Explainable Artificial Intelligence in Medical Diagnostics: Bridging Accuracy and Trust

Affiliation: INAI

Date: December 13, 2025

Abstract

Artificial Intelligence (AI) systems are increasingly deployed in medical diagnostics; however, their practical adoption remains constrained by limited transparency and interpretability. This paper presents an author-defined analytical framework for Explainable Artificial Intelligence (XAI) in medical diagnostics, focusing on the relationship between explanation type, clinical task, and decision risk. Unlike purely descriptive surveys, this work introduces a structured taxonomy that categorizes XAI methods according to their clinical applicability and potential sources of misinterpretation. The study synthesizes recent research and highlights design trade-offs between diagnostic accuracy and explanation usability. The proposed framework is intended to support clinicians, researchers, and system designers in selecting appropriate XAI techniques for real-world medical applications.

Keywords

Explainable AI, medical diagnostics, healthcare AI, interpretability, clinical decision support, trust in AI, machine learning, ethical AI.

Introduction

The rapid integration of AI into healthcare has transformed diagnostic processes, enabling faster and more accurate detection of diseases. Deep learning models, particularly convolutional and transformer-based architectures, are widely used for analyzing medical images, electronic health records, and genomic data. Despite their effectiveness, these models often operate as opaque systems, offering little insight into how conclusions are reached.

In clinical environments, where decisions can have life-critical consequences, the inability to explain AI predictions poses ethical, legal, and practical

challenges. Regulatory frameworks such as GDPR and emerging medical AI standards increasingly emphasize the "right to explanation." Explainable AI seeks to address these issues by providing human-understandable justifications for model outputs. This paper explores the transition from performance-driven AI to trust-centered diagnostic systems powered by XAI methodologies.

Core XAI Techniques in Medical Diagnostics

Author Contribution and Taxonomy

The primary contribution of this paper is a clinically oriented taxonomy of XAI techniques, grouping methods based on their mode of explanation and diagnostic relevance rather than purely algorithmic properties. This taxonomy emphasizes how explanations are consumed by clinicians during decision-making.

The application of XAI in healthcare relies on several foundational techniques:

Model-Agnostic Explanation Methods: Tools such as LIME and SHAP generate local explanations by approximating complex models with simpler interpretable representations, helping clinicians understand feature importance for individual predictions.

Intrinsic Interpretability Models: Instead of explaining black-box systems, some approaches favor inherently interpretable models such as decision trees, rule-based systems, or generalized additive models, particularly for structured clinical data.

Visual Explanations for Imaging: In radiology and pathology, saliency maps, Grad-CAM, and attention heatmaps highlight regions of medical images that most influence AI predictions, enabling visual validation by specialists.

Counterfactual Explanations: These methods describe how minimal changes in input data could alter a diagnostic outcome, supporting clinical reasoning and treatment planning.

Table 1. Taxonomy of XAI Techniques in Medical Diagnostics

XAI Category	Typical Methods	Data Type	Clinical Purpose	Key Limitations
Feature-based explanations	SHAP, LIME	Structured clinical data	Risk factor analysis	Local instability
Visual attribution	Grad-CAM, saliency maps	Medical imaging	Lesion localization	Sensitivity to noise
Rule-based reasoning	Decision rules, trees	EHR, lab data	Transparent decision support	Limited scalability
Counterfactual explanations	What-if analysis	Mixed data	Treatment planning	Clinical plausibility

Benefits and Clinical Impact

Explainable AI delivers multiple advantages in diagnostic settings. First, it improves clinician trust by aligning AI reasoning with medical knowledge. Second, it enhances error detection by allowing experts to identify spurious correlations or data bias. Third, XAI supports education and training by providing insight into complex diagnostic patterns. Finally, transparent models facilitate compliance with healthcare regulations and ethical guidelines.

Empirical evidence suggests that XAI-assisted diagnostic tools reduce decision uncertainty and improve collaboration between AI systems and medical professionals. Rather than replacing clinicians, XAI reinforces a human-in-the-loop paradigm where AI acts as an intelligent assistant.

Challenges and Limitations

Despite its promise, XAI faces significant challenges. Explanations may oversimplify complex model behavior, leading to false confidence. Excessive or

poorly designed explanations can overwhelm clinicians and reduce usability. Additionally, there is no universal standard for evaluating explanation quality, making comparison across systems difficult.

Another concern is the trade-off between accuracy and interpretability. Highly interpretable models may underperform compared to deep neural networks, particularly in complex diagnostic tasks. Balancing these factors remains an open research problem.

Conclusion

This paper proposed a clinically grounded framework for understanding and applying Explainable Artificial Intelligence in medical diagnostics. By introducing an author-defined taxonomy and explicitly linking explanation types to clinical tasks, the study goes beyond a general literature overview and offers practical guidance for system design and evaluation. While Explainable AI does not eliminate all risks associated with automated diagnostics, it provides mechanisms for accountability, error detection, and informed human oversight. Future research should focus on empirical validation of explanation effectiveness in clinical workflows and the development of standardized evaluation metrics for medical XAI.

References

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11).

European Commission. (2021). Ethics guidelines for trustworthy AI. *Digital Strategy Publications*.

15.E. Usupova and A. Khan, "Optimizing ML Training with Perturbed Equations," 2025 6th International Conference on Problems of Cybernetics and Informatics (PCI), Baku, Azerbaijan, 2025, pp. 1-6, doi: 10.1109/PCI66488.2025.11219819.