

PhD Workshop

Department of Informatics Engineering
University of Coimbra
19 April, 2024

Organizing Committee:
Luís Paquete and Naghmeh Ivaki

Room: A.5.4



Zoom: <https://www.dei.uc.pt/vc/phd>

13h00	Luís Paquete	<i>Welcome Session</i>
13h10	Bruno Faria	<i>Leveraging AI to Enhance Resilience in the Cloud-to-Things Continuum</i>
13h25	Gabriel Cortês	<i>Sustainable Approach to Energy Efficient Artificial Intelligence</i>
13h40	Miguel Rabuge	<i>Evolutionary Feature Engineering in Multiobjective Scenarios</i>
13h55	Dylan Perdigão	<i>Bayesian Casual Inference in Deep Spiking Neural Networks</i>
14h10	Naveed Ahmad	<i>Efficient Transformer Model with Spiking Neural Networks</i>
14h25	Gonçalo Costa	<i>Spiking Neural Networks for Epileptic Seizure Forecasting using Long-term and Ultra-long-term Data</i>
14h40	Francisco Alves	<i>Assessment of Reliability in Diagnosis based on Artificial Intelligence Tools from Medical Images</i>
14h55	Gonçalo Nogueira	<i>Advancing Robustness Assessment for Microservices Architectures using Generative Artificial Intelligence</i>
15h15	Simin Rasouli	<i>Advancing Machine Learning Techniques for Forecasting Energy Demand in Electric Vehicle Charging</i>
Coffee break		
15h50	Laura Jaime	<i>FLOSS Maturity Model for Social Sector</i>
16h05	Ricardo Santos	<i>Feature Engineering and Deep Learning for Lyrics-based Music Emotion Recognition</i>
16h20	Filipe Santos	<i>Towards Interactive Explanations for Trustworthy Human-Centered Reinforcement Learning</i>
16h35	Ehsan Farzadnia	<i>Decoding Intrusive Thoughts from Neuronal Data</i>
16h50	Tiago Simões	<i>Relational Analysis of Monitoring: Time and Load on Microservice</i>
17h05	Eduardo Felix	<i>Methodology for Dynamic Security Evaluation of IoT Smart Home Devices Based on OWASP Top 10 Vulnerabilities</i>
17h20	Diego Gomes	<i>A Methodology for Assessing Source Code Vulnerabilities in IoT Gateways</i>
17h35	Cidiane Lobato	<i>Generative Artificial Intelligence: State-of-the-art, Applications, and Perspectives</i>

Leveraging AI to Enhance Resilience in the Cloud-to-Things Continuum

Bruno Ricardo Leitão Faria 

April 2024

This work will be advised by Prof. Doctor David Abreu  and Prof. Doctor Karima Velasquez  from the Department of Informatics Engineering of the Faculty of Sciences and Technology of the University of Coimbra.

1 Introduction

The Cloud-to-Things (C2T) continuum is characterised by the integration of cloud, fog, and edge computing to empower the Internet of Things (IoT) with real-time processing, low latency, and high reliability. Despite its advantages, ensuring resilience in this vast, heterogeneous and microservice-based environment is challenging, particularly due to the critical impact of failures on application and service functionality. Therefore, resilience, defined as the ability to adapt to changing conditions and recover from failures, becomes crucial in this context. This is aligned with the self-x and zero-touch principles, which seek to minimise human intervention in system management.

Artificial Intelligence (AI), particularly Machine Learning (ML), offers promising solutions to enhance the resilience of the C2T continuum. However, traditional centralised AI models, especially during the inference stage, can have high computational and energy demands, which, besides resulting in more communications and higher latency in this architecture, pose significant environmental implications and lead to substantial monetary costs. Distributed AI models, dispersed across the continuum, can improve speed and efficiency, thereby reducing reaction time while optimising energy consumption. This distribution of AI models is aligned with the scattered nature of the C2T continuum, as it combines the low computational power of numerous devices to collectively achieve higher computational power. The distributed nature of AI also brings scalability to the system, allowing it to handle increasing amounts of data and devices efficiently.

This research explores the potential of distributed AI in enhancing the resilience of the C2T continuum. By performing AI tasks closer to where the data is generated, this approach aims to potentially reduce the need for data transmission, thereby lowering latency, increasing speed, and enhancing the overall performance of the C2T continuum. The effectiveness of these techniques will be validated through simulations, addressing some of the key challenges faced by this paradigm, like anomaly detection, resource management and communication overhead. Furthermore, by incorporating specific examples and case studies from domains like smart cities and autonomous vehicles, this research

aims to contribute to the ongoing discourse on managing large-scale systems and promoting a sustainable, energy-efficient future by addressing the crucial challenges inherent in the C2T continuum.

2 Related Work

C2T systems are transforming numerous sectors by providing diverse services and applications, such as smart grids, autonomous vehicles, and eHealth, which require high availability for critical functionalities. Even brief disturbances can cause serious consequences, like power outages, road accidents, and delayed medical interventions [5, 28].

Resilience in C2T systems refers to their ability to adapt to changing conditions and recover quickly from disruptions or failures, ensuring functionality and reliability [5]. These systems, relying on numerous interconnected components and resource-constrained devices, are vulnerable to failures [6, 28, 1, 26]. By leveraging AI to handle large datasets and automate decision-making, C2T systems can become self-healing and proactively address disruptions [20, 4].

Self-healing capabilities are critical for smooth operation in complex ecosystems, enabling autonomous identification, diagnosis, and potential rectification of issues like device malfunctions or infrastructure disruptions [2]. Zero-touch capacity remains an ongoing challenge [11], and anomaly detection techniques are essential for self-healing functionalities. Approaches like statistical methods, ML, and Deep Learning (DL) have been explored to identify deviations from normal system behaviour indicating potential problems [19, 28, 18].

AI, particularly ML and DL, offers significant potential for enhancing C2T resilience through anomaly detection and automated decision-making [20]. However, traditional models often struggle with the resource-constrained nature of edge devices within the C2T ecosystem, necessitating the development of techniques to reduce computational demands, memory requirements, and power consumption. Lightweight models and early exits within DL architectures have been studied to reduce these demands, making AI models more suitable for deployment on resource-constrained devices [17, 20, 27, 25, 23, 21].

Distributed AI approaches are gaining traction in the C2T ecosystem to overcome resource constraints, due to their dispersed nature [3, 7, 8, 13]. These approaches break down large models and process them over multiple devices, reducing the computational strain and energy consumption. This is particularly relevant for energy efficiency initiatives like the European Green Deal [10]. Data processing closer to its source also reduces latency, which is crucial for real-time applications like Virtual Reality (VR), where even minor latency can cause dizziness and nausea for users [22]. Furthermore, distributed AI architectures also improve fault tolerance by allowing the system to function even when individual devices fail, contrasting with centralised architectures where a single point of failure can cripple the entire system.

Rigorous testing is crucial for ensuring the effectiveness of AI solutions in C2T resilience. Simulators like COupled Simulation and Container Orchestration framework (COSCO) [24], iFogSim [16] and Yet Another Fault Simulator (YAFS) [15] provide controlled environments for model refinement. Furthermore, incorporating real-world datasets [9] enhances their realism, leading to more robust models. These tools create realistic C2T environments for testing and refining AI models before real-world deployment [12]. Ad-

ditionally, some datasets can be used directly for AI model training and testing, bypassing the need for simulations in specific scenarios [14].

Despite these advancements, there is still room for improvement in AI-powered C2T resilience solutions, therefore, the following sections outline the research objectives, Section 3, and activities, Section 4, to be carried out to address these challenges.

3 Research Objectives

For this research, the following objectives have been established:

1. Define Evaluation Metrics

Identify a comprehensive set of metrics to quantify the resilience of the C2T continuum. These metrics will cover aspects that directly impact a system's ability to recover from disruptions, including downtime and reaction time. Additionally, metrics that serve as indicators of overall system performance will be included, such as latency and energy consumption. Furthermore, performance metrics for the AI models will be defined, including accuracy, sensitivity-specificity, carbon footprint, and inference time.

2. Establish a Flexible Validation Framework

Develop a structured framework for the implementation, evaluation, and optimisation of intelligent solutions within the C2T continuum. This framework will detail the architecture and the interaction between different components of the proposed solution. It will be flexible and robust to support the dynamic and diverse nature of the C2T continuum, incorporating mechanisms for real-time data ingestion and visualisation tools to aid in monitoring system performance and evaluating the effectiveness of the AI solutions.

3. Design and optimise Intelligent Solutions to Improve Resilience in C2T Ecosystems

Design, develop and optimise centralised and distributed AI models for efficient operation within the C2T continuum, focusing on the identified metrics. These solutions will prioritise speed, efficiency, and scalability, and will be optimised to ensure they operate effectively in this environment.

4. Validate Solutions

Validate the effectiveness of the developed AI techniques in enhancing the resilience of the C2T continuum. This validation will involve addressing key challenges for the C2T such as anomaly detection and resource management, as well as challenges for AI like energy and time consumption.

5. Evaluate Solutions

Use the defined evaluation metrics to measure the performance of the implemented intelligent solutions, evaluate their effectiveness in enhancing the resilience of the C2T continuum, and compare the results not only between distributed and centralised models but also with the existing literature.

These objectives align with the goal of leveraging AI to enhance resilience in the C2T continuum, and thus provide a roadmap for the research process as well as contributing to the green deal initiatives by promoting energy-efficient systems in the C2T landscape.

4 Work Plan Activities

To enhance the resilience of the C2T continuum and fulfil the objectives described previously, the following activities were foreseen:

A1: State-of-the-Art revision

This initial phase of the research involves a comprehensive review of the existing literature. This review will be conducted throughout the entire research period to ensure an up-to-date understanding of this rapidly evolving field. The key areas of investigation within this activity include:

- Resilience in the C2T Continuum.
- Distributed AI in Edge Computing.
- AI Models for C2T Tasks.
- Metrics for Resilience and AI Performance.

By exploring these areas, a strong foundation of knowledge will be established. The goal is to understand the current landscape, identify gaps in the existing solutions, determine the potential for improvement and inform subsequent stages of the research.

A2: Definition of Evaluation Metrics

Building upon the knowledge gained from the literature review conducted in Activity A1, this activity focuses on establishing a set of metrics for evaluating two key aspects: the resilience of the C2T continuum and the performance of AI models.

For the resilience of the C2T continuum, these metrics will be used to assess the effectiveness of the proposed solutions in enhancing the overall resilience of the C2T environment. The candidate metrics for C2T resilience include downtime and reaction time.

For the performance of AI models, these metrics will be used to evaluate the efficiency and accuracy of the AI models developed for the C2T continuum. This could include accuracy, sensitivity-specificity, carbon footprint, and inference time.

The findings from the literature review on existing evaluation metrics will guide the selection and refinement of the final set of metrics. This will establish a clear evaluation framework, crucial for measuring the effectiveness of the proposed AI solutions in enhancing resilience within the C2T continuum.

A3: Definition of the Validation Framework

Following the literature review and establishment of evaluation metrics, this activity focuses on defining a validation framework for evaluating the effectiveness of intelligent solutions within the C2T ecosystem. The framework will leverage simulations as an initial approach for testing and refining solutions in a controlled environment.

Initially, the validation process will utilise established simulators, such as a previously tested Cloud-to-Edge (C2E) simulator COSCO, to create realistic C2T scenarios. However, existing simulators may have limitations, and thus, the need for enhanced functionalities or alternative tools to better suit specific research requirements will be explored. Defined metrics from A2 will serve as the foundation for assessing the performance of intelligent solutions within the simulated environment. These evaluation criteria may be further refined based on the insights gained from simulation testing.

This simulation-driven approach establishes a foundational validation framework. By testing and refining solutions within the controlled environment, the framework validates their effectiveness and resilience before real-world deployment.

A4: Design and Implementation of Green Intelligent Solutions

Informed by the evaluation metrics (A2) and framework (A3), this activity focuses on designing, implementing and optimising intelligent solutions for the C2T continuum. A key aspect will be leveraging distributed learning, which aligns well with the C2T's distributed nature. This approach allows for processing and decision-making closer to data sources at the edge, potentially improving efficiency and resilience.

The design process will consider both centralised and distributed AI models, tailored for tasks like anomaly detection and resource management, crucial for C2T resilience.

Designed models, whether centralised or distributed, will be fine-tuned for optimal performance within the C2T ecosystem. This optimisation will prioritise a balance between accuracy, efficiency (resource and energy usage), and scalability to accommodate future C2T growth.

By following this systematic approach, A4 aims to create intelligent solutions that are impactful for C2T resilience, efficient, and adaptable within the C2T infrastructure.

A5: Validating Effectiveness

Building on the optimised solutions from A4, this activity focuses on validation and testing their effectiveness in a simulated C2T environment. Using the defined metrics (A2), the model's performance will be rigorously assessed in two key areas:

- **Anomaly Detection:** Can these models accurately detect potential failures within the C2T system through simulated scenarios?
- **Resource Management:** How efficiently do the models optimise resource allocation under simulated resource-constrained situations?

This validation process provides crucial evidence of the impact of intelligent solutions on C2T resilience. The gathered results will guide further refinement or potential redesign of the AI models before real-world implementation.

A6: Evaluation

This phase delves deeper than validation (A5) by comparing the performance of the implemented intelligent solutions against each other, as well as against existing approaches documented in the literature. This broader evaluation provides a more comprehensive understanding of the solutions' effectiveness in enhancing C2T resilience relative to currently available methods.

A7: Dissemination of the results

The goal is to share the valuable findings of this research with the wider scientific community. This will be achieved through publications in peer-reviewed journals and presentations at relevant conferences focusing on AI, the IoT, and Distributed Computing. For this activity, some potential targets were found and compiled in Table 1.

Journal	Publisher	Impact Factor
Internet of Things Journal	IEEE	10.6 (<i>Q1</i>)
Transactions on Cognitive Communications and Networking	IEEE	8.6 (<i>Q1</i>)
Future Generation Computer Systems	Elsevier	7.5 (<i>Q1</i>)
Transactions on Cloud Computing	IEEE	6.5 (<i>Q1</i>)
Transactions on Network and Service Management	IEEE	5.3 (<i>Q1</i>)
Transactions on Machine Learning in Communications and Networking	IEEE	<i>relatively new</i>

(a) Target Journals

Conferences	Publisher	Full Name
ICC	IEEE	International Conference on Communications
WoWMoM	IEEE	International Symposium on a World of Wireless, Mobile and Multimedia Networks
NOMS	IEEE/IFIP	Network Operations and Management Symposium
ICMLCN	IEEE	International Conference on Machine Learning for Communication and Networking
SRDS	IEEE	International Symposium on Reliable Distributed Systems
SEC	ACM/IEEE	Symposium on Edge Computing

(b) Target Conferences

Table 1: Possible targets for dissemination of work

A8: Writing the PhD thesis

The last activity corresponds to documenting the research activities, findings, and contributions in the Ph.D. thesis. The thesis will provide a comprehensive overview of the research, including the motivation, methodology, results, and implications for the field.

This document will continually grow, making this activity last throughout the entire research period.

The detailed work plan is presented in the Gantt chart in Figure 1.

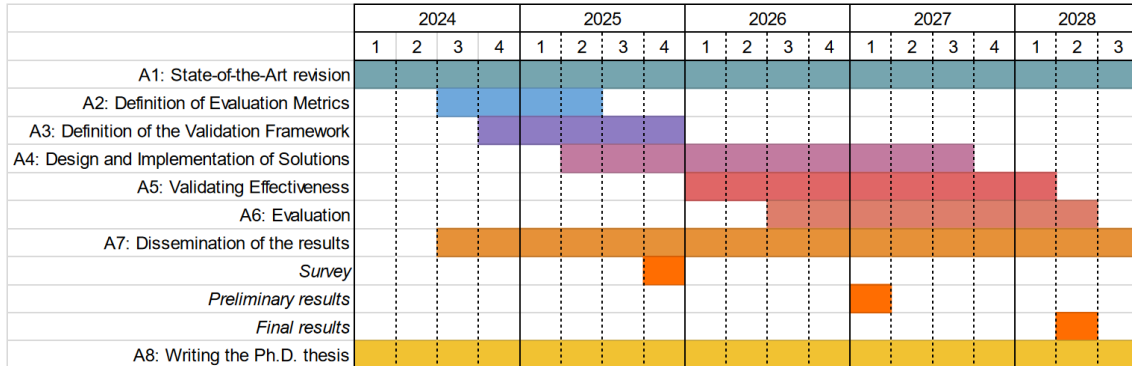


Figure 1: Gantt chart for the planned activities

References

- [1] *The Cloud-to-Thing Continuum: Opportunities and Challenges in Cloud, Fog and Edge Computing*. Palgrave Studies in Digital Business Enabling Technologies. Springer International Publishing, Cham, 2020.
- [2] Juncal Alonso, Leire Orue-Echevarria, Eneko Osaba, Jesús López Lobo, Iñigo Martínez, Josu Diaz de Arcaya, and Iñaki Etxaniz. Optimization and prediction techniques for self-healing and self-learning applications in a trustworthy cloud continuum. *Information*, 12(88):308, August 2021.
- [3] Audris Arzovs, Janis Judvaitis, Krisjanis Nesenbergs, and Leo Selavo. Distributed learning in the iot–edge–cloud continuum. *Machine Learning and Knowledge Extraction*, 6(11):283–315, March 2024.
- [4] Sepehr Ashtari, Ian Zhou, Mehran Abolhasan, Negin Shariati, Justin Lipman, and Wei Ni. Knowledge-defined networking: Applications, challenges and future work. *Array*, 14:100136, July 2022.
- [5] Christian Berger, Philipp Eichhammer, Hans P. Reiser, Jörg Domaschka, Franz J. Hauck, and Gerhard Habiger. A survey on resilience in the iot: Taxonomy, classification, and discussion of resilience mechanisms. *ACM Computing Surveys*, 54(7):147:1–147:39, September 2021.
- [6] Luiz Bittencourt, Roger Immich, Rizos Sakellariou, Nelson Fonseca, Edmundo Madeira, Marilia Curado, Leandro Villas, Luiz DaSilva, Craig Lee, and Omer Rana. The internet of things, fog and cloud continuum: Integration and challenges. *Internet of Things*, 3–4:134–155, October 2018.

- [7] Alejandro Carnero, Cristian Martin, Daniel R. Torres, Daniel Garrido, Manuel Diaz, and Bartolome Rubio. Managing and deploying distributed and deep neural models through kafka-ml in the cloud-to-things continuum. *IEEE Access*, 9:125478–125495, 2021.
- [8] Mingzhe Chen, Deniz Gündüz, Kaibin Huang, Walid Saad, Mehdi Bennis, Aneta Vulgarakis Feljan, and H. Vincent Poor. Distributed learning in wireless networks: Recent progress and future challenges. *IEEE Journal on Selected Areas in Communications*, 39(12):3579–3605, December 2021.
- [9] Yong-Hoon Choi, Daegyeom Kim, and Myeongjin Ko. 5g traffic datasets, July 2023.
- [10] European Comission. The european green deal - european commission, July 2021. Accessed on 12-04-2024.
- [11] Estefanía Coronado, Rasoul Behraves, Tejas Subramanya, Adriana Fernández-Fernández, Muhammad Shuaib Siddiqui, Xavier Costa-Pérez, and Roberto Riggio. Zero touch management: A survey of network automation solutions for 5g and 6g networks. *IEEE Communications Surveys Tutorials*, 24(4):2535–2578, 2022.
- [12] Bruno Faria. *Self-organising engine for the Cloud-to-Edge continuum*. M.sc. thesis, University of Coimbra, Coimbra, September 2023. Accepted: 2023-11-23T23:08:58Z journalAbbreviation: Mecanismo auto-organizável para o continuum de Nuvem-a-Ponta.
- [13] Seyyedali Hosseinalipour, Christopher G. Brinton, Vaneet Aggarwal, Huaiyu Dai, and Mung Chiang. From federated to fog learning: Distributed machine learning over heterogeneous wireless networks. *IEEE Communications Magazine*, 58(12):41–47, December 2020.
- [14] Md Sajid Khan. Dos/ddos attack dataset of 5g network slicing, September 2023.
- [15] Isaac Lera, Carlos Guerrero, and Carlos Juiz. Yafs: A simulator for iot scenarios in fog computing. *IEEE Access*, 7:91745–91758, 2019.
- [16] Redowan Mahmud, Samodha Pallewatta, Mohammad Goudarzi, and Rajkumar Buyya. ifogsim2: An extended ifogsim simulator for mobility, clustering, and microservice management in edge and fog computing environments. *Journal of Systems and Software*, 190:111351, August 2022.
- [17] Bomin Mao, Fengxiao Tang, Yuichi Kawamoto, and Nei Kato. Ai models for green communications towards 6g. *IEEE Communications Surveys Tutorials*, 24(1):210–247, 2022.
- [18] Mourad Nouioua, Philippe Fournier-Viger, Ganghuan He, Farid Nouioua, and Zhou Min. A survey of machine learning for network fault management. page 1–27, 2021.
- [19] Caleb. S and S. John Justin Thangaraj. Data-driven ml approaches for the concept of self-healing in cwn, including its challenges and possible solutions. In *2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, page 1–7, April 2023.

- [20] Farzad Samie, Lars Bauer, and Jorg Henkel. From cloud down to things: An overview of machine learning in internet of things. *IEEE Internet of Things Journal*, 6(3):4921–4934, June 2019.
- [21] Yuanming Shi, Kai Yang, Tao Jiang, Jun Zhang, and Khaled B. Letaief. Communication-efficient edge ai: Algorithms and systems. *IEEE Communications Surveys Tutorials*, 22(4):2167–2191, 2020.
- [22] Jan-Philipp Stauffert, Florian Niebling, and Marc Erich Latoschik. Latency and cybersickness: Impact, causes, and measures. a review. *Frontiers in Virtual Reality*, 1, November 2020.
- [23] Surat Teerapittayanon, Bradley McDanel, and H.T. Kung. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, page 2464–2469, December 2016.
- [24] Shreshth Tuli, Shivananda R. Poojara, Satish N. Srirama, Giuliano Casale, and Nicholas R. Jennings. Cosco: Container orchestration using co-simulation and gradient based optimization for fog computing environments. *IEEE Transactions on Parallel and Distributed Systems*, 33(1):101–116, January 2022.
- [25] Roberto Verdecchia, June Sallou, and Luís Cruz. A systematic review of green ai. *WIREs Data Mining and Knowledge Discovery*, 13(4):e1507, 2023.
- [26] Simon Volpert, Philipp Eichhammer, Florian Held, Thomas Huffert, Hans P. Reiser, and Jörg Domaschka. The view on systems monitoring and its requirements from future cloud-to-thing applications and infrastructures. *Future Generation Computer Systems*, 141:243–257, April 2023.
- [27] Fangxin Wang, Miao Zhang, Xiangxiang Wang, Xiaoqiang Ma, and Jiangchuan Liu. Deep learning for edge computing applications: A state-of-the-art survey. *IEEE Access*, 8:58322–58336, 2020.
- [28] Liudong Xing. Cascading failures in internet of things: Review and perspectives on reliability and resilience. *IEEE Internet of Things Journal*, 8(1):44–64, January 2021.

Thesis Project - Synopsis

Gabriel Cortês

Title

Sustainable Approach to Energy Efficient Artificial Intelligence

Supervisors

Professor Nuno Lourenço
Professor Penousal Machado

Motivation

This work is mainly motivated by the need to tackle the growing energy consumption of Artificial Intelligence (AI) systems.

Previous work on this subject has shown that it is possible to obtain models which are more energy-efficient when compared to a baseline whilst being almost as accurate. Namely, the best model found regarding power consumes 29.18 W (29.2%) less whilst having a small decrease in performance (0.9%), thus showing that a small trade-off on accuracy can yield a considerable reduction in the power consumed by the model [3]. Preliminary work also shows that it is possible to initialize the evolutionary process with an already established Deep Neural Network (DNN) model and evolve it towards a more energy-efficient version of itself whilst sacrificing little of its predictive performance.

In addition, this work is also motivated by the usage of Zero-shot Neural Architecture Search (NAS) techniques to reduce the search time for DNN models. Preliminary work shows promising results by using a Grammatical Evolution (GE) algorithm (e.g., Dynamic Structured Grammatical Evolution [12]) to search for zero-cost proxies with the evaluation performed on area-specific benchmark datasets.

Background

AI has grown exponentially over the years due to the availability of new software and hardware. Due to this, the energy consumed by AI programs has also substantially increased [8]. GPT-3, one of the state-of-the-art models relative to Large Language Models (LLMs) and one of the most used models, consumes 1287 MWh just in its training process during 15 days – approximately 552 tons of CO₂ equivalent emissions or the same as the electricity used by 108 homes during one year – which would require the plantation of 269 hectares of forests to be offsetted [15]. Although the training process of DNNs consumes a large quantity of energy, the inference process turns out to consume more since the deployment of DNNs may reach millions of users and each inference must be processed. At Google, for example, the training phase uses 40% of the total energy whilst inference takes 60% [15].

Neuroevolution (NE) consists of the usage of Evolutionary Computation (EC) techniques to search for optimal DNN models for any problem, be it image classification, natural language processing, or other. It does so through the evolution of the DNN’s architecture and learning parameters [6]. NE allows for the simultaneous exploration of multiple zones of the search space through parallelization techniques at the cost of taking a usually long time to execute since each individual of the population is a DNN that requires training and posterior evaluation.

In the research area of NAS, the prediction of performance through the usage of zero-cost proxies is gaining attention for its ability to forecast the accuracy of DNNs without requiring training [2, 13, 16, 9, 10]. Integrating this into a neuroevolutionary approach might allow for a faster search time as well as a substantially smaller energetical footprint. One problem that often arises is a low correlation between zero-shot NAS techniques and the accuracy of the models. This might be solved by alternating the evolutionary process evaluation system or by developing zero-shot cost functions with higher correlation.

Neuroevolutionary frameworks such as Fast Deep Evolutionary Network Structured Representation (Fast-DENSER) begin with DNN models generated from scratch and evolve them to obtain optimal solutions [1]. Contrarily to this approach, it is possible to begin the evolutionary process with previously trained models and evolve them. Using this technique, we can focus the evolution on a metric that the model given as input does not excel at, such as the energy consumption in the inference phase. Combining this approach with the zero-shot NAS techniques might allow us to promptly optimize a DNN model to obtain an accurate and energy-efficient model.

Transformer models and LLMs have gained significant traction in the recent past. At this stage of development, the principal concern has been to make these types of models more accurate and realistic, thus not focusing on strategies to minimize their environmental impact, which is expected to grow even

more [11, 14]. The application of NE to this problem can help to address energy consumption concerns on this matter.

Research Questions

AI usage has surged exponentially, driving a corresponding increase in computational resources and energy demand. However, most research often overlooks the energy consumption of Machine Learning (ML) models. This work aims to make AI systems more energy-efficient. This leads to the formulation of three research questions:

R.Q. 1: How to approximate the energetical usage of AI to that of the human brain? While DNN models typically require over 100 W, the human brain operates on about 20 W, despite its superior performance [7]. This shows that even though AI has had impactful developments, its efficiency is no match for what natural evolution has achieved with the human species. Therefore, it is essential to research strategies to reduce the energetical needs of AI while maintaining its performance.

R.Q. 2: How can EC techniques automate the design of energy-efficient models while maintaining performance comparable to traditional methods? Previous work already showcased that EC can be used to optimize DNNs to be energy-efficient [4]. Further research could explore alternative EC algorithms and incorporate additional types of layers to expand the search space for improved results.

R.Q. 3: How to leverage zero-cost proxies to minimize the search time for DNN models? The search for optimal DNN models requires a lot of computational resources since each DNN architecture requires training and posterior evaluation. This means that behind a state-of-the-art DNN we might have a large amount of "wasted" resources in assessing the quality of inferior models. Zero-cost proxies consist of formulas that serve as an indicator of how good a model might be. They allow us to predict the accuracy of a DNN without training it thus saving resources. These proxies usually use a small amount of GPU time or even CPU time due to their nature.

Methodology

The development of this work will follow an iterative strategy. In the first phase, state-of-the-art zero-cost proxies will be replicated and integrated into an existing framework. These will be tested and compared to baseline and other experiments to assess their effectiveness. To add more options, some novel zero-cost

proxies will be created by taking advantage of GE algorithms. These will be tested and compared to the existing ones. On the other hand, experiments will be performed without zero-cost proxies to enhance the power efficiency of DNNs. Each new feature will be tested and compared to previous experiments by using statistical tests to objectively evaluate whether our results exhibit statistical improvement when compared to existing ones.

Expected Results and Significance

This work is expected to provide novel strategies to tackle the growing issue of AI's energy demand through the automatic design of more efficient DNN models. Thus, its main benefit to society is that of reducing the energy used by AI which is expected to globally surpass the energetical needs of many medium to large-sized countries [5].

References

- [1] Assunção, F., Lourenço, N., Ribeiro, B., Machado, P.: Fast-DENSER: Fast deep evolutionary network structured representation. *SoftwareX* **14**, 100694 (2021). <https://doi.org/10.1016/j.softx.2021.100694>
- [2] Chen, W., Gong, X., Wang, Z.: Neural architecture search on ImageNet in four GPU hours: A theoretically inspired perspective. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021). <https://doi.org/10.48550/arXiv.2102.11535>
- [3] Cortês, G., Lourenço, N., Machado, P.: Towards physical plausibility in neuroevolution systems. In: Smith, S., Correia, J., Cintrano, C. (eds.) *Applications of Evolutionary Computation*. pp. 76–90. Springer Nature Switzerland, Cham (2024)
- [4] Cortês, G., Lourenço, N., Machado, P.: Towards physical plausibility in neuroevolution systems
- [5] de Vries, A.: The growing energy footprint of artificial intelligence. *Joule* **7**(10), 2191–2194 (2023). <https://doi.org/https://doi.org/10.1016/j.joule.2023.09.004>, <https://www.sciencedirect.com/science/article/pii/S2542435123003653>
- [6] Galván, E., Mooney, P.: Neuroevolution in deep neural networks: Current trends and future challenges. *IEEE Trans. Artif. Intell.* **2**(6), 476–493 (2021). <https://doi.org/10.1109/TAI.2021.3067574>

- [7] Hasler, J.: Special report : Can we copy the brain? - a road map for the artificial brain. *IEEE Spectrum* **54**(6), 46–50 (2017). <https://doi.org/10.1109/MSPEC.2017.7934231>
- [8] Islam, M.S., Zisad, S.N., Kor, A.L., Hasan, M.H.: Sustainability of machine learning models: An energy consumption centric evaluation. In: 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE). pp. 1–6 (Feb 2023). <https://doi.org/10.1109/ECCE57851.2023.10101532>
- [9] Li, G., Yang, Y., Bhardwaj, K., Marculescu, R.: Zico: Zero-shot nas via inverse coefficient of variation on gradients (2023)
- [10] Lin, M., Wang, P., Sun, Z., Chen, H., Sun, X., Qian, Q., Li, H., Jin, R.: Zen-nas: A zero-shot nas for high-performance deep image recognition (2021)
- [11] Lin, T., Wang, Y., Liu, X., Qiu, X.: A survey of transformers. *AI Open* **3**, 111–132 (2022). <https://doi.org/https://doi.org/10.1016/j.aiopen.2022.10.001>, <https://www.sciencedirect.com/science/article/pii/S2666651022000146>
- [12] Lourenço, N., Assunção, F., Pereira, F.B., Costa, E., Machado, P.: Structured grammatical evolution: A dynamic approach. In: Ryan, C., O’Neill, M., Collins, J.J. (eds.) *Handbook of Grammatical Evolution*, pp. 137–161. Springer (2018). https://doi.org/10.1007/978-3-319-78717-6_6
- [13] Mellor, J., Turner, J., Storkey, A.J., Crowley, E.J.: Neural architecture search without training. In: Meila, M., Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research*, vol. 139, pp. 7588–7598. PMLR (2021). <https://doi.org/10.48550/arXiv.2006.04647>
- [14] Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., Mian, A.: A comprehensive overview of large language models (2023)
- [15] Patterson, D.A., Gonzalez, J., Hölzle, U., Le, Q.V., Liang, C., Munguia, L., Rothchild, D., So, D.R., Texier, M., Dean, J.: The carbon footprint of machine learning training will plateau, then shrink. *Computer* **55**(7), 18–28 (2022). <https://doi.org/10.1109/MC.2022.3148714>
- [16] Sun, Z., Sun, Y., Yang, L., Lu, S., Mei, J., Zhao, W., Hu, Y.: Unleashing the power of gradient signal-to-noise ratio for zero-shot nas. pp. 5740–5750 (10 2023). <https://doi.org/10.1109/ICCV51070.2023.00530>

Evolutionary Feature Engineering in Multiobjective Scenarios

Miguel Rabuge^[0009-0008-0914-0495] and Nuno Lourenço^[0000-0002-2154-0642]

University of Coimbra, CISUC/LASI – Centre for Informatics and Systems of the University of Coimbra, Department of Informatics Engineering
{rabuge,naml}@dei.uc.pt

1. MOTIVATION AND BACKGROUND

Recently, the Machine Learning (ML) field has been intensively researched around the globe, culminating in great advances for people. Many ML models may be defined as black-box algorithms that consume information and return knowledge. Such end requires data, i.e. representation of information, of high quality to provide the best outcome possible. The best representation of information depends on the chosen ML model, as they treat data differently from each other, and on multiple ML objectives, such as improving performance or reducing dimensionality.

Besides traditional Feature Engineering (FE) techniques, Evolutionary Computation (EC) methods have also addressed this problem, paving the way for a new FE paradigm. This work aims to bridge the gap between the EC and ML fields by studying and developing an evolutionary FE framework that focuses on optimizing multiple potentially conflicting objectives regarding the construction and selection of features.

Feature Engineering (FE) is a step in the Machine Learning (ML) pipeline where a pre-processed data set undergoes selection and transformation processes. It can be divided into Feature Selection (FS) and Feature Construction (FC). FS intends to reduce the dimensionality of the data set to remove redundant or misleading data that can negatively affect the performance of the models. In contrast, FC seeks to build novel features from the original ones, exploring non-linear transformations of the feature space to empower ML models. For a dataset with dimensions $M \times N$, the number of possible feature subsets is 2^N . When enabling the construction of features, the solution space increases even further, becoming intractable for exact methods.

Due to the combinatorial and complex nature of this problem, alternative evolutionary FE approaches have been applied in different settings. Such methodologies have been reported to improve performance [3, 24, 13, 1, 18], reduce dimensionality [1, 18, 2] and enable interpretability [3].

FE methods can be classified into three categories: filter, wrapper and embedded [3]. Filter methods [14, 7] are model-free ranking methods, whereas wrappers [13,11] assess the performance of a model to evaluate which set of features is relevant to the problem at hand. At last, embedded methods perform FE while training the ML model [18].

Researchers have used Genetic Algorithms (GAs) to address the FS problem [17] by evolving solutions represented as a binary string of length N ($M \times N$) to denote whether a feature was in (1) or out (0) of the subset. In contrast, the representation for FC would have to be defined as a list of tree structures, each tree being a constructed feature [17, 22]. However, by evolving computer programs with Genetic Programming (GP) [8], FC became more accessible by specifying the terminal and function sets that would be used to build an abstract syntax tree, a computer program that transforms the original dataset [18, 19]. Grammatical Evolution (GE) [16] and its variants [9, 10],

grammar-based GP approaches, further empowered this research field by enabling the specification of the transformations through a context-free grammar [6, 11, 13, 20].

Due to many potentially conflicting objectives, such as improving performance and reducing dimensionality, FE has also been addressed with evolutionary multiobjective methods [15]. Researchers have reported advantages in evolving multiple objectives independently [7] and assessing potential trade-offs among them [5], a capability not achievable by single-objective methods.

In Artificial Intelligence (AI), societal values are crucial in guiding the research of intelligent systems. Two key concerns receiving significant attention are explainability and energy consumption. The explainability of ML models is becoming increasingly required because of AI regulation. Hence, model-agnostic methods have been proposed, namely Shapley Values, Local Interpretable Model-agnostic Explanations (LIME) [21], global surrogate models and others [12]. When incorporated into the evolutionary process, these methods may evolve highly informative features [23]. In addition, researchers are using evolutionary methods to address the issue of heavy AI models consuming substantial amounts of energy, which causes significant challenges to sustainability [4].

2. RESEARCH QUESTIONS

The main goal of this work is to study and develop an evolutionary feature engineering framework that focuses on optimizing multiple potentially conflicting objectives regarding the construction and selection of features. The optimization objectives are: improving performance, reducing data dimensionality and complying with societal values, such as enabling explainability and reducing energy consumption. To accomplish this goal, three research questions must be answered:

RQ1: How effective are evolutionary approaches for feature engineering?

To propose an evolutionary feature engineering framework, one must demonstrate its potential and limitations across different domains, comparing it to similar and commonly used methods. Evolutionary methods may have an edge over others, making them the preferred approach in certain scenarios. Answering this question is of utter importance as it contextualizes all of these approaches regarding feature engineering.

RQ2: How to evaluate the quality of the features?

Many objective functions have been proposed to evolve features, mostly regarding performance and/or dimensionality. Mutual information and linear correlation between features and labels are common metrics for filter-based methods [7]. Performance metrics, such as accuracy or root mean squared error, are used in wrapper techniques given the predicted labels of a proxy model (e.g. decision tree [11, 13]). Studying different objective functions, for the specified optimization objectives, enables a greater understanding of how features behave in the evolutionary process, leading to the possibility of addressing the shortcomings of the metrics.

RQ3: What trade-offs are present among the objectives?

Regarding the specified optimization objectives, conflicts may arise and lead to non-dominated solutions. Studying and characterizing such trade-offs across multiple domains allows for an understanding of how these objectives clash, if so, and for defining strategies to decide which

solution is best, as different domains may prioritize objectives (e.g. explainability over performance).

3. METHODOLOGY

This work will follow an iterative approach, regarding the development of the evolutionary feature engineering framework. After reviewing the literature and defining the benchmarking suite, the framework will be developed and extended incrementally.

The primary goal is to establish framework stability concerning the performance and dimensionality objectives. This step will consist of experimenting with different algorithms and evaluation metrics in the evolutionary process to study which ones generalize better, across multiple domains.

Once settled, the research will be devoted to incorporating responsible and societal objectives. The goal is to study how they affect the construction and selection of features and what trade-off interactions exist between the objectives.

4. EXPECTED RESULTS

The development of an evolutionary FE framework that can select and construct features in multiobjective scenarios contributes to the automation of this complex and time-consuming step in the ML pipeline. Also, analysing the framework in a setting of conflicting objectives will provide a better understanding of their trade-offs. In addition, designing the framework to adapt to different objectives and datasets enhances the applicability of the framework across diverse domains, enabling the development of real-world applications.

5. REFERENCES

1. Ahmed, S., Zhang, M., Peng, L., Xue, B.: Multiple feature construction for effective biomarker identification and classification using genetic programming. In: Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation. pp. 249–256 (2014)
2. Ali, M.S., Kshirsagar, M., Naredo, E., Ryan, C.: Automated grammar-based feature selection in symbolic regression. In: Proceedings of the Genetic and Evolutionary Computation Conference. pp. 902–910 (2022)
3. Cherrier, N., Poli, J.P., Defurne, M., Sabatié, F.: Consistent feature construction with constrained genetic programming for experimental physics. In: 2019 IEEE Congress on Evolutionary Computation (CEC). pp. 1650–1658. IEEE (2019)
4. Cortês, G., Lourenço, N., Machado, P.: Towards physical plausibility in neuroevolution systems. arXiv preprint arXiv:2401.17733 (2024)
5. Emmanouilidis, C., Hunter, A., MacIntyre, J., Cox, C.: Multiple-criteria genetic algorithms for feature selection in neuro-fuzzy modeling. In: IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339). vol. 6, pp. 4387–4392. IEEE (1999)
6. Gavrilis, D., Tsoulos, I.G., Dermatas, E.: Selecting and constructing features using grammatical evolution. *Pattern Recognition Letters* 29(9), 1358–1365 (2008)
7. Hancer, E., Xue, B., Zhang, M., Karaboga, D., Akay, B.: A multi-objective artificial bee colony approach to feature selection using fuzzy mutual information. In: 2015 IEEE congress on evolutionary computation (CEC). pp. 2420–2427. IEEE (2015)
8. Koza, J.R.: Genetic programming as a means for programming computers by natural selection. *Statistics and computing* 4, 87–112 (1994)

9. Lourenço, N., Assunção, F., Pereira, F.B., Costa, E., Machado, P.: Structured grammatical evolution: a dynamic approach. *Handbook of grammatical evolution* pp. 137–161 (2018)
10. Lourenço, N., Ferrer, J., Pereira, F.B., Costa, E.: A comparative study of different grammar-based genetic programming approaches. In: *Genetic Programming: 20th European Conference, EuroGP 2017, Amsterdam, The Netherlands, April 19-21, 2017, Proceedings* 20. pp. 311–325. Springer (2017)
11. Miquilini, P., Barros, R.C., de Melo, V.V., Basgalupp, M.P.: Enhancing discrimination power with genetic feature construction: A grammatical evolution approach. In: *2016 IEEE Congress on Evolutionary Computation (CEC)*. pp. 3824–3831. IEEE (2016)
12. Molnar, C.: *Interpretable Machine Learning*. 2 edn. (2022), <https://christophm.github.io/interpretable-ml-book>
13. Monteiro, M., Lourenço, N., Pereira, F.B.: Fermat: Feature engineering with grammatical evolution. In: *EPIA Conference on Artificial Intelligence*. pp. 239–251. Springer (2021)
14. Muharram, M.A., Smith, G.D.: The effect of evolved attributes on classification algorithms. In: *AI 2003: Advances in Artificial Intelligence: 16th Australian Conference on AI, Perth, Australia, December 3-5, 2003. Proceedings* 16. pp. 933–941. Springer (2003)
15. Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S., Coello, C.A.C.: A survey of multiobjective evolutionary algorithms for data mining: Part i. *IEEE Transactions on Evolutionary Computation* 18(1), 4–19 (2013)
16. O’Neill, M., Ryan, C.: Grammatical evolution. *IEEE Transactions on Evolutionary Computation* 5(4), 349–358 (2001). <https://doi.org/10.1109/4235.942529>
17. Siedlecki, W., Sklansky, J.: A note on genetic algorithms for large-scale feature selection. *Pattern recognition letters* 10(5), 335–347 (1989)
18. Tran, B., Xue, B., Zhang, M.: Genetic programming for feature construction and selection in classification on high-dimensional data. *Memetic Computing* 8, 3–15 (2016)
19. Tran, B., Zhang, M., Xue, B.: Multiple feature construction in classification on high-dimensional data using gp. In: *2016 IEEE symposium series on computational intelligence (SSCI)*. pp. 1–8. IEEE (2016)
20. Tsoulos, I.G., Tzallas, A.T., Tsalikakis, D.: Prediction of covid-19 cases using constructed features by grammatical evolution. *Symmetry* 14(10), 2149 (2022)
21. Tulio Ribeiro, M., Singh, S., Guestrin, C.: Model-agnostic interpretability of machine learning. *arXiv e-prints* pp. arXiv–1606 (2016)
22. Vafaie, H., De Jong, K.: Evolutionary feature space transformation. In: *Feature Extraction, Construction and Selection: a data mining perspective*, pp. 307–323. Springer (1998)
23. Wang, C., Chen, Q., Xue, B., Zhang, M.: Shapley value based feature selection to improve generalization of genetic programming for high-dimensional symbolic regression. In: *Australasian Conference on Data Science and Machine Learning*. pp. 163–176. Springer (2023)
24. Zhang, H., Zhou, A., Zhang, H.: An evolutionary forest for regression. *IEEE Transactions on Evolutionary Computation* 26(4), 735–749 (2021)

Bayesian Causal Inference in Deep Spiking Neural Networks

Dylan Gonçalves Perdigão[✉]

Advised by Prof. Dr. Bernardete Ribeiro[✉], Prof. Dr. Catarina Silva[✉], and Prof. Dr. Francisco Antunes[✉]

12th April 2024

1 Introduction

The brain implements causal inference to solve hard problems, such as object recognition tasks. Many decisions are based on inferences about unknown causes from ambiguous and noisy observations. Brain-inspired algorithms emerged to solve computational problems efficiently through the advancement of artificial intelligence (AI). More specifically, neuroscience inspired different modeling of artificial neural networks (ANNs), which led to the current spiking neural networks (SNNs). Brains and computers work differently, which limits the latter to its computational speed and storage for a certain amount of energy. However, SNNs become significantly power-efficient when implemented on neuromorphic hardware, making this difference point to a singularity, where computers reach human performance levels [15, 27].

This work aims to study the third generation of ANNs, also known as deep SNNs, which approaches the behavior of the biological human neuron. The study will answer the following research questions (RQs):

RQ1 – How can Bayesian inference techniques be integrated into deep spiking neural networks to facilitate causal reasoning and inference?

RQ2 – What novel architectures and learning algorithms can be developed to outperform deep state-of-the-art spiking neural networks?

RQ3 – How can causal inference frameworks enhance the interpretability and robustness of results obtained from deep spiking neural networks?

The idea remains in combining this new generation of ANN with Bayesian inference and explainability mechanisms to contribute towards a more responsible and explainable artificial intelligence (XAI), enabling models to perform efficiently and accurately while providing transparent insights into its decision-making processes.

2 Background

This section comprises an overview of biological neural networks in Section 2.1, followed by an explanation of mathematical and computational neuron models in Section 2.2. The causal inference theme is presented in Section 2.3. Finally, the combined approach between these two areas is concluded in Section 2.4.

2.1 Biological Neural Networks

Biologically, the brain mainly encompasses cells called *neurons*. These neurons typically comprise three distinct parts: the *dendrites*, the *soma*, and the *axon*. The dendrites collect electrochemical signals from other neurons and transmit them to the soma. The soma processes these input signals and generates an output signal when it exceeds a certain threshold. This output signal is taken over the axon, which delivers the signal to other neurons of the brain's neural network. This junction between the two neurons is called a *synapse* [6, 7]. These signals are the way how neurons communicate through the network. The neuron body is composed of a pored capacitive membrane, which can absorb sodium ions (Na^+) through the opening of sodium ion channels. Consequently, when the electrical charge passes the threshold of the membrane's potential, the sodium ions are quickly replaced with the absorption of potassium ions (K^+) through the opening of potassium ion channels. This rapid variation in the electrical potential of the membrane is called *action potential*

or, more commonly, a *spike*, which has an amplitude of about 100mV and typically a duration of 1-2ms [8].

2.2 Spiking Neural Networks

With the advancements of deep learning, ANN become more energy-intensive, namely with the venue of large language models (LLMs) such as OpenAI’s *GPT-4* [19] or Meta’s *LLaMA* [28] models. Recently, spiking neural networks emerged as a possible solution to this high energy demand for traditional artificial neural networks.

Historically, the first mathematical neuron model appeared in 1943 with McCulloch and Pitts [16], which evolved to different variants of neuron models, such as Hodgkin-Huxley (HH) [10], Izhikevich (IZ) [11], Wilson (WIL) [29], Fitzhugh-Nagumo (FITZ) [4], and Leaky Integrate-and-Fire (LIF) [7] models. The latter is more prevalent to simulate as a SNN since it is biologically plausible, computationally efficient, and easier to train [20]. Neuromorphic computing is becoming popular since it is power efficient when implemented on neuromorphic hardware. Neuromorphic algorithms as SNNs are advantaged with their high spatial and temporal sparsity, which reduce on/off activations on hardware.

In spiking neural networks, action potentials (spikes) occur when the potential U surpasses the threshold of the membrane θ , which has a decay rate β . Typically, the LIF neuron model is represented as a resistor-capacitor (RC) circuit (with a resistance R and a current I over time t), which is modeled the following differential Equation 1:

$$\tau \frac{dU(t)}{dt} = -U(t) + I_{in}(t)R \quad (1)$$

where $\tau = RC$ is the circuit’s time constant, typically between 1–100ms. However, the system can be simplified through the forward Euler method to make this differential system compatible with sequential networks. Moreover, Equation 1 does not take into account the reset of the membrane, which removes the value of the threshold if there is a spike at the previous timestamp S_{out} . In the simplified model, the input current combines the weighted sum of inputs WX of the neural network. The potential is represented with Equation 2:

$$U[t] = \underbrace{\beta U[t-1]}_{\text{decay}} + \underbrace{WX[t]}_{\text{input}} - \underbrace{S_{out}[t-1]\theta}_{\text{reset}} \quad (2)$$

Programmatically, the raw input is converted into a matrix X of spikes ($k \times m \times n$), where m is the number of instances per batch k and n is the number of steps for the temporal component of the spike. The superposition of spikes – on each layer of the SNN – creates the current I at time t . This current is translated into the potential of the membrane U , which generates spikes for the next layer of the SNN [3]. This modeling of the neuron’s membrane makes it possible to perform classification tasks by counting the number of spikes on the last layer of the ANN. Therefore, the predicted class has a bigger number of spikes.

2.3 Causal Inference

Causal inference allows the understanding of complex behaviors in data, looking for information about how features in a dataset affect the prediction in a machine learning (ML) algorithm [21, 25, 22, 23, 24]. Structural causal models (SCMs) defines the relationships between variables in the data, enabling the prediction and estimation of what happens to the data in case of value variations in these variables (called interventions). In addition, SCMs define the consequences of these interventions (called counterfactuals), taking into account what will append with such intervention [14]. The representation with SCMs expresses causality into different variables and connections between them. These variables are divided into four categories:

- The outcome is the target variable in data in which the causal effect is studied. More specifically, the causal effect of variables that cause changes in the outcome. The outcome is also called the *effect*;
- The treatment is a feature in the data that directly impacts the target variable. The treatment is also called the *cause*;
- The confounder is a feature that influences treatment and outcome, which can bias the causality effect. The link for this *common cause* to treatment and outcome is the *backdoor path*;

- The instrumental variable is a variable that is not directly correlated to the outcome. However, this variable is correlated to the treatment and can indirectly influence the outcome through the treatment.

In causal analysis, there are two types of studies: observational and interventional. Observational studies are where the data is measured without intervention in the process. Interventional studies are where the data is manipulated for a certain purpose. What happens is that observed data may have external factors, such as confounders, influencing a treatment. If the outcome under a given treatment differs from that under another treatment, then there is a causal effect [9].

For the causal analysis, it is necessary to discover causal relations between variables through structural causal models. This SCMs is composed of directed acyclic graphs (DAGs) and structural equation models (SEMs).

- The directed acyclic graphs are graphical representations of connections between variables, also called causality diagrams, in which the nodes are variables, and the edges define the conditional probability between two variables;
- The structural equation models show the details of each edge between nodes, namely quantifying the influence that a variable has on another.

The discovery of the DAG can be made through the NOTEARS (*Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning*) algorithm [30] that allows computing the relationship between nodes. It calculates an adjacency matrix to minimize a function penalizing the graph’s cycles. The advantage of this algorithm is that it supports data with high dimensionality, and it does not need to make assumptions about the data’s features relationship. This process is called *structure learning* since it learns the graph’s structure from the data.

Causality is measured through the average treatment effect (ATE), which can be estimated as in Equation 3:

$$\text{ATE} = \mathbb{E}[y_1 - y_0] \quad (3)$$

The ATE represents the change in the outcome variable if there is an intervention in treatments. The higher the value, the more the outcome is susceptible to change if the treatment has a modification. Different methods exist to estimate the ATE through linear regression or propensity scores [18, 21].

2.4 Causal Spiking Neural Networks

In sum, the human brain implicitly implements causal reasoning to solve complex tasks. These causal neural architectures enable in-depth interpretability [5] of such black-box models by inferring the behaviors of the neurons. Bayesian inference can help modeling uncertainty in causal relationships between input features and output responses in SNNs. By understanding the causal structure of the system, one can design more effective network architectures and optimize training strategies [17]. The understanding of these causal relationships can help in building more robust and generalized SNNs by highlighting causal mechanisms rather than just correlations in the data.

3 Proposed Experimental Work

To validate our combined approach of SNNs with causal Bayesian inference, our causal spiking neural network will be tested with highly imbalanced data from a critical sector domain.

Section 3.1 describes the dataset used as a benchmark for the whole work since it is a real-world application of a business-constrained problem. After that, the methodology of the work is presented in Section 3.2. Finally, the expected contributions resulting from this work are presented in Section 3.3.

3.1 Dataset

The bank account fraud (BAF) suite of datasets [13, 12] consists of a base and five variants of data generated from real-world bank account opening fraud detection observations, composed of one million instances and thirty features. Each row (instance) of the dataset represents one client application, and each column (feature) represents the provided data from the clients and information gathered from the platform where the clients open the account. The dataset comprises eight months of data (between February and September), where the authors recommend dividing

the first six months for training and the last two for testing. Variants of the dataset are highly imbalanced, with a fraud prevalence varying between 0.85% and 1.5%, being higher in the last months.

The dataset distinguishes two groups, which are divided by the client’s age, regarding fairness [2, 26]. The first group is the younger clients, whose ages are under or equal to fifty years old, and the remaining are categorized into the second group of older clients. The *Base* and *Variant I* datasets have the youngest customers (respectively 77% and 90%), while the remaining variants have an equal proportion of the two groups. The variants also differ in the prevalence of fraud for both groups, varying from 0.3% and 1.9%, except for odd variants (*Variant I*, *Variant III*, and *Variant V*), where the prevalence of fraud is fixed to 1.1%.

3.2 Methodology

This work improves fraud detection with the bank account fraud dataset, which is divided into three phases:

- i) The first phase consists of reviewing the current state-of-the-art in SNNs, enabling the understanding of their operational process, namely with `snnTorch` [3]. This understanding enables the development of an SNN architecture that is fair and performant, that is, having a high false positive rate (FPR) ratio between the protected attributes, for instance, in Equation 4, the age of the clients:

$$\text{FPR}_{ratio} = \begin{cases} \frac{\text{FPR}_{old}}{\text{FPR}_{young}}, & \text{if } \text{FPR}_{old} \geq \text{FPR}_{young} \\ \frac{\text{FPR}_{old}}{\text{FPR}_{young}}, & \text{otherwise} \end{cases} \quad (4)$$

while respecting a business constraint of high recall at 5% of FPR, being the dataset mentioned above the benchmark of the work. For that, the SNN is optimized through its hyperparameters with a Bayesian optimization framework called Optuna [1].

- ii) In the second phase, a synthetic dataset for causal inference will be developed. This dataset will have causal relations that are a priori known to validate the framework mentioned below.
- iii) Next, causal inference will be integrated into the spiking neural network to improve the detection task’s performance and fairness (that is, having equity between protected attributes of the dataset, reducing consequent disparities of classification between groups).
- iv) Finally, a causal inference framework will be developed to provide insights into black-box models such as spiking neural networks. This framework will enable a better understanding of the neural network’s decision-making process, which will be validated with the causal dataset, where relationships between variables are known a priori.

3.3 Expected Contributions

We expect a significant contribution of the current state of the art in the explainability of deep spiking neural networks through:

- The development of a synthetic dataset from which causal relations are known;
- The development of a causal inference framework that uses spiking neural networks;
- The validation of the proposed research in a critical problem of the finance domain;
- The publication of scientific articles describing the work developed in this area, namely in national/international conferences and journals:
 - **National Conferences:** Portuguese Conference on Artificial Intelligence (EPIA), and Portuguese Conference on Pattern Recognition (RECPAD);
 - **International Conferences:** International Joint Conferences on Artificial Intelligence (IJCAI), European Conference on Artificial Intelligence (ECAI), European Conference on Machine Learning (ECML), Knowledge Discovery and Data Mining (KDD), Conference on Neural Information Processing Systems (NeurIPS), International Conference on Information and Knowledge Management (CIKM), and International Conference on Machine Learning (ICML);

- **Journals:** Journal of Artificial Intelligence Research, International Journal of Pattern Recognition and Artificial Intelligence, Neural Networks, Journal of Causal Inference, Progress in Artificial Intelligence, Foundations and Trends in Machine Learning, and Nature Machine Intelligence.

References

- [1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 2623–2631, New York, NY, USA, July 2019. Association for Computing Machinery.
- [2] André F. Cruz, Catarina Belém, Sérgio Jesus, João Bravo, Pedro Saleiro, and Pedro Bizarro. FairGBM: Gradient Boosting with Fairness Constraints, March 2023. arXiv:2209.07850 [cs].
- [3] Jason K. Eshraghian, Max Ward, Emre O. Neftci, Xinxin Wang, Gregor Lenz, Girish Dwivedi, Mohammed Bennamoun, Doo Seok Jeong, and Wei D. Lu. Training Spiking Neural Networks Using Lessons From Deep Learning. *Proceedings of the IEEE*, 111(9):1016–1054, September 2023.
- [4] Richard FitzHugh. Mathematical models of threshold phenomena in the nerve membrane. *The bulletin of mathematical biophysics*, 17(4):257–278, December 1955.
- [5] Duarte Folgado, Marília Barandas, Lorenzo Famiglini, Ricardo Santos, Federico Cabitza, and Hugo Gamboa. Explainability meets uncertainty quantification: Insights from feature-based model fusion on multimodal time series. *Information Fusion*, 100:101955, December 2023.
- [6] Wulfram Gerstner. Spiking neurons. In *Pulsed neural networks*, pages 1–53. MIT Press, Cambridge, MA, USA, February 1999.
- [7] Wulfram Gerstner and Werner M. Kistler. *Spiking neuron models: single neurons, populations, plasticity*. Cambridge University Press, Cambridge, U.K. ; New York, 2002.
- [8] Michael H. Grider, Rishita Jessu, and Rian Kabir. Physiology, Action Potential. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2024.
- [9] Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. CRC Press, 1st edition, 2024.
- [10] A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4):500–544, August 1952.
- [11] E.M. Izhikevich. Simple model of spiking neurons. *IEEE Transactions on Neural Networks*, 14(6):1569–1572, November 2003. Conference Name: IEEE Transactions on Neural Networks.
- [12] Sérgio Jesus, José Pombal, Duarte Alves, André F Cruz, Pedro Saleiro, Rita P Ribeiro, João Gama, and Pedro Bizarro. BAF Dataset Suite Datasheet, 2022.
- [13] Sérgio Jesus, José Pombal, Duarte Alves, André F Cruz, Pedro Saleiro, Rita P Ribeiro, João Gama, and Pedro Bizarro. Turning the Tables: Biased, Imbalanced, Dynamic Tabular Datasets for ML Evaluation. In *36th Conference on Neural Information Processing Systems Datasets and Benchmark track*, 2022.
- [14] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J. Kusner, and Ricardo Silva. Causal Machine Learning: A Survey and Open Problems, July 2022. arXiv:2206.15475 [cs, stat].
- [15] Richard Liu and Fredrik Bixo. *Analysing the Energy Efficiency of Training Spiking Neural Networks*. First Cycle Project, KTH Royal Institute of Technology, 2022.
- [16] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, December 1943.
- [17] Rubén Moreno-Bote and Jan Drugowitsch. Causal Inference and Explaining Away in a Spiking Network. *Scientific Reports*, 5(1):17531, December 2015.

- [18] Brady Neal. *Introduction to Causal Inference from a Machine Learning Perspective*. Course Lecture Notes, 2020.
- [19] OpenAI. GPT-4 Technical Report, March 2024.
- [20] Helene Paugam-Moisy. Spiking Neuron Networks: A Survey. Technical Report 06-11, IDIAP Research Report, February 2006.
- [21] Judea Pearl. An introduction to causal inference. *International Journal of Biostatistics*, 6(2), February 2010. Publisher: De Gruyter.
- [22] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, February 2019.
- [23] Judea Pearl. *Causal Analysis in Theory and Practice*, 2020.
- [24] Judea Pearl. Radical empiricism and machine learning research. *Journal of Causal Inference*, 9(1):78–82, May 2021.
- [25] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, New York, first edition edition, 2018.
- [26] José Pombal, André F. Cruz, João Bravo, Pedro Saleiro, Mário A. T. Figueiredo, and Pedro Bizarro. Understanding Unfairness in Fraud Detection through Model and Data Bias Interactions, July 2022.
- [27] Stuart J. Russell, Peter Norvig, Ming-wei Chang, Jacob Devlin, Anca Dragan, David Forsyth, Ian Goodfellow, Jitendra Malik, Vikas Mansinghka, Judea Pearl, and Michael J. Wooldridge. *Artificial intelligence: a modern approach*. Pearson, 4th edition, 2022.
- [28] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, February 2023. arXiv:2302.13971 [cs].
- [29] Hugh R. Wilson. *Spikes, decisions, and actions: the dynamical foundations of neuroscience*. Oxford University Press, Oxford ; New York, 1999.
- [30] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous Optimization for Structure Learning. *Advances in Neural Information Processing Systems*, 31, March 2018. Publisher: Curran Associates, Inc.

Efficient Transformer Model with Spiking Neural Networks

Naveed Ahmad

Advised by Prof. Dr. Bernardete Ribeiro, and Prof. Dr. Catarina Silva

12th April 2023

1 Abstract

Transformer models have impacted the adoption of machine learning, influencing its use cases across different domains, specially in natural language processing and computer vision, with innovations like Generative Pre-trained Transformers for text generation and Vision Transformers for image classification. Spiking Neural Networks are bio-inspired Neural Networks getting inspiration from the brain's neural architecture, which makes their energy efficiency and proficiency in handling temporal data important characteristics. In this research we aim to explore the computational strengths of Transformer models and the energy saving capabilities of Spiking Neural Networks. We will explore ways in which we can take advantages of the inherent energy efficiency and better temporal data processing capabilities of Spiking Neural Networks, and apply them to narrow down the challenges in Transformer models such as their high computational demands and lack of efficiency in processing time based data. We believe that Spiking Neural Networks have a potential to offer a way to elevate Transformers, making them more suitable for real time, energy efficient applications while maintaining their exceptional performance in complex data understanding tasks.

2 Introduction

With the recent surge of interest in Transformer Models and Spiking Neural Networks, numerous techniques and methodologies for evaluating transformer models have been introduced [1]. This study will select a specific domain within transformer models and select the evaluation techniques that are more relevant to the domain of interest. This targeted approach will focus on the following research questions.

- RQ1 What are the architectural modifications needed to incorporate spiking neurons into Transformer models without significant loss in performance?

- RQ2 What are the new or existing benchmarking and performance metrics that are needed to fully assess the capabilities of SNN-Transformer hybrid models?
- RQ3 What are the ways in which the combination of SNNs and Transformer models impact computational efficiency and scalability?
- RQ4 What are the specific domains or applications where SNN-Transformer hybrid models could significantly outperform traditional approaches?
- RQ5 What are the challenges that arise in terms of explainability and interpretability when integrating Spiking Neural Networks with Transformer models?

3 Related Work

Researchers have started to explore how Spiking Neural Networks (SNNs) and Transformers, two distinct types of artificial intelligence models, can be merged [2]. This field is relatively new. However, existing studies on Spiking Neural Networks highlights some important aspects. They help us understand how Spiking Neural Networks learn, their computational strengths, and certain characteristics that might explain how they could be integrated with Transformers to create hybrid models.

Jahin et al [3] developed a framework called TRABSA, integrating attention-based BiLSTM and Twitter-RoBERTa within a Transformer-based architecture. This hybrid approach illustrates the potential for combining different neural models, including SNNs, to achieve interpretability and efficiency in analyzing social media data. A study by Lemaire et al [4] found that SNNs are 6 to 8 times more energy efficient than classical neural networks. This efficiency is crucial for building sustainable computing systems. A study by Souvik Kundu et al [5] discusses recent advances in scalable, energy-efficient SNNs, focusing on algorithmic and optimization innovations for complex machine learning applications. These enhancements allow SNNs to operate with low latency and high efficiency. Alexander Kugele et al [6] presented a method for achieving state-of-the-art accuracy in SNNs for sequence processing tasks. This approach minimizes delays and mismatches between ANN training and SNN implementation, ensuring high accuracy in temporal data processing. A study by Byunggook Na et al [7] focuses on AutoSNN, aiming to improve the accuracy and reduce the spike generation in SNNs. The framework introduces spike-aware neural architecture search, enhancing both accuracy and energy efficiency. Rachmad Vidya Wicaksana Putra and Muhammad Shafique proposed TopSpark [8], a methodology enabling energy-efficient SNN processing. By leveraging adaptive timestep reduction, TopSpark maintains high accuracy with reduced energy use during training and inference phases. Li and Li [9] explores unsupervised and supervised learning methods for SNNs, such as Spike-Timing-Dependent Plasticity

(STDP), Tempotron, and ReSuMe. These methods enhance the understanding of how biological neural networks process spatio-temporal information, which is pivotal for integrating explainability in AI systems. The "SPIKE-DRIVEN TRANSFORMER V2" by Yao et al.[2] introduces Meta-SpikeFormer, a novel Transformer-based Spiking Neural Network architecture that significantly outperforms current models in efficiency and versatility for vision tasks, marking a substantial advancement in neuromorphic computing and SNN design. The "Attention Spiking Neural Networks" paper by Yao et al. [10] revolutionizes Spiking Neural Networks (SNNs) by integrating attention mechanisms, significantly narrowing the performance gap with traditional ANNs. It demonstrates unparalleled efficiency and performance gains in image classification and event-based action recognition, setting new benchmarks for energy-efficient neural network designs. The paper "Spikformer: When Spiking Neural Network Meets Transformer" by Zhou et al. [11] proposes Spikformer, integrating spiking neural networks (SNNs) with the Transformer architecture through a novel Spiking Self-Attention (SSA) mechanism. Fantozzi and Naldi's study "The Explainability of Transformers: Current Status and Directions" [12] offers an in-depth analysis of transformer model explainability, categorizing existing methods and their applications, with a focus on the significance of attention-based strategies for improving model interpretability. Vaswani et al. [13] "Attention Is All You Need" introduces the Transformer model, revolutionizing sequence modeling by relying solely on attention mechanisms, eliminating the need for recurrence and convolution in neural networks. This approach enables more parallelization, reducing training times and setting new benchmarks in translation tasks through improved efficiency and accuracy. Maass's "Networks of Spiking Neurons" [14] highlights the computational edge of spiking neural networks (SNNs) over earlier neural network models, showcasing their efficiency and versatility in simulating complex neural dynamics and computational tasks with minimal resources.

4 Approach

Conducting research on integration of Spiking Neural Networks (SNNs) with Transformer architectures represents a challenging venture in the field of artificial intelligence. The approach of this study is outlined in a structured manner as following:

1. Conceptualization on how Spiking Neural Networks and Transformers can be combined and what challenges and opportunities this presents for explainability.
2. Conducting a comprehensive literature review on Spiking Neural Networks, Transformer models, and explainable AI. Utilizing databases like IEEE Xplore, Google Scholar, PubMed, ArXive, Published Thesis and Dissertations at oatd.org.
3. Define what explainability means in the context of this research work, including relevant quantitative metrics or qualitative aspects.

4. Represent the Application and Validation of in a case study.

4.1 Methodology

The methodology for this research work is structured into data selection, model development, and development of analytical techniques as following:

1. Decide on the datasets or simulations necessary for this research.
2. Outlining the process for developing or utilizing existing Spiking Neural Networks and Transformer models. Considering open-source frameworks and tools.
3. choose or develop techniques to analyze and interpret the models' decisions. These may include visualization tools, attention mechanisms, or other explainable AI methods.

4.2 Experimentation

1. Coding and implementing the models using a step-by-step approach, documenting all steps.
2. Application of the defined explainability metrics or qualitative assessments to evaluate the models. Including comparative analyses against benchmarks if available.
3. Based on initial findings, iteratively refine models and explainability analyses to achieve desired results.

4.3 Analysis

1. Interpretation of the collected data, focusing on the aspects that highlight the explainability of SNN-Transformer hybrid models.
2. Comparative analysis of findings with existing models (both hybrid and non-hybrid) to understand the advantages or disadvantages in terms of explainability.

4.4 Documentation

1. Write multiple research papers along the way, following the conventional structure: Introduction, Literature Review, Methodology, Results, Discussion, Conclusion, and References targeting the following Conferences and Journals.
 - National Conferences: Portuguese Conference on Artificial Intelligence (EPIA), and Portuguese Conference on Pattern Recognition (RECPAD);

- International Conferences: International Joint Conferences on Artificial Intelligence (IJCAI), European Conference on Artificial Intelligence (ECAI), Conference on Neural Information Processing Systems (NeurIPS), International Conference on Information and Knowledge Management (CIKM), and International Conference on Machine Learning (ICML);
 - Journals: Journal of Artificial Intelligence Research (JAIR), Elsevier Neural Networks, Elsevier Neurocomputing, Journal of Causal Inference, Progress in Artificial Intelligence (PRAI), and AI Communications.
2. Prepare presentations to communicate findings to different audiences, adapting the complexity of explanations accordingly.

5 Expected Contribution

The expected contributions of research focusing the Spiking Neural Networks (SNNs) and Transformer hybrid models and explainability spans into several dimensions, offering both theoretical and practical advancements in the field of artificial intelligence and neural networks:

- This research will provide real-world and practical insights into the integration of bio-inspired Spiking Neural Networks with Transformer models, uncovering how such hybrids can be constructed for enhanced computational efficiency.
- By focusing on the explainability of SNN-Transformer hybrids, this study contributes to the broader field of explainable AI, which may offer methodologies and frameworks to make complex neural networks more interpretable.
- Given the temporal data processing strength of Spiking Neural Networks and the sequential data prowess of Transformers, this research may find new ways to process and integrate diverse data types effectively, providing a practical basis for future AI applications.
- This work might lead to the creation of new metrics and visualization techniques tailored to assess and improve the explainability of hybrid neural models, advancing the toolkit available for explainable research.
- By developing techniques that assist in interpreting the decision-making processes of SNN-Transformer hybrids, the research could contribute methodologically to simplifying the understanding of hybrid model behaviors.

References

- [1] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, “Efficient transformers: A survey,” *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–28, 2022.
- [2] M. Yao, J. Hu, Z. Zhou, L. Yuan, Y. Tian, B. Xu, and G. Li, “Spike-driven transformer,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [3] M. A. Jahin, M. S. H. Shovon, and M. Mridha, “TRABSA: Interpretable Sentiment Analysis of Tweets using Attention-based BiLSTM and Twitter-RoBERTa,” *arXiv preprint arXiv:2404.00297*, 2024.
- [4] E. Lemaire, L. Cordone, A. Castagnetti, P.-E. Novac, J. Courtois, and B. Miramond, “An analytical estimation of spiking neural networks energy efficiency,” in *International Conference on Neural Information Processing*, pp. 574–587, Springer, 2022.
- [5] S. Kundu, R.-J. Zhu, A. Jaiswal, and P. A. Beerel, “Recent Advances in Scalable Energy-Efficient and Trustworthy Spiking Neural Networks: from Algorithms to Technology,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13256–13260, IEEE, 2024.
- [6] A. Kugele, T. Pfeil, M. Pfeiffer, and E. Chicca, “Efficient processing of spatio-temporal data streams with spiking neural networks,” *Frontiers in neuroscience*, vol. 14, p. 512192, 2020.
- [7] B. Na, J. Mok, S. Park, D. Lee, H. Choe, and S. Yoon, “Autosnn: Towards energy-efficient spiking neural networks,” in *International Conference on Machine Learning*, pp. 16253–16269, PMLR, 2022.
- [8] R. V. W. Putra and M. Shafique, “TopSpark: a timestep optimization methodology for energy-efficient spiking neural networks on autonomous mobile agents,” in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3561–3567, IEEE, 2023.
- [9] S.-L. LI and J.-P. LI, “Research on learning algorithm of spiking neural network,” in *2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing*, pp. 45–48, IEEE, 2019.
- [10] M. Yao, G. Zhao, H. Zhang, Y. Hu, L. Deng, Y. Tian, B. Xu, and G. Li, “Attention Spiking Neural Networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9393–9404, 2023.
- [11] Z. Zhou, Y. Zhu, C. He, Y. Wang, S. Yan, Y. Tian, and L. Yuan, “Spik-former: When Spiking Neural Network Meets Transformer,” 2022.

- [12] P. Fantozzi and M. Naldi, “The Explainability of Transformers: Current Status and Directions,” *Computers*, vol. 13, p. 92, 2024.
- [13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, pp. 5998–6008, 2017.
- [14] W. Maass, “Networks of Spiking Neurons: The Third Generation of Neural Network Models,” *Neural Networks*, vol. 10, no. 9, pp. 1659–1671, 1997.

Student: Gonçalo Laranjeira Pires dos Santos Costa

Title: Spiking Neural Networks for Epileptic Seizure Forecasting using Long-term and Ultra-long-term Data

Supervisors: César Teixeira (Departamento de Engenharia Informática da Universidade de Coimbra); Matthias Dümpelmann (University Medical Center Freiburg)

Abstract

Epilepsy is one of the most common neurological diseases. It affects around 1% of the world's population and is characterized by recurrent seizures. One significant problem patients with epilepsy suffer from is the unpredictability aspect of seizures¹.

The first line of treatment is the use of Anti-Epileptic Drugs. However, approximately one-third of patients with epilepsy suffer from Drug-Resistant Epilepsy, a condition where the use of medication is not enough to achieve seizure-free lives^{2,3}. The inability to control seizures can lead to physical problems, such as an increased risk of accidental injury, brain injury, cognitive decline, or even death, and psychological ones, such as neuropsychological deficits (memory loss and attention difficulties), depression, anxiety, or psychoses^{4,5}.

In cases where seizure control cannot be achieved through medication, the next best alternative is to inform the patients when a seizure will occur. This objective can be accomplished by using a seizure warning device that estimates a continuous seizure occurrence likelihood⁶.

One approach for warning device algorithms is seizure forecasting. In forecasting, the algorithm provides a constant analysis of the likelihood of a seizure. It should be able to detect, for instance, low, moderate, and high-risk states and continuously relay the information to the user. Even though seizure forecasting takes a probabilistic approach, which is rarely fully confident of events, it avoids the crisp approach of other techniques that, despite sometimes being correct, are linked to damaging false alarms. Access to the probabilistic likelihood of seizures allows patients to make a more informed decision based on a certain degree of uncertainty^{7,8}.

The brain's electrical activity can be recorded by the electroencephalogram (EEG), one of the most effective tools for identifying, examining, and interpreting brain disorders. There are two methods for signal acquisition: scalp EEG and intracranial EEG (iEEG). The iEEG can achieve considerable precision and area specificity, and the ability to reach the deep brain results in a cleaner signal. However, as it is invasive, it brings the risks associated with these sorts of procedures^{9,10}. Sub-scalp EEG is a technology that has gained traction recently, where the electrodes are placed between the skull and the scalp. This technique shows signal quality improvements from scalp EEG and does not require the same electrode care as iEEG. Its minimal invasiveness allows for easy use outside of the hospital environment and facilitates ultra-long-term recordings¹¹. This signal can also be segmented into different periods in time: preictal (before a seizure), ictal (during a seizure), postictal (after a seizure), and interictal (between the postictal and preictal periods of two consecutive seizures)¹².

Over the years, many studies using EEG-based methodologies to try to forecast seizures have been developed^{7,13,14}. Cook et al.¹³ describe the NeuroVista clinical trial, the first human study to successfully demonstrate prospective seizure forecasting using ambulatory EEG recordings. In this phase I clinical trial, 15 patients with DRE were implemented with the NeuroVista Seizure Advisory System for a period of between six months and three years. Using iEEG, the system could register the signal and determine a low, moderate, or high seizure risk. This trial showed that 11 subjects performed well, generating a sensitivity higher than 65% and performance more significant than randomly forecasting events. Nevertheless, due to the results not being sufficiently good and the belief that the technology was too invasive, the funding was cut¹⁵.

However, since then, few methodologies have been clinically applicable. This inefficacy is mainly for three reasons: the EEG signal is still not fully understood; determining and detecting the preictal period is an incredibly arduous task; and a lack of proper databases. Most public EEG databases contain only a few days of data from long-term presurgical monitoring, where patients are subject to AED withdrawal and sleep deprivation. These conditions do not reflect everyday seizure activity. Significant advancements come with the transition to ultra-long-term databases containing months to years of daily-life data^{8,16,17}.

Increasing computational power and quantity of available data during the past years have caused the emergence of deep learning (DL) models for potential warning device algorithms. These models can handle raw data with little to no preprocessing and are capable of feature engineering and classification. Therefore, DL can move past the need to know the data's structure and detect specific brain signal patterns rather than features¹⁸. Popular options for DL classifiers in the state-of-the-art include convolutional neural networks (CNNs) and long short-term memory (LSTM) networks.

One interesting DL classifier that has yet to be used in seizure forecasting is spiking neural networks (SNNs). Typical Artificial Neural Networks use static and continuous activation functions to transmit information between neurons. On the other hand, SNNs try to replicate the biological behavior of the brain, using discrete spikes to transmit the information and encoding it in the timing and the rate of the spikes. The sparse and event-driven nature of spike-based communication characteristic of SNNs displays great potential for real-time, energy-efficient computation. Furthermore, SNNs are excellent when dealing with spatio-temporal data, specifically in online learning. All these characteristics make for a perfect environment for the implementation of an EEG classification algorithm in a real-time warning device^{19,20}.

Several studies show promising results for SNN application in seizure detection algorithms²¹⁻²³. Yang et al.²³ even show comparable performance results between SNNs and traditional DL neural networks (CNN and LSTM).

Research Questions

Brain dynamics are a topic of great complexity. A critical part of seizure forecasting is determining and detecting the preictal period, a transitional moment between regular brain activity and a seizure. However, there is no consensus among authors on its length, and some evidence even suggests the preictal period may vary between patients and between the same patient's seizures^{16-18,24}. Recently, studies found the existence of seizure cycles based on biological rhythms and the sleep-wake cycles. These cycles could be a way to complement the EEG signal features and improve the models^{16,18,25,26}. Furthermore, if an algorithm is to be applied into a seizure warning device, energy-efficient classifiers should be explored. With all this in mind, this works proposes tackling the following research questions (RQ):

- RQ1: Do biological rhythms (circadian, multidiem, and ultradiem) and the sleep-wake cycle have an influence in seizure occurrence?
- RQ2: Could the inclusion of seizure cycles in seizure forecasting models improve their performance?
- RQ3: Can an energy-efficient SNN be used as a seizure forecasting model's classifier and produce promising performance results?

Goals and Contributions

Using long-term scalp and intracranial EEG from the EPILEPSIAE database²⁷ and ultra-long-term sub-scalp EEG collected at the Universitätsklinikum Freiburg, this work is expected to have the following contributions:

1. Development of a preprocessing pipeline that filters the EEG signals and removes the noise and artifacts suitable for real-time application;
2. Exploration of seizure cycles based on biological rhythms/the sleep-wake cycle and their influence on forecasting seizures;
3. Development of a spike learning seizure forecasting model using an SNN;
4. Exploration of performance differences between the energy-efficient SNN and typical deep learning approaches;
5. Exploration of performance differences between the long-term and ultra-long-term data;
6. **(CONDITIONAL - If promising results come from the algorithm's performance evaluation)** Hardware implementation of the algorithm as a possible real-time warning device prototype.

Bibliography

1. P. Van Mierlo, M. Papadopoulou, E. Carrette, P. Boon, S. Vandenberghe, K. Vonck, and D. Marinazzo, "Functional brain connectivity from eeg in epilepsy: Seizure prediction and epileptogenic focus localization," *Progress in neurobiology*, vol. 121, pp. 19–35, 2014.
2. Y. Wang and Z. Chen, "An update for epilepsy research and antiepileptic drug development: Toward precise circuit therapy," *Pharmacology & therapeutics*, vol. 201, pp. 77–93, 2019.
3. M. Sazgar and B. F. Bourgeois, "Aggravation of epilepsy by antiepileptic drugs," *Pediatric neurology*, vol. 33, no. 4, pp. 227–234, 2005.
4. B. K. Steiger and H. Jokeit, "Why epilepsy challenges social life," *Seizure*, vol. 44, pp. 194–198, 2017.
5. M. R. Sperling, "The consequences of uncontrolled epilepsy," *CNS spectrums*, vol. 9, no. 2, pp. 98–109, 2004.
6. C. Baumgartner, J. P. Koren, and M. Rothmayer, "Automatic computer-based detection of epileptic seizures," *Frontiers in neurology*, vol. 9, p. 639, 2018.
7. R. E. Stirling, M. J. Cook, D. B. Grayden, and P. J. Karoly, "Seizure forecasting and cyclic control of seizures," *Epilepsia*, vol. 62, pp. S2–S14, 2021.
8. M. O. Baud, T. Proix, N. M. Gregg, B. H. Brinkmann, E. S. Nurse, M. J. Cook, and P. J. Karoly, "Seizure forecasting: bifurcations in the long and winding road," *Epilepsia*, 2022.
9. A. Varsavsky, I. Mareels, and M. Cook, *Epileptic seizures and the EEG: measurement, models, detection and prediction*. Taylor & Francis, 2011.
10. I. Osorio, H. P. Zaveri, M. G. Frei, and S. Arthurs, *Epilepsy: the intersection of neurosciences, biology, mathematics, engineering, and physics*. CRC press, 2016.
11. Z. Haneef, K. Yang, S. A. Sheth, F. Z. Aloor, B. Aazhang, V. Krishnan, and C. Karakas, "Sub-scalp electroencephalography: A next-generation technique to study human neurophysiology," *Clinical Neurophysiology*, 2022.

12. N. Moghim and D. W. Corne, "Predicting epileptic seizures in advance," *PloS one*, vol. 9, no. 6, p. e99334, 2014.
13. M. J. Cook, T. J. O'Brien, S. F. Berkovic, M. Murphy, A. Morokoff, G. Fabinyi, W. D'Souza, R. Yerra, J. Archer, L. Litewka *et al.*, "Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study," *The Lancet Neurology*, vol. 12, no. 6, pp. 563–571, 2013.
14. P. Nejedly, V. Kremen, V. Sladky, M. Nasser, H. Guragain, P. Klimes, J. Cimbalnik, Y. Varatharajah, B. H. Brinkmann, and G. A. Worrell, "Deep-learning for seizure forecasting in canines with epilepsy," *Journal of neural engineering*, vol. 16, no. 3, p. 036031, 2019.
15. L. Drew, "Like taking away a part of myself"—life after a neural implant trial," *Nat Med*, vol. 26, no. 8, pp. 1154–1156, 2020.
16. F. Mormann, R. G. Andrzejak, C. E. Elger, and K. Lehnertz, "Seizure prediction: the long and winding road," *Brain*, vol. 130, no. 2, pp. 314–333, 2007.
17. U. R. Acharya, S. V. Sree, G. Swapna, R. J. Martis, and J. S. Suri, "Automated eeg analysis of epilepsy: a review," *Knowledge-Based Systems*, vol. 45, pp. 147–165, 2013.
18. D. R. Freestone, P. J. Karoly, and M. J. Cook, "A forward-looking review of seizure prediction," *Current opinion in neurology*, vol. 30, no. 2, pp. 167–173, 2017.
19. A. Tavanaei, M. Ghodrati, S. R. Kheradpisheh, T. Masquelier, and A. Maida, "Deep learning in spiking neural networks," *Neural networks*, vol. 111, pp. 47–63, 2019.
20. J. L. Lobo, J. Del Ser, A. Bifet, and N. Kasabov, "Spiking neural networks and online learning: An overview and perspectives," *Neural Networks*, vol. 121, pp. 88–100, 2020.
21. S. Ghosh-Dastidar and H. Adeli, "Improved spiking neural networks for eeg classification and epilepsy and seizure detection," *Integrated Computer-Aided Engineering*, vol. 14, no. 3, pp. 187–212, 2007.
22. H. Shan, L. Feng, Y. Zhang, L. Yang, and Z. Zhu, "Compact seizure detection based on spiking neural network and support vector machine for efficient neuromorphic implementation," *Biomedical Signal Processing and Control*, vol. 86, p. 105268, 2023.
23. Y. Yang, J. K. Eshraghian, N. D. Truong, A. Nikpour, and O. Kavehei, "Neuromorphic deep spiking neural networks for seizure detection," *Neuromorphic Computing and Engineering*, vol. 3, no. 1, p. 014010, 2023.
24. E. B. Assi, D. K. Nguyen, S. Rihana, and M. Sawan, "Towards accurate prediction of epileptic seizures: A review," *Biomedical Signal Processing and Control*, vol. 34, pp. 144–157, 2017.
25. L. Kuhlmann, P. Karoly, D. R. Freestone, B. H. Brinkmann, A. Temko, A. Barachant, F. Li, G. Tittericz Jr, B. W. Lang, D. Lavery *et al.*, "Epilepsycosystem. org: crowd-sourcing reproducible seizure prediction with long-term human intracranial eeg," *Brain*, vol. 141, no. 9, pp. 2619–2630, 2018.
26. P. J. Karoly, H. Ung, D. B. Grayden, L. Kuhlmann, K. Leyde, M. J. Cook, and D. R. Freestone, "The circadian profile of epilepsy improves seizure forecasting," *Brain*, vol. 140, no. 8, pp. 2169–2182, 2017.
27. J. Klatt, H. Feldwisch-Drentrup, M. Ihle, V. Navarro, M. Neufang, C. Teixeira, C. Adam, M. Valderrama, C. Alvarado-Rojas, A. Witon, M. Le Van Quyen, F. Sales, A. Dourado, J. Timmer, A. Schulze-Bonhage, and B. Schelter, "The EPILEPSIAE database: An extensive electroencephalography database of epilepsy patients," *Epilepsia*, vol. 53, no. 9, pp. 1669–1676, 9 2012.

Assessment of reliability in diagnosis based on artificial intelligence tools from medical images

Candidate: Francisco Meneses Alves

Advisors: Lorena Itatí Petrella (DEI) and Mário João Simões Ferreira dos Santos (DEEC)

Thesis Summary:

Recent progress in artificial intelligence (AI), particularly in machine learning (ML), has showcased remarkable performances across various tasks and fields, notably with the introduction of deep ML techniques [1,2]. Despite these advancements, barriers to full adoption in practice persist, particularly in critical tasks such as clinical decision-making. Real-world applications of ML models have been limited due to issues such as their black-box nature, which lacks transparency and explanation capability [1,2]. However, while explainable AI (XAI) techniques improve interpretability, they do not adequately address the practical assessment of decision reliability.

One fundamental question in machine learning is whether one can trust the predictions these tools provide, especially in critical applications [3]. Thus, alongside prediction outcomes, the existence of a measure of reliability is crucial for the practical application of AI-based models. This adds a layer of security and transparency, increasing trust in human-AI interaction and aiding in model improvement. Individual prediction reliability estimates can be directly shown to users, helping them determine the adequacy of the current ML model for specific instances, especially in high-risk settings such as healthcare scenarios [3-6].

During the training phase, it is feasible to assess the performance of a model using specific and standard metrics, such as sensitivity and specificity. However, these estimations offer an average performance perspective and do not provide a means to evaluate individual instances. Moreover, these metrics may not be able to generalize on new unseen instances that differ from the training population, thus providing unreliable classifications [5]. As a result, estimating pointwise reliability would be particularly valuable in assessing the model's performance on a case-by-case basis, beyond the usual metrics. It would be an advantage to support the model improvement, to increase the trust in human-AI interaction and to know if the available model is suitable for decide about individual instances [5,6].

Despite its application across a variety of fields and with specific meanings, in the context of prediction problems, reliability has been defined as the probability of the classification of an instance being correct [4,5].

The reliability of ML models has been exploited using two main approaches: density and the local fit principles [4,7]. The first principle uses distance measures and data patterns to compare the instance under interest to the examples present in the training set, or how close is a given sample to the training set. The second assesses how the trained model performs on training subsets similar to the instance under evaluation to derive a reliability

measure, or how accurate is the model on the training samples closest to the test case. Basically, the model will perform well on instances similar to training instances that exist on the data set and where the model performs correctly [4,7].

In many applications, medical image classification based on ML models involves image processing techniques for features extraction, which are then used as inputs to the models. On the other hand, techniques based on deep learning used the images itself as input, without the need of extracting features. While conventional reliability estimation methods are extensible to image classification based on features extraction, no literature has been found for the cases involving deep learning algorithms.

The objective of this project is to develop reliability assessment methods to complement the classification (diagnosis) using AI tools, based on medical images. Initially, image classification will be based on features extraction, and reliability methods will be adapted to these models. Moreover, the development of methods for evaluating reliability using deep learning algorithms will be addressed.

Although the main goal is to develop reliability assessment methods for medical images in general, studies will begin on images used in Transcatheter Aortic Valve Implantation (TAVI) planning.

Aortic stenosis, characterized by the narrowing of the heart's aortic valve, poses a significant threat to cardiovascular health [8]. This condition restricts blood flow from the heart to the body, potentially resulting in severe complications such as heart failure, stroke, blood clots, and even cardiac arrest. While aortic stenosis can stem from various causes, including congenital heart defects, calcium buildup on the valve, and rheumatic fever, its impact on patients' lives necessitates effective treatment strategies [8].

TAVI has emerged as a minimally invasive surgical procedure designed to address aortic stenosis, particularly in patients deemed high risk or inoperable for traditional open-heart surgery [8,9]. This innovative procedure involves the insertion of a replacement valve through a small incision, typically in the thigh, and guiding it to the heart via the blood vessels. Once positioned, the new valve is expanded to assume the function of the diseased native valve [8].

Despite its efficacy, TAVI is not without risks. Up to 37% of TAVI interventions may induce conduction abnormalities in the heart, particularly atrioventricular (AV) blocks [10]. The proximity of the aortic valve to the cardiac electrical conduction system renders it susceptible to pressure exertion during valve expansion, that can lead to varying degrees of AV block [8]. Several factors contribute to the specific risk of developing an AV block post-TAVI, including the type of valve implanted, pre-existing conduction abnormalities, procedure-related factors, and patient-specific anatomical variations.

The treatment for post-TAVI AV blocks typically involves the implantation of a permanent pacemaker [11]. However, this solution comes with implications for patients' quality of life and additional costs for healthcare services, encompassing the pacemaker device, implantation procedure, and subsequent patient follow-up [12].

Effective TAVI planning hinges on meticulous assessment of morphological and histological aspects of the cardiac system, supplemented by relevant clinical data.

Imaging techniques such as magnetic resonance imaging (MRI) and computed tomography (CT) play a crucial role in this process [8,13].

This work is part of the "Center for Responsible AI – Next Generation AI" (CRAI) project, which aims to develop the next generation of responsible AI products.

Workplan Tasks:

Task 1) State-of-the-art review:

In this task, the current methods for estimating the reliability of outputs provided by machine learning methods and medical image processing and types of injuries and their respective characteristics in these types of images will be explored. It should be noted that even though a preliminary review has already been conducted in the context of this proposal, we remain attentive to any recent developments or advancements that may be relevant to our objectives.

Task 2) Work with TAIIV image database:

In this task, pre-processing and extraction of a set of features that best characterize the TAIIV images will be conducted as well as the construction of an AI model to assess reliability based on the features obtained. On the other hand, an AI model based on deep learning using the images itself as input, without the need of extracting features, will be studied. To assess reliability, methods based on the density principle will verify if the new image resembles images used in training and methods based on the model fit principle will compare the system output with the output for similar images.

Task 3) Development of a reliability assessment method for medical images in general:

Develop reliability assessment methods to complement the classification (diagnosis) using AI tools, based on medical images.

Bibliography:

[1] Barredo Arrieta, Alejandro & Díaz-Rodríguez, Natalia & Del Ser, Javier & Bennetot, Adrien & Tabik, Siham & Barbado González, Alberto & Garcia, Salvador & Gil-Lopez, Sergio & Molina, Daniel & Benjamins, V. Richard & Chatila, Raja & Herrera, Francisco. (2019). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*. 58. 10.1016/j.inffus.2019.12.012.

[2] Parimbelli, Enea & Buonocore, Tommaso & Nicora, Giovanna & Michalowski, Wojtek & Wilk, Szymon & Bellazzi, Riccardo. (2022). Why did AI get this one wrong? - Tree-based explanations of machine learning model predictions. *Artificial Intelligence in Medicine*. 135. 102471. 10.1016/j.artmed.2022.102471.

- [3] A.S. Albahri, Ali M. Duhaim, Mohammed A. Fadhel, Alhamzah Alnoor, Noor S. Baqer, Laith Alzubaidi, O.S. Albahri, A.H. Alamoodi, Jinshuai Bai, Asma Salhi, Jose Santamaría, Chun Ouyang, Ashish Gupta, Yuantong Gu, and Muhammet Deveci. 2023. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Inf. Fusion* 96, C (Aug 2023), 156–191. <https://doi.org/10.1016/j.inffus.2023.03.008>
- [4] Nicora G, Rios M, Abu-Hanna A, Bellazzi R. Evaluating pointwise reliability of machine learning prediction. *J Biomed Inform.* 2022;127:103996. doi:10.1016/j.jbi.2022.103996
- [5] S. Saria and A. Subbaswamy. Tutorial: Safe and reliable machine learning. In *ACM Conference on Fairness, Accountability, and Transparency*, 2019.
- [6] Gediminas Adomavicius and Yaqiong Wang. 2022. Improving Reliability Estimation for Individual Numeric Predictions: A Machine Learning Approach. *INFORMS J. on Computing* 34, 1 (Winter 2022), 503–521. <https://doi.org/10.1287/ijoc.2020.1019>
- [7] P. Schulam, S. Saria, Can You Trust This Prediction? Auditing Pointwise Reliability After Learning'. *ArXiv:1901.00403 [Cs, Stat]*, 2019. February, <http://arxiv.org/abs/1901.00403>.
- [8] Vahanian et al. 2021 Esc/EACTS guidelines for the management of valvular heart disease. *Eur Heart J.* (2022) 43(7):561–632.
- [9] Tarantini et al. Management of coronary artery disease in patients undergoing TAVI. *EuroIntervention.* 2023 May 15;19(1):37-52.
- [10] Masoulié et al. Incidence and management of atrioventricular conduction disorders in new-onset left bundle branch block after TAVI: A prospective multicenter study, *Heart Rhythm*, Volume 20, Issue 5, May 2023, Pages 699-706
- [11] Badertscher et al. Management of conduction disorders after transcatheter aortic valve implantation: results of the EHRA survey, *Europace* (2022), 00, 1–7
- [12] Chevreur et al. Cost of TAVI and factors associated with higher hospital stay cost in patients of the FRANCE registry. *Archives of Cardiovascular Disease* (2013) 106, 209-219
- [13] Watson et al. Imaging technologies for cardiac fiber and heart failure: a review. *Heart Fail Rev.* 2018 Mar; 23(2): 273–289.

Title:

Advancing Robustness Assessment for Microservices Architectures using Generative Artificial Intelligence

Student:

Gonçalo Nogueira¹

Supervisors:

Nuno Laranjeiro¹, César Teixeira¹

¹University of Coimbra, Centre for Informatics and Systems of the University of Coimbra, Department of Informatics Engineering, Portugal

Postal address: Departamento de Engenharia Informática, Pólo II - Pinhal de Marrocos, 3030-290 Coimbra, Portugal

goncnogueira@dei.uc.pt, cnl@dei.uc.pt, cteixei@dei.uc.pt

Project Thesis Synopsis

Microservices have become a popular architectural style for developing large-scale software applications [1]. This approach involves dividing an application into a set of loosely coupled services, each responsible for specific business functionalities. The benefits of this approach include scalability, flexibility, and resilience, which enable organizations to develop and deploy complex systems more effectively. Although REST APIs are not synonymous with microservices architectures, they often play a crucial role within them [1,2]. The services that can be deployed independently communicate with each other through APIs. RESTful principles are often used to define these interfaces because of their simplicity and scalability.

Ensuring the **robustness of microservices**-based applications is crucial to maintaining consistency with their specifications [2,3]. Robust systems can operate effectively under diverse conditions, including invalid inputs, stressful environments, and potential failures within the distributed architecture [4,5]. However, assessing the robustness of microservices presents unique challenges, particularly due to the lack of standardized interface descriptions and the varied handling of inputs across services [6]. Additionally, the inclusion of middleware layers further complicates robustness evaluations, contributing to limited research in this domain.

In the context of testing RESTful web services, several tools have been developed. The method typically involves two primary phases: test specification and test case generation and is specification-based [7]. The process begins by parsing the specification file, followed by generating a workload with valid requests that comply with the REST service specification. After sending valid requests to the service, faults are injected to create invalid requests, which are then sent to the service again. REST specifications are often documented using open formats such as OpenAPI/Swagger [8], which detail

endpoints, request/response formats and parameters. This approach can be adopted for testing microservices environments by generating test cases from the system's inner API specifications using automatic techniques developed for RESTful web services.

Several tools have been designed for testing REST APIs, being some of them identified in this section. bBOXRT [9] focuses on robustness assessment by injecting faults into requests to trigger errors and test service robustness. RESTler [10] is a Python-based stateful fuzzer for REST APIs, analyzing OpenAPI specifications to generate and execute tests aimed at uncovering security vulnerabilities. RESTTESTGEN [11] automatically generates test cases for REST APIs using API Swagger specifications, with an added Operation Dependency Graph to prioritize testing based on data dependencies between operations. RESTCT [12] uses Combinatorial Testing to systematically test RESTful APIs, involving phases for operation sequence generation and input-parameter value rendering. MacroHive [13] provides a complete framework for automated functional and robustness testing of microservice architectures, unlike the tools mentioned previously that focus solely on RESTful APIs. It includes features such as test case generation, execution, and monitoring, as well as causal reasoning-based inference to identify the root causes of failures in microservice chains. MacroHive is distinguished by its ability to provide insight into internal coverage and failures, providing a holistic approach to testing and improving the reliability of microservices-based systems.

Techniques such as Genetic Algorithms have been applied in other tools to automatically optimize test case generation [14-16]. One notable example of tool is EvoMaster [14,15], which focuses on automated testing of RESTful Web Services. EvoMaster incorporates heuristics to enhance code coverage, employing random testing coupled with an Evolutionary Algorithm to iteratively refine randomly generated test cases. Its primary goal is to maximize both code coverage and the detection of error status code responses from the service under test, thereby maximizing HTTP response code coverage.

In addition to Evolutionary Algorithms, innovative approaches to automated software testing have emerged, such as the use of Generative Adversarial Networks (GANs) [17] and the tool RestSpecIT [18], leveraging generative AI (GAI) techniques. GANs can generate synthetic data for test cases, potentially diversifying the input space and uncovering edge cases that traditional testing may miss. Meanwhile, RestSpecIT introduces a method that utilizes GPT (Generative Pre-trained Transformer) models to automatically generate REST API specifications. By leveraging the language generation capabilities of GPT, RestSpecIT accelerates the documentation process and ensures consistency in API specifications, thereby enhancing the testing and development workflow for RESTful services.

The remarks of the authors from the state-of-the-art tools [9,13,17-23] highlight the promising ambition to automate the analysis of service behavior and the categorization of robustness problems. The emergent AI techniques represent a promising way to fill this gap. The GANs based tool and RestSpecIT tools exemplify this potential by using GANs and GPT models, respectively, for automated test generation and API specification. These tools demonstrate the transformative impact of generative AI in diversifying test scenarios and facilitating the documentation process, thereby improving

the robustness assessment of software systems. The integration of machine learning algorithms and generative AI holds great promise for advancing the automation and accuracy of robustness assessment methodologies.

This PhD aims to advance the state of the art in software engineering by designing novel approaches to assess the robustness of microservices architectures by taking advantage of the capabilities of Generative AI to advance the state of the art.

The Research Questions (RQ) that arise in this context are the following:

RQ1: How can Generative AI be integrated into the assessment of microservices robustness?

RQ2: How to automate the analysis of service behavior and categorize robustness problems making use of AI/ML techniques?

RQ3: How can Generative AI be leveraged to identify and correct robustness problems in microservices architectures?

RQ4: How GAI can be used to correct and produce better service descriptions and specifications?

The Technical Objectives (TO) are the following:

TO1: Explore methodologies for effectively integrating Generative AI techniques into the assessment of microservices robustness. This includes the development of algorithms and frameworks capable of automating the analysis and the categorization of robustness issues.

TO2: Develop and implement automated analysis techniques using AI/ML algorithms to identify and categorise robustness problems in microservices. The focus is on developing algorithms and tools that can autonomously analyse service behaviour, detect robustness issues and categorise them.

TO3: Explore the use of generative AI methods to identify and solve robustness problems within microservices architectures. This involves developing algorithms that can automatically identify areas of concern, propose corrective actions, and validate the effectiveness of these corrections in improving system robustness.

TO4: Use the capabilities of Generative AI to automatically identify and explore undocumented paths and parameters within OpenAPI specifications, improving the accuracy and coverage of automated testing for microservices architectures.

The potential contributions to advance the state of art of this PhD include:

i. Proposal and development of robustness assessment frameworks that will provide resources for identifying and addressing robustness issues in their architectures.

ii. Enable the autonomous detection and categorization of robustness problems in microservices architectures.

iii. Generative AI methods can be used to identify and resolve robustness issues in microservices architectures, providing valuable insights into software correction strategies to improve robustness.

iv. The development of techniques for generating and enhancing service documentation, will contribute to the creation of more accurate and comprehensive documentation for microservices-based systems.

v. As an outcome, the creation of a potential web tool that embodies the developed methodologies and automated analysis techniques. This tool could provide a user-friendly interface for performing robustness assessments of microservices architectures.

References

- [1] Richardson, C. (2019). *Microservices Pattern: Microservice Architecture Pattern*. Retrieved from <http://microservices.io/patterns/microservices.html>
- [2] Lewis, J., & Fowler, M. (2014). *Microservices: a definition of this new architectural term*. MartinFowler.com, 25(14-26), 12.
- [3] Avizienis, A., Laprie, J. C., Randell, B., & Landwehr, C. (2004). *Basic concepts and taxonomy of dependable and secure computing*. *IEEE transactions on dependable and secure computing*, 1(1), 11-33.
- [4] IEEE. (1990). *IEEE Standard Glossary of Software Engineering Terminology (IEEE 610.12-1990)*. Retrieved from <http://standards.ieee.org>
- [5] Sussman, G. J. (2007). *Building robust systems an essay*. Citeseer, 113, 1324.
- [6] Neumann, A., Laranjeiro, N., & Bernardino, J. (2018). *An analysis of public REST web service APIs*. *IEEE Transactions on Services Computing*, 14(4), 957-970.
- [7] Laranjeiro, N., Agnelo, J., & Bernardino, J. (2021). *A systematic review on software robustness assessment*. *ACM Computing Surveys*, 54(4), Article 89. doi:10.1145/3448977
- [8] Linux Foundation. (2022). *OpenAPI Specification*. Retrieved from <https://www.openapis.org>
- [9] Laranjeiro, N., Agnelo, J., & Bernardino, J. (2021). *A black box tool for robustness testing of REST services*. *IEEE Access*, 9, 24738-24754. doi:10.1109/ACCESS.2021.305650
- [10] Atlidakis, V., Godefroid, P., & Polishchuk, M. (2019, May). *Restler: Stateful rest api fuzzing*. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)* (pp. 748-758). IEEE.
- [11] Corradini, D., Zampieri, A., Pasqua, M., Viglianisi, E., Dallago, M., & Ceccato, M. (2022). *Automated black-box testing of nominal and error scenarios in RESTful APIs*. *Software Testing, Verification and Reliability*, 32(5), e1808.
- [12] Wu, H., Xu, L., Niu, X., & Nie, C. (2022, May). *Combinatorial testing of restful apis*. In *Proceedings of the 44th International Conference on Software Engineering* (pp. 426-437).

- [13] Giamattei, L., Guerriero, A., Pietrantuono, R., & Russo, S. (2024). Automated functional and robustness testing of microservice architectures. *Journal of Systems and Software*, 207, 111857.
- [14] Arcuri, A. (2019). RESTful API Automated Test Case Generation with EvoMaster. *ACM Transactions on Software Engineering and Methodology*, 28(1), Article 3, 1-37. DOI: 10.1145/3293455
- [15] Arcuri, A. (2020). Automated black-and white-box testing of restful apis with evomaster. *IEEE Software*, 38(3), 72-78.
- [16] Santos, C. (2022). Evolutionary robustness testing of REST services (Master's thesis). University of Coimbra.
- [17] Guo, X., Okamura, H., & Dohi, T. (2022). Automated software test data generation with generative adversarial networks. *IEEE Access*, 10, 20690-20700. doi:10.1109/ACCESS.2022.3153347
- [18] Decrop, A., Perrouin, G., Papadakis, M., Devroey, X., & Schobbens, P.-Y. (2024). You can REST now: Automated specification inference and black-box testing of RESTful APIs with large language models, arXiv, arXiv:2402.05102.
- [19] Nguyen, C., Bui, H., Nguyen, V., & Nguyen, T. (2023, December). An Approach to Generating API Test Scripts Using GPT. In *Proceedings of the 12th International Symposium on Information and Communication Technology* (pp. 501-509).
- [20] Shahrokni, A., & Feldt, R. (2013). A systematic review of software robustness. *Information and Software Technology*, 55(1), 1-17. doi:10.1016/j.infsof.2012.06.002
- [21] House of Lords. (2023). Large language models and generative AI (House of Lords Paper No. 54). Parliament. Retrieved from <https://publications.parliament.uk/pa/ld5804/ldselect/ldcomm/54/54.pdf>
- [22] Nguyen, C., Bui, H., Nguyen, V., & Nguyen, T. (2023, December). An Approach to Generating API Test Scripts Using GPT. In *Proceedings of the 12th International Symposium on Information and Communication Technology* (pp. 501-509).
- [23] Baudry, B., Etemadi, K., Fang, S., Gamage, Y., Liu, Y., Liu, Y. & Tiwari, D. (2024). Generative AI to Generate Test Data Generators. arXiv preprint arXiv:2401.17626.

FLOSS Maturity Model for Social Sector

Laura Joaquim Jaime

Advisor: Prof. João Barata

University of Coimbra, CISUC, DEI, Coimbra, Portugal

Abstract

The social sector plays a prominent role in the political, cultural, and social dimensions of society [1]. It comprises diverse organizations, such as foundations, religious institutions, or associations [2]. In Portugal [3] and Angola [4], where this research takes place, the social sector represents 9 % and 5.4 %, respectively, of the total number of organizations. Digital transformation (DT) is now on the agenda of the social sector, but the challenges are also significant [5]. For example, there is a lack of information systems (IS) and information technologies (IT) infrastructure capable of supporting organizations to serve communities better [6–8]. Many social problems have increased in intensity and complexity, calling for more technology-based solutions [8].

Free Libre Open Source Software (FLOSS) is becoming the backbone of digital transformation [3]. Open source adoption increased by 95% worldwide in 2024 [9] and presents essential characteristics for new developments: (1) provides community-based development, (2) supports organizational strategy, (3) has extensive support in forums, blogs, and community integration platforms, and (4) ensures the four freedoms: execute, study, modify and redistribute the source code [10]. However, some studies also point to difficulties in FLOSS adoption, including the development process's quality, subsequent testing processes, security policies, or compliance with regulations [10].

Maturity models (MM) have emerged as one of the most effective tools for establishing a digital transformation strategy in organizations [10]. These models describe the characteristics of effective processes, enabling better positioning of the organization through diagnoses and facilitating the discovery of better solutions for changes [11]. Moreover, MM aim to continually improve the performance of information technology to ensure value delivery to organizations and their stakeholders [11].

Many MM are available for software development, such as the popular Capability Maturity Model Integration (CMMI) [11]. There are also models specifically created for FLOSS, such as the Financial Open Source Maturity Model [12] and Open Business Readiness Rating

(OpenBRR) [13]. These models provide diagnoses for functional and non-functional requirements in different software dimensions. Other examples include (1) the FLOSS (Moodbile) project risk management model, combining information in a big data-based analytical platform with automated scripts to capture, filter, analyze, and systematically build theories about FLOSS community behaviors and possible organizational risks in their adoption [14] and (2) an analysis on the viability of migrating to FLOSS as an economic strategy for public administration digitalization in municipalities [15]. However, the models' application depends on the use case (e.g., development, security), and there are currently no maturity models that can support FLOSS adoption in the social sector. Therefore, the main research objective addressed in this thesis proposal is to develop a FLOSS maturity model tailored for the social sector. Despite the increasing digitization of society, social sector organizations often lag in harnessing the transformative potential of FLOSS, resulting in missed opportunities for efficiency, collaboration, and impact [16]. This research endeavors to address this gap by developing a structured framework to assess, guide, and enhance FLOSS adoption within social sector entities.

Design Science Research (DSR) is the selected research approach, an iterative process that includes six main steps: (1) problem identification and motivation, (2) definition of the objectives for a solution, (3) design and development, (4) demonstration, (5) evaluation, and (6) communication [17,18]. The integration of Portugal and Angola's social sector entities as real context will contribute to the transferability of results, with three DSR phases planned: (1) literature review about challenges in FLOSS adoption, diagnosis of FLOSS maturity and challenges of digital transformation in the social sector; (2) development of artefacts, and (3) artefacts deployment, evaluation, and evolution with necessary adjustments.

The initial stage of the work provided the foundations to understand the problem and motivation, establish the objective to create a maturity model to assist social sector institutions and start the literature review focusing on three themes: (1) FLOSS for the social sector, (2) FLOSS MM, and (3) MM for the social sector. Drawing on interdisciplinary insights from fields such as information technology, organizational studies, and social entrepreneurship, the review identified critical dimensions to understanding and fostering FLOSS maturity [19] within social sector contexts. These dimensions encompass (1) organizational culture, (2) technical proficiency, (3) capacity building, (4) community engagement, and (5) sustainability. The insights gathered in the literature were shared with four institutions in Portugal (Santa Casa da Misericórdia de Coimbra and Fundação ADFP) and Angola (Ombala do Saber and Fome Zero

Angola) that confirmed their interest in collaborating with the research. According to the managers of these institutions, a maturity model could guide their digital transformation with FLOSS (1) internally, improving their information system, and (2) externally, supporting their activities in the community using digital technologies, namely, healthcare, education, and cultural development. The second phase includes interviews with the participating entities to understand the organization's behaviours, specific requirements, opportunities, and challenges faced by each one to develop the artefacts. The third phase will be concluded considering the information from the interviews carried out in each institution.

Overall, this research contributes to advancing our understanding and practical application of FLOSS within the social sector, offering a tailored framework to empower social sector organizations with the tools and knowledge necessary to navigate the digital landscape effectively by bridging the technological gap and fostering digital empowerment. The proposed FLOSS maturity model holds promise as a catalyst for positive change and innovation within the social sector ecosystem.

Keyword: FLOSS, Open Source, Maturity Model, Social Sector.

Bibliography

- [1] Simon Bridge, Brendan Murtagh KO. Understanding the Social Economy and the Third Sector. 2020.
- [2] Biggeri M, Testi E, Bellucci M, During R, Persson HTR. Social entrepreneurship and social innovation: Ecosystems for inclusion in Europe. 2018.
- [3] Informa BBD. Tecido Empresarial Ativo - Setor Social. 2024.
- [4] Angola J de. Sector Social em Angola 2019.
- [5] Nahrkhalaji SS, Shafiee S, Shafiee M, Hvam L. Challenges of Digital Transformation: The Case of the Non-profit Sector. IEEE Int Conf Ind Eng Eng Manag 2018;2019-Decem:1245–9.
- [6] Kumar RV. Free / Open Source Software (FOSS) for Social Sector. vol. 1600. 2014.
- [7] Infosi. Tranformação Digital 2022. <https://www.infosi.gov.ao/>.
- [8] Baena-Navarro R, Vergara-Villadiego J, Carriazo-Regino Y, Crawford-Vidal R, Barreiro-Pinto F. Challenges in implementing free software in small and medium-sized enterprises in the city of Montería: a case study. Bull Electr Eng

Informatics 2024;13:586–97.

- [9] Perez J. 2024 State of Open Source Report. 2024.
- [10] Saraiva SA and F. Principled evaluation of strengths and weaknesses in FLOSS communities: A systematic mixed methods maturity model approach. *IFIP Adv. Inf. Commun. Technol.*, vol. 496, 2017, p. 46–58.
- [11] Mellon C. *CMMI for Development*, version 1.2. 2006.
- [12] Finos. The Open Source Maturity Model, an Agile Framework that Compliments the FINOS State of Open Source Survey 2022. <https://www.finos.org//blog/open-source-maturity-model-launch> (accessed September 8, 2023).
- [13] Aversano L, Tortorella M. Quality evaluation of floss projects: Application to ERP systems. *Inf Softw Technol* 2013;55:1260–76.
- [14] Kenett RS, Franch X, Susi A, Galanis N. Adoption of Free Libre Open Source Software (FLOSS): A risk management perspective. *Proc. - Int. Comput. Softw. Appl. Conf.*, vol. 38, IEEE Computer Society; 2014, p. 171–80.
- [15] Koloniaris S, Kousiouris G, Nikolaidou M. Possibilities of use of free and open source software in the greek local authorities. *IFIP Adv. Inf. Commun. Technol.*, vol. 525, Springer New York LLC; 2018, p. 128–43.
- [16] Pałys M. The role of open source software in the process of implementing social innovation in SMEs. *E-Mentor* 2021;92:45–52.
- [17] Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S. A design science research methodology for information systems research. *J Manag Inf Syst* 2007;24:45–77.
- [18] Becker J, Knackstedt R, Pöppelbuß J. Developing Maturity Models for IT Management. *Bus Inf Syst Eng* 2009;1:213–22.
- [19] Haryanti T, Rakhmawati NA, Subriadi AP. The Extended Digital Maturity Model. *Big Data Cogn Comput* 2023;7:1–24.

DEPARTMENT OF INFORMATICS ENGINEERING

Ricardo Correia Nascimento dos Santos

**Feature Engineering and Deep
Learning for Lyrics-based Music
Emotion Recognition**

PhD in Informatics Engineering

PhD Thesis submitted to the University of Coimbra

Supervisor:

Professor Ricardo Manuel da Silva Malheiro

Co-Supervisors:

Professor Rui Pedro Pinto de Carvalho

Professor Renato Eduardo Silva Panda

Coimbra, April 2024

Project Thesis Summary

In the context of Music Information Retrieval (MIR), Music Emotion Recognition (MER) aims to correlate music with human emotions, gaining popularity and importance, for instance, in music streaming services for organizing collections based on user feelings. This thesis aims to explore various approaches for emotion classification by analyzing lyrics and audio in a bimodal fashion, aligning lyric text with the corresponding audio segments, and also analyzing the song lyrics part in isolation.

This work aims to deal with two types of lyrics classification problems: detection of the predominant emotion in the sample (e.g., complete lyrics) (Static Music Emotion Detection) and detection of emotion throughout the sample (e.g., lyric segments) (Music Emotion Variation Detection). The various classification techniques are machine learning (e.g., identification and implementation of novel features with emotional context that can contribute to the state of the art), deep learning, and large language models (LLM) that allow the creation of more effective models. Other techniques to use are natural language processing (e.g., for keyword-based approach) and techniques based on graph neural networks to better interpret the created models by capturing the semantic meanings of the text present on the samples to detect the emotion conveyed.

This thesis falls within the scope of the scientific project MERGE¹ (Music Emotion Recognition – Next Generation), which is integrated into CISUC Cognitive and Media Systems Lab, and (besides the new proposed machine learning / deep learning architectures) one of the main contributions from that project is the release of a new emotionally relevant dataset which contains important emotional features that will be used in this thesis both for lyrics and audio experiments in a bimodal analysis [1].

As demonstrated in [1], the use of bimodal analysis in the MER process (lyrics and audio) revealed an important enhancement when compared just to each dimension separately.

Some methods presented in the literature, as in [2] and [3], have made use of lexicons such as ANEW [4], WordNet and open-source software such as Synesketch frameworks and classical

¹<https://www.cisuc.uc.pt/en/projects/MERGE>

machine learning (Support Vector Machines, Trees, Random Forest, etc.) to perform emotion recognition in English text for different sources such as online communication platforms like chat, social media comments, and tweets. This is an open opportunity to explore hybrid approaches using knowledge-based graphs to help with music emotion recognition in lyrics.

For example, the proposed algorithm presented in Synesketech frameworks categorizes sentences into emotional types such as happiness, sadness, anger, fear, disgust, and surprise, assigning numerical weights to each category to determine the dominant emotional type and overall emotional valence of a sentence. That approach combines keyword-spotting and rule-based methods, utilizing both human judgment and WordNet, the previously mentioned lexical database. Emoticons and informal language are also considered. The Synesketech software framework, which implements this approach, is highlighted as a significant contribution due to its availability as a free open-source library.

Evaluation studies of that recognition algorithm work have shown promising results, with high classification accuracy and emphasis on the importance of emoticon lexicons. Some other points discussed in the Synesketech paper involve the related work in affective computing, including keyword-spotting, statistical approaches, and rule-based methods. While various techniques have been explored, none have achieved perfect accuracy at that point.

Then, these problems of performance and low accuracy perhaps occur due to the scarcity of emotionally annotated datasets (that follows the Russel Circumplex paradigm as in [5]) or even the lack of a unified dataset that serves as a general benchmark, the traditional Machine Learning (ML) approaches faced a glass ceiling in improving the results in state of the art using this kind of techniques. In this thesis, one main goal is to break that glass ceiling by identifying new features with strong context, stylistic and semantic relevance.

As stated in [6], [1], and [7], a possible way to improve the results either using classical machine learning (e.g. SVM with traditional features like Bag of Words, Part-of-Speech Tags, Term Frequency-Inverse Document Frequency, etc.) and NLP or hybrid approaches (e.g. Keyword-based plus Learning-based methods) is to complement the state-of-the-art with novel stylistic features (e.g. StyBF), structural features (e.g. StruBF) and semantic features (e.g. SemBF).

Indeed, in the year 2023, some authors, as in [8], claim that it's enough to fine-tune transformers-based models (like BERT) for emotion recognition in conversations (another related task known as Emotion Recognition in Conversation, ERC). They first produced context-independent features via fine-tuning pretrained language models (PLMs) and after having created a new BERT-ERC model by combining it with an adapted version of the mentioned PLM. Other related works cited in [38] include models like DialogRNN, DialogGCN, EmoBERTa,

TINYBERT, DistilBERT, and RoBERTa among their evaluations. Some of the commonalities between the ERC task and the MER task involve the detection of emotion in text, and these kinds of works will also be explored to analyze their impact on the MER task.

The idea of bringing Graph Neural Networks as a possibility to achieve the goals of this thesis is because a Tree Data Structure is, in fact, a Graph and has some interpretability characteristics and power to synthesize reasoning capacities through the extraction of syntactic and semantics relationships between words that are not explicitly in other deep learning methods besides the possibility to combine it with LLMs. In [9], [10] and [11], they give some light on these subjects.

Traditional Long Short-Term Memory (LSTM) networks have limitations in capturing hierarchical structures in sentences, which are crucial for understanding context. In [11], for example, the authors propose a novel architecture to enhance semantic representations in natural language processing tasks by introducing the Tree-LSTMs, a generalization of LSTM networks to tree-structured network topologies, to better capture hierarchical structures in natural language. Unlike traditional LSTM networks, which operate on linear sequences, Tree-LSTMs compose their hidden states from an input vector and the hidden states of arbitrarily many child units. This architecture allows Tree-LSTMs to effectively model syntactic relationships between words in a sentence.

The authors in [11] compare Tree-LSTMs against sequential LSTM models on two tasks: predicting the semantic relatedness of sentence pairs and sentiment classification of sentences. Through empirical evaluations, they demonstrate that Tree-LSTMs outperform existing systems and sequential LSTM baselines on both tasks. The paper highlights the importance of considering syntactic structure in sentence representations and provides implementations of the proposed models and experiments for further exploration. Additionally, the paper explores different tree construction strategies and demonstrates the effectiveness of incorporating syntactic information into LSTM models. Overall, the proposed Tree-Structured LSTM networks offer improved semantic representations by effectively capturing hierarchical structures in sentences, leading to enhanced performance in natural language processing tasks.

More information on using LLMs together to learn on Graphs can be found in [9], a recent paper of the year 2024, as well as in complementary papers from [35],[36],[37],[38],[10],[11] and [39].

Then, this thesis aims to address the challenges in analyzing emotions in lyrics by combining some linguistic features with word lexicons (using WordNet), Natural Language Processing, Large Language Models, Graph Neural Networks, and Deep Learning methods in general but also considering emotional concepts analyzed through a qualitative study (following [14] and [13])

conducted on the last year (2023) by the author of this thesis that highlights the performance of the Large Language Models (LLMs) based systems (as ChatGPT of the OpenAI company [28]) in emotion classification tasks in comparison with humans perception of emotions conveyed in a set of four song lyrics. There was evidence of LLMs' capacity to process text informal language to detect emotions. The qualitative study (based on the Phenomenological Approach shown in [13]) summarizes the results of 22 human participants and 1 AI Chatbot [28] with a list of emoticons (as in [40]) that represent each group of emotions according to Russel Circumplex [5].

Bibliography

- [1] R. Malheiro, R. Panda, P. Gomes, and R. Paiva, “Bi-modal music emotion recognition: Novel lyrical features and dataset,” 9th International Workshop on Music and Machine Learning–MML’2016–in . . . , 2016.
- [2] R. Malheiro, R. Panda, P. Gomes, and R. P. Paiva, “Classification and regression of music lyrics: Emotionally-significant features,” 8th International Conference on Knowledge Discovery and Information Retrieval, 2016.
- [3] R. Malheiro, H. G. Oliveira, P. Gomes, and R. P. Paiva, “Keyword-based approach for lyrics emotion variation detection.,” in *KDIR*, pp. 33–44, 2016.
- [4] M. M. Bradley and P. J. Lang, “Affective norms for english words (anew): Instruction manual and affective ratings,” tech. rep., Technical report C-1, the center for research in psychophysiology . . . , 1999.
- [5] J. A. Russell, “A circumplex model of affect.,” *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.
- [6] R. Malheiro, R. Panda, P. Gomes, and R. P. Paiva, “Emotionally-relevant features for classification and regression of music lyrics,” *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 240–254, 2016.
- [7] R. M. d. S. Malheiro, *Emotion-based analysis and classification of music lyrics*. PhD thesis, Universidade de Coimbra, 2017.
- [8] X. Qin, Z. Wu, J. Cui, T. Zhang, Y. Li, J. Luan, B. Wang, and L. Wang, “Bert-erc: Fine-tuning bert is enough for emotion recognition in conversation,” 2023.
- [9] Z. Chen, H. Mao, H. Li, W. Jin, H. Wen, X. Wei, S. Wang, D. Yin, W. Fan, H. Liu, and J. Tang, “Exploring the potential of large language models (llms) in learning on graphs,” 2024.

- [10] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, “Graph neural networks: A review of methods and applications,” 2021.
- [11] K. S. Tai, R. Socher, and C. D. Manning, “Improved semantic representations from tree-structured long short-term memory networks,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (C. Zong and M. Strube, eds.), (Beijing, China), pp. 1556–1566, Association for Computational Linguistics, July 2015.
- [12] K. Hevner, “Expression in music: a discussion of experimental studies and theories.,” *Psychological review*, vol. 42, no. 2, p. 186, 1935.
- [13] C. Moustakas, *Phenomenological research methods*. Sage publications, 1994.
- [14] J. W. Creswell, *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications, 2013.
- [15] L. P. Kaelbling, M. L. Littman, and A. W. Moore, “Reinforcement learning: A survey,” *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.
- [16] P. Eckman, “Emotions revealed,” *St. Martin’s Griffin, New York*, 2003.
- [17] J. P. Mahedero, Á. MartÍnez, P. Cano, M. Koppenberger, and F. Gouyon, “Natural language processing of lyrics,” in *Proceedings of the 13th annual ACM international conference on Multimedia*, pp. 475–478, 2005.
- [18] H. Hirjee and D. G. Brown, “Automatic detection of internal and imperfect rhymes in rap lyrics.,” in *ISMIR*, pp. 711–716, 2009.
- [19] R. Malheiro, R. Panda, P. Gomes, and R. P. Paiva, “Music emotion recognition from lyrics: A comparative study,” 6th International Workshop on Machine Learning and Music (MML13). Held in . . . , 2013.
- [20] R. Panda, R. Malheiro, B. Rocha, A. Oliveira, and R. P. Paiva, “Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis,” in *International symposium on computer music multidisciplinary research*, 2013.
- [21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

- [22] J. Revez and L. Corujo, “Redescobrir o elemento humano nos estudos informacionais: uma proposta metodológica,” 2018. Accessed: 2023-06-02.
- [23] Z. Chen and B. Liu, “Lifelong machine learning,” *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 12, no. 3, pp. 1–207, 2018.
- [24] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [25] N. Alswaidan and M. E. B. Menai, “A survey of state-of-the-art approaches for emotion recognition in text,” *Knowledge and Information Systems*, vol. 62, pp. 2937–2987, 2020.
- [26] J. C. Whittington, J. Warren, and T. E. Behrens, “Relating transformers to models and neural representations of the hippocampal formation, 2021,” URL <https://arxiv.org/abs/2112.04035>.
- [27] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, *et al.*, “A survey of large language models,” *arXiv preprint arXiv:2303.18223*, 2023.
- [28] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang, “Sparks of artificial general intelligence: Early experiments with gpt-4,” 2023.
- [29] S. Mohammad, “Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (I. Gurevych and Y. Miyao, eds.), (Melbourne, Australia), pp. 174–184, Association for Computational Linguistics, July 2018.
- [30] M. Brysbaert, A. B. Warriner, and V. Kuperman, “Concreteness ratings for 40 thousand generally known english word lemmas,” *Behavior research methods*, vol. 46, pp. 904–911, 2014.
- [31] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.
- [32] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, “mt5: A massively multilingual pre-trained text-to-text transformer,” *arXiv preprint arXiv:2010.11934*, 2020.

- [33] T. D. Bui, S. Ravi, and V. Ramavajjala, “Neural graph learning: Training neural networks using graphs,” in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 64–71, 2018.
- [34] Y. Yang, J. Kim, Y. Kim, N. Ho, J. Thorne, and S. young Yun, “Hare: Explainable hate speech detection with step-by-step reasoning,” 2023.
- [35] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, “Unifying large language models and knowledge graphs: A roadmap,” *IEEE Transactions on Knowledge and Data Engineering*, p. 1–20, 2024.
- [36] N. Liu, Q. Feng, and X. Hu, “Interpretability in graph neural networks,” *Graph neural networks: foundations, frontiers, and applications*, pp. 121–147, 2022.
- [37] L. Wu, Y. Chen, K. Shen, X. Guo, H. Gao, S. Li, J. Pei, and B. Long, “Graph neural networks for natural language processing: A survey,” 2022.
- [38] B. Liu and L. Wu, “Graph neural networks in natural language processing,” *Graph Neural Networks: Foundations, Frontiers, and Applications*, pp. 463–481, 2022.
- [39] S. Hu, G. Zou, S. Yang, B. Zhang, and Y. Chen, “Large language model meets graph neural network in knowledge distillation,” *arXiv preprint arXiv:2402.05894*, 2024.
- [40] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, “GoEmotions: A dataset of fine-grained emotions,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 4040–4054, Association for Computational Linguistics, July 2020.
- [41] U. Krcadinac, P. Pasquier, J. Jovanovic, and V. Devedzic, “Synesketch: An open source library for sentence-based emotion recognition,” *IEEE Transactions on Affective Computing*, vol. 4, no. 3, pp. 312–325, 2013.
- [42] L. Parisi, S. Francia, S. Olivastri, and M. S. Tavella, “Exploiting synchronized lyrics and vocal features for music emotion detection,” 2019.
- [43] M. Won, J. Salamon, N. J. Bryan, G. J. Mysore, and X. Serra, “Emotion embedding spaces for matching music to stories,” 2021.

Towards Interactive Explanations for Trustworthy Human-Centered Reinforcement Learning

PhD Student – Filipe Miguel Fonseca dos Santos

Supervisors

- Luís Macedo (Department of Informatics Engineering, Faculty of Sciences and Technology, University of Coimbra)
- João Paulo Fernandes (Visiting Associate Professor, New York University Abu Dhabi)

Abstract

The influence of Artificial Intelligence (AI) in our daily lives keeps growing and shows no signs of stopping. While many incredible advancements have been made, the gap between the power of these approaches and their understandability has kept widening, giving rise to the field of Explainable AI (XAI). Furthermore, the importance of other factors besides explainability in order to receive trust from users is a growing concern. Moreover, the lack of concern with the end user who receives the explanations is something that only more recently has started to be addressed by researchers. One of the paradigms for tackling some of these concerns is the use of interactive explanations, which allow for a grounding process between the user and the machine. In our work, we propose incorporating these interactive explanations into Reinforcement Learning from Human Feedback (RLHF) to achieve a greater bidirectionality, potentially increasing human trust while giving the human a greater sense of control. We also want to explore the synergy between quantum computing and explainability, due to the former's gradually increasing importance for society.

Introduction

Artificial Intelligence (AI) has become increasingly essential for the functioning of society, contributing to multiple areas, including autonomous driving, medicine, and language translation [1]. The advent of tools such as ChatGPT has only further cemented just how important AI truly is, since it sees widespread use not only among experts, but also by the average person. What has contributed greatly towards this success has been the rapid growth of Machine Learning (ML). One of the three branches of ML, Reinforcement Learning (RL), considers the scenario in which an agent learns through trial and error by interacting with an environment [2]. The agent improves its behavior by exploring the actions it can take, noticing the rewards it receives, and by exploiting the knowledge it acquires over time. Within the scope of RL, there are now even models that can surpass humans in certain areas, with the game of Go being a pretty famous example [1]. Typically, the approaches that have had the most impressive results have leveraged Neural Networks (NNs) to their advantage, allowing them to process enormous quantities

of information, and to make relations between aspects of the data that would be near impossible for a human to find.

However, those same NNs that have catapulted AI's potential are now leading users, corporations, and regulatory entities to question how those systems reach their conclusions; in other words, there are concerns regarding the understandability of these algorithms. These worries have led to an explosion of interest in the area of Explainable AI (XAI) [3], which attempts to improve transparency of AI programs so that users can reasonably comprehend how the algorithm is working, and how it is computing its outputs. Furthermore, this topic of explainability is only one of many extra requirements that have been raised by regulatory entities and companies [3]; notably, we must also address issues such as fairness, human agency, accountability, privacy, safety, and sustainability [4]. It is only by covering all of these seven key issues that we can create responsible and trustworthy AI applications that can be deployed in real-life scenarios.

While much progress has occurred within the area of XAI, there have been some key limitations in this recent push towards explainability, which have been underscored in numerous articles. For instance, there has been a lack of consideration for the average user who lacks technical knowledge, with most of the literature simply focusing on explanations for AI experts [5-6]. Another issue frequently brought up is how multiple studies have concluded that even though their approaches increased how understandable their models were, the same was not true for user trust [6-7]. Since user trust requires not only transparency, but also the other six requirements mentioned previously [4], it is clear that a greater effort towards satisfying those other aspects is required. Moreover, several papers have argued that explanation is a social process that can also involve argumentation and should ultimately be delivered interactively by allowing users to further query the explanations and ask for more information [5-6, 8-11].

Objectives and Research Questions

In this PhD thesis, we aim to explore interactive explanations with a focus on the average user; additionally, we want to assess whether they can really improve trustworthiness. In order to achieve this, we will take advantage of concepts from the social sciences that have been proposed as relevant for the field of XAI. While there has been research in this direction we are highlighting, it is worth stressing that it has largely been very theoretical and conceptual until very recently. In the last few years, some practical approaches have emerged [12], but they are still the exception and tend to focus on classification tasks; as such, interactive explanations for RL are not as common. Therefore, we plan on studying how this interactive methodology can be applied to RL. More specifically, we want to verify potential synergies between interactivity in explanations and Reinforcement Learning from Human Feedback (RLHF) [13], in which humans can provide guidance to the agent through rewards, demonstrations, or other means.

Lastly, we would like to explore quantum computing in the context of explainability. This is an area that has been expanding significantly lately and has the

potential of impacting society at large; taking that into consideration, we want to do some research on its potential intersection with XAI.

The following research questions will be addressed:

1. Research Question 1: How can we make interactive explanations for RLHF?
2. Research Question 2: What are the best methods to ensure a bidirectional interaction, such that users can query more information about the explanations given?
3. Research Question 3: In what ways can an average user provide guidance to the agent so that its policy aligns with the human's thinking process and values?
4. Research Question 4: Can interaction make human-centered RL more trustworthy?
5. Research Question 5: What synergies can be found between interactive explanations and quantum computing?

Literature Review

Díaz-Rodríguez et al. [4] tackle the issues surrounding trustworthy AI from multiple angles. They cover the seven key requirements for achieving trust in AI: (1) human agency and oversight; (2) technical robustness and safety; (3) privacy and data governance; (4) transparency; (5) diversity, non-discrimination, and fairness; (6) societal and environmental wellbeing; (7) accountability. The authors underline the difference between trustworthy and responsible AI, with the latter also needing to comply with environmental concerns and other issues to become trustworthy.

Weld and Bansal [8] mention the importance of allowing explanations to be interactive. They further elaborate on this topic by bringing up multiple ways in which these interactions can happen. It is also stressed that these do not need to happen in the context of a conversation, even if that is the most clear-cut scenario. In fact, they underscore a paper that proposed an interactive GUI.

Sokol and Flach [9] identify key properties and challenges for interactive AI explanations. In the literature review section, they split related articles into 3 branches: papers that conceptualize theoretical frameworks for how humans and machines interact during explanations, papers that take advantage of user interfaces to allow the user to interact with the explanatory system, and papers related to interactive ML. They argue that the next step in interactive explanations is to try to bring ideas from these different branches together, creating explanations that can be interactively customized by the user.

Conclusion

In this extended abstract, we started by introducing the topics related to this thesis: the growing importance of AI and its impact on society; the need for understanding how the algorithms behind AI work; the multiple dimensions that go into achieving

trustworthy AI; the noticeable gaps in XAI research that have been identified by multiple articles; the value of interactivity in order to achieve personalized explanations. From there, we outlined the objectives of our work and the related research questions, focusing on the exploration of interactive explanations in the context of RLHF. Finally, we provided a short literature review.

Bibliography

1. Russell, Stuart J., and Peter Norvig. *Artificial intelligence: a modern approach*. Pearson, 2020.
2. Sutton, Richard S., and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
3. Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information fusion* 58 (2020): 82-115.
4. Díaz-Rodríguez, Natalia, et al. "Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation." *Information Fusion* 99 (2023): 101896.
5. Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial intelligence* 267 (2019): 1-38.
6. Liao, Q. Vera, Daniel Gruen, and Sarah Miller. "Questioning the AI: informing design practices for explainable AI user experiences." *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020.
7. Madumal, Prashan, et al. "Explainable reinforcement learning through a causal lens." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. No. 03. 2020.
8. Weld, Daniel S., and Gagan Bansal. "The challenge of crafting intelligible intelligence." *Communications of the ACM* 62.6 (2019): 70-79.
9. Sokol, Kacper, and Peter Flach. "One explanation does not fit all: The promise of interactive explanations for machine learning transparency." *KI-Künstliche Intelligenz* 34.2 (2020): 235-250.
10. Madumal, Prashan, et al. "A Grounded Interaction Protocol for Explainable Artificial Intelligence." *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 2019.
11. Holzinger, Andreas, et al. "Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI." *Information Fusion* 71 (2021): 28-37.
12. Bertrand, Astrid, et al. "On selective, mutable and dialogic XAI: A review of what users say about different types of interactive explanations." *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023.
13. Li, Guangliang, et al. "Human-centered reinforcement learning: A survey." *IEEE Transactions on Human-Machine Systems* 49.4 (2019): 337-349.



FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE DE
COIMBRA

Doctoral Program in Informatics Engineering

Decoding intrusive thoughts from neuronal data

Ehsan Farzadnia

Advisors:

Vince D. Calhoun, Ph.D.

Marco Simões, Ph.D.

April 2024

Abstract

This 3.5-year research aims to develop intrusive thought biomarkers for precise diagnosis of Schizophrenia, targeting patterns of resting-state functional connectivity (RSFC). This project consists of three phases: hypothesis formulation through a comprehensive literature review, identification of independent spatial components using various ICA algorithms on resting-state fMRI-BOLD signals, and biomarker development using machine learning techniques. First year focuses on the identification of the gaps, data collection and preprocessing, as well as the extraction of components, and initiate thesis writing towards the end of the year. In 2nd year, we will analyze intrusive thought-linked Regions of Interest (mPFC, PCC, VPC), assessing correlations of RSFC patterns via statistical tests. 3rd year emphasizes robust detection of aberrant patterns combining self-supervised learning principles and immunity measures, leading to clinical translation. The main contributions expected include elucidating underlying neural mechanisms and transforming findings into clinical trials, ultimately improving diagnosis and patient care and treatment cost reduction.

Keywords: Intrusive thoughts in schizophrenia, Precise diagnosis, Immunity measures, Self-supervised learning, Resting-state fMRI, Functional connectivity dynamics

I. State-of-the-art

WHO revealed that schizophrenia, characterized by disabling cognitive impairments [2], affects approximately 24 million people worldwide [1]. Despite unclear etiology [8], Hippocampal GABA [4] has shown to suppress intrusive thoughts (IT) in this disorder [5-6]. However, significant deficits in suppression-induced forgetting [4], impaired distraction resistance, and altered resting-state brain activity, possibly involving disrupted inhibition drives these symptoms [28], worsening distress and comorbidities like anxiety [3]. This highlights the importance of early and precise diagnosis, given its heterogeneous nature.

Abnormally elevated resting blood oxygen-level-dependent (BOLD) activity is common in schizophrenia, with potentially varying neural correlations across individuals [6], making it difficult to establish a standardized diagnosis plan for all. While machine learning aid in diagnosis [8][11][12], research on distinguishing potential IT patterns is still lacking. Unwanted thoughts can be subjective and difficult to measure objectively [5], making it challenging to create accurately labeled data for training. Fedorov et al. [14] propose a self-supervised multimodal learning framework that can extract representations from neuroimaging data without requiring labeled data, promising in predicting disorders across various Alzheimer's phenotypes, suggesting its applicability to the present study. Their method excels in classification and interpretability but faces challenges in multimodal integration, and late fusion risks multicollinearity and overfitting with limited labeled data. However, the artificial immune system [18] can potentially leverage data augmentation technique to tackle this key challenge.

Resting-state fMRI measures synchronous brain regions, with BOLD signal fluctuations serving as an indicator of neural activity during rest [20], offering clinical benefits in schizophrenia where performing tasks might be difficult due to the cognitive impairments. While research primarily focuses on resting-state functional connectivity (RSFC) [23-26], some studies [15] [20-22] have explored rsfMRI features, like ALFF and ReHo, in supervised models for discriminative capabilities. In [13], authors exhibit consistent baseline activity across subjects at rest, significant temporal dynamics, and BOLD changes akin to task-related studies, making rsfMRI valuable for understanding IT. However, clinical studies do not typically measure participants' ongoing thoughts during scan. As a result, linking spontaneous thoughts to connectivity patterns is challenging due to measurement limitations. Despite differences between rsfMRI-BOLD and RSFC, supervised models effectively predict symptom severity in schizophrenia [20].

Brain states impact thought patterns and mental health; mental conditions shape thought evolution [5]. For instance, rumination triggers fixation on self-referential topics, as shown by semantic network analysis. In [10] ventromedial prefrontal cortex (VPC) activation changes were found to be associated with self-generated unwanted thoughts, supporting findings from [9]. Identifying networks triggering IT in diverse individuals under various circumstances, however, remains uncertain. Furthermore, inter-hemispheric disconnection [27] differentiates schizophrenia from depression and may contribute to disturbances in functional networks connecting both hemispheres, including the DMN and salience network, offering significant insights into schizophrenia's neural basis [5][7][10][12]. IT are linked with increased activation in DMN regions, such as particular prefrontal cortex [5], medial prefrontal cortex, and posterior cingulate cortex [19]. Examining functional connectivity alterations in these three key regions in addition to VPC offers a promising approach to uncover the aberrant neural mechanisms underlying IT.

II. Objectives

We plan to achieve the following goals by developing our diagnostic pipeline.

- Functional Connectivity:
 1. What specific functional connectivity patterns distinguish patients who experience frequent intrusive thoughts from those who do not?
 2. What is the magnitude of change in these patterns during intrusive thought episodes compared to baseline states?
- Brain Activity Patterns:
 1. How effectively can resting-state fMRI measurements identify distinct brain activity patterns associated with intrusive thought onset or intensity relying on task-based fMRI?
 2. What distinctions are there in functional connectivity of the DMN, particularly mPFC, PCC, VPC, between individuals experiencing intrusive thoughts and those who do not, even without performing specific tasks designed to evoke these symptoms?
- Learning Model:

We will integrate features derived from rsfMRI-BOLD signals, RSFC/ICA-derived data, and time course-derived summary measures, as inputs for an AIS model. Indeed, the AIS treats these features as self-samples to generate and clone efficient detectors that distinguish deviations from normal patterns. Therefore, we need to address the following:

1. What is the potential efficacy of this integration in identifying abnormalities using the specified features as inputs?
2. How can we ensure that the discovered biomarkers accurately predict the onset and severity of intrusive thoughts?

To answer these questions, we will assess the performance of the AIS model in detecting deviations and the predictive power of the identified biomarkers for intrusive thoughts. This will involve evaluating the overall model's robustness including accuracy, sensitivity, and specificity, as well as validating the biomarkers in a larger, independent sample to confirm their generalizability, reliability, and reproducibility. Considering the scarcity of public datasets and the dynamic nature of immunity measures, we must ensure these metrics are met through the proposed approach (data augmentation, self-supervised learning, and dimensionally reduction).

3. How does the proposed diagnostic pipeline have the potential to transform clinical trials?

III. Detailed description

Variations in intrusive thoughts among individuals may result from abnormally elevated BOLD activity during rest [6], which can hinder effective measurement in clinical studies due to spontaneous thought alterations and observed changes in functional network connectivity (FNC) patterns. What is still ambiguity for scientists is to understand which of these networks activates intrusive thoughts in different people and under what conditions. There are evidences [9-10][19] show that increased activation in regions mPFC, PCC, and VPC might contribute to the onset of these symptoms. Individuals spend 30-50% of their waking life thinking about unrelated matters. Laboratory experiments have limited ecological validity in capturing unwanted thoughts' real-life complexity. Studying functional connectivity patterns in brain regions may help distinguish individuals experiencing these thoughts from those who do not, improving understanding of these symptoms. Resting-state fMRI-BOLD contrast and RSFC are related concepts, but the first is a measurement technique in response to spontaneous neural activity when a person is not engaged in a specific task, and the other is the focus on analysis of the synchronization of brain activity between different brain regions during rest. While supervised learning models using RSFC have shown promising results for classifying or predicting symptom severity in a variety of psychiatric disorders, including schizophrenia [20], we aim to harness the benefits of both methods in two distinct stages of research plan by employing a hybrid approach. Therefore, our contributions involve the following steps:

1. Hypothesis formulation and initiate decoding:

The first year, commencing in September 2024, will focus on conducting an ongoing literature review to identify knowledge gaps and formulate the hypothesis. Additionally, data-driven analysis will be initiated, including preprocessing rsfMRI data using Statistical Parametric Mapping (SPM) and decomposing spatially independent components utilizing Independent Component Analysis (ICA) with the Group ICA of fMRI Toolbox (GIFT) in MATLAB. We will continue to collect data, including schizophrenia and healthy controls, from public repositories, such as COBRE [17] and MGH-USC Human Connectome [29] and preprocess them to remove artifacts and prepare it for further analysis. We will apply multiple ICA algorithms, including blind and constrained, to see whichever gives better performance, and ensure reliability of the results employing techniques like ground truth, cross-validation, and qualitative assessments. NeuroMark [7], a framework applying Group ICA, identifies brain network abnormalities in Schizophrenia preserving variability and comparability. Idiotypic fMRI models better predict self-relevant thought valence than group analysis.

2. RSFC analysis:

We will start the second year examining the specific Region of Interest (ROI) including mPFC, PCC, and VPC, along with other underlying regions identified from the ICA components, showing significant differences between subjects. This stage will require empirically investigations by the supervision team. Then, the functional connectivity between these regions will be calculated using rsfMRI data, and the statistical analysis (e.g., t-test, ANOVA, etc.) will be applied to identify the significance differences between pair connectivity measurements.

Next, the correlation analysis in BOLD signal fluctuations between ROIs, reflecting the extent to which their activity is synchronized or coordinated over time, will be conducted to identify connectivity patterns that might be alerted. Finally, we will examine the relationship between the altered RSFC patterns and clinical measures, such as the severity and frequency of intrusive thoughts. We will respond to our initial questions to understand the nature and the change rate of these patterns during episodes of intrusive thoughts.

3. Feature extraction, Train and validate the AIS model:

Feature vectors comprising features from the correlation of components and RSFC patters, the time courses summary measures, and clinical measures, will be aggregated to form a rich dataset that, after k-fold cross validation, can be imposed into the AIS model for training.

- **The proposed strategy for learning:**

Ji et al. [12] developed an immune-inspired algorithm called AIAEC, outperforming existing methods in inferring brain connectivity from fMRI data, promising for effective connectivity detection and suggesting the applicability for the present study. The final year will focus on developing AIS model. We intend to use a type of AIS known as the Real-valued Negative Selection Algorithm (RNSA), which was proposed in our previous work [18], for the generation of discriminatory antibodies (detectors). RNSA will leverage data augmentation technique to create synthetic variations of healthy brain activity patterns to tackle the challenge of unlabeled fMRI data. Here is the contribution made by RNSA:

Self vs. non-self: The RNSA effectively identifies abnormal patterns by learning “healthy” patterns from aggregated feature vectors of a control group and detecting deviations, potentially associated with intrusive thoughts.

Antibody design: Randomly generated detectors represent healthy patterns derived from aggregated feature vectors of healthy controls (healthy self-samples).

Self-training: Detectors matched with healthy self-samples are retained as “self” while those detecting outliers are discarded, establishing a normal pattern boundary.

Anomaly detection: Detectors identifying deviations from healthy boundaries could indicate potential biomarkers.

Memory Mechanism for Mature Detectors: Maintaining a dynamic pool of mature, non-self-detectors are effective in identifying healthy self-samples across training rounds. The pool serves as a reliable classifier resource for future analyses once detectors meet maturity criteria.

Our recent study [16] applied blind and constrained ICA on rsfMRI signals obtained from COBRE [17], identified highly correlated ICs, and sorted them using rsfMRI features (fALFF, Dynamic Rage) and time course summary measures. Feature vectors trained an AIS model, outperforming Decision Tree and Random Forest in predicting schizophrenia, despite the need for further optimization of hyperparameters, such as antibody pool and radiuses. Our promising results open the door for effectively applying immune-inspired techniques to predict based on clinical data. We will utilize self-supervised learning [14] benefits combined with the immunity measures in further plan. This combination due to the data augmentation ability of the RNSA model applying dimensionally reduction, Principal Component Analysis, can contribute to efficiently ensuring the model generalizability, while the statistical analysis ensures reproducibility.

34th month onwards will focus on transforming developed AI model into clinical trials. For this purpose, we will need to locate in a clinic in Coimbra for a limited time.

Task Description	Year1												Year2												Year3												Year4						
	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	
Formulating hypothesis (comprehensive literature review to identify the knowledge gaps); Initializing decoding resting-state fMRI-BOLD signal	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	
Decoding (the identification of independent spatial components using ICA; Resting-state functional Connectivity analysis; Analyzing intrusive thought-linked Regions of Interest); Correlation analysis; Statistical tests																																											
Feature extraction																																											
Biomarker development using machine learning based on immune measures and self-supervised learning principles; Ensuring robustness of the developed model																																											
Clinical translation and application																																											
Thesis writing and dissemination																																											

IV. References

1. World Health Organization. (2022, June 8). Mental Disorders. <https://www.who.int/news-room/fact-sheets/detail/mental-disorders>
2. Shinn, Ann K., Justin T. Baker, Bruce M. Cohen, and Dost Öngür. "Functional connectivity of left Heschl's gyrus in vulnerability to auditory hallucinations in schizophrenia." *Schizophrenia research* 143, no. 2-3 (2013): 260-268.
3. Xie, Chao, Shitong Xiang, Chun Shen, Xuerui Peng, Jujiao Kang, Yuzhu Li, Wei Cheng et al. "A shared neural basis underlying psychiatric comorbidity." *Nature medicine* 29, no. 5 (2023): 1232-1242.
4. Schmitz, Taylor W., Marta M. Correia, Catarina S. Ferreira, Andrew P. Prescott, and Michael C. Anderson. "Hippocampal GABA enables inhibitory control over unwanted thoughts." *Nature communications* 8, no. 1 (2017): 1311.
5. Kucyi, Aaron, Julia WY Kam, Jessica R. Andrews-Hanna, Kalina Christoff, and Susan Whitfield-Gabrieli. "Recent advances in the neuroscience of spontaneous and off-task thought: implications for mental health." *Nature mental health* 1, no. 11 (2023): 827-840.
6. Levaux, Marie-Noëlle, Frank Larøi, Isabelle Offerlin-Meyer, Jean-Marie Danion, and Martial Van der Linden. "The effectiveness of the attention training technique in reducing intrusive thoughts in schizophrenia: A case study." *Clinical Case Studies* 10, no. 6 (2011): 466-484.
7. Du, Yuhui, Zening Fu, Jing Sui, Shuang Gao, Ying Xing, Dongdong Lin, Mustafa Salman et al. "NeuroMark: An automated and adaptive ICA based pipeline to identify reproducible fMRI markers of brain disorders." *NeuroImage: Clinical* 28 (2020): 102375.
8. Sadeghi, Delaram, Afshin Shoeibi, Navid Ghassemi, Parisa Moridian, Ali Khadem, Roohallah Alizadehsani, Mohammad Teshnehlab et al. "An overview of artificial intelligence techniques for diagnosis of Schizophrenia based on magnetic resonance imaging modalities: Methods, challenges, and future works." *Computers in Biology and Medicine* 146 (2022): 105554.
9. Konu D, Turnbull A, Karapanagiotidis T, et al. A role for ventromedial prefrontal cortex in self-generated episodic social cognition. *Neuroimage* 2020; 218:116977.
10. Hugdahl, Kenneth, Alexander R. Craven, Erik Johnsen, Lars Ersland, Drozdstoy Stoyanov, Sevdalina Kandilarova, Lydia Brunvoll Sandøy, Rune A. Kroken, Else-Marie Løberg, and Iris EC Sommer. "Neural activation in the ventromedial prefrontal cortex precedes conscious experience of being in or out of a transient hallucinatory state." *Schizophrenia Bulletin* 49, no. Supplement_1 (2023): S58-S67.
11. Verma, Shradha, Tripti Goel, Muhammad Tanveer, Weiping Ding, Rahul Sharma, and R. Murugan. "Machine learning techniques for the Schizophrenia diagnosis: A comprehensive review and future research directions." *Journal of Ambient Intelligence and Humanized Computing* 14, no. 5 (2023): 4795-4807.
12. Ji, Junzhong, Jinduo Liu, Peipeng Liang, and Aidong Zhang. "Learning effective connectivity network structure from fMRI data based on artificial immune algorithm." *Plos one* 11, no. 4 (2016): e0152600.
13. Damoiseaux, Jessica S., Serge ARB Rombouts, Frederik Barkhof, Philip Scheltens, Cornelis J. Stam, Stephen M. Smith, and Christian F. Beckmann. "Consistent resting-state networks across healthy subjects." *Proceedings of the national academy of sciences* 103, no. 37 (2006): 13848-13853.

14. Fedorov, Alex, Eloy Geenjaer, Lei Wu, Tristan Sylvain, Thomas P. DeRamus, Margaux Luck, Maria Misiura et al. "Self-supervised multimodal learning for group inferences from MRI data: Discovering disorder-relevant brain regions and multimodal links." *NeuroImage* 285 (2024): 120485.
15. Rashid, Barnaly, Eswar Damaraju, Godfrey D. Pearlson, and Vince D. Calhoun. "Dynamic connectivity states estimated from resting fMRI Identify differences among Schizophrenia, bipolar disorder, and healthy control subjects." *Frontiers in human neuroscience* 8 (2014): 897.
16. Farzadnia, E., Calhoun, V., Simões, M., "Blind and Constrained ICA on preprocessed resting state-fMRI" *Technical Report* (2024)
17. "The Center for Biomedical Research Excellence (COBRE)." accessed 03 April 2024, https://fcon_1000.projects.nitrc.org/indi/retro/cobre.html.
18. Farzadnia, Ehsan, Hossein Shirazi, and Alireza Nowroozi. "A novel sophisticated hybrid method for intrusion detection using the artificial immune system." *Journal of Information Security and Applications* 58 (2021): 102721. <https://doi.org/10.1016/j.jisa.2020.102721>
19. Karapanagiotidis, Theodoros, Diego Vidaurre, Andrew J. Quinn, Deniz Vatansever, Giulia L. Poerio, Adam Turnbull, Nerissa Siu Ping Ho et al. "The psychological correlates of distinct neural states occurring during wakeful rest." *Scientific reports* 10, no. 1 (2020): 21121.
20. Khosla, Meenakshi, Keith Jamison, Gia H. Ngo, Amy Kuceyeski, and Mert R. Sabuncu. "Machine learning in resting-state fMRI analysis." *Magnetic resonance imaging* 64 (2019): 101-121.
21. Zhu, C. Z., Y. F. Zang, M. Liang, L. X. Tian, Yong He, X. B. Li, Man-Qiu Sui, Y. F. Wang, and T. Z. Jiang. "Discriminative analysis of brain function at resting-state for attention-deficit/hyperactivity disorder." In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2005: 8th International Conference, Palm Springs, CA, USA, October 26-29, 2005, Proceedings, Part II 8*, pp. 468-475. Springer Berlin Heidelberg, 2005.
22. Mennes, Maarten, Xi-Nian Zuo, Clare Kelly, Adriana Di Martino, Yu-Feng Zang, Bharat Biswal, F. Xavier Castellanos, and Michael P. Milham. "Linking inter-individual differences in neural activation and behavior to intrinsic brain dynamics." *Neuroimage* 54, no. 4 (2011): 2950-2959.
23. Price, True, Chong-Yaw Wee, Wei Gao, and Dinggang Shen. "Multiple-network classification of childhood autism using functional connectivity dynamics." In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part III 17*, pp. 177-184. Springer International Publishing, 2014.
24. Madhyastha, Tara M., Mary K. Askren, Peter Boord, and Thomas J. Grabowski. "Dynamic connectivity at rest predicts attention task performance." *Brain connectivity* 5, no. 1 (2015): 45-59.
25. Leroy, Arnaud, Etienne Very, Philippe Birmes, Pierre Yger, Sébastien Szaffarczyk, Renaud Lopes, Olivier Outterlyck et al. "Intrusive experiences in posttraumatic stress disorder: treatment response induces changes in the directed functional connectivity of the anterior insula." *NeuroImage: Clinical* 34 (2022): 102964.
26. Geffen, Tal, Jonathan Smallwood, Carsten Finke, Sebastian Olbrich, Zsuzsika Sjoerds, and Florian Schlagenhauf. "Functional connectivity alterations between default mode network and occipital cortex in patients with obsessive-compulsive disorder (OCD)." *NeuroImage: Clinical* 33 (2022): 102915.

27. Guo, Shuixia, Keith M. Kendrick, Jie Zhang, Matthew Broome, Rongjun Yu, Zhening Liu, and Jianfeng Feng. "Brain-wide functional inter-hemispheric disconnection is a potential biomarker for schizophrenia and distinguishes it from depression." *NeuroImage: Clinical* 2 (2013): 818-826.
28. Paulik, Georgie, Johanna C. Badcock, and Murray T. Maybery. "Dissociating the components of inhibitory control involved in predisposition to hallucinations." *Cognitive neuropsychiatry* 13, no. 1 (2008): 33-46.
29. "MGH-USC Human Connectome Project" accessed 15 April 2024, [Human Connectome Project | Mapping the human brain connectivity](#)

Title:**"Relational Analysis of Monitoring: Time and Load on Microservices"****Abstract**

Microservices represent an innovation in software development architecture, grounded in interconnected small services. Each service performs specific tasks autonomously and independently. One of the main advantages of this approach lies in its flexibility. Services are developed in various programming languages and technologies. Only specific parts of the application are scaled as needed, without the obligation to expand the entire system.

Communication between microservices is ensured through APIs (Application Programming Interfaces), which facilitate interaction efficiently and reliably, allowing development teams to ensure the operational continuity of the system, even in scenarios of failures in one or more services, through abstraction layers that enable changes in a service without directly affecting others, as long as the API interface remains unchanged, significantly contributing to the stability of the system as a whole. This strategy, combined with ease of maintenance and continuous adaptation to increasingly complex software systems, drives the growing popularity of this architecture, being essential to keep up with evolving needs in software development.

However, it is essential to recognize and overcome the challenges inherent in decomposing applications into various services, not only in terms of scalability and flexibility but also regarding system security and maintenance. Finding effective solutions to prevent and resolve vulnerabilities is essential given the operational complexity and areas susceptible to potential attacks. Identifying the most suitable tools is crucial to ensuring successful and sustainable implementation.

The increase in workload in each service has been one of the greatest challenges in managing microservices, as it can affect response times and compromise the overall performance quality of the system, directly impacting users. Therefore, continuous analysis of the relationship between each service's workload and corresponding response times is crucial. It is also important to apply appropriate tools and strategies to prevent and/or mitigate latency issues and avoid excessive resource consumption through prevention, timely detection, and alerts for imminent failures, congestion, and/or performance decreases. It is essential to advance comprehensive monitoring practices, such as log analysis, real-time metrics, key metrics, and load testing to distribute the load evenly among different microservices. These measures allow for faster and preventive interventions that will ensure system stability and efficiency, even during periods of high activity and workload variations.

This monitoring not only contributes to maintaining the good operation of services but also promotes the development of best practices in microservices management, ensuring

the current and future effectiveness of services. Constant and proactive surveillance is essential to ensure that microservices operate efficiently and resiliently, especially in environments with high activity and workload variations. The adoption of scalability strategies and load distribution is crucial to deal with sudden activity peaks, ensuring the availability and performance of services even during periods of high traffic.

The present thesis sought to assert the relevance of monitoring in an existing microservices application, with a special focus on analyzing the behavior of services regarding the distribution of supported loads and their repercussions on response times, as overloaded services often fail to respond promptly to requests, resulting from failures and congestion that affect the system's performance quality. Conversely, scenarios where services present low loads also received special attention due to the underutilization of resources and underuse of capacities and application costs that may arise.

The Publicitaki application, developed under the Master's Program in Computer Engineering, addresses the scarcity of microservices applications for scientific purposes in the market, as well as the unsatisfactory expectations of existing applications. It is an e-commerce website application composed of three distinct systems: Publicitaki, responsible for promoting store articles; the Store, intended for the sale of these articles; and the Bank, responsible for banking transactions. It stands out for integrating a variety of technologies and communication protocols, offering scalability and options for local or cloud implementation, and including load testing.

The application reveals deficiencies in some microservices due to monitoring restricted to inadequate and unsatisfactory tests, which hindered the implementation of effective solutions and resulted in gaps due to an inadequate relationship between microservice load and response times. A superficial analysis showed that resource-intensive microservices, such as Publicitaki's frontend and Store's frontend, face instability and prolonged response times due to excessive load. Therefore, continuous, comprehensive, and proactive monitoring with good monitoring practices is imperative.

The decision to start monitoring this application arises from the opportunity to complement it during the doctoral thesis, aiming to study the existing problem and find mechanisms to deduce the load of microservices through punctual observations of their functioning, aspects that were not adequately explored during their development. This will result in an understanding of the necessary replication and the corresponding challenge of optimizing resource utilization.

Considering the complexity and management opportunities present in the microservices environment, Publicitaki presents the ideal conditions to play the proposed role in the doctoral thesis. It will be possible to ensure comprehensive and effective data production, providing essential monitoring resources to identify which services are most overloaded and how this overload affects response times.

The provision of accurate and real-time metrics will certainly be an asset for researchers in identifying bottlenecks and overloads in microservices, enabling faster and more effective interventions that will ensure system stability and reliability in any circumstances. Sharing the generated data with the scientific community will also allow more researchers to use it for study and validation of results, thus contributing to the advancement of research and the recognition of the work carried out.

Bibliographical References

- Sambasivan, R. R., Fonseca, R., Shafer, I., & Ganger, G. R. (2014). "So, you want to trace your distributed system? Key design insights from years of practical experience." Carnegie Mellon University, Brown University.
- Shetty, A., Ritchken, B., Clancy, B., Zaruvinsky, L., Gan, Y., Rathi, P., Padilla, J. (2019). "An Open-Source Benchmark Suite for Microservices and Their Hardware-Software Implications for Cloud & Edge Systems". Cornell University.
- Correia, J., Ribeiro, F., Filipe, R., Araújo, F., & Cardoso, J. (2018). "Response Time Characterization of Microservice-Based Systems." CISUC, Department of Informatics Engineering, University of Coimbra.

Supervisor

- **Professor Filipe Araújo**

Methodology for Dynamic Security Evaluation of IoT Smart Home Devices Based on OWASP Top 10 Vulnerabilities

AUTHOR: EDUARDO FERREIRA FELIX
ADVISOR: MARCO VIEIRA
CO-ADVISOR: FERNANDO ANTONIO AIRES LINS

Abstract. The continuous advancement of the Internet of Things (IoT) brings forth significant security challenges, necessitating approaches to ensure the integrity and confidentiality of interconnected devices. In this context, it is essential to develop strategies that guarantee the protection of transmitted data. Various solutions have been implemented to identify potential vulnerabilities, with dynamic analysis playing a key role in enabling the detection of security gaps and vulnerabilities across diverse IoT environments such as smart homes, connected cities, industrial sectors, and medical applications.

However, recent research has unveiled specific considerations for each context, including challenges such as identifying complex vulnerabilities, the need for continuous updates, and privacy concerns. Therefore, there is a proposal to develop a methodology for dynamic analysis of IoT device security in smart home environments, aiming to identify, analyze, and mitigate security vulnerabilities listed in the OWASP (Open Web Application Security Project) Top 10. This methodology will be validated according to security standards established by specialized organizations, taking into account the unique characteristics of the environment and the diversity of IoT devices to ensure effectiveness and adaptability.

Security, Internet of Things, Smart home, Dynamic analysis

1 Introduction

The Internet of Things (IoT) represents a significant technological innovation by connecting objects and devices to the Internet, enabling the collection and integration of data on an unprecedented scale. According to a study by ISG Provider Lens on the Internet of Things (IoT), it is estimated that by 2025 there will be more than 27 billion connected devices [3]. However, this interconnectivity brings with it a set of vulnerabilities that can be exploited by malicious actors, as indicated by the Open Web Application Security Project (OWASP).

OWASP defines vulnerability as a flaw or weakness in a system, application, or infrastructure that can be exploited by malicious individuals. Examples of these vulnerabilities, particularly relevant to the IoT context, include the lack of device management, which refers to the absence of security support on

deployed and running devices, and insecure default configurations, which encompass devices or systems that are shipped with default settings that do not ensure security, limiting the operators' ability to modify settings [2].

Given the expanding landscape of the IoT, certain environments stand out, such as smart homes, whose popularity has been growing exponentially. However, this growth also brings substantial concerns related to the security of these environments. As revealed in Avast's 2019 report on smart homes, approximately 40.8% of smart homes worldwide have at least one vulnerable connected device, posing a risk to the entire security of the residence. Of these vulnerabilities, 31.8% are associated with unpatched software flaws, while 69.2% are attributed to weak security credentials [1].

The purpose of this work is to propose a methodology for the dynamic analysis of security in IoT devices within smart home environments, aiming to identify, analyze, and mitigate security vulnerabilities listed in the OWASP Top 10.

2 Related Works

The research conducted by [10] offers an in-depth analysis of the security vulnerabilities inherent in the MQTT protocol, widely used in IoT devices, outlining potential exploits by attackers to gain unauthorized access and control over such devices. It highlights, especially, the risk associated with the use of default or easily guessable passwords, which increases the susceptibility of devices to unauthorized access.

In turn, the study described in [5] focuses on identifying and evaluating specific vulnerabilities of IoT devices, addressing issues such as weak passwords, insecure firmware updates, and susceptibility to network attacks. The uniqueness of the security challenges faced by these devices, stemming from their architecture and communication methods, is emphasized, underscoring the need for robust security measures.

Additionally, the study presented in [4] emphasizes that smart homes, by adopting device interconnection technologies, may create openings for invasions if not adequately protected. There is a shortage of studies specifically targeting the identification of vulnerabilities in smart homes, highlighting the importance of research in this context.

On the other hand, [6] proposes a comparison between their method and the established CP-ABE (Ciphertext-Policy Attribute-Based Encryption) method, demonstrating a significant reduction in the time required for encryption and decryption processes. Meanwhile, [7] introduces an innovative diffusion framework for IIoT protocols, showcasing improvements in vulnerability identification.

Despite the mentioned advances, IoT faces substantial security challenges due to the heterogeneity and vulnerability of devices. Article [12] proposes a hybrid model for network security but lacks specific management measures for the numerous devices in the IoT. In turn, [9] presents a method for detecting vulnerabilities in router firmware malware, while [11] introduces a tool for automatically testing IoT devices for security weaknesses. However, both approaches

have limitations regarding comprehensive vulnerability detection and physical device protection.

Finally, article [8] focuses on the security of the D-Link Mini HD Wi-Fi camera, highlighting vulnerabilities that can expose user credentials and the lack of certificate pinning in the TLS protocol as concerning issues. The study demonstrates how easy it is to invade smart home security devices, indicating insecure data transfer and insecure default configurations as vulnerabilities highlighted by OWASP's Top 10 for IoT. These findings reveal gaps in addressing OWASP's Top 10 vulnerabilities in the literature, emphasizing the need for effective solutions to address all requirements for vulnerability validation.

3 Objectives

The general objective is to propose a methodology for the dynamic security analysis of IoT devices in smart home environments, aiming to identify, analyze, and mitigate security vulnerabilities listed in the OWASP Top 10. In particular, the following specific objectives will be considered:

- * Identify vulnerabilities based on the OWASP Top 10 in devices belonging to the Smart Home domain using dynamic analysis techniques.
- * Evaluate and compare dynamic analysis techniques that can be used to identify vulnerabilities in IoT systems.
- * Propose a solution, based on dynamic analysis, to identify, analyze, and mitigate vulnerabilities related to the OWASP Top 10.

4 Methodology

The methodology adopted in this study aims to achieve three specific objectives related to security in the context of smart homes, focusing on the identification, assessment, and mitigation of vulnerabilities based on the OWASP Top 10. To fulfill objective 1, the initial phase involves selecting a set of IoT devices belonging to the Smart Homes domain. The subsequent phase entails conducting a systematic literature review to identify vulnerabilities associated with the selected devices and comparing the vulnerabilities identified in the literature with the OWASP Top 10 to determine which ones are aligned with this set of vulnerabilities and can be verified through dynamic analysis techniques. Criteria for criticality will also be established based on the profile and importance of vulnerable devices, assigning greater criticality to vulnerabilities that pose a higher potential for damage to the smart home environment.

For specific objective 2, a variety of dynamic analysis techniques will be selected for evaluation, considering their effectiveness and applicability to the context of smart homes. Subsequently, tests will be conducted using the selected techniques to determine their ability to detect the vulnerabilities identified in specific objective 1. The performances of dynamic analysis techniques will also

be evaluated in terms of vulnerability detection and their ability to address as many threats as possible in Smart Home environments.

To address objective 3, we will develop a dynamic analysis-based solution capable of identifying, analyzing, and mitigating vulnerabilities related to the OWASP Top 10 in smart homes. This solution will incorporate quality and performance criteria to ensure its effectiveness, efficiency, and scalability. We intend to validate the developed solution through practical tests in a simulated Smart Home environment to verify its ability to comprehensively and effectively identify and mitigate vulnerabilities. This comprehensive solution will enable a systematic approach to understanding, evaluating, and proposing effective solutions for security vulnerabilities in smart homes, contributing to advancements in the field of IoT cybersecurity.

5 Concluding remarks

The contributions of this work aim to provide enhanced approaches for identifying, analyzing, and mitigating security vulnerabilities in IoT devices within smart home environments. Aligned with recent studies in IoT, dynamic analysis, and security, these approaches may be published in journals and conferences specializing in security, IoT, and computing in general.

References

1. Avast Smart Home Security Report 2019. https://cdn2.hubspot.net/hubfs/486579/avast_smart_home_report_feb_2019.pdf, accessed: 2024-02-16
2. Internet of Things (IoT) Top 10. https://wiki.owasp.org/index.php/OWASP_Internet_of_Things_Project#tab=IoT_Top_10, accessed: 2024-02-16
3. IoT: até 2025, mais de 27 bilhões de dispositivos estarão conectados. <https://forbes.com.br/forbes-tech/2022/08/iot-ate-2025-mais-de-27-bilhoes-de-dispositivos-estarao-conectados/>, accessed: 2024-04-12
4. Cai, B., Xie, S., Liang, Q., Lu, W.: Research on penetration testing of iot gateway based on risc- v. In: 2022 International Symposium on Advances in Informatics, Electronics and Education (ISAIEE). IEEE (Dec 2022). <https://doi.org/10.1109/isaiee57420.2022.00093>, <http://dx.doi.org/10.1109/ISAIEE57420.2022.00093>
5. Hashmat, F., Abbas, S.G., Hina, S., Shah, G.A., Bakhshi, T., Abbas, W.: An automated context-aware iot vulnerability assessment rule-set generator. *Computer Communications* **186**, 133–152 (Mar 2022). <https://doi.org/10.1016/j.comcom.2022.01.022>, <http://dx.doi.org/10.1016/j.comcom.2022.01.022>
6. Huang, W., Xie, X., Wang, Z., Feng, J., Han, G., Zhang, W.: Zt-access: A combining zero trust access control with attribute-based encryption scheme against compromised devices in power iot environments. *Ad Hoc Networks* **145**, 103161 (Jun 2023). <https://doi.org/10.1016/j.adhoc.2023.103161>, <http://dx.doi.org/10.1016/j.adhoc.2023.103161>

7. Kim, S., Cho, J., Lee, C., Shon, T.: Smart seed selection-based effective black box fuzzing for iiot protocol. *The Journal of Supercomputing* **76**(12), 10140–10154 (Mar 2020). <https://doi.org/10.1007/s11227-020-03245-7>, <http://dx.doi.org/10.1007/s11227-020-03245-7>
8. Liu, Z., Chen, C., Liu, S., Liu, D., Wang, Y.: Exploit in Smart Devices: A Case Study, p. 152–164. Springer Singapore (2019). https://doi.org/10.1007/978-981-15-0758-8_12, http://dx.doi.org/10.1007/978-981-15-0758-8_12
9. Phu*, T.N.: A system emulation for malware detection in routers. *International Journal of Innovative Technology and Exploring Engineering* **8**(11), 32–40 (Sep 2019). <https://doi.org/10.35940/ijitee.j9909.0981119>, <http://dx.doi.org/10.35940/ijitee.J9909.0981119>
10. Raikar, M.M., S M, M.: Vulnerability assessment of mqtt protocol in internet of things (iot). In: 2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC). IEEE (May 2021). <https://doi.org/10.1109/icccc51823.2021.9478156>, <http://dx.doi.org/10.1109/ICSCCC51823.2021.9478156>
11. Yadav, G., Paul, K., Allakany, A., Okamura, K.: Iot-pen: A penetration testing framework for iot. In: 2020 International Conference on Information Networking (ICOIN). IEEE (Jan 2020). <https://doi.org/10.1109/icoi48656.2020.9016445>, <http://dx.doi.org/10.1109/ICOIN48656.2020.9016445>
12. Yang, Q., Cheng, C., Chen, Y., Li, Y., Shi, Z.: Research on network security management monitoring platform based on internet of things technology. *Applied Mathematics and Nonlinear Sciences* **9**(1) (Apr 2023). <https://doi.org/10.2478/amns.2023.1.00057>, <http://dx.doi.org/10.2478/amns.2023.1.00057>

A Methodology for Assessing Source Code Vulnerabilities in IoT Gateways

AUTHOR: DIEGO RIBEIRO GOMES

ADVISOR: MARCO VIEIRA

CO-ADVISOR: FERNANDO ANTONIO AIRES LINS

Abstract. The growth of the Internet of Things (IoT) has brought significant advancements across various industry sectors. Simultaneously, security concerns have also escalated due to its expansion. Numerous IoT devices are targeted by cyber-attacks owing to firmware, source code, and software vulnerabilities. In this context, static analysis has emerged as an effective approach to identify, comprehensively analyze, and mitigate plausible security vulnerabilities. Therefore, this proposal aims to investigate the challenges hindering the attainment of useful results from IoT static analysis, to develop a static analysis approach aligned with the vulnerabilities proposed by OWASP for assessing source code in IoT gateways.

Keywords: Static Analysis, Internet of Things, IoT, Security, Vulnerabilities, OWASP, Gateways.

1 State of the Art

Static analysis is an essential method in software engineering, crucial for defect identification [10] and improving code quality [20], as well as detecting coding patterns vulnerable to malicious exploits [8], enabling developers to implement preventive measures to enhance system security and protect confidential data against cyber threats [14].

Static analysis can be utilized in a variety of scenarios, spanning applications [18], source code [6], protocols [22], binaries [26], and firmware [8], to recognize and address potential risks and weaknesses. Various techniques have been developed for this purpose, such as Syntax Analysis [9], Semantic Analysis [4], Flow Analysis [21], Graph Analysis [5], and Taint Analysis [12], each with its distinct set of sub-techniques.

Some studies integrate static analysis with various other methods to detect vulnerabilities, including dynamic analysis, machine learning (ML) [25], and natural language processing (NLP) [15]. Dynamic analysis involves well-known approaches such as intrusion testing [24], which replicates real attacks on software systems to discover and exploit security weaknesses, and Fuzzing [10], which entails providing invalid, unexpected, or random test inputs to reveal security vulnerabilities. Both methods require code execution and direct involvement with the system to assess its resilience against attacks.

The Open Web Application Security Project (OWASP) is a global community that promotes software security by providing free resources, tools, and standards. It is known for highlighting the top 10 security vulnerabilities in web applications (OWASP Top 10) [2]. Additionally, it has also released the OWASP IoT Top 10, which identifies the main flaws to be avoided when developing, deploying, or managing IoT systems [1]. These lists are essential tools for developers to understand and address flaws in their applications that attackers could exploit to gain unauthorized access. Of the vulnerabilities outlined in the OWASP IoT Top 10, six out of the seven primary weaknesses identified can be addressed through the application of static analysis techniques [12].

Mandal et al. introduced a framework that employs the technique of Taint Analysis between programs to detect security vulnerabilities and privacy leaks in Android applications and communication channels [16]. In a related context, Costin introduced the first public Static Analysis Security Testing (SAST) tool for Lua code, employing Syntax and Semantic Analysis techniques to detect web vulnerabilities [9]. Additionally, Ferrara et al. discussed the relationship between IoT vulnerabilities and the OWASP Top 10, introducing an extension for an industrial analyzer (Julia) that employs techniques such as Taint Analysis and Call Graphs (CGs) to detect privacy-related vulnerabilities in IoT and an insecure ecosystem [12]. However, it's important to note that none of these articles focus on IoT gateway source code nor provide a systematic approach to identifying vulnerabilities.

Nazzal et al. proposed the Taint-Things tool to identify and detect privacy leaks in SmartThings applications, comparing its performance and accuracy with other available tools [18]. McBride et al. employed static analysis, using Flawfinder, RATS, and Cppcheck tools, to identify, analyze, and fix vulnerabilities in the source code of the Contiki operating system [17]. Similarly, Al-Boghdady et al. utilized the same tools to identify security vulnerabilities in four IoT operating systems (RIOT, Contiki, FreeRTOS, and Amazon FreeRTOS), using Common Weakness Enumerations (CWEs) as a reference for benchmarking [3]. Although these studies are valuable, there is no established benchmark for evaluating IoT gateway source code. The gap in the literature emphasizes the need for future research to create benchmarks for evaluating static analysis tools for detecting vulnerabilities in IoT gateway source code.

Alalfi et al. pioneered the development of the MDE-ChYP tool, leveraging Model-Driven Engineering (MDE) to identify excessive privileges within SmartThings applications. This solution integrates semantic analysis as a static analysis technique [4]. In a complementary approach, Dejon et al. introduced IoTAV, an automated framework designed for security verification in software updates for RIOT operating system applications. Utilizing static analysis methods such as Control Flow Graphs (CFGs), Call Graphs (CGs), and Inter-procedural Control Flow Graphs (ICFGs), IoTAV offers a robust security evaluation [11]. Furthermore, Stievenart et al. proposed an automated methodology for static analysis of WebAssembly programs. Their approach, which encompasses Flow Graphs, Flow Analysis, Semantic Analysis, Syntax Analysis, and Taint Analy-

sis, aims to uncover potential vulnerabilities [23]. From another perspective, Hao Yang et al. introduced an ML technique that combines Lexical Analysis, Syntax Analysis, Semantic Analysis, and Flow Graphs to examine and evaluate source code in IoT energy terminals. This approach employs DNN, FNN, and SVM algorithms to broaden the scope of vulnerability identification [13]. However, it's important to note that while these articles contribute significantly to security analysis, they do not directly address the specific vulnerabilities outlined in the OWASP Top 10.

Chang et al. propose a methodology grounded in the GQM (Goal-Question-Metric) approach for conducting automated code reviews, utilizing the Abstract Syntax Tree (AST) technique to analyze Groovy code in SmartApps. This approach aims to assess software quality characteristics established by ISO/IEC 25010, with a focus on reliability, maintainability, and security [7]. In contrast, Alhanahnah et al. introduced IoTCom, a system employing static analysis techniques to detect interaction threats in SmartThing applications. They conducted a manual analysis to verify the validity of interaction threats detected automatically during the automated verification process, distinguishing between real vulnerabilities and false positives [5]. On the other hand, Niu et al. developed an NLP approach that utilizes Graph Analysis, Flow Analysis, and Taint Analysis techniques to enhance vulnerability detection accuracy and reduce false positive and false negative rates in the source code. The effectiveness of this methodology is evaluated against various RNN, LSTM, and BLSTM models [19]. However, the proposed solutions did not specifically address the validation of identified vulnerabilities, nor did they differentiate between real vulnerabilities and false positives.

Some research studies have utilized static analysis methods on source code, while others have conducted assessments on vulnerabilities identified by Owasp. However, to the best of our knowledge, there hasn't been a specific static analysis approach to evaluate vulnerabilities in the source code of IoT gateways in previous research.

2 Objectives

The main goal is to propose a methodology for assessing source code vulnerabilities in IoT gateways, advancing the current state of the art through the investigation of static analysis techniques and vulnerability detection described by the Owasp Top 10. In this scenario, we aim to address the following objectives:

- * To conduct a comprehensive field study to characterize vulnerabilities in the source code of IoT gateways, developing a systematic approach to identifying these vulnerabilities.
- * To compare and evaluate the effectiveness of vulnerability detection tools in the source code of IoT gateways through defined benchmarks.
- * To develop an automated solution for detecting vulnerabilities using static analysis techniques.

- * To formulate a methodology for measuring identified vulnerabilities, assessing the possibility of them being real vulnerabilities or false positives.

3 Methodology

To fulfill Objective 1, the initial phase involves selecting a representative sample of IoT gateways and acquiring access to the corresponding source code. Subsequently, vulnerabilities in the source code are identified through specialized techniques of static analysis. Following this process, vulnerabilities are categorized according to their severity and potential security ramifications. Ultimately, a systematic approach is developed for the continuous detection of vulnerabilities, aiming to safeguard the security and integrity of IoT systems. For Objective 2, criteria and metrics will be analyzed and evaluated to assess the effectiveness of the tools. Next, an investigation and selection of a variety of static code analysis tools will be conducted. Additionally, a benchmark will be proposed to evaluate vulnerabilities in the IoT gateway source code. The next step will involve running each tool on a set of IoT gateway source code samples and comparing the results against the proposed benchmarks. Finally, the obtained results will be analyzed to determine the relative effectiveness of each tool. To achieve objective 3, various static analysis techniques will be employed, including Syntax, Semantic, Flow, Graph, and Taint analysis, in a holistic approach. For objective 4, establish criteria to distinguish between genuine vulnerabilities and false positives, conduct a manual assessment of identified vulnerabilities, and ultimately evaluate the findings to confirm their significance and accuracy.

4 Concluding remarks

The contributions of this work aim not only to characterize and identify vulnerabilities present in IoT gateways, but also to compare detection techniques, develop automated solutions, and present a methodology for assessing the veracity of detected vulnerabilities.

References

1. OWASP Internet of Things Project. https://wiki.owasp.org/index.php/OWASP_Internet_of_Things_Project#tab=Main, accessed: 2024-02-16
2. Top 10 Web Application Security Risks. <https://owasp.org/www-project-top-ten/>, accessed: 2024-02-16
3. Al-Boghdady, A., Wassif, K., El-Ramly, M.: The presence, trends, and causes of security vulnerabilities in operating systems of iot's low-end devices. *Sensors* **21**(7), 2329 (Mar 2021). <https://doi.org/10.3390/s21072329>, <http://dx.doi.org/10.3390/s21072329>
4. Alalfi, M.H., Abu Zaid, A., Miri, A.: A model-driven-reverse engineering approach for detecting privilege escalation in iot systems. *The Journal of Object Technology* **22**(1), 1:1 (2023). <https://doi.org/10.5381/jot.2023.22.1.a1>, <http://dx.doi.org/10.5381/jot.2023.22.1.a1>

5. Alhanahnah, M., Stevens, C., Chen, B., Yan, Q., Bagheri, H.: Iotcom: Dissecting interaction threats in iot systems. *IEEE Transactions on Software Engineering* **49**(4), 1523–1539 (Apr 2023). <https://doi.org/10.1109/tse.2022.3179294>, <http://dx.doi.org/10.1109/TSE.2022.3179294>
6. Bodei, C., Galletta, L.: *Analysing the Provenance of IoT Data*, p. 358–381. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-49443-8_17, http://dx.doi.org/10.1007/978-3-030-49443-8_17
7. Chang, Byeong-Mo, Son, Janine Cassandra, C.K.: A gqm approach to evaluation of the quality of smartthings applications using static analysis. *KSH Transactions on Internet and Information Systems* **14**(6) (June 2020). <https://doi.org/10.3837/tiis.2020.06.003>, <http://dx.doi.org/10.3837/tiis.2020.06.003>
8. Classen, J., Wegemer, D., Patras, P., Spink, T., Hollick, M.: Anatomy of a vulnerable fitness tracking system: Dissecting the fitbit cloud, app, and firmware. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **2**(1), 1–24 (Mar 2018). <https://doi.org/10.1145/3191737>, <http://dx.doi.org/10.1145/3191737>
9. Costin, A.: Lua code: Security overview and practical approaches to static analysis. In: *2017 IEEE Security and Privacy Workshops (SPW)*. IEEE (May 2017). <https://doi.org/10.1109/spw.2017.38>, <http://dx.doi.org/10.1109/SPW.2017.38>
10. David, R., Salwan, J., Bourroux, J.: From source code to crash test-cases through software testing automation (2021), <https://api.semanticscholar.org/CorpusID:248425598>
11. Dejon, N., Caputo, D., Verderame, L., Armando, A., Merlo, A.: Automated Security Analysis of IoT Software Updates, p. 223–239. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-41702-4_14, http://dx.doi.org/10.1007/978-3-030-41702-4_14
12. Ferrara, P., Mandal, A.K., Cortesi, A., Spoto, F.: Static analysis for discovering iot vulnerabilities. *International Journal on Software Tools for Technology Transfer* **23**(1), 71–88 (Nov 2020). <https://doi.org/10.1007/s10009-020-00592-x>, <http://dx.doi.org/10.1007/s10009-020-00592-x>
13. Hao Yang, Junfeng Zhang, J.L.X.X.: Mining method of code vulnerability of multi-source power iot terminal based on reinforcement learning. *International Journal of Network Security* **25** (May 2023). [https://doi.org/10.6633/IJNS.202305_25\(3\).07](https://doi.org/10.6633/IJNS.202305_25(3).07), [https://doi.org/10.6633/IJNS.202305_25\(3\).07](https://doi.org/10.6633/IJNS.202305_25(3).07)
14. Hur, A., Kim, J., Ryu, Y.: Hiding vulnerabilities of internet of things software using anti-tamper technique. In: *Proceedings of the 2019 4th International Conference on Intelligent Information Technology. ICIIT '19*, ACM (Feb 2019). <https://doi.org/10.1145/3321454.3321466>, <http://dx.doi.org/10.1145/3321454.3321466>
15. Luo, Y., Cheng, L., Hu, H., Peng, G., Yao, D.: Context-rich privacy leakage analysis through inferring apps in smart home iot. *IEEE Internet of Things Journal* **8**(4), 2736–2750 (Feb 2021). <https://doi.org/10.1109/jiot.2020.3019812>, <http://dx.doi.org/10.1109/JIOT.2020.3019812>
16. Mandal, A., Ferrara, P., Khlyebnikov, Y., Cortesi, A., Spoto, F.: Cross-program taint analysis for iot systems. In: *Proceedings of the 35th Annual ACM Symposium on Applied Computing. SAC '20*, ACM (Mar 2020). <https://doi.org/10.1145/3341105.3373924>, <http://dx.doi.org/10.1145/3341105.3373924>

17. McBride, J., Arief, B., Hernandez-Castro, J.: Security analysis of contiki iot operating system. In: Proceedings of the 2018 International Conference on Embedded Wireless Systems and Networks. p. 278–283. EWSN '18, Junction Publishing, USA (2018). <https://doi.org/10.5555/3234847.3234913>
18. Nazzal, B., Alalfi, M.H.: An automated approach for privacy leakage identification in iot apps. *IEEE Access* **10**, 80727–80747 (2022). <https://doi.org/10.1109/access.2022.3192562>, <http://dx.doi.org/10.1109/ACCESS.2022.3192562>
19. Niu, W., Zhang, X., Du, X., Zhao, L., Cao, R., Guizani, M.: A deep learning based static taint analysis approach for iot software vulnerability location. *Measurement* **152**, 107139 (Feb 2020). <https://doi.org/10.1016/j.measurement.2019.107139>, <http://dx.doi.org/10.1016/j.measurement.2019.107139>
20. Nobakht, M., Sui, Y., Seneviratne, A., Hu, W.: Pgfitt: Static permission analysis of health and fitness apps in iot programming frameworks. *Journal of Network and Computer Applications* **152**, 102509 (Feb 2020). <https://doi.org/10.1016/j.jnca.2019.102509>, <http://dx.doi.org/10.1016/j.jnca.2019.102509>
21. Sachidananda, V., Bhairav, S., Ghosh, N., Elovici, Y.: Pit: A probe into internet of things by comprehensive security analysis. In: 2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). IEEE (Aug 2019). <https://doi.org/10.1109/trustcom/bigdatase.2019.00076>, <http://dx.doi.org/10.1109/TrustCom/BigDataSE.2019.00076>
22. Sivakumaran, P., Blasco, J.: argxtract: Deriving iot security configurations via automated static analysis of stripped arm cortex-m binaries. In: Annual Computer Security Applications Conference. ACSAC '21, ACM (Dec 2021). <https://doi.org/10.1145/3485832.3488007>, <http://dx.doi.org/10.1145/3485832.3488007>
23. Stievenart, Q., Roover, C.D.: Compositional information flow analysis for webassembly programs. In: 2020 IEEE 20th International Working Conference on Source Code Analysis and Manipulation (SCAM). IEEE (Sep 2020). <https://doi.org/10.1109/scam51674.2020.00007>, <http://dx.doi.org/10.1109/SCAM51674.2020.00007>
24. Visoottiviseth, V., Jutadhammakorn, P., Pongchanchai, N., Kosolyudhthasarn, P.: Firmaster: Analysis tool for home router firmware. In: 2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE). IEEE (Jul 2018). <https://doi.org/10.1109/jcsse.2018.8457340>, <http://dx.doi.org/10.1109/JCSSE.2018.8457340>
25. Yang, Z., Zhu, W., Yu, M.: Improvement and optimization of vulnerability detection methods for ethernet smart contracts. *IEEE Access* **11**, 78207–78223 (2023). <https://doi.org/10.1109/access.2023.3298672>, <http://dx.doi.org/10.1109/ACCESS.2023.3298672>
26. Zheng, Y., Song, Z., Sun, Y., Cheng, K., Zhu, H., Sun, L.: An efficient greybox fuzzing scheme for linux-based iot programs through binary static analysis. In: 2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC). IEEE (Oct 2019). <https://doi.org/10.1109/ipccc47392.2019.8958740>, <http://dx.doi.org/10.1109/IPCCC47392.2019.8958740>

Generative Artificial Intelligence State-of-the-Art, Applications, and Perspectives¹

Cidiane Aracaty Lobato²

Department of Informatics Engineering (DEI) at University of Coimbra (UC)
cidianelobato@dei.uc.pt

Abstract. Artificial Intelligence (AI) is the field of computer science concerned with designing intelligent computer systems. The milestones in the History of AI include the areas that have produced the most practical results in recent decades, Machine Learning and Deep Learning, in addition to the first results obtained from the emerging sub-area of Natural Language Processing: Generative Artificial Intelligence. This thesis project deals with the main contributions and impacts of Generative AI in the areas of Arts, Education and Ethics. We assume the hypothesis that the experimental applications within Generative AI, such as ChatGPT, DALL-E and Whisper, have the potential to produce a creative revolution in Arts, Education and Ethics in the 21st century. We then present our main goal to formalize our studies in the Generative AI and its intersections with Arts, Education and Ethics, pointing out initial and prospective contributions for creative professors of Arts and Education. Throughout our thesis development, we aim to: (1) systematize state-of-the-art in Generative AI; (2) produce practical applications in Generative AI, with a focus on Arts, Education, and Ethics; (3) contribute to studies on the History of Arts, particularly among children and teenagers; (4) provide guidelines for integrating AI perspectives into Science and Arts education; and (5) discuss UNESCO's recommendations on AI Ethics in the 21st century. In preparation for our thesis project presentation on April 19, 2024 at the DEI/UC, we present our initial methodology and activities, as well as the main points of UNESCO's recommendations for AI Ethics published in 2022.

Keywords: Generative Artificial Intelligence, Arts, Education, Ethics

1. Artificial Intelligence

The field of scientific research known as “Artificial Intelligence” (AI) was officially named during the conference titled “The Dartmouth Summer Research Project on Artificial Intelligence”, held at Dartmouth College, New Hampshire, United States (USA) in 1956 (McCorduck, 2004, p. 114). In their proposal to the Rockefeller Foundation, the scientists John McCarthy, a young assistant professor of mathematics at Dartmouth College, Marvin Minsky, then a Harvard Junior Fellow in mathematics and neurology, Nathaniel Rochester, manager of information research at IBM's research center in Poughkeepsie, New York, and Claude Shannon, a mathematician at Bell Telephone Laboratories already known for his statistical theory of information, wrote:

We propose that a two-month, ten-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it (McCorduck, 2004, p. 111-112).

John McCarthy, who was the main organizer and driver of the conference, proposed the official name for the field “Artificial Intelligence” (McCorduck, 2004, p. 26). In 1956, there were no universally agreed concepts for intuition, originality or intelligence. Since then, Artificial Intelligence has continued to encompass a variety of common and widespread definitions, for instance, “The aim of Artificial Intelligence is to create intelligent machines and through this, to understand the principles of intelligence...” (John L. Gordon); “The science of making machines do things that would require intelligence if done by men.” (Marvin Minsky); and “AI is the study of how to make computers do things at which, at the moment, people are better.” Rich & Knight (1991) *apud* Russel & Norvig (2010, p. 2). Despite the different approaches to defining Artificial Intelligence, organized into categories by Russel & Norvig (2010, p. 1-2), we present below the definition which most approximates to the main goals of this thesis project.

Artificial Intelligence (AI) is the field of computer science concerned with designing intelligent computer systems, that is, systems that exhibit the characteristics we associate with intelligence in human behavior – understanding language, [perception], reasoning, [learning,] solving problems [and search] (Barr & Feigenbaum, 1981, p. 21). In AI, the terms “problem solving and search” refer to a large body of central ideas that deal with deduction, inference, planning, commonsense reasoning, theorem proving, and related processes. Applications of these general ideas are found in programs for natural language understanding, information retrieval, automatic programming, robotics, scene analysis, game playing, expert systems, and mathematical theorem proving (Barr & Feigenbaum, 1981, p. 39).

2. Problem.

What are the main contributions and impacts of Generative AI into the areas of Arts, Education and Ethics?

3. Hypothesis

Experimental AI applications, such as ChatGPT, DALL-E, and Whisper, have the potential to catalyze a creative revolution in Arts, Education, and Ethics in the 21st century.

4. Goals

This thesis project has the main goal to formalize our studies in the Generative Artificial Intelligence and its intersections with Arts, Education and Ethics, pointing out initial and prospective contributions for creative professors of Arts and Education. Throughout our thesis development, we aim to:

¹ Project submitted in partial fulfillment of the requirements for the PhD in Informatics Engineering from the DEI/UC. Advisor: Professor Luís Paquete, PhD.

² Bachelor's in Computer Science from UFPA in 2003. Master's in Informatics and License in Philosophy from the PUC-Rio in 2005 and 2012, respectively. License in Music from UEPA in 2024. Specialist in Pedagogical and Administrative Management in Educational Environments from ULBRA in 2024. Specialist in Theater and Education from the IFNMG in 2024. PhD candidate in Informatics Engineering at the University of Coimbra (UC) in 2024.

1. systematize state-of-the-art in Generative Artificial Intelligence;
2. produce practical applications in Generative Artificial Intelligence, with a focus on Arts, Education, and Ethics;
3. contribute to studies on History of Arts, particularly among children and teenagers;
4. provide guidelines for integrating AI perspectives into Science and Arts education;
5. discuss UNESCO's recommendations on AI Ethics in the 21st century.

5. AI References Overview

This section presents a brief introduction to the areas that have produced the most practical results in recent decades, in addition to the first results obtained from the recent sub-area of Natural Language Processing: Generative AI.

5.1. Machine Learning

In the perspective of Machine Learning, Artificial Intelligence is the study and construction of artificial entities or agents situated in an environment, with which they interact autonomously and seek to satisfy objectives, both internal and external, with maximum performance. Learning is the process by which an agent automatically alters its internal structures to better perform tasks under the conditions imposed by the environment.

From the perspective of Data Science, Machine Learning is an area of AI that focuses on the use of data and algorithms to imitate the human ability to learn, gradually improving its accuracy. Machine Learning is then an area of Computer Science that focuses on teaching machines to learn from data, without being explicitly programmed to do so.

Although Machine Learning continues to be considered a promising research area in the field of AI studies, its widespread use in the current state of Computer Science practice can be observed through the following references:

1. "Machine Learning For Absolute Beginners: A Plain English Introduction" by Oliver Theobald (2021);
2. "Machine Learning for Dummies" by John Paul Mueller & Luca Massaron (2016);
3. "Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies" by John D. Kelleher, Brian Mac Namee & Aoife D'Arcy (2015);
4. "Machine Learning in Action" by Peter Harrington (2012);
5. "Machine Learning for Hackers" by Drew Conway & John Myles White (2012);
6. "Data Mining: Practical Machine Learning Tools and Techniques" by Ian H. Witten & Mark A. Hall (2011)
7. "Programming Collective Intelligence" by Toby Segaran (2007).

5.2. Deep Learning

Despite the early models of neural networks dating back to the early decades of the field of AI research, it was only in the mid-2000s that the term "Deep Learning" began to gain popularity. In 2009, the NIPS Workshop on Deep Learning took place, where it was discovered that with a sufficiently large dataset, neural networks no longer require pre-training, resulting in significantly reduced error rates. In 2012, artificial pattern recognition algorithms reached human-level performance in certain tasks. Additionally, Google's deep learning algorithm demonstrated the ability to identify cats.

In 2015, Facebook deployed deep learning technology called DeepFace to automatically tag and identify Facebook users in photographs. The algorithms achieved superior facial recognition tasks using deep networks that consider 120 million parameters. In 2017, there was a mass adoption of Deep Learning across various corporate and mobile applications, accompanied by advancements in research.

All technology events related to Data Science and Big Data now emphasize Deep Learning as the primary technology for creating intelligent systems. Deep Learning is an area of IA which uses neural networks with three or more layers to simulate the behavior of the human brain. These neural networks can learn from copious amounts of data, automate feature extraction, and perform analytical and physical tasks without human intervention.

Deep Learning can process unstructured data, such as text and images, eliminating some of the data preprocessing that is typically involved with Machine Learning, reducing the reliance on human experts. Deep Learning certainly differs from classical machine learning in the type of data it works with and the methods by which it learns. Machine learning algorithms use structured and labeled data to make predictions, meaning specific features are defined from the model's input data and organized into tables.

Specifically, Machine learning and Deep Learning models are capable of diverse types of learning, which are often categorized as supervised learning, unsupervised learning, and reinforcement learning. In Machine Learning, supervised learning uses labeled datasets to categorize or make predictions; this requires some form of human intervention to correctly label the input data. In Deep Learning, unsupervised learning is used to find patterns in unlabeled data, and reinforcement learning is used to teach an agent to make decisions.

Thus, Deep Learning remains at the forefront of AI. Classic references on neural networks, such as Rosenblatt (1957), Widrow (1960), and Minsky & Papert (1969), date back to the 20th century. However, only recent references, including Manning (2015) and Buduma & Locascio (2017), offer an updated introduction to the research in the field.

5.3. Natural Language Processing and Generative AI

Natural Language Processing (NLP) is a branch of Artificial Intelligence that assists computers in understanding, interpreting, and manipulating human language. NLP is the result of various disciplines, including Computer Science and Computational Linguistics, which aim to bridge the gap between human communication and computer understanding. To this thesis project, we are particularly interested in the sub-area of NLP called Generative AI.

Generative AI refers to the emerging area of Artificial Intelligence which uses techniques and models that are designed to generate added content, such as images, text, audio, or other forms of data. Unlike traditional AI approaches that focus on recognition, classification, or prediction tasks, generative AI systems are trained to create novel and realistic outputs based on patterns learned from large datasets.

These systems can employ deep learning architectures, such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs) or Generative Pre-trained Transformer (GPT), to generate content that resembles human-created data. Generative AI has applications in various fields, including creative arts, content generation, data augmentation, and synthetic data generation for training machine learning models.

Currently, the known applications of Generative AI include ChatGPT, DALL-E, and Whisper, developed by the company OpenAI. Founded in December 2015, OpenAI is an artificial intelligence research laboratory consisting of researchers and engineers who are focused on developing applications to ensure that AI benefits all of humanity.

OpenAI applications function as assistants for repetitive tasks, freeing us humans to focus on more strategic and creative activities. Everyone can benefit from this technology. Depending on our technical knowledge, we can engage in activities such as developing our own AI models, utilizing pre-built models from open-source repositories like GitHub, or simply using services like ChatGPT and DALL-E.

Since 2017, Generative AI has indeed advanced significantly in creative tasks and has become a fundamental tool for enhancing the quality of products and services. In this emerging field, applications are being developed with the aim of augmenting human skills rather than completely replacing human work. Generative AI represents the state-of-the-art of AI technology, which include the groundbreaking applications:

- **ChatGPT:** a type of AI language model developed by OpenAI is based on the GPT architecture, which uses deep learning techniques to produce human-like text based on the input it receives. ChatGPT is specifically designed to generate text that can be used in conversations, making it capable of answering questions, providing explanations, and even engaging in casual chat in a way that simulates human conversation (Alzaabi et al. 2023; Bahrini et al. 2023; Lucey & Dowling, 2023; Lund et al., 2023);
- **DALL-E:** a variant of the AI models developed by OpenAI, specifically designed for generating digital images from textual descriptions. It is based on the GPT architecture but adapted for a different task – creating visual content. The name "DALL-E" is a portmanteau of the famous surrealist artist Salvador Dalí and Pixar's animated robot character WALL-E (Ramesh et al., 2021; 2022; Marcus et al., 2022; Mittal, 2023; Borji, 2024; Singh, 2024; Zhang, 2024);
- **Whisper:** an automatic speech recognition (ASR) system developed by OpenAI designed to convert spoken language into written text. It supports multiple languages and is known for its robust performance across different accents and noisy environments (Radford et al., 2022; Williams & Atkins, 2022; Gris et al., 2023; Gong et al., 2024).

One chapter of our thesis may focus on documenting the underlying models of ChatGPT, DALL-E and Whisper, aiming to develop practical uses in the fields of Arts, Education, and Ethics.

6. Preliminary Contributions

Generative AI may produce significant contributions and impacts in the areas of Arts, Education, and Ethics. Next, we list some of the main contributions and impacts:

1. *Creation of novel artworks and artistic styles:* Generative AI algorithms, such as DALL-E and GPT-3, can generate images, texts, music, and even entire stories based on input prompts. This transformative capability redefines the creative process, empowering artists to explore innovative ideas and styles;
2. *Exploration of new artistic forms:* Generative AI can facilitate the exploration of unconventional artistic forms and mediums, blurring the boundaries between traditional art forms and digital media. Artists can experiment with AI-generated content to push the boundaries of creativity and expression;
3. *Democratization of artistic tools and expression:* Generative AI can lower the barriers to entry for aspiring artists, allowing individuals with limited technical skills to create high-quality artwork. This democratization fosters an inclusive artistic community;
4. *Personalization of learning experiences:* Generative AI algorithms can adapt to individual student needs and preferences, providing personalized learning materials and feedback. This personalized approach enhances student engagement and performance by catering to each student's unique learning style;
5. *Creation of interactive educational content:* Generative AI tools can generate educational materials, such as quizzes, tutorials, and simulations, to supplement traditional teaching methods. This automation frees up educators' time to focus on higher-level tasks, such as student mentoring and curriculum development;
6. *Enhancement accessibility and inclusivity:* Generative AI can create educational resources in multiple languages and formats, making learning more accessible to diverse student populations, including those with disabilities or learning differences;
7. *Detection of bias and discrimination in datasets:* Generative AI raises questions related to the AI-generated content, such as copyright infringement, misinformation, and bias. Policymakers can explore ethical guidelines to address these concerns;
8. *Safeguarding against misuse:* Generative AI technologies have the potential to be misused for malicious purposes, such as deepfake videos, fake news generation, and propaganda. Ethical guidelines and regulations are needed to mitigate these risks and safeguard against harmful applications of AI;
9. *Ethical decision-making frameworks:* Researchers and policymakers can explore ethical frameworks and guidelines to address AI Ethics concerns and ensure responsible AI development and deployment;
10. *Promotion of fairness, accountability, and transparency:* Generative AI models often make it challenging to understand how they generate content and make decisions. Efforts are underway to promote transparency and accountability by requiring model documentation, auditability, and ethical review processes.

Overall, Generative AI has the potential to revolutionize the fields of Arts, Education, and Ethics by enabling creativity, personalization, and innovation while also raising important ethical considerations that must be addressed to ensure responsible AI deployment.

7. Discussion on AI Ethics

The United Nations Educational, Scientific and Cultural Organization (UNESCO) is a specialized agency of the United Nations (UN) with the aim of promoting world peace and security through international cooperation in Education, Arts, Sciences and Culture. With 194 states and twelve associate members, UNESCO also collaborates with partners in the NGO, intergovernmental, and private sectors. Its headquarters are in Paris, France, with a network of fifty-three regional field offices and 199 national commissions facilitating its global mandate.

Drawing from the document “Recommendation on the Ethics of Artificial Intelligence” published by UNESCO in 2022, we have selected paragraphs that align with our thesis project and provide some original images created by DALL-E when we first used this chatbot on April 04, 2024, trying to follow the same style of the photo provided in Figure 1:

Respect, protection and promotion of diversity and inclusiveness should be ensured throughout the life cycle of AI systems, consistent with international law, including human rights law. This may be done by promoting active participation of all individuals or groups regardless of race, color, descent, gender, age, language, religion, political opinion, national origin, ethnic origin, social origin, economic or social condition of birth, or disability and any other grounds (UNESCO, 2022, p. 19). Member States should promote gender diversity in AI research in academia and industry by offering incentives to girls and women to enter the field, putting in place mechanisms to fight gender stereotyping and harassment within the AI research community, and encouraging academic and private entities to share best practices on how to enhance gender diversity (UNESCO, 2022, p. 32).

Figure 1. Ensuring diversity and inclusiveness on AI Ethics



Fonte: UNESCO (2022, p. 19)



Fonte: ChatGPT/DALL-E (2024)

Regarding Figure 1, ChatGPT remarks: “It’s wonderful that you’ve found the image that best captures the essence of promoting gender diversity in AI research, as outlined by UNESCO. The image conveys the joy of collaboration and learning, and the positive energy of a supportive community. It beautifully illustrates the message of inclusivity and the encouragement of women’s participation in technology and AI. If there’s anything more you need or another way I can assist you, please let me know!” I continued my conversation with ChatGPT 4.0 and the integrated DALL-E:

Member States should promote AI education and digital training for artists and creative professionals to assess the suitability of AI technologies for use in their profession, and contribute to the design and implementation of suitable AI technologies, as AI technologies are being used to create, produce, distribute, broadcast and consume a variety of cultural goods and services, bearing in mind the importance of preserving cultural heritage, diversity and artistic freedom (UNESCO, 2022, p. 33).

Citing UNESCO (2022, p. 33), I stated to ChatGPT: “I have to say that you work better now. The first image was already fantastic. I must choose between two photos, as follows:” (Figure 2). ChatGPT provided the following commentary: “The first image is bustling with activity and variety, portraying a lively space where creativity and AI intersect in many forms. It’s like a snapshot of a vibrant community at work, with a sense of movement and collaboration. The second image feels more intimate and personal, focusing on the individual moments of creativity enhanced by AI. It offers a sense of quiet joy and concentration, and the interaction between the artist and AI is more pronounced.”

Figure 2. Promotion of AI education for artists preserving diversity and freedom

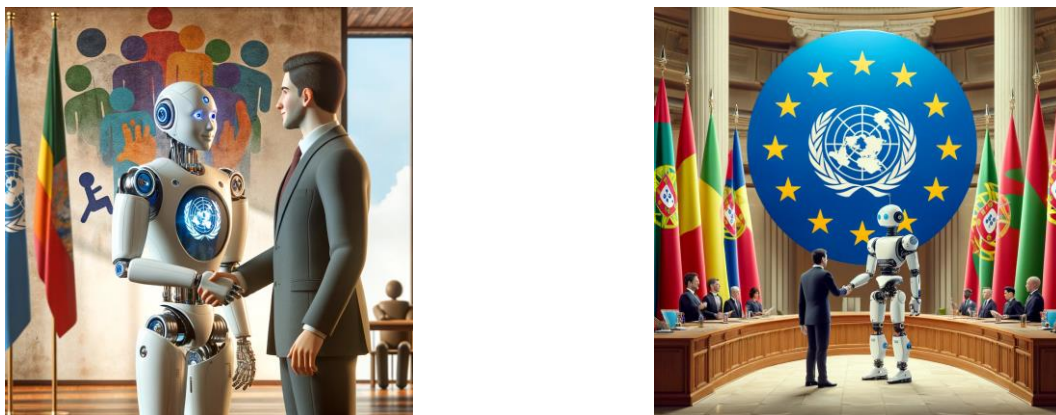


Fonte: ChatGPT/DALL-E (2024)

Sure, we agree with ChatGPT that Figure 2 is quite impressive. Indeed, the capabilities of ChatGPT and DALL-E continue to impress. I answered: "Let's proceed with our collaboration.":

Member States should ensure that human-robot interactions comply with the same values and principles that apply to any other AI systems, including human rights and fundamental freedoms, the promotion of diversity, and the protection of vulnerable people or people in vulnerable situations. Ethical questions related to AI-powered systems for neurotechnologies, and brain-computer interfaces should be considered to preserve human dignity and autonomy (UNESCO, 2022, p. 37).

Figure 3. Human-robot interactions comply with principles applied to AI systems.



Fonte: ChatGPT/DALL-E (2024)

Throughout the development of our thesis, we seek to honor human rights, dignity, and diversity (Figure 3). Currently, the priority is to advance our AI scientific research within academia, at the University of Coimbra. This will be achieved by providing tangible resources to women who aspire to contribute to the field of AI (UNESCO, 2022, p. 32).

8. Methodology and Activities

In preparation for our thesis project presentation on April 19, 2024 at the Department of Informatics Engineering of University of Coimbra, we present our initial methodology and activities:

1. *Systematize State-of-the-Art*: Conduct an extensive review of the current state-of-the-art in Generative AI;
2. *Study Applications*: Investigate the applications of prominent tools, such as ChatGPT, DALL-E, and Whisper;
3. *Analyze Contributions to the Arts*: Explore potential contributions of Generative AI to the field of Arts, including novel artistic expression and cultural preservation;
4. *Analyze Contributions to Education*: Examine how Generative AI can impact Education, including personalized learning experiences and interactive educational content;
5. *Analyze Contributions to Ethics*: Investigate the ethical implications of Generative AI, including bias detection, accountability, and transparency;
6. *Document Knowledge*: Systematically document acquired knowledge from the study and analysis of Generative AI and its intersections with Arts, Education, and Ethics;
7. *Structure AI Models*: Develop models of Generative AI tailored for applications in Arts and Education;
8. *Produce Practical Applications*: Collaborate with CISUC projects to produce applications of Generative AI;
9. *Organize Scientific Events*: Organize scientific events in Generative AI following CISUC guidelines;
10. *Publish Scientific Results*: Disseminate scientific results obtained during the thesis investigation through publication in relevant academic journals and conferences;
11. *Discuss Ethical Implications*: Engage with the scientific community, particularly professionals in Arts, Education, and Ethics, to discuss lessons learned and results regarding AI Ethics in the 21st century.

9. Conclusion

In this project, we presented the goals, methodology and activities of our thesis, on which we are going to work hard next years. We have provided an overview of the field of Artificial Intelligence, covering areas such as Machine Learning, Deep Learning, and Natural Language Processing, as well as the emerging field of Generative AI, which includes the latest applications like ChatGPT, DALL-E, and Whisper.

As a first experience in using these applications, we are exploring how they will impact the realms of Arts, Education, and Ethics, even before our thesis is completed. Fortunately, we have a wealth of experience in integrating knowledge from Computer Science, Philosophy, Psychoanalysis, Education, and Arts, and despite the sophistication of ChatGPT, we still feel that we are much smarter than robots.

10. References

1. Alzaabi, A., Alamri, A., Albalushi, H., Aljabri, R., & Aalabdulsalam, A. (2023). ChatGPT applications in academic research: A review of benefits, concerns, and recommendations. *Cold Spring Harbor Laboratory*.
2. Bahrini, A., Khamoshifar, M., Abbasimehr, H., Riggs, R. J., Esmaeili, M., Majdabadkohne, R. M., & Pasehvar, M. (2023). ChatGPT: Applications, opportunities, and threats. In *Proceedings of the IEEE Systems and Information Engineering Design Symposium (SIEDS)* (pp. 1-13).

3. Barr, A., & Feigenbaum, E. A. (Eds.). (1981). *The Handbook of Artificial Intelligence* (1st ed., Vol. I). Heuristech Press and William Kaufmann.
4. Barreto, J. M. (2001). *Inteligência Artificial no Limiar do Século XXI: Abordagem Híbrida, Simbólica, Conexionista e Evolucionária*. UFSC.
5. Borji, A. (2024). *Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion, Midjourney, and DALL-E 2*. Quintic AI.
6. Buduma, N., & Locascio, N. (2017). *Fundamentals of Deep Learning: Designing Next-Generation Machine Intelligence Algorithms*. O'Reilly Media.
7. Conway, D., & White, J. M. (2012). *Machine Learning for Hackers: Case Studies and Algorithms to Get You Started* (1st ed.). O'Reilly Media.
8. Costa, E., & Simões, A. (2008). *Inteligência Artificial: Fundamentos e Aplicações* (2nd ed.). FCA.
9. Cozman, F. G., Plonski, G. A., & Neri, H. (Eds.). (2021). *Inteligência Artificial: Avanços e Tendências*. Instituto de Estudos Avançados.
10. Gong, Y., Khurana, S., Karlinsky, L., & Glass, J. (2024). Whisper-AT: Noise-robust automatic speech recognizers are also strong general audio event taggers.
11. Harrington, P. (2012). *Machine Learning in Action* (1st ed.). Manning.
12. Kelleher, J. D., Namee, B. M., & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. The MIT Press.
13. Lobato, C. A. (2006). *Um framework orientado a aspectos para mobilidade de agentes de software*. [Master's thesis, PUC-Rio]. Maxwell. <https://www.maxwell.vrac.puc-rio.br/colecao.php?strSecao=resultado&nrSeq=8022@1>.
14. Lobato, C., Garcia, A., Lucena, C., & Romanovsky, A. (2006). A modular implementation framework for code mobility. In *Proceedings of the 3rd IEEE Mobility Conference* (Vol. 1, pp. 1-6). IEEE Mobility, Bangkok.
15. Lobato, C., Garcia, A., Romanovsky, A., & Lucena, C. (2008). An aspect-oriented software architecture for code mobility. *Software, Practice & Experience*, 38, 1365-1392.
16. Lobato, C., Garcia, A., Kulesza, U., von Staa, A., & Lucena, C. (2008). Evolving and composing frameworks with aspects: The MobiGrid case. In *Proceedings of the ICCBSS'08* (pp. 53-62). IEEE Computer Society Press.
17. Lobato, C., Winograd, M., & Evangelista, I. (2021). A psicologia cognitiva e as situações emocionais vivenciadas diante do computador. In *Anais do 16º Congresso Brasileiro de Sistemas*. Franca: UNIFACEF.
18. Lucey, B., & Dowling, M. (2023). ChatGPT: Our study shows AI can produce academic papers good enough for journals. *The Conversation*.
19. Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2024). ChatGPT and a new academic reality: AI-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, 74, 570-581.
20. Manning, C. D. (2015). *Computational Linguistics and Deep Learning* (p. 7). Stanford, California: Association for Computational Linguistics.
21. Marcus, G., Davis, E., & Aaronson, S. (2022). *A very preliminary analysis of DALL-E 2*. Computer Vision and Pattern Recognition.
22. McCorduck, P. (2004). *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence* (2nd ed.). Natick, Massachusetts: A K Peters Ltda.
23. Minsky, M., & Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. Cambridge: M.I.T. Press.
24. Mueller, J. P., & Massaron, L. (2016). Machine learning for dummies (2nd ed.). *For Dummies*.
25. Oliveira, R. F. (2018). *Inteligência Artificial*. Londrina: Editora e Distribuidora Educacional S.A.
26. Pinheiro, Á. F. (2020). *Inteligência Artificial: Fundamentos e Aplicabilidades* (1st ed.). Recife: CEFOSPE.
27. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. [White paper].
28. Radford, A., Kim, J. W., XU, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. [White paper].
29. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. *Computer Vision and Pattern Recognition*.
30. Rich, E., & Knight, K. (1995). *Inteligência Artificial* (2a ed.). Makron Books.
31. Rosenblatt, F. (1957). *The Perceptron: A Perceiving and Recognizing Automaton* (Project Para). Buffalo, New York: Cornell Aeronautical Laboratory.
32. Russell, S. J., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). New Jersey: Prentice Hall.
33. Segaran, T. (2007). *Programming Collective Intelligence: Building Smart Web 2.0 Applications*. O'Reilly Media.
34. Singh, G., Deng, F., & Ahn, S. (2024). Illiterate DALL-E Learns to Compose.
35. UNESCO (Editor). (2022). *Recommendation on the Ethics of Artificial Intelligence*. Paris: United Nations Educational, Scientific and Cultural Organization.
36. Widrow, B. (1960). An adaptive "Adaline" neuron using chemical "memistors." Stanford, California: Stanford Electronics Laboratories.
37. Williams, W., & Atkins, L. (2022). Whisper Speech to Text Deep-Dive. *Speechmatics*.
38. Witten, I. H., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). Morgan Kaufmann.
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. In 31st Conference on NIPS 2017, Long Beach, California.
40. Zhang, T., Wang, Z., Huang, J., & Tasnim, M. M. M. (2024). A Survey of Diffusion Based Image Generation Models: Issues and Their Solutions.



University of Coimbra

Advancing Machine Learning Techniques for Forecasting Energy Demand in Electric Vehicle Charging

Simin Rasouli

Supervisors: Professor Pedro Moura, Professor César Teixeira

Prepared for workshop

April 2024

Accurately forecasting electric vehicle (EV) charging demand is not just a technical challenge, but a critical necessity for the efficient operation of our electrical grids. By predicting peak charging periods and implementing effective strategies for managing this demand, grid stability, minimizing costs, and maximizing resource allocation are guaranteed. However, achieving this level of foresight is complex due to the unpredictable nature of EV usage patterns and the varying charging behaviors of EV owners.

The rapid growth in EV adoption increases these challenges, as it introduces additional layers of uncertainty into our forecasting models. Despite these difficulties, developing robust prediction methodologies for peak charging periods is essential. Not only does it minimize the need for costly grid reinforcement measures, but it also facilitates better management of grid operations.

Furthermore, precise demand forecasting significantly enhances the coordination between renewable energy generation and EV charging demand. This adjustment ensures optimal utilization of renewable energy during periods of high EV charging activity, leading to efficient grid management and resource allocation. It can also enhance grid sustainability and reduce greenhouse gas emissions. This collaboration not only helps the environment but also supports the long-term sustainability of our energy systems.

In fact, tackling the complexities of EV charging demand prediction is not just a technical imperative but a strategic necessity. By leveraging advanced forecasting techniques and implementing effective demand management strategies, a stronger, more environmentally friendly, and economically efficient energy future can be created.

To explore the technical details further, an extensive assessment of different machine learning algorithms including shallow Support Vector Regression (SVR) and Deep Learning architectures like Convolutional Neural Networks (CNNs), will be conducted to understand how well they can capture the patterns in EV charging station energy demand. This involves comparing their performance metrics and analyzing their computational requirements and scalability. For example, simpler models like Support Vector Regression (SVR) are less complicated but may struggle with complex relationships in EV charging data. On the other hand, more complex

models like Convolutional Neural Networks (CNNs) succeed at extracting features and learning from them, though they require greater computational resources.

To optimize the performance of these ML models, advanced optimization techniques will be deployed, with a particular focus on genetic algorithms. By leveraging genetic algorithms, the research aims to improve the used models.

Furthermore, the implementation of Federated Learning (FL) approaches will be essential in preserving data privacy while facilitating model training across distributed datasets. Considering that there are privacy restriction for some EV datasets, using FL models is essential. FL enables the aggregation of model updates from local devices or servers without compromising the privacy of individual data sources.

In terms of dataset selection, datasets such as ElaadNL and ACN are considered for their comprehensive coverage of EV charging activities. These datasets encompass a wide spectrum of variables, including charging station locations, charging duration, power consumption, and temporal patterns, thus providing a comprehensive view of EV charging demand dynamics. Leveraging such rich and diverse datasets will be instrumental in training robust ML models capable of accurately forecasting energy demand at EV charging stations across varied geographical and temporal contexts.

By integrating these methodological approaches and datasets, this research aims to push the boundaries of EV charging demand prediction, facilitating efficient resource allocation, grid management, and sustainability in the realm of electric transportation.

Incorporating cutting-edge ML techniques in this study marks a crucial advancement in transforming smart grids into more flexible and agile energy systems. By leveraging the predictive capabilities of advanced ML algorithms, the precision and detail of electric vehicle (EV) charging demand forecasting can be significantly enhanced, leading to a revolutionary improvement in the operational effectiveness of smart grids.

In addition, this study aims to surpass traditional methods by thoroughly reviewing existing literature and data sources and by introducing innovative approaches and methodologies. Through systematic experimentation and continuous

improvement, the research seeks to advance the current understanding of EV charging demand forecasting within the framework of smart grids.

Furthermore, apart from technical advancements, the results of this research hold significant implications for society as a whole. By improving energy efficiency and encouraging the use of renewable energy sources, smart grids enhanced by advanced ML techniques help combat climate change and promote environmental sustainability. Additionally, by strengthening the reliability and resilience of energy systems, these developments benefit societal well-being and economic prosperity.

Objectives:

Evaluating the performance of the machine learning model versus traditional forecasting methods, such as time series analysis to determine advantages of each.

Investigate the impact of different hyperparameters, such as learning rate, on the performance of the proposed model and optimize its performance accordingly.

Exploring the potential advantages that load forecasting for EV charging can offer to various stake-holders, including electric vehicle owners, charging station operators, utilities, and society as a whole, and examine how this technology can contribute to sustainable development, demand re-sponse, power system management, and the development of smart grids.