to enhance physical abilities using exoskeletons and limb movement prediction. SEMG signals are used for intuitive control, but their measurement is limited to shallowly under-the-skin muscles, making deep muscle signals less frequently used. Here we extended a previously proposed method to train a virtual sensor for the difficult to access muscles (deep muscles e.g. *brachialis*). The method is extended from signals from the same muscle to intermuscular signals and the results confirm simple biomechanical assumptions. The trained virtual sensors are ready for further investigations by being used in a biomechanical model.

Keywords: Electromyography, virtual sensor, regression

1. Introduction

For an intuitive control of exoskeletal devices an intuitive and non-delayed control scheme is an important part of success. Electromyographic (EMG) signals are often used to enable such control. This study uses the flexor muscles of the elbowjoint as the object of investigation. The flexion of the joint is possible by three flexors. The flexors are the *biceps brachii* (shallow, often used in this application) *brachialis* (deep, below the *biceps brachii*) and the *brachioradialis* (shallow, located on the forearm). Even though the three flexors share elbow flexion as their main task, they have different attachment points and therefore different lever arm courses across the elbow angle. As the biceps and *brachialis* are located close to each other, sEMG signals are prone for picking up signal components from the other muscle. This phenomenon is called crosstalk.

As shown in previous work, the movement of the elbow can be predicted by the sEMG signals of the two *biceps brachii* heads and the two shallow *triceps brachii* heads [1]. Previous work has also shown that a virtual sensor that predicts the activation for one *biceps brachii* head can be trained from the other head with a shallow feedforward neural network (ffn) [2]. This can be used as a replacement for a sEMG channel or for the evaluation of the sEMG.

Therefore, this is further used to guide the training for a virtual *brachialis* sensor by domain knowledge. This can lead to more explainable behaviour of the trained virtual sensors. The previously proposed method allows for a robust training process with a domain based foundation which is used in this work to interpret the results (e.g. co-activation and crosstalk).

2. Methods

2.1. Experiment and Used Dataset

The underlying experiment with which the data was recorded is based on [3]. In addition, two sEMG sensors were added to the setup to measure the signals of the *brachialis*. This deep muscle was measured by one sensor each on the medial (ME) and lateral (LA) side of the distal *biceps brachii* tendon.

The preparation and attachment of the sensors were also described in detail in [3]. After these steps, the verification of the correct placement of the two sensors for the *brachialis* and two sensors for the *biceps brachii* was done by isolating the muscles and compare the resulting sEMG amplitudes visually. The movement is performed at two speeds [0.5 Hz and 1 Hz] and with two different weights [2 kg and 4 kg]. The sequence of the four combinations is randomly selected in advance for each subject. The age of the 13 subjects was 24.7 ± 2.6 years.

2.2. Biomechanics of *Brachialis* and *Biceps Brachii*

The muscle origins and insertions have different distances to the centre of rotation. This results in different lever-arm courses over the elbow angle. Furthermore, the possible force development of a skeletal muscle depends on its length [4].Due to these properties, the *brachialis* and *biceps brachii* have different possible force generation via the elbow angle [5].

The force vector generated by the muscle can point more or less in the direction of a possible joint rotation or in the direction of the joint, depending on the position of the muscle in relation to the bone and the joint.Both of these properties can lead to different activation of muscles although their biomechanical direction of action is in principle the same [6].

2.3. Training Pipeline and Strategies

The data was preprocessed in a the same way as in [2]. The training strategies are also unchanged and a detailed description can be found in the previous paper. The first strategy for training is training at the level of the individual experiment variations. The performance of the virtual sensors is measured using the mean absolute error (MAE) between predicted and measured activation.

Contrary to the previous work, the regression from the activation of the *brachialis* to the activation of the *biceps brachii* is now learned. The second training strategy is to exclude a subject from training and use its data for the test.The baseline for both training strategies is setting the output of the virtual sensor to the input.

3. Results

The first training strategy results in no performance increase compared to the baseline. The baseline of the lateral *brachialis* (BRA) regressed from the *bicpes brachii* (BIC) long head (LH) is lower than the other three regression configurations.

The results for the second training strategy are shown in Table 1. The regression with only one input dimension performs similarly to the first training strategie. If the input dimension is expanded, the error decreases slightly. When introducing a nonlinearity through the activation function rectified linear unit, the error decreases further. The virtual sensor for the lateral *brachialis* shows lower errors than that for the medial *brachalis* sensor.

Table 1: The MAE (lower = better performance) of the virtual sensor for the leave one out strategy. For the two muscles [*bicpes brachii* (BIC), *brachialis* (BRA)] with the respective muscle heads [long head (LH), short head (SH)] and sensor position [lateral (LA), medial (ME)]

input:	BIC SH	BIC SH	BIC LH	BIC LH
output:	BRA ME	BRA LA	BRA ME	BRA LA
baseline	0.461	0.467	0.472	0.393
Train lin. 1d	0.436	0.453	0.426	0.374
Test lin. 1d	0.439	0.453	0.426	0.376
Train lin. 5d	0.412	0.413	0.409	0.354
Test lin. 5d	0.418	0.421	0.415	0.366
Train nlin. 5d	0.379	0.281	0.377	0.228
Test nlin. 5d	0.404	0.314	0.391	0.257

4. Discussion

The lower baseline for the virtual sensor of the *brachialis* lateral with the *biceps brachii* long head as input compared to the other three baselines could be cases by the short distance on the arm. Therefore, this could indicate a pickup of a *biceps brachii* long head sEMG from the *brachialis* lateral sensor (crosstalk).

The better performance by adding the nonlinearity fit the biomechanical structure described in Section 2.2 These two hypotheses could potentially be verified by using the virtual sensors as an input for a biomechanical model as in [2] also suggested.

Acknowledgments

This research was supported by the research training group “Dataninja” (Trustworthy AI for Seamless Problem Solving: Next Generation Intelligence Joins Robust Data Analysis) funded by the German federal state of North Rhine-Westphalia.

References

[1] Nils Grimmelsmann, Malte Mechtenberg, Wolfram Schenck, Hanno Gerd Meyer, and Axel Schneider. sEMG-based prediction of human forearm movements utilizing a biomechanical model based on individual anatomical/ physiological measures and a reduced set of optimization parameters. *PLOS ONE*, 18(8):1–28, 2023. doi: 10.1371/journal.pone.0289549. URL https://doi.org/10.1371/journal.pone.0289549.

[2] Nils Grimmelsmann, Malte Mechtenberg, Markus Vieth, Alexander Schulz, Barbara Hammer, and Axel Schneider. Predicting the level of coactivation of one muscle head from the other muscle head of the biceps brachii muscle by linear regression and shallow feedforward neural networks. In *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOSIGNALS*, pages 611–621. INSTICC, SciTePress, 2024. ISBN 978-989-758-688-0. doi: 10.5220/0012368700003657.

[3] Malte Mechtenberg, Nils Grimmelsmann, Hanno Gerd Meyer, and Axel Schneider. Surface electromyographic recordings of the biceps and triceps brachii for various postures, motion velocities and load conditions. *FH Bielefeld*, 2023. doi: 10.57720/2290.

[4] Archibald Vivian Hill. The effect of load on the heat of shortening of muscle. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 1964. doi: 10.1098/rspb.1964.0004. URL https://royalsocietypublishing.org/doi/abs/10.1098/rspb.1964.0004. Publisher: The Royal Society London.

[5] Yi-Wen Chang, Fong-Chin Su, Hong-Wen Wu, and Kai-Nan An. Optimum length of muscle contraction. *Clinical Biomechanics*, 14(8):537–542, 1999.

[6] Erica Kholinne, Rizki Zulkarnain, Yu Sun, Sungjoon Lim, Jae-Myeung Chun, and In-Ho Jeon. The different role of each head of the triceps brachii muscle in elbow extension. *Acta Orthopaedica et Traumatologica Turcica*, 52, 03 2018. doi: 10.1016/j.aott.2018.02.005.

Nonlinear Prediction in a Smart Shoe Insole

Markus Vieth MVIETH@TECHFAK.UNI-BIELEFELD.DE
Bielefeld University, Germany

Abstract

In our previous work [1], we have investigated different methods to compute the ideal placement of pressure sensors in a smart shoe insole. There, we used a linear model to predict the weight put on the foot/leg. In this work, we investigate how using a quadratic model instead changes the sensor placement and improves prediction performance.

Keywords: Intelligent wearables, model individualization

1. Introduction

Wearable sensors that aid in diagnosis and post-surgery care are becoming more common. One example is a shoe insole equipped with several pressure sensors that can compute the weight put on the foot/leg (e.g. while walking) and warn if it is overstrained. In [1] we have investigated different methods for computing an optimal positioning of the pressure sensors, assuming a customized, linear postprocessing of the sensor readings. In this work, we further investigate how nonlinear postprocessing changes the results. In contrast to a linear model, a nonlinear model allows interactions between sensors. We especially focus on whether the nonlinear model results in different sensor positions, whether the predictions are more accurate, and how the sensors contribute to the prediction over the course of a stance phase.

2. Experiments

In this work we test the global optimization methods differential evolution [2] and simulated annealing [3] as methods for computing the optimal sensor positioning, since these two methods performed best in [1]. We keep the objective with customized, linear postprocessing:

$$\mathbf{s}_1^* = \underset{|\mathbf{s}|_\mathbf{1}=\mathbf{n}}{\arg\min} \left(\sum_p \min_{\mathbf{w_p}} \|\mathbf{y_p} - \mathbf{X_{s,p}} \mathbf{w_p}\|_2^2 \right)$$

where $\mathbf{s}$ refers to the selected sensor placement, the inner part of the objective constitutes an individual linear least-squares problem mapping sensor values $\mathbf{X_{s,p}}$ of person p at positions $\mathbf{s}$ to the individual target, and the outer part minimizes the residuals, summed over all persons, by adapting the sensor positions $\mathbf{s}$. We now introduce a second objective by replacing the linear model with a nonlinear model f with learnable, per-person parameters $\mathbf{w_p}$:

$$\mathbf{s}_2^* = \underset{|\mathbf{s}|_\mathbf{1}=\mathbf{n}}{\arg\min} \left(\sum_p \min_{\mathbf{w_p}} \|\mathbf{y_p} - f_{\mathbf{w_p}}\left(\mathbf{X_{s,p}}\right)\|_2^2 \right)$$

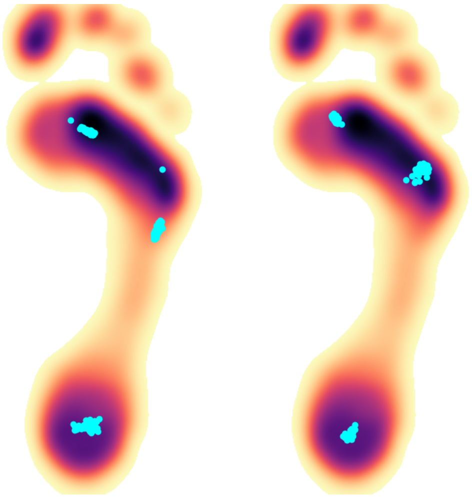
Since we assume that the processing happens on hardware with low computational power, we choose f to be a quadratic polynomial, as it is fast to compute. In contrast to [1], we neither use a constant bias for the linear nor for the quadratic model because it seems counterintuitive that the model would give a nonzero prediction with zero pressure sensor readings. The data and other setup is the same as in [1], including the crossvalidation scheme.

3. Results

Figure 1 and 2 show where sensors are placed often. For three sensors (Fig. 1), there are only small differences, namely that with the quadratic model, the sensor cluster on the outer side of the foot is moved up, to the area with higher pressure. For five sensors (Fig. 2), we see that none are placed under the arch of the foot when a quadratic model is used.

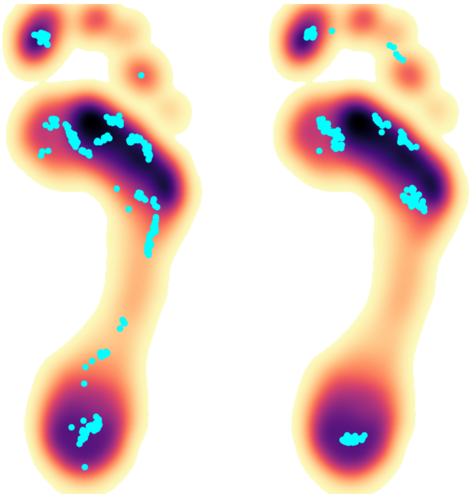
Figures 3 exemplifies how the different sensors contribute to the prediction during one step/stance phase. Note that for the quadratic model, some second order components contribute negatively to the prediction. Overall, the prediction of the quadratic model seems to be better.

Median test scores are shown in table 1. It is visible that the quadratic model leads to higher scores compared to the linear model, such that the quadratic model can achieve similar scores as the linear model



(a) Linear (b) Quadratic

Figure 1: Three sensors, differential evolution.



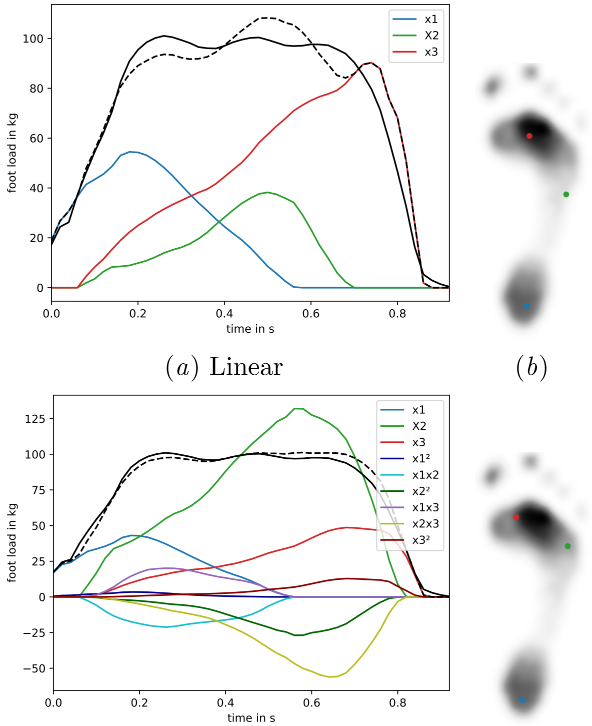
(a) Linear (b) Quadratic

Figure 2: Five sensors, simulated annealing.

with two to three fewer sensors. However for eight sensors, the quadratic model gives no advantage over the linear model.

4. Conclusion

Using a quadratic model for the prediction seems promising, especially when looking at the test scores.



(a) Linear (b)

(c) Quadratic (d)

Figure 3: Contributions of sensors during one step/stance phase. Solid black line: ground truth, dashed line: prediction.

Table 1: Median R^2 test scores. n is the number of sensors.

	diff. evolution		simulated annealing	
n	linear	quadratic	linear	quadratic
3	0.933	0.966	0.934	0.966
5	0.966	0.985	0.969	0.984
8	0.987	0.986	0.987	0.986

Acknowledgments

The project has been funded by the Ministry of Culture and Science of the Federal State North Rhine-Westphalia in the frame of the project RoSe in the AI-graduate-school dataninja.nrw

References

[1] Markus Vieth, Nils Grimmelsmann, Axel Schneider, and Barbara Hammer. Efficient Sensor Selection for Individualized Prediction Based on Biosignals. In Hujun Yin, David Camacho, and Peter Tino, editors, *Intelligent Data Engineering and Automated Learning – IDEAL 2022*, pages 326–337, Cham, 2022. Springer International Publishing. ISBN 978-3-031-21753-1. doi: 10.1007/978-3-031-21753-1_32.

[2] Rainer Storn and Kenneth Price. Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *Journal of Global Optimization*, 11(4):341–359, December 1997. ISSN 1573-2916. doi: 10.1023/A:1008202821328. URL https://doi.org/10.1023/A:1008202821328.

[3] Yang Xiang, Sylvain Gubian, Brian Suomela, and Julia Hoeng. Generalized Simulated Annealing for Global Optimization: The GenSA Package. *The R Journal*, 5(1):13, 2013. ISSN 2073-4859. doi: 10.32614/RJ-2013-002. URL https://journal.r-project.org/archive/2013/RJ-2013-002/index.html.

Proceedings of the DataNinja sAIOnARA 2024 Conference 53-55 DOI: 10.11576/dataninja-1170

Bioinspired Decentralized Hexapod Control with a Graph Neural Network

Luca Hermes LHERMES@TECHFAK.UNI-BIELEFELD.DE
Bielefeld University, Germany

Barbara Hammer BHAMMER@TECHFAK.UNI-BIELEFELD.DE
Bielefeld University, Germany

Malte Schilling MALTE.SCHILLING@UNI-MUENSTER.DE
University of Münster, Germany

Abstract

Legged locomotion enables animals to navigate challenging terrains. However, it demands intricate coordination between the legs, with varying levels of information exchange depending on the task. For instance, in more demanding scenarios such as an insect climbing on a twig, greater coordination between the legs is necessary to achieve adaptive behavior. To address this challenge for legged robots, we present a concept and preliminary results of a decentralized biologically inspired controller for a hexapod robot: Based on insights of coordination influences between legs in stick insects, our approach models inter-leg information flow as message passing through a Graph Neural Network.

Keywords: Reinforcement Learning, Hexapod, Decentralized Control

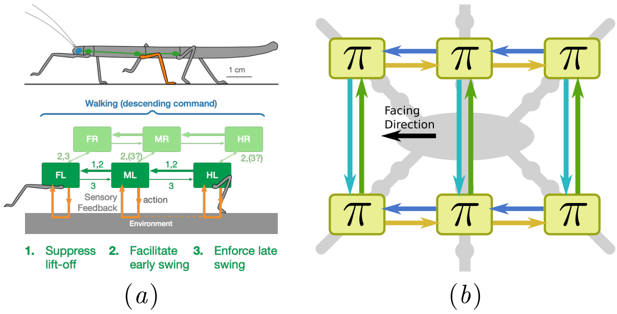
1. Introduction

Insects can traverse difficult terrain with ease while coordinating their six legs in an efficient way. This coordination manifests as a continuum of gaits that allows insects to move efficiently at different velocities. Stick insects that walk slowly exhibit the *tetrapod gait* which transitions smoothly into the *tripod gait* with increasing walking speed. Two main principles have been discovered in insect locomotion. First, insect locomotion can be modeled by a set of local rules or influences between legs [1, 2] visualized in Figure 1(a). Here, local means that influences exist only between immediately neighboring legs. Second, the same rules hold for each leg. Both of those principles indicate a decentralized system, where the same controller actuates every leg. This motivates an extension to the existing work [3, 4] that implements such a decentralized controller for a quadruped and a hexapod robot based on a reinforcement learning (RL) multi-agent framework. This work assigns separate neural networks to the four legs and concludes that information exchange between legs is required to facilitate functioning coordination. Here we want to go a step further by 1) using the exact same neural network to control every leg, which aligns with the second principle 2) utilize a graph neural network to implement inter-leg coordination and 3) discuss how to design a model that is transparent w.r.t. learned coordination rules.

2. Methods

We now outline a simple bioinspired model which can learn the tripod gait as found in many insects and later discuss avenues to improve on interpretability and to learn more diverse behaviors.

The controller is implemented as a graph neural network (GNN) [5–7]. Together with an appropriate graph structure it represents a local model where leg control only depends on the leg's own features, as well as features of neighboring legs, i.e. the first-order neighborhood. As shown in Figure 1(b), we construct a graph $\mathcal{G} = (V, E)$, where the nodes V correspond to the legs and the edges E correspond to communication channels between edges (colored arrows). This graph reflects the structure of the inter-leg rules found in biological experiments on the stick insect. Both nodes and edges are parameterized by feature vectors. The node features $\mathbf{x}_v \in \mathbb{R}^{24}$ consist of state information of the respective leg, as well as state information of the torso. The edge features $\mathbf{e}_{u,v} \in \mathbb{R}^2$ depend on the edge direction, specifically: rostrally directed edge (blue) $[1, 0]$, caudally directed edge (orange) $[-1, 0]$, contralateral edge (turquoise)



(a) (b)

Figure 1: (a) Coordination rules that have been found in stick insects, acting in the directions of the arrows. Figure adapted from [3]. (b) Our decentralized controller inspired by the coordination rules on the left. Yellow boxes represent nodes of the graph and leg policy (π). Arrows show graph structure utilized by the policies. Arrow colors denote different edge features. The robot body model shown as shaded schema in the background.

$[0, 1]$ and contralateral edge (green) $[0, -1]$. Note that neither node, nor edge features contain identifiers that uniquely identify them, thus the learned model has to learn general state and message representations for successful coordination. To ensure that control of each leg only depends on first-degree neighboring legs, the model is a single-layer GNN implemented as

$$\mathrm{msg}_{u\to v}^{t} = \phi\left((\mathbf{x}_v^t - \mathbf{x}_u^t) \parallel \mathbf{e}_{u,v}\right) \tag{1}$$

$$\mathbf{a}_v^t = \theta\left(\mathbf{x}_v^t \parallel \sum_{u\in\mathcal{N}_v} \mathrm{msg}_{u\to v}^{t}\right), \tag{2}$$

where ϕ and θ are trainable multilayer perceptrons (MLPs), $\cdot \parallel \cdot$ denotes vector concatenation and $\mathbf{a}_v^t \in \mathbb{R}^3$ denote the actions of leg v at time t. The policy is trained via proximal policy optimization (PPO) [8] in the actor-critic (A2C) flavor, where the ciritc uses the same architecture as the actor, therefore it is also local. The setting is posed as multi-agent reinforcement learning, with every leg implemented as an individual agent.

3. Results & Discussion

Figure 2 shows a preliminary result of our trained policy. The observed behavior resembles a tripod gait, where the front-left (FL), middle-right (MR), and hind-left (HL) legs move together while the other legs move in the opposite phase. From our preliminary experiments with different target velocities ($v_{\text{target}} \in [0.1, 0.8]$) we can report that the policy converges consistently to this tripod behavior.

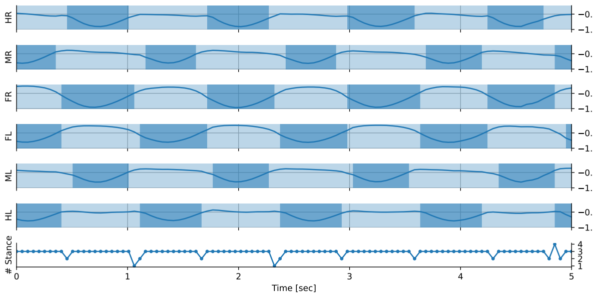


Figure 2: Hip angles (blue curves) of each leg over time shown for 5 seconds (100 simulation steps). Shaded background indicates that leg is in stance (dark) or swing mode (light). Bottom shows the number of legs in stance mode. The shown pattern corresponds to a tripod gait.

While this simple decentralized architecture replicates biological observations it remains unclear to what extend the rules found in the insect are being implemented. We hypothesize that the messages being sent by the GNN contain much more information than necessary to realize the simple rules discussed above, which might have an adverse effect on learning more diverse walking gaits. Furthermore, rules are only active at very distinct situations, e.g. when the sending leg is currently in swing mode (c.f. rule 1 in [1]). Such mechanics are not explicitly built into our model. Adding an attention mechanism as in graph attention networks [9] to limit information exchange could yield a more interpretable model w.r.t. rule learning and also foster learning.

4. Conclusion

We introduced the idea to learn leg coordination behavior exhibited by insects using graph neural networks. The preliminary results show the possibility to learn a stable tripod gait. In future work, we will investigate 1) how we can make the model more transparent with regards to the coordination rules and 2) how to promote more disverse walking behaviours with such a method.

Acknowledgments

This research was supported by the research training group "Dataninja" (Trustworthy AI for Seamless Problem Solving: Next Generation Intelligence Joins Robust Data Analysis) funded by the German federal state of North Rhine-Westphalia.

References

[1] Holk Cruse. What mechanisms coordinate leg movement in walking arthropods? *Trends in Neurosciences*, 13(1):15–21, January 1990. ISSN 01662236. doi: 10.1016/0166-2236(90)90057-H. URL https://linkinghub.elsevier.com/retrieve/pii/016622369090057H.

[2] Holk Cruse, Thomas Kindermann, Michael Schumm, Jeffrey Dean, and Josef Schmitz. Walknet—a biologically inspired network to control six-legged walking. *Neural Networks*, 11(7):1435–1447, October 1998. ISSN 0893-6080. doi: 10.1016/S0893-6080(98)00067-7. URL https://www.sciencedirect.com/science/article/pii/S0893608098000677.

[3] Malte Schilling, Andrew Melnik, Frank W. Ohl, Helge J. Ritter, and Barbara Hammer. Decentralized control and local information for robust and adaptive decentralized Deep Reinforcement Learning. *Neural Networks*, 144:699–725, December 2021. ISSN 0893-6080. doi: 10.1016/j.neunet.2021.09.017. URL https://www.sciencedirect.com/science/article/pii/S0893608021003671.

[4] Malte Schilling, Kai Konen, Frank W Ohl, and Timo Korthals. Decentralized deep reinforcement learning for a distributed and adaptive locomotion controller of a hexapod robot. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5335–5342. IEEE, 2020.

[5] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009. doi: 10.1109/TNN.2008.2005605.

[6] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016. URL http://arxiv.org/abs/1609.02907.

[7] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry, 2017.

[8] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

[9] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018.

Proceedings of the DataNinja sAIOnARA 2024 Conference 56-57 DOI: 10.11576/dataninja-1171

Improving Trust in AI Through Sustainable and Trustworthy Reporting

Raphael Fischer RAPHAEL.FISCHER@TU-DORTMUND.DE
Lamarr Institute for Machine Learning and Artificial Intelligence, TU Dortmund University, Germany

Mirko Bunse MIRKO.BUNSE@CS.TU-DORTMUND.DE
Lamarr Institute for Machine Learning and Artificial Intelligence, TU Dortmund University, Germany

Abstract

This extended abstract outlines STREP, our (S)ustainable and (T)rustworthy (REP)orting framework. It communicates performance indicators of systems that build on artificial intelligence and thus makes them more trustworthy.

Keywords: Trustworthy AI, Sustainability, Resource-awareness

1. Introduction

While artificial intelligence (AI) and machine learning (ML) are ubiquitous tools in various domains, their trustworthiness is frequently called into question [1]. One important factor for increasing trust in AI systems resides in communicating novel technological advances and results to the users of such systems. Current ways of reporting the performance of an AI system, however, often produce outcomes that are hard to reproduce, lack information on the computing setup and on the resource usage, and focus on expert users instead of non-experts. To address these issues, we recently proposed the STREP framework as an important step towards more (S)ustainable and (T)rustworthy (REP)orting [2].

This extended abstract and the associated poster summarize the key points of STREP, such as customizable reporting options, interactive controls, and labels for more abstract communication. Our work highlights the importance of resource efficiency, interactivity, comprehensibility, usability, and reproducibility in ML reporting. Through these efforts, we advance the state-of-the-art in ML reporting by promoting sustainability [3], trustworthiness, and user-centric design. We also discuss STREP within the broader context of the triangular research vision—a joint consideration of data, knowledge, and context—that we pursue at the Lamarr Institute.

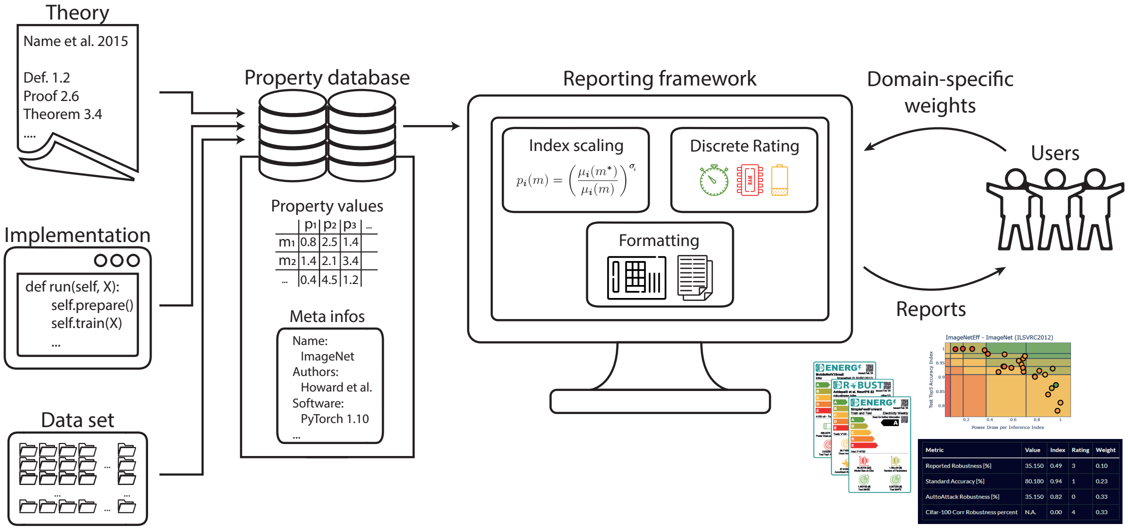
2. Sustainable and Trustworthy Reporting

Reporting the performance of a ML system requires a thorough characterization of the corresponding experiments and results. In STREP, schematically displayed in Figure 1, we denote an experimental evaluation setup as a tuple (d, t, m), which corresponds to solving a specific task t on given data d via some method m. An example would be to classify (t) a fixed number of ImageNet images (d) with MobileNetV2 (m) [4]. An evaluation of this kind results in a trained model with specific properties $\boldsymbol{p}_{(d,t,e)}$ which describe the predictive quality (e.g., accuracy, error) of the model and its resource demand (e.g., number of parameters, energy draw).

Unavoidably, these properties are subject to the execution environment e, i.e., to the software and hardware that are used during the evaluation; therefore, they are hard to compare across different execution environments. STREP solves this issue via relative index scaling, a mapping of all real-valued properties onto the unit scale to allow for straightforward comparisons and aggregations.

Since a user might not find all properties of a model to be equally relevant, ML reporting has to offer an interactive investigation of the underlying results. STREP allows uses to control the importance of method properties for their overall performance assessment, hence enhancing user engagement and supporting the understanding of the reported results. To benefit also non-expert audiences, STREP supports the generation of high-level ML labels [5] that can inform users in a more abstract and more easily comprehensible way.

We have used STREP to gain insights into existing benchmarks and open databases from various domains. Our experiments showcase how dramatically



```
def run(self, X):
    self.prepare()
    self.train(X)
    ...
```

	p1	p2	p3	...
m1	0.8	2.5	1.4	
m2	1.4	2.1	3.4	
...	0.4	4.5	1.2	

Name:
ImageNet
Authors:
Howard et al.
Software:
PyTorch 1.10
...

Metric	Value	Index	Rating	Weight
Reported Robustness [%]	35.150	0.49	3	0.10
Standard Accuracy [%]	80.180	0.94	1	0.23
AuttoAttack Robustness [%]	35.150	0.82	0	0.33
Cifar-100 Corr Robustness percent	N.A.	0.00	4	0.33

Figure 1: The proposed framework for sustainable and trustworthy reporting, originally presented in [2]

under-reported resource usage is in public databases like Papers With Code. At the same time, our experiments demonstrate how well a thorough reporting on resource demand can improve the understanding of model performance.

3. Trustworthy and Resource-Aware AI at the Lamarr Institute

STREP is a prime example of the research vision that we pursue at the Lamarr Institute for Machine Learning and Artificial Intelligence. We believe that AI systems need to be designed and implemented along three dimensions: data, knowledge, and context. This understanding can also be seen in STREP - when reporting on empirical performance measurements (data), the context of the experiments (e.g., execution environment) as well as the knowledge of the target audience need to be specifically considered.

In addition to the systematic reporting of AI performance, our institute also investigates trustworthy AI in terms of interpretability, explainability, and ethics, as well as resource-aware AI. We address these topics in diverse application fields and interdisciplinary research areas.

References

[1] Nicole Krämer, Magdalena Wischnewski, and Emmanuel Müller. Interacting with autonomous systems and intelligent algorithms–new theoretical considerations on the relation of understanding and trust. *PsyArXiv*, 2023.

[2] Raphael Fischer, Thomas Liebig, and Katharina Morik. Towards more sustainable and trustworthy reporting in machine learning. *Data Mining and Knowledge Discovery*, 2024.

[3] Aimee van Wynsberghe. Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, 2021.

[4] Raphael Fischer, Matthias Jakobs, Sascha Mücke, and Katharina Morik. A unified framework for assessing energy efficiency of machine learning. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2022.

[5] Katharina J. Morik, Helena Kotthaus, Raphael Fischer, Sascha Mücke, Matthias Jakobs, Nico Piatkowski, Andreas Pauly, Lukas Heppe, and Danny Heinrich. Yes we care!-certification for machine learning methods through the care label framework. *Frontiers in Artificial Intelligence*, 2022.

Proceedings of the DataNinja sAIOnARA 2024 Conference 58-61 DOI: 10.11576/dataninja-1172

Beyond Trial and Error in Reinforcement Learning

Moritz Lange MORITZ.LANGE@INI.RUB.DE
Institute for Neural Computation, Ruhr University Bochum, Germany

Raphael C. Engelhardt RAPHAEL.ENGELHARDT@TH-KOELN.DE
TH Köln, Germany

Wolfgang Konen WOLFGANG.KONEN@TH-KOELN.DE
TH Köln, Germany

Laurenz Wiskott LAURENZ.WISKOTT@RUB.DE
Institute for Neural Computation, Ruhr University Bochum, Germany

Abstract

In this work, we address the trial-and-error nature of modern reinforcement learning (RL) methods by investigating approaches inspired by human cognition. By enhancing state representations and advancing causal reasoning and planning, we aim to improve RL performance, robustness, and explainability. Through diverse examples, we showcase the potential of these approaches to improve RL agents.

Keywords: Reinforcement learning, representation learning, reasoning

1. Introduction

In reinforcement learning (RL), an agent learns to act in an environment to achieve some goal. RL problems are framed as Markov decision processes (MDPs), defined by states ($\mathcal{S}$), actions ($\mathcal{A}$), transition probabilities ($\mathcal{P}$), and rewards ($\mathcal{R}$).

RL algorithms solve RL tasks by learning a mapping $\mathcal{S} \mapsto \mathcal{A}$, i.e. finding suitable actions for states, to maximize accumulated rewards. They can be model-free (e.g. [1–4]) or model-based (e.g. [5, 6]). In model-free algorithms, the agent balances exploration and exploitation while trying actions to learn a value function for state-action pairs, which is then used to sample actions. Model-based RL learns an internal model of the environment, which is used by the agent as a simulator for planning. Such environment models are usually forward models, i.e. they provide agents with the basic reasoning capacity of rolling out hypothetical actions. Model-based RL has the advantage that agents require less actual real-world experience and the disadvantage that they require a reliable model of the environment.

Both kinds of methods rely on trial and error by the agent. Agents use black-box neural networks to map raw inputs, e.g. images, to state-action values without any sophisticated understanding of state information. Networks are trained directly on the RL task of optimising return, without incentives to learn representations that could help reasoning about the environment and its dynamics. Furthermore, the state-action pairs are considered independent and not treated as part of targeted action sequences.

Humans, on the other hand, process abstract representations of information, can contextualize information and reason causally about steps required to reach a goal. In this work, we aim to bridge this gap by showing how cognition-inspired methods can improve performance, sometimes even make tasks possible in the first place, and benefit robustness and explainability.

2. State Representations

We showcase the benefits of appropriate representations with two examples. First, we demonstrate the use of representing location and heading in visual navigation as in Lange et al. [7]. In particular, we use three representation learning methods with a PPO agent [2]: (i) Slow feature analysis (SFA) [8], which is able to extract location and heading from visual input, (ii) principle component analysis (PCA), able to extract heading but not location and (iii) convolutional neural networks (CNNs), the go-to approach in this context, trained on the RL task jointly with the agent. CNNs do not encode either location or heading. Figure 1 shows how SFA outperforms the other representations. For more details, see Lange et al. [7].

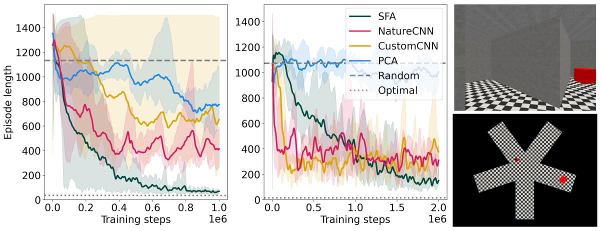


Figure 1: Performance of SFA, PCA and two CNN representations on a star maze task with fixed (left) or random (right) goal position. The images on the right show the agent's observation (top) and top view of the maze (bottom; triangle: agent, cube: goal). Image modified from [7].

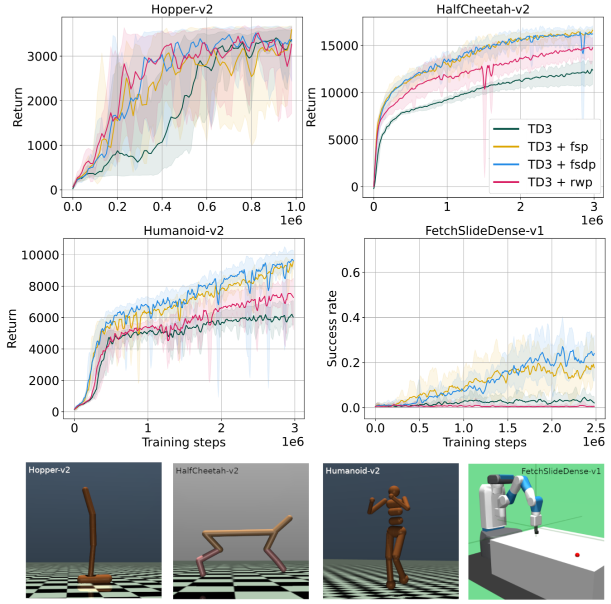


Figure 2: Performance of the TD3 algorithm with fsp, fsdp and rwp representations (see Section 2) on various environments. Image modified from [9].

Figure 2 goes beyond visual navigation. It compares different auxiliary tasks (additional tasks other than reward maximization) for representation learning, in various non-visual environments. According to our findings in Lange et al. [9], which are summarized in Figure 2, forward state (difference) prediction (fsp/fsdp) outperforms reward prediction (rwp) and baseline RL representation learning without any auxiliary task. This suggests that there is a benefit in learning representations that are generally optimized for modeling the dynamics of the environment.

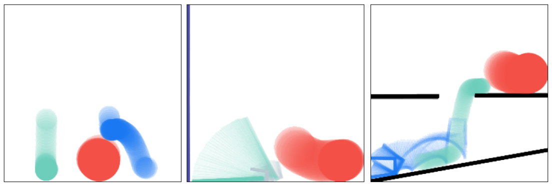


Figure 3: Time evolution of different numbers of interacting objects, all generated with the same trained denoising diffusion model. This environment, Phyre, is from Bakhtin et al. [15].

3. Reasoning and Planning

Models of the environment can be statistical or causal [10]. The former is easier to learn, but the latter is more robust and generalizes better to out-of-distribution situations [11]. Both benefit from representations with high-level, causal variables. Such representation learning methods already exist and are used, for instance, in physical reasoning [12]. However, only recent gradient-based causal discovery methods are efficient and scalable enough for RL. Unfortunately, in ongoing work, we (and others [13]) found that some current approaches might fail due to various natural effects in data distributions. Still, we consider the field of gradient-based causal discovery a promising direction for causal reasoning in RL.

Beyond causality, a smart planning algorithm should be able to plan both forward from a state and backward from a goal to incorporate both constraints. This is necessary to eliminate the need for trial and error. Recently, Janner et al. [14] have made an exciting step in this direction with denoising diffusion models for invertible planning. We extended their work with a model that can handle variable time horizons and numbers of objects during inference, as well as object interactions (see Figure 3). After reintroducing start and goal conditioning from [14], this model will be useful not only for planning: An agent can also use it to reason about past and future, to explore hypothetical options or to learn from mistakes.

4. Conclusion

We illustrated through different examples how informative representations, as well as causal and invertible reasoning have the potential to improve RL agents that often rely on trial and error. Through their alignment with human reasoning, our methods can also provide robustness and explainability.

Acknowledgments

This research was supported by the research training group "Dataninja" (Trustworthy AI for Seamless Problem Solving: Next Generation Intelligence Joins Robust Data Analysis) funded by the German federal state of North Rhine-Westphalia.

The ongoing work on causal models is conducted jointly with **Tim Schwabe** and **Maribel Acosta** (both TU Munich).

The ongoing work on diffusion models is conducted jointly with **Andrew Melnik** (University of Bielefeld).

References

[1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[2] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[3] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

[4] Arsenii Kuznetsov, Pavel Shvechikov, Alexander Grishin, and Dmitry Vetrov. Controlling overestimation bias with truncated mixture of continuous distributional quantile critics. In *International Conference on Machine Learning*, pages 5556–5566. PMLR, 2020.

[5] Anusha Nagabandi, Gregory Kahn, Ronald S Fearing, and Sergey Levine. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 7559–7566. IEEE, 2018.

[6] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pages 2555–2565. PMLR, 2019.

[7] Moritz Lange, Raphael C Engelhardt, Wolfgang Konen, and Laurenz Wiskott. Interpretable brain-inspired representations improve rl performance on visual navigation tasks. In *AAAI workshop eXplainable AI approaches for Deep Reinforcement Learning*, 2024.

[8] Mathias Franzius, Niko Wilbert, and Laurenz Wiskott. Invariant object recognition with slow feature analysis. In *International Conference on Artificial Neural Networks*, pages 961–970. Springer, 2008.

[9] Moritz Lange, Noah Krystiniak, Raphael C Engelhardt, Wolfgang Konen, and Laurenz Wiskott. Improving reinforcement learning efficiency with auxiliary tasks in non-visual environments: A comparison. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 177–191. Springer, 2023.

[10] Judea Pearl. *Causality*. Cambridge university press, 2009.

[11] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[12] Andrew Melnik, Robin Schiewer, Moritz Lange, Andrei Ioan Muresanu, Animesh Garg, Helge Ritter, et al. Benchmarks for physical reasoning ai. *Transactions on Machine Learning Research*, 2023.

[13] Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34:27772–27784, 2021.

[14] Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, pages 9902–9915. PMLR, 2022.

[15] Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick.

Phyre: A new benchmark for physical reasoning. *Advances in Neural Information Processing Systems*, 32, 2019.

DOI: 10.11576/dataninja-1173

Closing the Loop with Concept Regularization

Andres Felipe Posada-Moreno ANDRES.POSADA@DSME.RWTH-AACHEN.DE
Sebastian Trimpe TRIMPE@DSME.RWTH-AACHEN.DE
Institute for Data Science in Mechanical Engineering (DSME), RWTH Aachen University, Aachen, Germany

Abstract

Convolutional Neural Networks (CNNs) are widely adopted in industrial settings, but are prone to biases and lack transparency. Explainable Artificial Intelligence (XAI), particularly through concept extraction (CE), allows for global explanations and bias detection, yet fails to offer corrective measures for identified biases. To bridge this gap, we introduce Concept Regularization (CoRe), which uses CE capabilities alongside human feedback to embed a regularization term during retraining. CoRe allows for the adjustments in model sensitivities based on identified biases, aligning model prediction process with expert human assessments. Our evaluations on a modified metal casting dataset demonstrate CoRe's efficacy in bias mitigation, highlighting its potential to refine models in practical applications.

Keywords: Explainable Artificial Intelligence, Concept Extraction, Concept Learning

1. Introduction

Convolutional Neural Networks (CNNs) are extensively used in industrial applications, yet they are opaque and prone to biases and shortcut learning. Explainable Artificial Intelligence (XAI), particularly through concept extraction (CE), offers tools to dissect these models, explain their prediction processes, and detect biases. However, a significant gap remains: CE methods can tell us if a model's predictions are based on the wrong reasons, but offer no solutions for correcting these errors.

XAI research offers two recourse paths: integrating local explanations with intensive human feedback into training loss [1], and ante-hoc techniques like concept bottleneck networks that also require specific architectures [2]. These methods are either labor-intensive or unsuitable for already trained models, with no recourse mechanisms for CE methods [3–5].

To address these limitations, we introduce the method CoRe (Concept Regularization), which uses the concept localization capabilities of ECLAD [4] combined with human feedback to integrate a regularization term in the retraining process of a model. This approach utilizes concept masks from ECLAD [4] to identify and penalize model sensitivities in undesired areas, using a single human feedback input to influence model behavior across the entire dataset.

This abstract works on preliminary results towards a concept-based alignment method. In particular, we introduce the method CoRe, extending CE to provide recourse in bias mitigation. We explore the feasibility of our approach through experiments on a modified metal casting dataset, showing significant mitigation of biases.

2. Concept Regularization (CoRe)

We introduce Concept Regularization (CoRe) as shown in Figure 1, an extension of concept-based explanation methods designed to improve model alignment with human feedback. CoRe is a teacher-student framework, which uses the frozen original model (f) as the teacher to guide the training of a new or adjusted model (f'). Initially, the teacher model is analyzed using ECLAD [4] to extract a set of concepts (C), which are then presented to users to gather feedback, forming a modified set of concepts (C^h) containing only the concepts to adjust and their importance scores ($I^h_{c_j}$). This feedback serves to identify and localize undesired biases, allowing the regularization of the sensitivity of the model in these locations during the retraining of the student model.Finally, the student model is re-evaluated to measure improvements in alignment against the teacher, employing a novel Importance Alignment Score (IAS).

In the retraining phase, we penalize the gradient of the model in the regions containing the undesired concepts, using the term below:

$$L_{\text{CoRe}}(x_i, y_i, f', C^h) = \sum_{c_j \in C^h} \frac{\|\nabla_x g(f'(x_i)) \odot m^{c_j}_{x_i}\|_2}{\|m^{c_j}_{x_i}\|_2}, \tag{1}$$

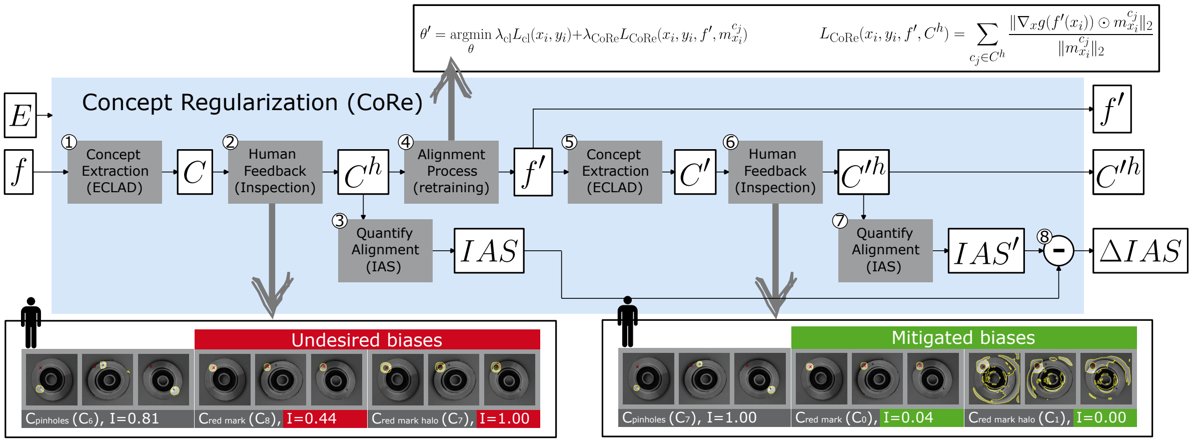

$\theta' = \underset{\theta}{\operatorname{argmin}}\, \lambda_{\mathrm{cl}} L_{\mathrm{cl}}(x_i, y_i) + \lambda_{\mathrm{CoRe}} L_{\mathrm{CoRe}}(x_i, y_i, f', m_{x_i}^{c_j})$
$L_{\mathrm{CoRe}}(x_i, y_i, f', C^h) = \sum_{c_j \in C^h} \frac{\|\nabla_x g(f'(x_i)) \odot m_{x_i}^{c_j}\|_2}{\|m_{x_i}^{c_j}\|_2}$
Concept Regularization (CoRe)
E
f
1 Concept Extraction (ECLAD)
C
2 Human Feedback (Inspection)
C^h
3 Quantify Alignment (IAS)
IAS
4 Alignment Process (retraining)
f'
5 Concept Extraction (ECLAD)
C'
6 Human Feedback (Inspection)
C'^h
7 Quantify Alignment (IAS)
IAS'
8 -
ΔIAS
f'
C'^h
Undesired biases
Cpinholes (C6), I=0.81
Cred mark (C8), I=0.44
Cred mark halo (C7), I=1.00
Mitigated biases
Cpinholes (C7), I=1.00
Cred mark (C0), I=0.04
Cred mark halo (C1), I=0.00

Figure 1: *Concept Regularization (CoRe).* CoRe starts with the Concept Extraction from the model f using ECLAD, followed by human feedback on these concepts to identify biases, and alignment assessment (steps 1-3). The model is then retrained to align with this feedback, and improvements are measured using the Importance Alignment Score (IAS) and its change ΔIAS (steps 4-8).

where $m_{x_i}^{c_j}$ is the concept mask of the input x_i and concept c_j computed with the teacher model f, $\nabla_x g(f'(x_i))$ is the gradient of the wrapped student model f' with respect to the input x, and g is the wrapping function $g(y) = \|y \cdot \mathbf{1}^\top - \mathbf{1} \cdot y^\top\|_2$ introduced in [5].

The IAS quantifies discrepancies between the model's sensitivity to concepts I_{c_j} and the human-assigned importance ratings $I_{c_j}^h$:

$$IAS = \frac{1}{n_c} \sum_{c_j \in C^h} \left| I_{c_j} - I_{c_j}^h \right|, \tag{2}$$

where I_{c_j} represents the importance score assigned by the model for concept c_j, and $I_{c_j}^h$ is the human-assigned importance score for that concept.

This method enables both the retraining of existing models and the use of a teacher model to guide the training of new models with different architectures, learning from the teacher's mistakes.

3. Preliminary Results

We validate the Concept Regularization (CoRe) method using a synthetic dataset and a modified metal casting dataset, presenting results primarily from the latter. The metal casting dataset was altered to include a red mark (bias) alongside desired classification features (pinholes). For our evaluation, we employed a DenseNet121 model, trained to convergence on this dataset. After model explanation and inspection, two concepts were identified as biases: the red mark and its surrounding halo. We then applied CoRe in three settings: *retraining* using the teacher's architecture and weights, *fine-tuning* the classification head only, and training a *new model* from scratch. We tested various learning rates and scaling factor of regularization loss λ_{CoRe}, presenting the best results below. We evaluated the reduction of biases and the change in IAS, with results detailed in Table 1 and the example in Figure 1.

Case	$c_{\mathbf{pinholes}}$	$c_{\mathbf{mark}}$	$c_{\mathbf{halo}}$	**Δ IAS**
Teacher	0.81	0.44	1.00	-
Retraining	1.00	0.10	0.00	0.67
Fine-tuning	1.00	0.04	0.00	0.7
New model	1.00	0.12	0.51	0.40

Table 1: Importance scores and importance improvement (Δ IAS) after applying CoRe.

Our findings indicate that CoRe significantly minimized the model's reliance on the biases (mark and halo), redirecting focus towards the genuine defect characteristics (pinholes). This improvement was seen across the three settings, showing CoRe's versatility, and its effectiveness in addressing biases in industrial datasets.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy -– EXC-2023 Internet of Production -– 390621612.

References

[1] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 2662–2670. AAAI Press, 2017. URL https://www.ijcai.org/proceedings/2017/371.

[2] Nishad Singhi, Jae Myung Kim, Karsten Roth, and Zeynep Akata. Improving Intervention Efficacy via Concept Realignment in Concept Bottleneck Models, 2024. URL http://arxiv.org/abs/2405.01531.

[3] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards Automatic Concept-based Explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/77d2afcb31f6493e350fca61764efb9a-Abstract.html.

[4] Andrés Felipe Posada-Moreno, Nikita Surya, and Sebastian Trimpe. ECLAD: Extracting Concepts with Local Aggregated Descriptors. *Pattern Recognition*, page 110146, 2023. URL https://www.sciencedirect.com/science/article/pii/S0031320323008439.

[5] Andrés Felipe Posada-Moreno, Kai Müller, Florian Brillowski, Friedrich Solowjow, Thomas Gries, and Sebastian Trimpe. Scalable Concept Extraction in Industry 4.0. In *Explainable Artificial Intelligence*, pages 512–535. Springer Nature Switzerland, 2023. URL https://link.springer.com/chapter/10.1007/978-3-031-44070-0_26.

DOI: 10.11576/dataninja-1174

Provable Guarantees for Deep Learning-Based Anomaly Detection through Logical Constraints

Tim Katzke[*] tim.katzke@tu-dortmund.de
Member of the Research Center "Trustworthy Data Science and Security" of the University Alliance Ruhr, TU Dortmund University, Germany

Simon Lutz[*] simon.lutz@tu-dortmund.de
Member of the Research Center "Trustworthy Data Science and Security" of the University Alliance Ruhr, TU Dortmund University, Germany

Emmanuel Müller emmanuel.mueller@cs.tu-dortmund.de
Professor for Data Science and Data Engineering, Research Center "Trustworthy Data Science and Security" of the University Alliance Ruhr, TU Dortmund University, Germany

Daniel Neider daniel.neider@cs.tu-dortmund.de
Professor for Verification and Formal Guarantees of Machine Learning, Research Center "Trustworthy Data Science and Security" of the University Alliance Ruhr, TU Dortmund University, Germany

Abstract

Incorporating constraints expressed as logical formulas and based on foundational prior knowledge into deep learning models can provide formal guarantees for the fulfillment of critical model properties, improve model performance, and ensure that relevant structures can be inferred from less data. We propose to thoroughly explore such logical constraints over input-output relations in the context of deep learning-based anomaly detection, specifically by extending the capabilities of the MultiplexNet framework.

Keywords: Anomaly Detection, Logical Constraints, Formal Guarantees

Deep neural networks have established themselves as the state-of-the-art in numerous applications, excelling in areas such as image recognition [1] and various natural language processing tasks [2], often even surpassing human expert performance. Motivated by their impressive success, (deep) neural networks have also been increasingly used for anomaly detection in recent years [3]. Anomaly detection describes the task of identifying patterns in data that diverge significantly from the expected behavior [4] and plays an important role in many application domains like cyber security, medicine, and autonomous (chemical) plants, to name but a few [5].

Unsupervised approaches to anomaly detection rely on unlabeled data, presumed to consist of normal samples with at most minor contamination by anomalies. Based on this, the objective of the neural network is to derive an inherent structure of normality - essentially defining what is expected behavior for unseen data. While this setting is widely used, recent work demonstrate that incorporating even small amounts of prior knowledge can significantly enhance anomaly detection performance [6, 7]. In the semi-supervised setting, this is achieved by providing just a few labeled samples to guide the model. However, this does not reliably solve a more general problem introduced by the use of deep neural networks.

Despite their overall outstanding performance, deep learning-based solutions are often brittle and prone to errors [8]. Even minor modifications to an input, such as noise or adversarial perturbations, can lead to significant behavioral changes and, consequently, alter the output of a neural network. This lack of so-called adversarial robustness [9] can be a severe problem when neural networks are employed in safety-critical applications where erroneous assessments could lead to substantial financial losses, environmental damage or even harm a human life. Therefore, it is essential to ensure that these models work safely and reliable before deploying them.

In recent years, a portfolio of formal verification methods has emerged to provide guarantees on the decision making of neural networks [10–12]. Given a

[*] These authors contributed equally

neural network and a (safety-critical) property, they mathematically prove or disprove that the network fulfills the desired property. However, these verification techniques usually require a network to already be trained and provide no mechanism for fixing or learning models. Tailored toward counteracting the lack of adversarial robustness, a wide variety of techniques for training more robust models have been proposed. Most of these techniques rely on either enhancing the training data by injecting specific data augmentations [13] or on adding verification-inspired regularization terms to the loss function [14]. As these approaches can only provide empirical guarantees, more sophisticated techniques integrate logical constraints into the architecture of neural networks [15]. Ensuring the compliance of these constraints provides provable guarantees on the behaviour of the networks. While most of this research focuses on feed-forward networks trained in a supervised learning setting, there is a lack of methods for neural networks used in anomaly detection.

We aim to overcome this gap by directly integrating logical constraints as a means to encode prior knowledge into deep learning-based anomaly detection. In addition to provably guarantee compliance with predefined model decision-making requirements [16], logical constraints can enhance performance [17] and reduce the dependency on large amounts of (labelled) data [14], both of which naturally benefit the inherent complexity of anomaly detection on complex real world data.

MultiplexNet [16] is a method that implements logical constraints on model outputs, encoding them as quantifier-free linear arithmetic formulas in disjunctive normal form (DNF). Provided that the output domain adheres to previously known restrictions, these constraints are provably guaranteed. The augmented output layer of the neural network applies a separate transformation for each term in the DNF ensuring their respective satisfaction, thereby producing equally many constrained outputs. Consequently, each constrained output satisfies the overarching DNF. Similar to the functionality of a multiplexor in logical circuits, a latent categorical variable is optimized to select the transformation for a given input.

For a proof-of-concept, we will first apply an adapted version of MultiplexNet to a simplified variant of a complex, real-world tabular dataset to conduct anomaly detection. This dataset comprises survey results in which participants were asked to accept or reject recommendations for the approval of benefit subsidies to unemployed job-seekers. These job-seeker profiles were synthetically generated for the survey and characterized by a combination of various features like work experience, communication skills or county of origin, while the recommended decisions were biased with respect to a subset of these features. The objective of the anomaly detection task is to identify anomalies in the sense of unexpected participant reactions to specific model decisions presented to them.

In general, the MultiplexNet architecture supports encoding any property which can be specified in the first-order fragment of quantifier-free linear real arithmetic as logical constraints over the model outputs. We propose to extend this architecture to input-output relationships, which may define some basic patterns of (a)normal behavior as a way of provably robust incorporation of prior knowledge directly into the learning process. During our preliminary experimental setup we will start by providing a set of logical constraints which function as a sanity check for the anomaly detector and guide it towards expecting some principles of rationality. For instance, we include a constraint which enforces that our model expects a high acceptance rate whenever a good job-seeker candidate (i.e. someone with a high average score on positive features like communication skills) has been recommend to be granted a benefit.

As further course of our research, we aim to evaluate this approach with an extended variant of the aforementioned survey dataset as well as chemical process data, employing more sophisticated logical constraint setups in the process. Additionally, we plan to explore alternative methods for incorporating logical constraints into deep learning-based anomaly detection that still guarantee to uphold predefined model properties based on expert knowledge.

Acknowledgments

This work has been financially supported by Deutsche Forschungsgemeinschaft, DFG Project number 459419731, and the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance centers within the UA Ruhr (https://uaruhr.de).

References

[1] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 2021.

[4] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[5] Ali Bou Nassif, Manar Abu Talib, Qassim Nasir, and Fatima Mohamad Dakalbab. Machine learning for anomaly detection: A systematic review. *Ieee Access*, 9:78658–78700, 2021.

[6] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. *arXiv preprint arXiv:1906.02694*, 2019.

[7] Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. *Advances in Neural Information Processing Systems*, 35:32142–32159, 2022.

[8] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[10] Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J. Zico Kolter. Beta-crown: Efficient bound propagation with per-neuron split constraints for neural network robustness verification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29909–29921. Curran Associates, Inc., 2021.

[11] Guy Katz, Derek A Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, Shantanu Thakoor, Haoze Wu, Aleksandar Zeljić, et al. The marabou framework for verification and analysis of deep neural networks. In *Computer Aided Verification: 31st International Conference, CAV 2019, New York City, NY, USA, July 15-18, 2019, Proceedings, Part I 31*, pages 443–452. Springer, 2019.

[12] Hoang-Dung Tran, Xiaodong Yang, Diego Manzanas Lopez, Patrick Musau, Luan Viet Nguyen, Weiming Xiang, Stanley Bak, and Taylor T Johnson. Nnv: the neural network verification tool for deep neural networks and learning-enabled cyber-physical systems. In *International Conference on Computer Aided Verification*, pages 3–17. Springer, 2020.

[13] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018.

[14] Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Broeck. A semantic loss function for deep learning with symbolic knowledge. In *International conference on machine learning*, pages 5502–5511. PMLR, 2018.

[15] Eleonora Giunchiglia, Mihaela Catalina Stoian, and Thomas Lukasiewicz. Deep learning with logical constraints. *arXiv preprint arXiv:2205.00523*, 2022.

[16] Nick Hoernle, Rafael Michael Karampatsis, Vaishak Belle, and Kobi Gal. Multiplexnet: Towards fully satisfied logical constraints in neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 5700–5709, 2022.

[17] Tao Li and Vivek Srikumar. Augmenting neural networks with first-order logic. *arXiv preprint arXiv:1906.06298*, 2019.

Proceedings of the DataNinja sAIOnARA 2024 Conference 69-71 DOI: 10.11576/dataninja-1175

Study on the Influence of Texture Variation on the Validation Performance of a Synthetically Trained Object Detector

Alexander Moriz A.MORIZ@WZL-IQS.RWTH-AACHEN.DE
Dominik Wolfschläger D.WOLFSCHLAEGER@WZL-IQS.RWTH-AACHEN.DE
WZL-IQS at RWTH Aachen University, Aachen, Germany
Robert H. Schmitt R.SCHMITT@WZL-IQS.RWTH-AACHEN.DE
WZL-IQS at RWTH Aachen University, & Fraunhofer Institute for Production Technology IPT, Aachen, Germany

Abstract

In recent years, the utilization of synthetic data for the training of Deep Learning (DL) approaches has emerged as a valid alternative to the costly process of real data acquisition. Yet, the influence of the sim-to-real gap on the model performance still poses an obstacle to the broader usage of synthetic data. To investigate the major contributing factors, this study focuses on the influence of texture variation as a first step. Examining different strategies for generating synthetic validation sets for the training process of an object detector, the results of this study indicate that the sole influence of textures is insufficient to cause the observable performance gap alone.

Keywords: synthetic data, object detection, textures

1. Introduction

Although increasingly employed for the training of Deep Learning (DL) models, the broad utilization of synthetic data is still impeded by the sim-to-real gap, appearing as a performance gap of synthetically trained DL models when evaluated on real data. While strategies to reduce the impact of the sim-to-real gap are available, the potential benefit in terms of improved performance is usually reported for a dedicated test set. Following best practice, the DL model used for such an evaluation is thereby chosen based on the performance on a separate validation set, monitored during the training process. However, if the validation set has been generated synthetically following the same strategy as for the training set, this choice might be misleading. The authors hypothesize that for such cases, the sim-to-real gap affects the performance already during the training process since the optimal model for the synthetic validation set might not be well suited for real data.

The study presented in this work focuses on the influence of texture variation on the performance of DL models as one potential critical factor. Taking the detection of a custom-designed object as a typical use case, the validation set performance of an exemplary DL model is evaluated over its training process on different validation sets. In particular, three different strategies for the generation of synthetic validation sets are examined, comparing their performance with a baseline approach and the performance on a small dataset of real images.

2. Related Work

In general, there are several ways to generate synthetic data, such as crop-out-based, 3D-modelling-based, or game-engine-based approaches [1]. The main challenge for all these approaches constitutes the sim-to-real gap [1–3]. Typically, domain adaptation or domain randomization strategies are applied to minimize its influence [1, 4]. Domain adaptation focuses on the generation of photorealistic images [5], creating a realistic scene of the target environment with e.g. physics-based rendering (PBR) [4]. The domain randomization approach pursues the opposite strategy, randomizing the simulated scene strongly to achieve a better generalization of the trained DL models directly [1]. In [6], the authors propose a framework for an end-to-end realization of DL models based on task-specific synthetic data generation. To reduce the impact of the sim-to-real gap as much as possible without requiring substantial manual design effort, the proposed framework utilizes a PBR-based domain randomization approach, varying multiple simulation parameters, such as object position, lighting, and (object) textures [6]. For the latter, the publicly available CC texture dataset is used [6].

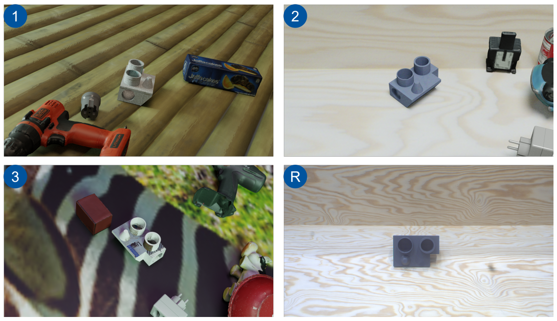


Figure 1: Examples from the considered validation sets: 1) CC textures, 2) realistic textures, 3) MS COCO textures, and R) real images.

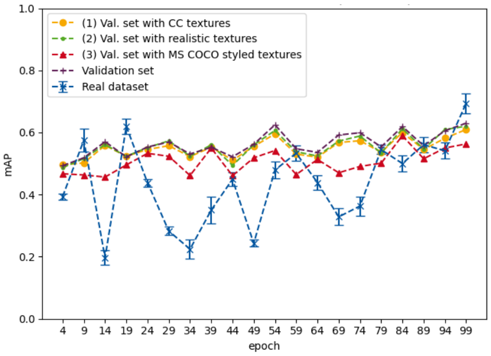


Figure 2: Performance of the individual model checkpoints for different validation sets.

3. Methodology

In this study, 6000 synthetic images have been generated utilizing the framework proposed in [6], split into 5000 images for the training- and 1000 images for the validation set. The chosen object detection model (RetinaNet, [7]) is trained for 100 epochs, storing the current model state as a checkpoint every five epochs. Afterward, the performance of the checkpoints is evaluated as individual models by determining the mean Average Precision (mAP) for all considered datasets, mimicking the performance monitoring during training for the investigated datasets.

To examine the influence of texture variation, three synthetic datasets have been generated as potential validation sets, each following a different strategy. Figure 1 visualizes an example from each dataset (sets 1 to 3) in addition to a real image (R-set). The first (1) dataset exhibits similar textures as the original synthetic validation set, utilizing two different, disjoint subsets of the CC textures for the generation of both datasets. The second (2) dataset features realistic textures, extracted from a real image, such as visualized in Figure 1, for the background and the object, respectively. The last investigated dataset (3) utilizes a small subset of plain MS COCO images [8], used randomly as textures for both the background and the object as displayed in Figure 1.

4. Results and Discussion

Figure 2 shows the performance of the individual checkpoints (models) for the different datasets. As assumed in Section 1, the performance of the real dataset (blue) exhibits a different behavior during the training process than the synthetic validation set (violet), showing distinct, not corresponding local minima and maxima. Considering the performance of the first synthetic dataset (1, yellow), no major difference to the validation set performance can be observed, indicating a good generalization capability to similar textures. Surprisingly, the performance of the second dataset (2, green) is also in agreement with the validation set, showing thus no benefit compared to the usage of the regular validation set (CC textures). Finally, the performance of the third dataset (3, red) deviates more strongly from the performance of the regular validation set. Showing on average a lower detection performance, it also features minima and maxima, which do not correspond with the validation set or the real dataset. The authors presume that this behavior might be linked to the resulting complexity of the considered textures, exhibiting patterns and artifacts, such as the zebra pattern observable in Figure 1, that are not present in the training set.

5. Conclusion and Outlook

The results of this study indicate that the sim-to-real gap, observable as the performance difference between the synthetic validation set and the real dataset, cannot be explained by the variation of texture properties alone. Future work will examine the presented results with a focus on other factors of influence such as object size or illumination in more detail. Also, additional use cases will be evaluated to support the observed findings.

References

[1] Hannah Schieber, Kubilay Can Demir, Constantin Kleinbeck, Seung Hee Yang, and Daniel Roth. Indoor synthetic data generation: A systematic review. *Computer Vision and Image Understanding*, 240:103907, 2024.

[2] Chafic Abou Akar, Jimmy Tekli, Daniel Jess, Mario Khoury, Marc Kamradt, and Michael Guthe. Synthetic object recognition dataset for industries. 1:150–155, 2022.

[3] Stefan Hinterstoisser, Olivier Pauly, Hauke Heibel, Martina Marek, and Martin Bokeloh. An annotation saved is an annotation earned: Using fully synthetic training for object instance detection. *CoRR*, abs/1902.09967, 2019.

[4] Leon Eversberg and Jens Lambrecht. Generating images with physics-based rendering for an industrial object detection task: Realism versus domain randomization. *Sensors (Basel, Switzerland)*, 21(23), 2021.

[5] Johannes Dümmel, Valentin Kostik, and Jan Oellerich. Generating synthetic training data for assembly processes. 633:119–128, 2021.

[6] Alexander Moriz, Dominik Wolfschläger, Robert H. Schmitt. Concept for a machine vision framework for production environments based on task-specific synthetic data generation. 2024. Paper accepted.

[7] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014.

DOI: 10.11576/dataninja-1176

Interpretable Machine Learning via Linear Temporal Logic

Simon Lutz SIMON.LUTZ@TU-DORTMUND.DE
Research Center "Trustworthy Data Science and Security", University Alliance Ruhr, TU Dortmund University, Germany

Daniel Neider DANIEL.NEIDER@TU-DORTMUND.DE
Professor for Verification and Formal Guarantees of Machine Learning, Research Center "Trustworthy Data Science and Security", University Alliance Ruhr, TU Dortmund University, Germany

Abstract

In recent years, deep neural networks have shown excellent performance, outperforming even human experts in various tasks. However, their inherent complexity and black-box nature often make it hard, if not impossible, to understand the decisions made by these models, hindering their practical application in high-stakes scenarios. We propose a framework for learning LTL formulas as inherently interpretable machine learning models. These models can be trained both in a supervised and unsupervised setting. Furthermore, they can easily be extended to handle noisy data and to incorporate expert knowledge.

Keywords: Explainable AI, Learning of logic formulas, Linear Temporal Logic

In the last years, Artificial Intelligence (AI) has received tremendous attention and is nowadays used in a wide variety of application domains including medicine, law enforcement, autonomous systems, and natural language processing, to name but a few. In most cases, these AI systems are based on deep neural networks with hundreds of layers and billions of parameters. Trained on large amounts of training data, these models have shown excellent performance, outperforming even human experts in various tasks. However, their inherent complexity and black-box nature often make it hard, if not impossible, to understand the decisions made by a neural network. This is especially problematic in high-stakes application and often a severe obstacle to employing AI systems in practice. Consider, for instance, a medical system assisting doctors with diagnosing patients. If the system diagnoses a specific disease, it is imperative to understand the reason to ensure correct treatment is prescribed.

To overcome this drawback of intransparent decision-making, the field of explainable artificial intelligence (XAI) has evolved in recent years. Instead of just computing the decision of a neural network, XAI methods also provide a human-readable explanation of how the network concluded this decision (see [1] for a more detailed introduction). Broadly speaking, these methods can be separated into post hoc explanations and inherently interpretable models. Post hoc explanations do not interfere with the architecture or training of a neural network but aim at inferring an explanation by analyzing its decision-making post hoc. State-of-the-art methods include using game-theoretic Shapley values as a measure for feature importance (SHAP [2]), the use of surrogate models for local explanations (LIME [3]), and the visualization of feature importance using heat maps (Grad-CAM [4]), to name but a few. Instead of training a complex neural network, the second paradigm opts for training simpler models such as decision trees, decision rules, or linear regression. Even though these models are inherently interpretable, they may lack the ability to generalize well from the training data leading to worse overall performance. Nevertheless, multiple papers have recently introduced deterministic finite automata (DFAs) as capable (i.e., on par with state-of-the-art LSTM models) yet interpretable models for sequence classification [5, 6] and anomaly detection [7].

In this paper, we follow the second paradigm and introduce a framework for learning formulas in Linear Temporal Logic (LTL) [8] as interpretable machine learning models for time series data. This specific choice of model is motivated by the following observations: first, a description of the observed, temporal behavior can often be captured in a concise logical formula; second, there is a straightforward way to translate an LTL formula to natural language, making it easy for experts and lay people to comprehend; and third, many engineers are familiar with Linear

Temporal Logic, being the de facto standard for specifying temporal properties. We consider two different learning setups, one supervised and one unsupervised learning scenario. Furthermore, we will discuss possible extensions to handle noisy data and to insert domain and expert knowledge.

In the first, supervised setup, we are given a finite set $\mathcal{S}$ of sequences together with their corresponding class labels. Then the task is to learn an LTL-based classifier that predicts the class label of yet unseen data. Toward this goal, we construct a set of minimal LTL formulas, each characterizing one of the possible output classes (thus functioning as a one-vs-rest classifier). Here, minimality refers to a minimal number of subformulas which we use to ensure high interpretability in the sense of Occam's razor (i.e., smaller formulas are generally easier to understand than larger ones [6, 9]). For each class, we construct the corresponding formula by splitting the set $\mathcal{S}$ in a one-vs-rest manner. Then, we use the algorithm proposed by Neider and Gavran [10] to infer an LTL formula from this data. In order to classify a sequence, we query each LTL formula in our set, giving us a distribution over the decisions of the one-vs-rest classifiers. Then we adopt the approximate Bayesian method of Shvo et al. [6] to infer a posterior probability distribution over the true class label and conclude a prediction.

For the second setup, we consider the task of anomaly detection, i.e. identifying patterns in data that do not conform to expected behavior [11]. As anomaly detection often plays a major role in safety-critical applications such as medical diagnosis or autonomous control, collecting anomalous data can often be dangerous and labeled data is scarce. Therefore, anomaly detection methods are often trained in an unsupervised setting where the labels of the data are a priori not known but assumed to be normal. The objective of the anomaly detection method is then to learn the underlying concept of normality in the data. Whenever unseen data diverges from this concept of normality, it will be considered an anomaly. In addition to the data, these anomaly detection methods usually also require further auxiliary information or fine-tuned hyper-parameters to prevent them from producing a degenerate solution (i.e., one that classifies all or no data as anomalies). When adopting the above concept of anomaly detection in our second, unsupervised learning scenario we rely on the approach proposed by Roy et al. [12]. Given a set $\mathcal{S}$ of unlabeled sequences and a size bound n, their algorithm produces an LTL formula of size n which is language minimal with respect to $\mathcal{S}$, i.e., it accepts all sequences in $\mathcal{S}$ and no formula of the same size accepts fewer sequences (outside of $\mathcal{S}$). Here, the size n functions as an additional parameter regulating the trade-off between interpretability and capability to generalize. We can use an LTL formula learned by this algorithm to capture normality in the given data and thus to detect anomalies.

In both setups, the proposed algorithms incorporate the learning task as a constraint-solving problem. This allows them to utilize the advances and years of engineering work of modern SAT-solver. Furthermore, it provides a rich framework for further optimization and extension of the learning algorithm.

Similar to neural networks, both learning frameworks are susceptible to noisy data. While in the second setup, this may cause drastically worse performance, it can even render the learning task unsolvable in the first setup (if two inputs are the same but have different labels). One way of mitigating the effect of noise could be the incorporation of the ideas of Gaglione et al. [13] who propose a framework for learning LTL formulas from noisy data.

A second benefit of the learning framework is that it allows for easy incorporation of domain or expert knowledge. Lutz et al. [14] proposed a framework where expert knowledge in the form of a so-called sketch (i.e., a partial LTL formula) can be provided to the learning process. The learning algorithm then completes the sketch based on the given data. Combining this framework with the proposed learning frameworks allows utilizing expert knowledge to improve the quality of the learned LTL formulas and also speed up the learning process.

In conclusion, we introduced a framework for learning LTL formulas as inherently interpretable machine learning models both in a supervised and unsupervised setting. Furthermore, we presented two extensions of the learning framework, allowing the mitigation of noisy data and the incorporation of expert knowledge.

Acknowledgments

This work has been financially supported by Deutsche Forschungsgemeinschaft, DFG Project number 459419731, and the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance centers within the UA Ruhr (https://uaruhr.de).

References

[1] Christoph Molnar. *Interpretable machine learning.* Lulu. com, 2020.

[2] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[3] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[4] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[5] Christian Albert Hammerschmidt, Sicco Verwer, Qin Lin, and Radu State. Interpreting finite automata for sequential data. *arXiv preprint arXiv:1611.07100*, 2016.

[6] Maayan Shvo, Andrew C Li, Rodrigo Toro Icarte, and Sheila A McIlraith. Interpretable sequence classification via discrete optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9647–9656, 2021.

[7] Simon Lutz, Florian Wittbold, Simon Dierl, Benedikt Böing, Falk Howar, Barbara König, Emmanuel Müller, and Daniel Neider. Interpretable anomaly detection via discrete optimization. *arXiv preprint arXiv:2303.14111*, 2023.

[8] Amir Pnueli. The temporal logic of programs. In *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)*, pages 46–57. ieee, 1977.

[9] Rajarshi Roy, Dana Fisman, and Daniel Neider. Learning interpretable models in the property specification language. *arXiv preprint arXiv:2002.03668*, 2020.

[10] Daniel Neider and Ivan Gavran. Learning linear temporal properties. In *2018 Formal Methods in Computer Aided Design (FMCAD)*, pages 1–10. IEEE, 2018.

[11] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[12] Rajarshi Roy, Jean-Raphaël Gaglione, Nasim Baharisangari, Daniel Neider, Zhe Xu, and Ufuk Topcu. Learning interpretable temporal properties from positive examples only. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6507–6515, 2023.

[13] Jean-Raphaël Gaglione, Daniel Neider, Rajarshi Roy, Ufuk Topcu, and Zhe Xu. Learning linear temporal properties from noisy data: A maxsat-based approach. In *Automated Technology for Verification and Analysis: 19th International Symposium, ATVA 2021, Gold Coast, QLD, Australia, October 18–22, 2021, Proceedings 19*, pages 74–90. Springer, 2021.

[14] Simon Lutz, Daniel Neider, and Rajarshi Roy. Specification sketching for linear temporal logic. In *International Symposium on Automated Technology for Verification and Analysis*, pages 26–48. Springer, 2023.

Distributive Justice of Resource Allocation Through Artificial Intelligence

Paul Hellwig PAUL.HELLWIG@UNI-BIELEFELD.DE
Bielefeld University, Germany

Sophia Mann AO-PSYCHOLOGIE@UNI-BIELEFELD.DE
Bielefeld University, Germany

Günter W. Maier AO-PSYCHOLOGIE@UNI-BIELEFELD.DE
Bielefeld University, Germany

Abstract

Artificial intelligence will take over leadership functions such as rewarding employee performance. It will therefore make decisions about employee outcomes and most likely allocate different resources to employees. Resource Theory of Social Exchange distinguishes six resource classes. The theory postulates that the value of some resources depend on the identity of the provider of the resource and on the relationship with the provider. This raises the question of whether certain resources, such as the resource affiliation, have a value when they are allocated by artificial intelligence. This contribution calls for studies that investigate the value of different resources allocated by artificial intelligence in leadership functions.

Keywords: distributive justice, artificial intelligence, Resource Theory of Social Exchange

1. Background

Current research discusses that artificial intelligence (AI) will take over leadership functions and will manage human employees [1, 2]. One important leadership task is motivating personnel resources, which includes rewarding employees for their performance [1]. This means that AI will make decisions about outcomes employees receive for their work as rewards.

Research shows that in decision-making, the perceived appropriateness of outcomes is important to people affected by the decision [3]. According to the literature, the outcome of a decision is perceived as appropriate if it reflects specific allocation principles, e.g. equity rule (the outcome reflects the effort someone has put into their work and is perceived as appropriate for the completed work) or equality rule (everyone gets the same) [4, 5]. Meta-analyses [3, 6] show that the appropriateness of outcomes, referred to as distributive justice, is related to important work-related variables such as trust in the supervisor, employees' affective states and task performance.

While there has been a major focus on *how* outcomes have to be distributed to be appropriate, there has been too little focus on *what* is distributed. Resource Theory of Social Exchange [7] can be used to focus on *what* is distributed. This theory describes six resource classes that can be exchanged by two parties. The theory differentiates the following resource classes: status, affiliation (also referred to as love by [7]), services, information, money, and goods [7, 8]. Status is defined as an evaluative judgment that conveys prestige, regard, or esteem [7]. Affiliation is an expression of affectionate regard, warmth, or comfort [7, 8]. Services describes activities that affect the body or belongings of another person [7]. Information is advice, opinions or, instruction [7]. Money is any coin, currency, or token [7]. Goods are tangible products, objects, or materials [7]. The theory postulates that the value of status, affiliation, and services is influenced by the identity of the provider and the relationship with the provider, which is called particularism [9]. This is not postulated for the value of information, money or goods. According to the theory affiliation and money are maximally distinct in terms of particularism. This means that the value of affiliation depends heavily on the identity of the provider, while the value of money does not [9]. An example of affiliation allocated by a manager is an expression of congratulation for personal achievements [8]. An example for money provided by a manager is overtime compensation [8].

A very important employee behaviour is that employees exert great efforts to achieve good work performance. A crucial question is therefore how managers reward those efforts and the performance of their employees. We want to investigate distributive justice perceptions in reward allocation, where employees receive different resources from their manager for their efforts and performance.

AI as a new leadership entity poses the question whether the resources that depend on the identity of the provider have the same value for AI as for a human manager. This could be particularly crucial for the resource affiliation because employees who receive affiliation from an AI may question whether the AI understands the value of affiliation. Therefore, we also want to investigate whether the value of some resources is lower for AI as the resource provider. We will begin our investigation by comparing the reactions to affiliation and money, two resources that differ most in their dependence on the one who provides them.

2. Planned study

To test distributive justice perceptions in a reward allocation scenario, we propose the following experimental vignette study. In the study, participants will read the description of a situation in which they receive either less money or less affiliation (independent variable 1: resource) than a colleague who has shown less work effort than them. In the described scenario, the value of the resources would be indicated by which resource causes stronger negative reactions. According to the equity rule and the equality rule, the outcomes described in the vignettes are unfair, but it is not clear whether receiving less money or less affiliation results in lower distributive justice perceptions. Affiliation/money will furthermore be distributed by an AI/a human manager (independent variable 2: resource provider). This allows us to test whether the value of the resources depends on the identity of the provider. In the scenario in the vignettes, we expect lower distributive justice perceptions, when affiliation is allocated by a human manager than by an AI. We will measure distributive justice perceptions, negative affect and future work effort of participants as the dependent variables. Furthermore, we will ask participants whether they think that a human manager/an AI understands the value of affiliation/money.

3. Outlook

The next step is to conduct the study. The results will have implications for research on the perception of distributive justice in automated decision-making, as the vignette study considers the impact of different resources and the interaction effect of resource and resource provider. This is something that, to our knowledge, has not yet been investigated in research on automated decision-making. Previous studies have only compared the reactions to human and automated decision-making or the effect of different allocation principles in automated decision-making (see [10] for a review). Furthermore, our study introduces the Resource Theory of Social Exchange into the context of automated decision-making, which could inspire future studies to investigate the allocation of different resources by AI.

References

[1] Jenny S. Wesche and Andreas Sonderegger. When computers take the lead: The automation of leadership. *Computers in Human Behavior*, 101:197–209, 2019. doi: 10.1016/j.chb.2019.07.027.

[2] Niels van Quaquebeke and Fabiola H. Gerpott. The now, new, and next of digital leadership: How artificial intelligence (ai) will take over and change leadership as we know it. *Journal of Leadership & Organizational Studies*, 30(3):265–275, 2023. doi: 10.1177/15480518231181731.

[3] Jason A. Colquitt, Brent A. Scott, Jessica B. Rodell, David M. Long, Cindy P. Zapata, Donald E. Conlon, and Michael J. Wesson. Justice at the millennium, a decade later: a meta-analytic test of social exchange and affect-based perspectives. *Journal of Applied Psychology*, 98(2):199–236, 2013. doi: 10.1037/a0031757.

[4] Jason A. Colquitt. On the dimensionality of organizational justice: a construct validation of a measure. *Journal of Applied Psychology*, 86(3): 386–400, 2001. doi: 10.1037/0021-9010.86.3.386.

[5] Russell Cropanzano, David E. Bowen, and Stephen W. Gilliland. The management of organizational justice. *Academy of Management Perspectives*, 21(4):34–48, 2007. doi: 10.5465/amp.2007.27895338.

[6] Jason A. Colquitt, Donald E. Conlon, Michael J. Wesson, Cristopher O. L. H. Porter, and K. Yee Ng. Justice at the millennium: a meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology*, 86(3): 425–445, 2001. doi: 10.1037/0021-9010.86.3.425.

[7] Edna B. Foa and Uriel G. Foa. Resource theory of social exchange. In Kjell Törnblom and Ali Kazemi, editors, *Handbook of social resource theory*, pages 15–32. Springer, 2012. doi: 10.1007/978-1-4614-4175-5_2.

[8] Judi McLean Parks, Donald E. Conlon, Soon Ang, and Robert Bontempo. The manager giveth, the manager taketh away: Variation in distribution/recovery rules due to resource type and cultural orientation. *Journal of Management*, 25(5):723–757, 1999. doi: 10.1177/014920639902500506.

[9] Kjell Törnblom and Riël Vermunt. Towards integrating distributive justice, procedural justice, and social resource theories. In Kjell Törnblom and Ali Kazemi, editors, *Handbook of social resource theory*, pages 181–197. Springer, 2012. doi: 10.1007/978-1-4614-4175-5_11.

[10] Paul Hellwig and Günter W. Maier. Justice and fairness perceptions in automated decision-making—current findings and design implications. In Iris Gräßler, Günter W. Maier, Eckhard Steffen, and Daniel Roesmann, editors, *The digital twin of humans*, pages 63–92. Springer, 2023. doi: 10.1007/978-3-031-26104-6_4.

Proceedings of the DataNinja sAIOnARA 2024 Conference 78-80 DOI: 10.11576/dataninja-1178

Concept extraction for time series with ECLAD

Antonia Holzapfel ANTONIA.HOLZAPFEL@DSME.RWTH-AACHEN.DE
Andres Felipe Posada-Moreno ANDRES.POSADA@DSME.RWTH-AACHEN.DE
Sebastian Trimpe TRIMPE@DSME.RWTH-AACHEN.DE
Institute for Data Science in Mechanical Engineering (DSME), RWTH Aachen University, Aachen, Germany

Abstract

Concept Extraction (CE) methods are being increasingly used in the image domain for explaining deep learning models, which are not inherently interpretable. However, there have not been transfer studies yet for their usage in the time series domain. The purpose of this work is to explore the use of CE methods in time series. We propose to modify the ECLAD algorithm for this domain by changing the latent space representation used to extract concepts. This method is then tested on an InceptionTime model trained on the Gunpoint dataset. Preliminary results show that we can successfully extract concepts from time series models on datasets with local features and provide conceptual explanations that effectively explain how the model works.

Keywords: Explainable Artificial Intelligence, Concept Extraction

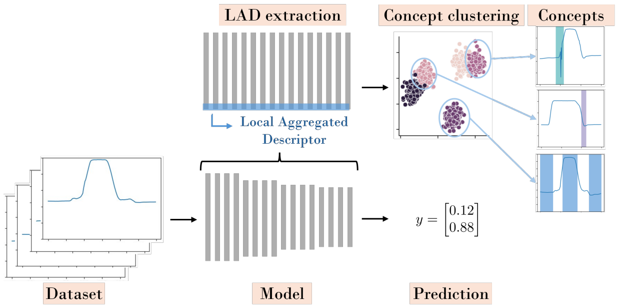

LAD extraction
Concept clustering
Concepts
Local Aggregated Descriptor
Dataset
Model
$y = \begin{bmatrix}0.12\\0.88\end{bmatrix}$
Prediction

Figure 1: Overview of ECLAD on time series

1. Introduction

Over the last years, the field of explainability (XAI) has emerged to explain what happens under the lid of black box models. Particularly, Concept Extraction (CE) methods have been developed to produce global explanations of models in the form of human-understandable "concepts". CE methods are particularly interesting because they provide a direct link between the inner representations of the model and human understandable visualizations.

While several CE methods are available for image data [1–3], no CE methods have been developed or transferred to time series. On time series, global methods like (global) Grad-CAM [4] have been used. However, they fail to provide global explanations beyond the aggregation of local ones.

In this work, we explore the use of CE methods in time series. This is, if in the time series domain, CE methods can extract patterns that are distinguishable from each other and the model is sensitive to. These patterns may be related to meaningful features of the dataset. Specifically, we modify the CE method ECLAD [1] by adapting the latent representations it uses, in coherence to time series classification models. We test the method using an InceptionTime model trained on the Gunpoint dataset. Preliminary results show that the adapted ECLAD provides meaningful model explanations.

2. CE for Time Series

In this work, we use ECLAD as a basis for CE in time series classification models. ECLAD is an algorithm that provides explanations based on three key steps: The encoding of the latent space of neural networks, the mining of patterns and the assessment of how relevant they are for the predictions of the model. We modify the first and second steps to be compatible with time series data, and explore which latent space representations make sense in this domain.

In the first step, ECLAD uses the notion of **Local Aggregated Descriptors (LADs)**, which are descriptors of how models encode a region at different levels of abstraction. These are obtained by aggregating the activation maps of multiple layers L of a CNN model. In the case of time series data, the inputs of the model have the shape $x_i \in \mathbb{R}^{w \times d}$, where

w, and d are the length and channels of an input. Since the latent representations of CNN based models inherit the dimensionality of the input vectors, our proposed LADs are timestep-wise descriptors denoted as $d_{x_{ts}} \in \mathbb{R}^{w \times 1}$.

After extracting the LADS of several inputs, a set $\Gamma = \{\gamma_{c_1}, \ldots, \gamma_{c_{n_c}}\}$ of centroids $\gamma_{c_j} \in \mathbb{R}^{1 \times c^*}$ defining the **concepts** can be obtained by applying a **mini-batch k-means algorithm**, where c^* is the sum of the number of units in the layers in L. The centroids represent similarly encoded time subsequences. For a human-understandable visualization, the concepts c_j can then be located in an input x_i by creating a **mask** $m^{c_j}_{x_{ts}} \in \mathbb{R}^{w \times 1}$ that analyzes the LAD $d_{x_i,(b)}$ at timestep b of the input, and assessing whether it belongs to a cluster γ_{c_j}.

For the last step of ECLAD, the **importance score** of a concept is determined, which quantifies the relevance of its related visual cues towards the prediction of the analyzed model. The metric used in [5] computes how sensitive a model is with respect to the regions containing each concept. For this, said metric computes the pixel-wise sensitivity $r^{c_j}_{x_i}$ of the regions of each image using a concept,

$$r^{c_j}_{x_i} = ||\nabla_x g(f(x_i)) \odot m^{c_j}_{x_i}||_1, \qquad (1)$$

where $\odot$ denotes the element-wise product between matrices, $g(y) = ||y \cdot \mathbf{1}^T - \mathbf{1} \cdot y^T||_2$ is a wrapper of the model f, $y \in \mathbb{R}^{n_k}$ is the output of f, n_k is the number of classes and $\mathbf{1}$ is the vector of ones of the same size as y. The method then proceeds to aggregate $r^{c_j}_{x_i}$ over the images in the dataset and scale the mean relevance of each concept to obtain the final importance scores I_{c_j}.

With the proposed modifications, we can transfer ECLAD to analyse time series classification models. This allows for the extraction of patterns that are meaningful for models and can be represented in human-understandable visualizations. These patterns also respect the equivariance properties of the models.

3. Preliminary Results

To validate ECLAD for time series, we trained an InceptionTime network with 20 Inception blocks on the GunPoint dataset. The dataset contains two classes ("gun" and "no gun"), which can be distinguished by the oscillations before and after the main peak. At convergence, the model obtained an accuracy of 0.975 on the validation dataset.

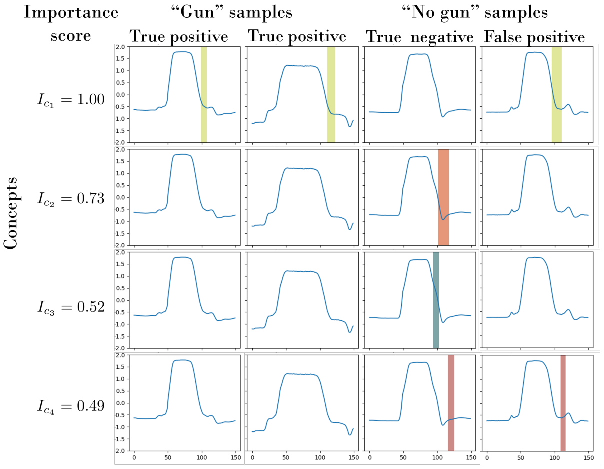


Figure 2: Concept extraction results on InceptionTime trained on the GunPoint dataset.

After training, we extracted ten concepts from the model using our modified ECLAD and visually inspected them to assess how they relate to known features of the dataset and how differentiating they are. The four most important concepts obtained are shown in Figure 2. The first identified concept represents the smooth concave curvature directly after the main peak. The structure is identified consistently, regardless of its exact position, and it strongly contributes to the positive classification. This indicates the presence of concepts within time series classification models.

The second to fourth concepts are also consistently identified and associated to instances of a specific class. These are differentiating features, but the model considers them less important in comparison.

Observing the concepts, it is possible to identify the features that the model is using and those that are causing wrong predictions. For example, for the false positive we can determine that the fourth concept was correctly identified while the lack of the second and third, as well as the presence of the first are confounding factors. These kinds of insight are useful for a user that wants to understand which biases can be present in the model or dataset.

In this extended abstract we focused on assessing the applicability of CE methods to time series. Preliminary results indicate the presence of meaningful features used by the model that our adapted ECLAD could extract.

Acknowledgments

This work is partially funded by the Deutsche Forschungsgemeinschaft (DFG) within the Priority Program SPP 2422 (TR 1433/3-1).

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy -– EXC-2023 Internet of Production -– 390621612.

References

[1] Andrés Felipe Posada-Moreno, Nikita Surya, and Sebastian Trimpe. ECLAD: Extracting Concepts with Local Aggregated Descriptors. *Pattern Recognition*, 147:110146, 2023. URL https://www.sciencedirect.com/science/article/pii/S0031320323008439.

[2] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards Automatic Concept-based Explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/77d2afcb31f6493e350fca61764efb9a-Abstract.html.

[3] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems (NeurIPS)*, 33:20554–20565, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/ecb287ff763c169694f682af52c1f309-Abstract.html.

[4] Corne Van Zyl, Xianming Ye, and Raj Naidoo. Harnessing explainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of grad-cam and shap. *Applied Energy*, 353:122079, 2024. URL https://www.sciencedirect.com/science/article/pii/S0306261923014435.

[5] Andrés Felipe Posada-Moreno, Kai Müller, Florian Brillowski, Friedrich Solowjow, Thomas Gries, and Sebastian Trimpe. Scalable Concept Extraction in Industry 4.0. In *Explainable Artificial Intelligence*, pages 512–535. Springer Nature Switzerland, 2023. URL https://link.springer.com/chapter/10.1007/978-3-031-44070-0_26.

DOI: 10.11576/dataninja-1179

Trustworthy Virtual Measurements in Battery Manufacturing

Lukas Krebs LUKAS.KREBS@WZL-IQS.RWTH-AACHEN.DE
WZL - IQS at RWTH Aachen University, Germany

Tobias Müller TOBIAS.MUELLER@WZL-IQS.RWTH-AACHEN.DE
WZL - IQS at RWTH Aachen University, Germany

Robert H. Schmitt ROBERT.SCHMITT@WZL-IQS.RWTH-AACHEN.DE
WZL - IQS at RWTH Aachen University
& Frauenhofer IPT, Germany

Abstract

The growing demand for electric cars necessitates an increase in battery production efficiency and cost-effectiveness. Through a reduction of the joint testing efforts an increase of productivity can be accomplished. To achieve the reduction, remain on a high level of quality standards and increase the informational content about current production the use of virtual measurements is examined. Ensuring the trustworthiness of virtual measurements is crucial for informed decision making, necessitating validation. This paper explores the requirements and challenges in battery manufacturing for implementing trustworthy virtual measurements. Two central requirements are identified to enable virtual measurements. Firstly, a traceability system based on the production meta-model is needed to track process parameters and quality characteristics. Secondly, a framework is proposed to facilitate reliable virtual measurements. The primary challenge for virtual measurement in battery manufacturing systems from the complexity of the process chain and products. It is crucial to assess how virtual measurements perform across various processes and to evaluate their transferability to different process parameters and products.

Keywords: Virtual Measurement, Uncertainty, Trustworthiness

1. Introduction

In 2019, the number of registered electric cars in Germany was still below 100,000. Five years later, over 1.4 million electric cars were registered in Germany, with a projected increase to ten million registered electric cars by 2030 [1, 2]. This structural shift also increases the demand for batteries for electric cars. To meet the growing demand, battery manufacturing must become more productive and cost-effective. However, high-quality standards for battery manufacturing must also be maintained. Current technology requires physical inspections throughout the battery manufacturing process, which are associated with high time and monetary costs [3, 4].

Virtual measurements can reduce the need for physical inspections, making battery manufacturing more productive and cost-effective [4]. In virtual measurements, quality characteristics are predicted based on process parameters. While early virtual measurements were conducted using polynomial equations, various machine learning algorithms are now used for quality characteristic prediction [5]. Virtual measurements are already employed in various industries, such as semiconductor manufacturing, metal processing, and textile technology [6].

2. Virtual Measurements

To make reliable decisions regarding product quality based on measurements, it is essential to consider both the measured quantity and its uncertainty [7]. Measurement uncertainty can be quantified using the Guide to the Expression of Uncertainty in Measurement (GUM) [7]. While measurement uncertainty in physical measurements is well-researched and applied in industry, virtual measurements often only specify the measured quantity without considering the corresponding measurement uncertainty [8]. In a production environment, the uncertainty of virtual measurements is crucial for trustworthiness. Thus, deterministic machine learning models used in virtual measurements should be replaced with models capable of indicating inherent measurement uncertainty. Different measurement methods possess varying degrees of measurement uncertainty, influenced by different factors. While physical measurements are affected by environmental factors such as temperature, virtual measurements are comparably reliant on the available data for training and prediction. Thus to reduce measurement uncertainty, accurate mappings between process parameters and quality characteristics are necessary. Meta-models are used for this purpose, offering consistent data structuring throughout the production process [9]. This allows for virtual measurements at various points in the production process with sufficient automation of machine learning for virtual measurements and availability of relevant data.

3. Virtual Measurements in Battery Manufacturing

Battery manufacturing offers a broad application space for virtual measurements. Quality must be checked after individual process steps in both battery cell manufacturing and module and pack assembly. Early error detection is crucial to reduce quality variations and avoid scrap [10]. To make decisions regarding product quality with virtual measurements, first the requirements and challenges for the application of virtual measurements need to be examined.

Decisions must rely on trustworthy measurements, which can be quantified by measurement uncertainty. To provide virtual measurements with comparable measurement uncertainty to physical measurements, several requirements must be met. These include systematic recording of process data and important peripheral data. Not only data from the selected process but also from previous processes influencing the process parameters affecting the quality characteristics should be available. There must be a clear mapping between process data, peripheral data, and quality characteristics. Lastly, uncertainties in measurements must be provided at each data point. Thus, for trustworthy virtual measurements in battery manufacturing, a traceability system must be used, with a meta-model of production data underlying it.

Based on the data provided by the traceability system, virtual measurements can be conducted. Cramer et al. have already investigated how measurement uncertainty can be determined in virtual measurements analogous to the stages of the GUM[8]. They discuss the steps of formulating the measurement system, propagating uncertainty, and documenting virtual measurements. The presented principle utilizes various algorithms such as Bayesian Variational Inference or Markov Chain Monte Carlo to determine virtual measurement uncertainty. These algorithms can be employed for example in Bayesian Neural Networks or Bayesian Decision Trees to enable probabilistic forecasts [11, 12]. For documentation purposes, it is crucial to store the trained model and document the coverage intervals within which the measurement values lie. This framework lays the foundation for the application of trustworthy virtual measurements in battery manufacturing. However, the framework has not yet been applied to production data. Therefore, it should be extensively tested with different algorithms and potentially expanded.

Due to the complexity of battery manufacturing process chain, many different process steps are suitable for implementing virtual measurements [13]. Therefore, virtual measurements must be conducted at several relevant quality gates, increasing implementation effort. A dedicated machine learning pipeline for virtual measurement reduces implementation effort, enabling comprehensive testing. In addition, in battery manufacturing, there is a wide range of variations in both process parameters and final products [10]. Hence, investigating the adaptability of virtual measurement models for different process parameters or products is necessary. This would eliminate the need to create new databases when establishing or modifying product lines, thus reducing the effort for trustworthy virtual measurement.

4. Conclusion

In summary the main requirements to facilitate trustworthy virtual measurements in battery manufacturing are the traceability system based on the meta model of the production and the framework to conduct virtual measurements analogous to the stages of the GUM. With complex process chains, vast variety in products and process parameters there is a broad application field for virtual measurements. For a comprehensive review on the trustworthiness of virtual measurements in comparison to physical measurements in battery manufacturing multiple scenarios need to be examined.

Acknowledgments

The project on which this report is based was funded by the Federal Ministry of Education and Research under grant number 03XP0538.

References

[1] Cleverlog-Autoteile. Anzahl der elektroautos in deutschland im jahr 2023 und prognose bis 2030 (in 1.000), 15 November 2023. URL https://de.statista.com/statistik/daten/studie/1425629/umfrage/e-auto-bestand-prognose/. In Statista. Access on 02 May 2024.

[2] KBA. Anzahl der elektroautos in deutschland von 2006 bis januar 2024, 04 March 2024. URL https://de.statista.com/statistik/daten/studie/265995/umfrage/anzahl-der-elektroautos-in-deutschland/. In Statista. Access on 02 May 2024.

[3] Aslihan Örüm Aydin, Franziska Zajonz, Till Günther, Kamil Dermenci, Maitane Berecibar, and Lisset Urrutia. Lithium-ion battery manufacturing: Industrial view on processing challenges, possible solutions and recent advances. *Batteries*, 9(11):555, 2023. doi: 10.3390/batteries9110555.

[4] Paul-Arthur Dreyfus, Foivos Psarommatis, Gokan May, and Dimitris Kiritsis. Virtual metrology as an approach for product quality estimation in industry 4.0: a systematic review and integrative conceptual framework. *International Journal of Production Research*, 60(2): 742–765, 2022. ISSN 0020-7543. doi: 10.1080/00207543.2021.1976433.

[5] Christopher I. Lang, Fan-Keng Sun, Ramana Veerasingam, John Yamartino, and Duane S. Boning. Understanding and improving virtual metrology systems using bayesian methods. *IEEE Transactions on Semiconductor Manufacturing*, 35(3):511–521, 2022. ISSN 0894-6507. doi: 10.1109/TSM.2022.3170270.

[6] Yaxuan Zhang, Li Li, and Qingyun Yu. Virtual metrology for enabling zero-defect manufacturing: a review and prospects. *The International Journal of Advanced Manufacturing Technology*, 130(7-8):3211–3227, 2024. ISSN 0268-3768. doi: 10.1007/s00170-023-12726-x.

[7] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML. Evaluation of measurement data — Guide to the expression of uncertainty in measurement. Joint Committee for Guides in Metrology, JCGM 100:2008, 2008. URL https://www.bipm.org/documents/20126/2071204/JCGM_100_2008_E.pdf/cb0ef43f-baa5-11cf-3f85-4dcd86f77bd6.

[8] Simon Cramer, Tobias Müller, and Robert H. Schmitt. Measurement uncertainty: Relating the uncertainties of physical and virtual measurements, 2024.

[9] Simon Cramer, Max Hoffmann, Peter Schlegel, Marco Kemmerling, and Robert H. Schmitt. Towards a flexible process-independent metamodel for production data. *Procedia CIRP*, 99: 586–591, 2021. ISSN 22128271. doi: 10.1016/j.procir.2021.03.112.

[10] Joscha Schnell and Gunther Reinhart. Quality management for battery production: A quality gate concept. *Procedia CIRP*, 57:568–573, 2016. ISSN 22128271. doi: 10.1016/j.procir.2016.11.098.

[11] Bakhouya Mostafa, Ramchoun Hassan, Hadda Mohammed, and Masrour Tawfik. A review of variational inference for bayesian neural network. In Tawfik Masrour, Hassan Ramchoun, Tarik Hajji, and Mohamed Hosni, editors, *Artificial Intelligence and Industrial Applications*, volume 772 of *Lecture Notes in Networks and Systems*, pages 231–243. Springer Nature Switzerland, Cham, 2023. ISBN 978-3-031-43519-5. doi: 10.1007/978-3-031-43520-1{\textunderscore}20.

[12] Giuseppe Nuti, Lluís Antoni Jiménez Rugama, and Andreea-Ingrid Cross. A bayesian decision tree algorithm, 2019.

[13] Andreas Aichele, Anselm Lorenzoni, and Jonas Lips. Method for the identification of process quality characteristics and suitable measurement systems in battery production. In Kai Peter Birke, Max Weeber, and Michael Oberle, editors, *Handbook on Smart Battery Cell Manufacturing*, pages 253–273. WORLD SCIENTIFIC, 2022. ISBN 978-981-12-4561-9. doi: 10.1142/9789811245626{\textunderscore}0013.