WAIT, I Can Explain! Introducing Weighted AI Integration for Tailored Explanations

Claudio Giovannoni[1,2]

[1]University of Pisa (Department of Computer Science), Largo B. Pontecorvo, 3, 56127 Pisa, Italy
[2]National Research Council, (ISTI-CNR), Area della Ricerca del CNR, via G. Moruzzi 1, 56124 Pisa, Italy

Abstract

Explainable Artificial Intelligence (XAI) aims to reduce the inherent opaqueness of modern Machine Learning (ML) systems and make it more interpretable. This has its highest potential benefits in critical societal application domains. However, current literature in XAI is yet unable to be fully effective to real world scenarios due to lack of expressiveness and easiness of understanding by domain experts outside AI. This PhD research proposal seeks to advance the field of Multimodal eXplainable AI (MulXAI) in a user-centric manner and achieve XAI concrete applicability in real-world scenarios. This through the introduction of WAIT, a MulXAI framework. WAIT is capable of profiling users generate enriched, adapted explanations in terms of data modality, explanation technique, explanation scope and verbosity. This will be accomplished through the implementation of an active profiling strategy which iteratively tailors the interpretable output according to user preferences, attitudes, domain expertise, and other relevant factors. The explanation process will be able to capture intra-modal and inter-modal relationships. WAIT's explainability module will produce enriched and comprehensive explanations adapting to diverse users and stakeholders trough external knowledge provided. This paper provides an overview of the work conducted since the start of the PhD, presents a brief but comprehensive review of the relevant MulXAI literature, outlines key challenges and foundational objectives of the research, and discusses the preliminary results obtained along with future steps.

Keywords

WAIT, Artificial Intelligence eXplainability, Explanation Enrichment, User Tailored Explanations, User Profiling

1. Context

Artificial Intelligence (AI) has experienced significant growth in recent years with the emergence of Transformer-based architectures, Large Language Models (LLMs), and large-scale AI in general. The focus now is on advancing the democratization of AI in important societal areas by prioritizing interpretable models, transparent and enriched explanations able to process useful data in a liquid, dynamic fashion rather than being specific to single purposes and data sources [1].

Healthcare, among others, is an area where the source of data is inherently multimodal [2]. Medical data comes in a wide range of formats, such as images, medical notes and signals, to name a few. There had been a number of efforts made to combine the inherent multimodality of the health sector for disease prediction tasks, such as Alzheimer's Disease (AD) and other significant conditions [3].

Despite recent AI models becoming more accurate and versatile in handling multiple data types, critical issues persist. An important one is the opaqueness and overall lack of transparency in modern deep learning AI systems, acting as one of the main barriers to their widespread adoption in socially valuable areas. This hampers the benefits that AI can reflect in society. Explainable AI aims to tackle these challenges through explanations and interpretable models but still falls short in integrating multimodal inputs and personalization. For XAI to be truly effective, it must be accessible to the end

Discovery Science - Late Breaking Contributions 2024, October 14–16, 2024, Pisa, IT
*Claudio Giovannoni
claudio.giovannoni@phd.unipi.it (C. Giovannoni)
https://github.com/cgiova/ (C. Giovannoni)
0009-0000-6582-9950 (C. Giovannoni)

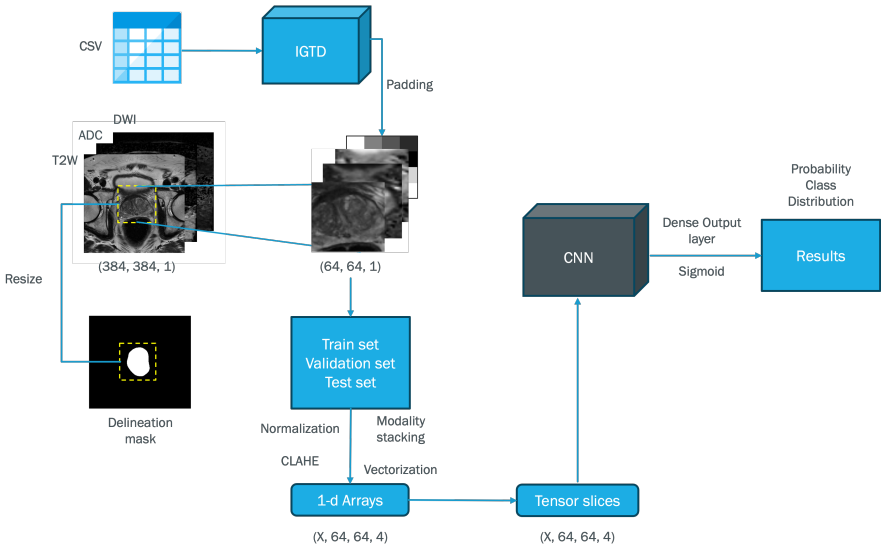

Figure 1: System workflow: from data pre-processing through the application of the explainers to the fine-tuned black box

user of different expertise than the computer scientist, tailored to individual user preferences and capable of managing multiple modalities [4].

2. Research Question and Challenges

This research introduces WAIT, a multimodal XAI framework using Weighted AI Integration for personalized explanations. The framework integrates diverse data modalities for predictions and explanations, tailoring the output based on user interactions. Key challenges and building blocks for WAIT's full implementation include:

1. **State-of-the-Art**: Assess and build on current methods in multimodal AI and XAI to enhance existing knowledge.
2. **Transparent Predictive Architecture**: Develop transparent models able to handle multiple data modalities trough an ad-hoc input fusion module.
3. **User Profiling Strategy**: Design a user-profiling system based on interaction data to customize explanations according to user preferences.
4. **Personalized Explanations**: Utilize user profiles to generate tailored explanations in various formats, such as counterfactuals, rule-based, saliency maps, and feature importance. Implement a dynamic, chatbot-based interface creating interactive, follow-up explanations.

3. Methods, Approach, and Evaluation

The first milestone of the project, the study *"Integrating Multimodal Deep Learning and Explainable AI for Enhanced Prostate Lesion Classification"*[5] is currently submitted to and under review by a renowned computer science journal. The work leverages a multimodal classifier on a medical dataset obtained from a public challenge [6] of prostate carcinoma for a binary classification task with the implementation of multiple XAI methods. The workflow of our study is reported in Figure 1.

The developed architecture is composed of a Convolutional Neural Network (CNN) integrating three different image modalities with tabular metadata, which was converted into image format through a specific algorithm [7]. The modalities are stacked along the width axis of each image. Furthermore, different explanation methods are applied to the model to generate multimodal explanations.

4. Preliminary Results

Our approach has two key aspects: integrating three distinct Magnetic Imaging modalities, each capturing critical prostate features, and incorporating tabular metadata, which is converted into images for training a Multi-Layer Perceptron (MLP). The final model combines these image modalities with the metadata, initialized using weights from the MLP's last layer.

Post-hoc explanation methods are then applied to the model to enhance interpretability leveraging both tabular and image data. Post-hoc XAI methods applied are SHAP [8] for quantitative interpretation of feature importance of tabular data and Grad-CAM [9] to images for visual identification of the most influential modalities for the classifier's output process. Our approach boosted performance by using comprehensive information from all modalities while enhancing transparency, making the model more accessible to medical domain experts.

Explainability methods applied across different modalities clarify each modality's role and importance in classification, while also providing the most important image areas and tabular features. The source code is available in the GitHub repository.

While increased transparency is crucial, it is yet insufficient by itself. Truly reliable AI explanation systems must also be personalized to fit users' specific needs and expertise.

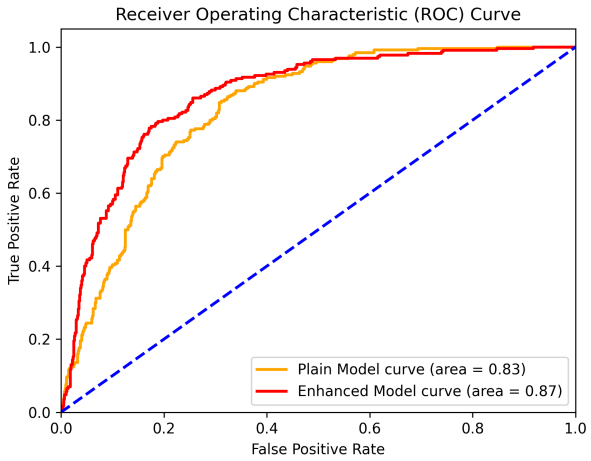

Figure 2: AUC-ROC curve demonstrates a significant improvement in classification performance with the inclusion of tabular data in the model.

5. Discussion and Future Work

Successive steps involve implementing personalization and fusion of explanations through user profiling, leveraging interaction logs. User profiles will enable the explainer to produce tailored explanations adjusted on the basis of users' domain expertise, along with specific attitude towards the explanation process.

As of today, we are conducting a literature review on the state-of-the-art in multimodal models that achieve interpretability trough XAI methods. This review will include a taxonomy to standardize and refine definitions of multimodality and multimodal explainability.

Then, we will integrate advanced architectures for better multimodal data processing and pair them with various explanation methods, such as counterfactuals and concept-based explanations [10], for more comprehensive results. Next, we will implement user log registration and develop a chatbot-based interface to allow accessibility in querying the model for domain experts.

Our ultimate goal is to deliver the WAIT system as a unified, user-friendly platform that combines classification and explanation, enhancing trust and usability for users with diverse attitudes and expertise.

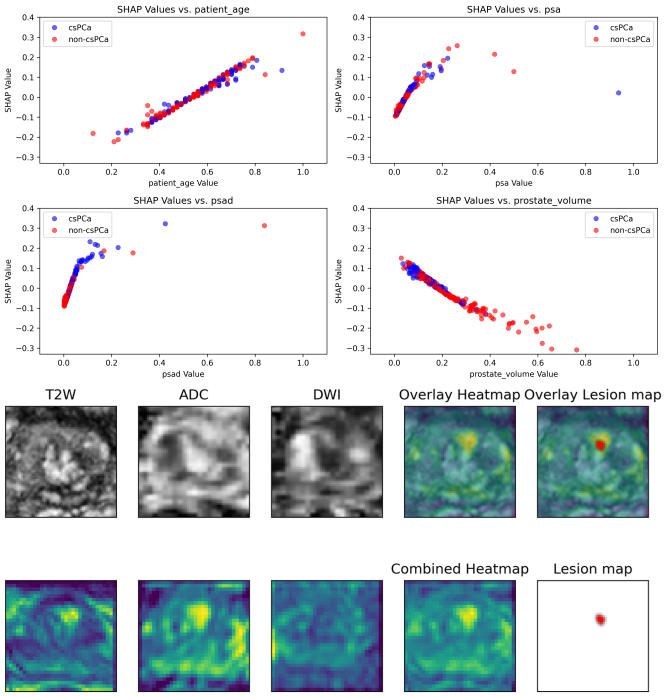

Figure 3: (Top): SHAP calculates Shapley values, representing the contribution of each tabular feature to the classification result. (Bottom): Grad-CAM generates a saliency map indicating the most important pixels for the image classification decision.

Acknowledgments

Funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. Grant Agreement no. 101120763 - TANGO. Research partly funded by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme. The author thanks their supervisors, Anna Monreale from the University of Pisa, Salvo Rinzivillo from the National Research Council, and Carlo Metta from the National Research Council, for their precious guidance and support throughout this initial part of its academic experience. Thanks to my surpervisors, Anna Monreale, Carlo Metta, Salvatore Rinzivillo.

References

[1] G. Joshi, R. Walambe, K. Kotecha, A review on explainability in multimodal deep neural nets, IEEE Access 9 (2021) 2169–3536. doi:10.1109/ACCESS.2021.3070212. arXiv:2105.07878 [cs].

[2] D.-Q. Wang, L.-Y. Feng, J.-G. Ye, J.-G. Zou, Y.-F. Zheng, Accelerating the integration of chatgpt and other large-scale ai models into biomedical research and healthcare, MedComm – Future Medicine 2 (2023) e43. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mef2.43. doi:https://doi.org/10.1002/mef2.43. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/mef2.43.

[3] M. Velazquez, Y. Lee, Multimodal ensemble model for alzheimer's disease conversion prediction from early mild cognitive impairment subjects, Computers in Biology and Medicine 151 (2022) 106201. doi:10.1016/j.compbiomed.2022.106201.

[4] V. Robbemond, O. Inel, U. Gadiraju, Understanding the role of explanation modality in AI-assisted decision-making, in: UMAP '22: 30th ACM Conference on User Modeling, Adaptation and

Personalization, ACM, 2022, pp. 223–233. URL: https://dl.acm.org/doi/10.1145/3503252.3531311. doi:10.1145/3503252.3531311.

[5] C. Giovannoni, C. Metta, A. Berti, S. Colantonio, A. Monreale, F. Pratesi, S. Rinzivillo, Integrating Multimodal Deep Learning and Explainable AI for Enhanced Prostate Lesion Classification, 2024. Under submission.

[6] The PI-CAI challenge - grand challenge, Official Website, 2024. URL: https://pi-cai.grand-challenge.org/.

[7] Y. Zhu, T. Brettin, F. Xia, A. Partin, M. Shukla, H. Yoo, Y. A. Evrard, J. H. Doroshow, R. L. Stevens, Converting tabular data into images for deep learning with convolutional neural networks, Nature Scientific Reports 11 (2021). doi:10.1038/s41598-021-90923-y.

[8] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, ACM, 2017, pp. 4768 – 4777. doi:1705.07874.

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626. doi:10.1109/ICCV.2017.74.

[10] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2018). URL: https://doi.org/10.1145/3236009. doi:10.1145/3236009.