Explainable Artificial Intelligence: Literature Review

May 31, 2024

Abstract

This literature review provides an overview of the research on Explainable Artificial Intelligence (XAI). The review covers various XAI methods in machine learning, human-centric neural explanations, the application of explainable AI in medical learning, user-centric XAI approaches, evaluations of explainability in large language models (LLMs), and different explainable AI models. The review concludes by discussing the implications of the research for transparency, trust, and the future direction of XAI methodologies.

1 Introduction

Explainable Artificial Intelligence (XAI) is a rapidly evolving field focused on making the decision-making processes of AI models transparent and interpretable. This is crucial for deploying AI in high-stakes domains such as healthcare, finance, and autonomous systems, where understanding model behavior can enhance trust and accountability. This review synthesizes recent advancements in XAI, covering diverse methods and applications, identifying research gaps, and discussing the broader implications for AI development and deployment.

2 Background

As machine learning models become increasingly complex, their opacity often undermines their deployment in critical areas. XAI seeks to bridge this gap by

developing techniques that elucidate how models make decisions, aiming to balance interpretability with performance. This involves a range of strategies from post-hoc explanations to inherently interpretable models. The need for XAI is underscored by the ethical, legal, and technical challenges posed by black-box models, prompting extensive research across multiple disciplines.

3 Literature Review

3.1 XAI Methods in ML

Recent studies highlight significant advancements and challenges in XAI techniques. [Bove et al., 2024] discuss the importance of addressing technical limitations and user-specific failures through a typological framework. Tools like SHAP and LIME are extensively used, yet face scrutiny for potentially misleading feature importance [Letoffe et al., 2024]. Efforts like the Explainable AI Comparison Toolkit (EXACT) aim to standardize XAI evaluation metrics [Clark et al., 2024]. Research continues to develop novel algorithms to handle large datasets effectively [Izza et al., 2024]. Techniques such as Model Parameter Randomisation Test (MPRT) address noise and evaluation biases, ensuring robust model explanations [Hedström et al., 2024]. Methods like counterfactual explanations [Suffian et al., 2024] and advancements in understanding network smoothness [Simpson et al., 2024] are also explored to provide actionable insights and accurate representations.

3.2 Human-Centric Neural Explanations

Enhancing the interpretability of neural networks for human users is critical. [Kohler et al., 2024] suggest using class expressions (CEs) for explaining heterogeneous graphs, while [Zaval and Ozer, 2024] improve surrogate models with nonlinear layers for better DNN interpretation. Frameworks like the Faithful Attention Explainer (FAE) [Rong et al., 2024] and concept-based explanations [Dalal et al., 2024] show promising results in image and text datasets. Integrating neural and symbolic systems with Knowledge Graphs [Zhu and Sun, 2024] and automated interpretability agents like MAIA [Shaham et al., 2024] further enhance reasoning and explanation generation.

3.3 Explainable AI in Medical Learning

Interpretable machine learning techniques have gained traction in healthcare, especially for critical tasks such as disease prognosis [Shen and Ma, 2024]. Studies highlight challenges posed by the complexity of models, such as reduced accuracy in deeper networks [Cedro and Chlebus, 2024]. Methods like LIME for brain tumor detection [Pasvantis and Protopapadakis, 2024] and the Information Bottleneck (IB) approach [Yu et al., 2024] are notable for enhancing model interpretability. Frameworks that combine CNNs with XAI techniques have been effective in breast cancer diagnosis [Ahmed et al., 2024] and lung cancer detection [Rafferty et al., 2024], emphasizing the need for transparent models in clinical settings.

3.4 User-Centric XAI

User requirements and experiences are central to the evolution of XAI. Tailored explanation experiences are advocated by [Wijekoon et al., 2024], while studies highlight the influence of personality traits on user agreement with AI suggestions [Pias et al., 2024]. The Human-Computer Interaction community emphasizes inclusive design and user-centered approaches to enhance engagement and understanding [Hamid et al., 2024]. The X Selector method [Fukuchi and Yamada, 2024] and Weight of Evidence (WoE) framework [Le et al., 2024] are proposed to strategically tailor explanations to user needs, enhancing trust and effective human-AI collaboration.

3.5 Evaluating LLM Explainability

Evaluating LLM explainability involves various methodologies and addresses significant challenges. Techniques like the Softmax Linked Additive Log Odds Model (SLALOM) [Leemann et al., 2024] and layer-wise relevance propagation [Vasileiou and Eberle, 2024] improve explanation quality. The use of LLMs in transforming ML explanations into human-readable narratives [Zytek et al., 2024] and refining explanations in NLI models [Quan et al., 2024] exemplifies ongoing advancements. Researchers emphasize the need to address challenges such as hallucinated explanations and computational costs [Singh et al., 2024].

3.6 Explainable AI Models

XAI models enhance transparency and trust, particularly in complex, black-box systems. Multi-modal LLMs exhibit advanced interpretive capabilities through techniques like saliency maps [Giulivi and Boracchi, 2024]. HiFAs and LoFAs are essential for understanding model decisions [Yoshikawa et al., 2024], while transformer models integrated with sparse attention mechanisms demonstrate efficacy in climate science predictions [Liu et al., 2024]. Methods like ONB-MACF [Pascual-Triana et al., 2024] and ECATS [Ferfoglia et al., 2024] provide accessible insights into AI decision-making, emphasizing the role of XAI in achieving contestable, transparent AI applications.

4 Discussion

The reviewed literature indicates significant progress in developing XAI methodologies, from enhancing interpretability in complex models to designing user-centric explanation approaches. Despite these advancements, challenges remain in balancing transparency with model performance and ensuring explanations are meaningful to diverse user groups. Emerging trends include the integration of neural-symbolic methods and the development of frameworks tailored to specific application domains. Substantial gaps persist, particularly in standardizing evaluation metrics and addressing the scalability of XAI techniques.

5 Conclusion

XAI remains a crucial field in AI research, striving to make AI systems more transparent, interpretable, and trustworthy. Continued development of innovative methods, inclusive design practices, and domain-specific applications will enhance the deployment of AI in critical areas. Future research should focus on addressing existing gaps, standardizing evaluation practices, and ensuring that XAI methodologies are robust, scalable, and user-centric.

References

Maryam Ahmed, Tooba Bibi, Rizwan Ahmed Khan, and Sidra Nasir. Enhancing breast cancer diagnosis in mammography: Evaluation and integration of convolutional neural networks and explainable ai, 2024. URL https://arxiv.org/abs/2404.03892.

Clara Bove, Thibault Laugel, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. Why do explanations fail? a typology and discussion on failures in xai, 2024. URL https://arxiv.org/abs/2405.13474.

Mateusz Cedro and Marcin Chlebus. Beyond the black box: Do more complex models provide superior xai explanations?, 2024. URL https://arxiv.org/abs/2405.08658.

Benedict Clark, Rick Wilming, Artur Dox, Paul Eschenbach, Sami Hached, Daniel Jin Wodke, Michias Taye Zewdie, Uladzislau Bruila, Marta Oliveira, Hjalmar Schulz, Luca Matteo Cornils, Danny Panknin, Ahcène Boubekki, and Stefan Haufe. Exact: Towards a platform for empirically benchmarking machine learning model explanation methods, 2024. URL https://arxiv.org/abs/2405.12261.

Abhilekha Dalal, Rushrukh Rayan, and Pascal Hitzler. Error-margin analysis for hidden neuron activation labels, 2024. URL https://arxiv.org/abs/2405.09580.

Irene Ferfoglia, Gaia Saveri, Laura Nenzi, and Luca Bortolussi. Ecats: Explainable-by-design concept-based anomaly detection for time series, 2024. URL https://arxiv.org/abs/2405.10608.

Yosuke Fukuchi and Seiji Yamada. Dynamic explanation emphasis in human-xai interaction with communication robot, 2024. URL https://arxiv.org/abs/2403.14550.

Loris Giulivi and Giacomo Boracchi. Explaining multi-modal large language models by analyzing their vision perception, 2024. URL https://arxiv.org/abs/2405.14612.

Md Montaser Hamid, Fatima Moussaoui, Jimena Noa Guevara, Andrew Anderson, and Margaret Burnett. Improving user mental models of xai sys-

tems with inclusive design approaches, 2024. URL https://arxiv.org/abs/2404.13217.

Anna Hedström, Leander Weber, Sebastian Lapuschkin, and Marina Höhne. A fresh look at sanity checks for saliency maps, 2024. URL https://arxiv.org/abs/2405.02383.

Yacine Izza, Xuanxiang Huang, Antonio Morgado, Jordi Planes, Alexey Ignatiev, and Joao Marques-Silva. Distance-restricted explanations: Theoretical underpinnings efficient implementation, 2024. URL https://arxiv.org/abs/2405.08297.

Hector Kohler, Quentin Delfosse, Paul Festor, and Philippe Preux. Towards a research community in interpretable reinforcement learning: the interppol workshop, 2024. URL https://arxiv.org/abs/2404.10906.

Thao Le, Tim Miller, Liz Sonenberg, and Ronal Singh. Towards the new xai: A hypothesis-driven approach to decision support using evidence, 2024. URL https://arxiv.org/abs/2402.01292.

Tobias Leemann, Alina Fastowski, Felix Pfeiffer, and Gjergji Kasneci. Attention mechanisms don't learn additive models: Rethinking feature importance for transformers, 2024. URL https://arxiv.org/abs/2405.13536.

Olivier Letoffe, Xuanxiang Huang, Nicholas Asher, and Joao Marques-Silva. From shap scores to feature importance scores, 2024. URL https://arxiv.org/abs/2405.11766.

Mingyu Liu, Nana Bao, Xingting Yan, Chenyang Li, and Kai Peng. A transformer variant for multi-step forecasting of water level and hydrometeorological sensitivity analysis based on explainable artificial intelligence technology, 2024. URL https://arxiv.org/abs/2405.13646.

José Daniel Pascual-Triana, Alberto Fernández, Javier Del Ser, and Francisco Herrera. Overlap number of balls model-agnostic counterfactuals (onbmacf): A data-morphology-based counterfactual generation method for trustworthy artificial intelligence, 2024. URL https://arxiv.org/abs/2405.12326.

Konstantinos Pasvantis and Eftychios Protopapadakis. Enhancing deep learning model explainability in brain tumor datasets using post-heuristic approaches, 2024. URL https://arxiv.org/abs/2404.19568.

Sabid Bin Habib Pias, Alicia Freel, Timothy Trammel, Taslima Akter, Donald Williamson, and Apu Kapadia. The drawback of insight: Detailed explanations can reduce agreement with xai, 2024. URL https://arxiv.org/abs/2404.19629.

Xin Quan, Marco Valentino, Louise A. Dennis, and André Freitas. Verification and refinement of natural language explanations through llm-symbolic theorem proving, 2024. URL https://arxiv.org/abs/2405.01379.

Amy Rafferty, Rishi Ramaesh, and Ajitha Rajan. Transparent and clinically interpretable ai for lung cancer detection in chest x-rays, 2024. URL https://arxiv.org/abs/2403.19444.

Yao Rong, David Sheerer, and Enkelejda Kasneci. Faithful attention explainer: Verbalizing decisions based on discriminative features, 2024. URL https://arxiv.org/abs/2405.13032.

Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A multimodal automated interpretability agent, 2024. URL https://arxiv.org/abs/2404.14394.

Jinzhi Shen and Ke Ma. Interpretable machine learning enhances disease prognosis: Applications on covid-19 and onward, 2024. URL https://arxiv.org/abs/2405.11672.

Lachlan Simpson, Kyle Millar, Adriel Cheng, Cheng-Chew Lim, and Hong Gunn Chew. Probabilistic lipschitzness and the stable rank for comparing explanation models, 2024. URL https://arxiv.org/abs/2402.18863.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. Rethinking interpretability in the era of large language models, 2024. URL https://arxiv.org/abs/2402.01761.

Muhammad Suffian, Jose M. Alonso-Moral, and Alessandro Bogliolo. Introducing user feedback-based counterfactual explanations (ufce), 2024. URL https://arxiv.org/abs/2403.00011.

Alexandros Vasileiou and Oliver Eberle. Explaining text similarity in transformer models, 2024. URL https://arxiv.org/abs/2405.06604.

Anjana Wijekoon, David Corsar, Nirmalie Wiratunga, Kyle Martin, and Pedram Salimi. Tell me more: Intent fulfilment framework for enhancing user experiences in conversational xai, 2024. URL https://arxiv.org/abs/2405.10446.

Yuya Yoshikawa, Masanari Kimura, Ryotaro Shimizu, and Yuki Saito. Explaining black-box model predictions via two-level nested feature attributions with consistency property, 2024. URL https://arxiv.org/abs/2405.14522.

Shujian Yu, Xi Yu, Sigurd Løkse, Robert Jenssen, and Jose C. Principe. Cauchy-schwarz divergence information bottleneck for regression, 2024. URL https://arxiv.org/abs/2404.17951.

Mounes Zaval and Sedat Ozer. Improving the explain-any-concept by introducing nonlinearity to the trainable surrogate model, 2024. URL https://arxiv.org/abs/2405.11837.

Shenzhe Zhu and Shengxiang Sun. Exploring knowledge graph-based neural-symbolic system from application perspective, 2024. URL https://arxiv.org/abs/2405.03524.

Alexandra Zytek, Sara Pidò, and Kalyan Veeramachaneni. Llms for xai: Future directions for explaining explanations, 2024. URL https://arxiv.org/abs/2405.06064.