Explainable AI in Neuroscience: Present Work and Future Directions

Oriana Presacan
AI Multimedia Lab
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
oriana.presacan@stud.etti.upb.ro

Bogdan Ionescu
AI Multimedia Lab
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
bogdan.ionescu@upb.ro

***Abstract*—This paper provides an overview of my past, current, and future work on artificial intelligence in medical research, with a focus on neuroscience. It begins with key projects that introduced me to artificial intelligence in clinical settings, including generating electrocardiogram data, developing a smart bracelet for stroke rehabilitation, and conducting a review of deep learning in electroencephalogram analysis. Recent projects include using synthetic data to improve embryo classification. Currently, I am researching artificial intelligence in neuroscience and gaining practical experience in clinical settings as a research volunteer at the Psychiatry Hospital in Sibiu. Looking ahead, I am interested in neural network mapping, scene reconstruction from fMRI, and developing metrics to make artificial intelligence more interpretable for clinical use. This work aims to advance artificial intelligence methods that are both innovative and practical for healthcare.**

***Index Terms*—AI, XAI, interpretabaility, neuroscience, EEG, machine learning**

I. Introduction

Artificial intelligence (AI) has experienced significant growth recently, marked by the emergence of powerful deep learning models. However, as these models become more complex their interpretability decreases [1], [2]. Early models, such as expert systems, are simple and easy to trace back [3]. In contrast, current deep learning models are so intricate that even their developers cannot fully understand why they produce certain outputs [4]. This raises an important question: can we truly trust and rely on these models for automated decision-making? While this may not be a significant issue in some use cases, it is crucial in medicine. Doctors need to understand why a model makes a certain decision due to the critical impact of medical decisions on patients. Furthermore, the application of research discoveries in clinical environments is hindered by skepticism towards AI algorithms [5].

In response to these concerns, explainable artificial intelligence (XAI) has emerged as a critical area of research, aiming to make complex AI models more transparent and interpretable. The European Commission's High-Level Expert Group on Artificial Intelligence has emphasized the importance of explainability alongside other ethical principles, including fairness, prevention of harm, and respect for human autonomy [6]. XAI encompasses a range of methods designed to clarify how AI models arrive at their decisions, often through visualizations like heat maps, saliency maps, and textual descriptions that reveal the model's inner workings.

This paper provides an overview of my journey in applying AI and machine learning techniques to healthcare. Beginning with several projects, such as generating electrocardiogram (ECG) data and developing wearable technology for stroke rehabilitation, XAI for electroencephalogram (EEG) analysis, it then transitions to my current work on reviewing AI in neuroscience. Finally, the paper outlines future research directions that will guide my PhD studies, including neural network mapping, fMRI scene reconstruction, and the development of interpretability metrics for clinical AI applications. Through this exploration, I aim to contribute to the advancement of AI models that are both powerful and ethically grounded, enhancing their impact on patient care.

II. Background

The section summarizes some of my past research project undertaken as part of my Master's program and my roles at various research centers in Oslo, Norway. These projects focused on applying XAI and machine learning (ML) techniques to complex datasets in medicine and related fields, advancing both theoretical understanding and practical applications in clinical settings.

A. 1-to-12-lead ECG Generation Using GANs

One of the primary projects involved the generation of 12-lead ECG data using Generative Adversarial Networks (GANs), starting with single-lead data (Lead I). This approach initially aimed to synthesize realistic multi-lead ECG signals from limited input data; however, the results demonstrated a limitation in GANs for this application. Specifically, the model generalized on the training data, converging toward the population mean rather than producing accurate, patient-specific ECG signals. This work, conducted during an internship at the Simula Metropolitan Research Center, was subsequently accepted for publication in Nature Communications Medicine.

1st Workshop on Artificial Intelligence for Multimedia, AI Multimedia Lab, CAMPUS Research Institute, POLITEHNICA Bucharest, Bucharest, Romania, 8 November, 2024

B. Stroke Rehabilitation Bracelet for Hand Activity Recognition

Another project focused on stroke rehabilitation, leveraging wearable technology and AI to support recovery and promote functional independence. In this project, we developed a smart bracelet, designed in collaboration with a rehabilitation hospital expert, specifically for monitoring rehabilitation activities. The bracelet's AI system achieved a 97% accuracy in distinguishing between daily hand activities—such as teeth brushing, hand washing, and flexion/extension movements. Its real-time monitoring allows both patients and healthcare professionals to track progress remotely, with detailed historical analytics accessible via a web application. We presented the paper online, at the International Conference on Electronics and Nanotechnology (ELANO) on Biomedical Engineering in Ukraine (May 2024).

C. Review on XAI in EEG Data

I recently worked on a survey paper, submitted to ACM Computing for Healthcare, focusing on the role of deep learning and XAI in EEG analysis. The paper reviewed commonly used algorithms, their applications in EEG analysis, and explores how XAI techniques can improve transparency and trust in these models. Additionally, it highlights current limitations, such as the need for clinical validation, and underscores the potential of XAI to transform neurological diagnosis and treatment, paving the way for greater adoption in healthcare.

D. Synthetic Embryo Images Improve Classification

In a recent project submitted to Scientific Reports, we used AI to improve embryo selection in assisted reproductive technology. Given the limited availability of embryo data, we trained two generative models to produce synthetic embryo images across various developmental stages, which were then used alongside real images to train a classification model. This approach improved the model's accuracy in predicting embryo cell stages compared to training on real data alone. Notably, the model trained solely on synthetic images also performed well on real data. A Turing test by embryologists revealed that synthetic images generated by the diffusion model appeared highly realistic, outperforming those created by the GAN model and further demonstrating the model's potential for clinical use.

III. Current Work

I am currently working on a review paper on AI in neuroscience to better understand the field and explore what research has been done before I choose a specific direction. This review allows me to survey existing methods and key developments, helping me identify important gaps and potential areas for future research. Alongside this, I have just enrolled as a research volunteer with the Scientific Research Group in Neuroscience at the Psychiatry Hospital in Sibiu. I hope that this role will deepen my understanding of the medical field and provide insight into real clinical settings. Working directly with clinicians, I will learn about the practical challenges involved in using AI XAI for patient care. In this environment, we often face critical issues related to interpretability and reliability—factors that are essential for ensuring AI tools are safe and effective for clinical use. By combining this hands-on experience with a thorough review of existing research, I believe I am building a strong foundation in AI applications for neuroscience, preparing myself for future specialized work in this field.

IV. Future Directions

Based on my review of the current literature, I have identified several potential areas for future research that I find intriguing. These include:

A. ANN-to-Brain Mapping

Exploring the parallels between artificial neural networks (ANNs) and human neural networks may yield insights into both fields. This line of research could deepen our understanding of brain-like structures in AI and assist in translating ANN patterns to a neurobiological context.

B. Scene Reconstruction from fMRI

A possible application of generative AI is reconstructing visual scenes from functional magnetic resonance imaging (fMRI) data. This technique would allow researchers to visualize brain activity patterns and correlate them with real-world scenes, providing a window into perceptual and cognitive processes.

C. Interpretability Metrics in XAI

A key focus of XAI research is developing quantitative metrics for model interpretability. Future work aims to establish robust metrics that can be applied to different AI models in neuroscience, enabling more standardized and reliable evaluations of their interpretability.

References

[1] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *European Journal of Operational Research*, vol. 247, no. 1, pp. 124–136, 2015.

[2] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "XAI—Explainable artificial intelligence," *Science Robotics*, vol. 4, no. 37, 2019.

[3] J. A. Oravec, "Expert systems and knowledge-based engineering (1984-1991): Implications for instructional systems research," *International Journal of Designs for Learning*, vol. 5, no. 2, 2014.

[4] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[5] M. Bucholc, C. James, A. A. Khleifat, A. Badhwar, N. Clarke, A. Dehsarvi, C. R. Madan, S. J. Marzi, C. Shand, B. M. Schilder *et al.*, "Artificial intelligence for dementia research methods optimization," *Alzheimer's & Dementia*, vol. 19, no. 12, pp. 5934–5951, 2023.

[6] E. Commission, "Ethics guidelines for trustworthy ai," Shaping Europe's Digital Future, 2019. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai