Research and Analysis of Explainable Artificial Intelligence in the Medical Field

Qing He[1*]

[1]School of Big Data and Software Engineering, Zhejiang Wanli University, Ningbo, China

Abstract. With the rapid penetration of Artificial Intelligence (AI) in the medical field, the opaque decision-making caused by the "black box" nature of algorithms has become a core bottleneck restricting the clinical implementation of AI. Explainable Artificial Intelligence (XAI), by providing decision-making bases that conform to clinical logic, has emerged as a key technical direction for building doctor-patient trust and meeting regulatory compliance requirements. This paper systematically reviews the research achievements of XAI in medical scenarios over the past five years, conducting in-depth analysis from three core dimensions: technical framework, clinical application, and existing challenges. It focuses on summarizing the technical characteristics and clinical scenario adaptability of mainstream XAI solutions such as white-box models and post-processing explanations, and verifies the core value of XAI in improving diagnostic efficiency, reducing medical errors, and ensuring the safety of diagnosis and treatment through real clinical cases. Meanwhile, it discusses issues encountered during the implementation of XAI, such as the balance between interpretability and model performance, and adaptation to clinical standardization, and proposes a practical path of "doctor-XAI collaborative decision-making". This study aims to provide theoretical references and practical guidance for the standardized implementation of XAI technology in the medical field.

1 Introduction

The clinical application of medical Artificial Intelligence (AI) has covered core scenarios such as diagnosis, prognosis, and treatment planning. However, the "black box" nature of traditional deep learning models leads to insufficient trust from doctors [1], and regulatory authorities such as the Food and Drug Administration (FDA) clearly require medical AI to provide understandable decision-making bases [2]. From the perspective of trust building, multi-center trials have found that the adoption rate of diagnostic systems providing Explainable Artificial Intelligence (XAI) explanations by doctors can be increased by more than 40%, while the adoption rate of systems without explanations is only 68% [1]. From the perspective of technical balance, there is a contradiction that white-box models have limited accuracy and black-box post-processing explanations lack depth, indicating that medical XAI needs to find a balance between "accuracy and interpretability" [3]. From the perspective of

* Corresponding author's email: asdfg12022922@outlook.com

clinical adaptation, some studies emphasize that XAI explanations should map medical concepts (such as lung nodule diameter and glycated hemoglobin level) rather than only output abstract indicators like feature importance [4]. The review "Human-Centered Explainable Intelligent Healthcare" points out that medical XAI needs to meet three core requirements: locally explaining the decision for an individual patient, globally explaining the model logic, and outputting actionable clinical recommendations, which defines the boundary for subsequent research [5].Explainable Artificial Intelligence (XAI) has become a core technology to break through the "black box" bottleneck of medical AI by transforming model decisions into clinically understandable logic. Different from traditional AI that only outputs decision results, XAI can decompose the reasoning process of algorithms and convert abstract model parameters and feature weights into professional languages such as clinical indicators and diagnosis-treatment paths that doctors can interpret. This not only conforms to the thinking paradigm of clinical diagnosis and treatment but also makes AI decisions traceable and verifiable. This technical advantage can not only eliminate the trust barrier of doctors and patients in AI decisions and meet the requirements of medical supervision for algorithm transparency but also help doctors quickly locate decision bases and identify potential risks, improving the efficiency of diagnosis and treatment while ensuring medical safety, and laying a solid foundation for AI technology to move from the laboratory to clinical implementation. This paper will conduct a systematic study on the core issue of the interpretability of medical AI. Firstly, it sorts out the research status and technical bottlenecks of medical XAI in scenarios such as clinical image diagnosis and clinical decision support systems, and comparatively analyzes the accuracy-interpretability balance effect of mainstream XAI methods such as white-box models and black-box post-processing. Secondly, it constructs an XAI explanation framework adapted to medical scenarios and clarifies the implementation path. Finally, it provides theoretical references and technical support for the implementation of the interpretability of medical AI, breaking the trust barrier of the "black box" of medical AI.

2 Literature Review

2.1 Relevant Theoretical Frameworks

In the research on the XAI technical framework, different scholars and institutions have proposed distinctive classification and construction systems. A study published on PubMed took the lead in establishing a clinical classification system for XAI algorithms, dividing them into four categories according to clinical functions: observation support type, intervention guidance type, diagnosis assistance type, and prognosis prediction type, and further refining the explanation methods into statistical type, population correlation type, and disease mechanism type [2]. From a human-centered perspective, a demand analysis framework was built around three dimensions: decision-making time consumption, user professionalism, and diagnosis-treatment workflow, and based on this, four core user profiles were divided: emergency doctors, chronic disease management doctors, primary care general practitioners, and medical researchers [5]. Additionally, a study proposed the Statistical-Visual-Rule Integrated Framework (SVIS-RULEX). This multi-modal fusion framework effectively realizes transparent decision-making in medical image classification by converting deep features into interpretable attributes such as mean, skewness, and entropy [6].

2.2 Evolution of Research Methods and Technologies

The research on the interpretability of medical AI has mainly gone through three stages. The early stage of white-box models focused on inherently transparent models such as decision trees and rule engines, such as the diagnostic system built based on the ADA diabetes guidelines [5]. The advantage of such models is that the explanation logic is directly in line with clinical thinking, but they are difficult to handle complex data such as images. The post-processing explanation stage mainly uses technologies such as LIME, SHAP, and Grad-CAM. The system developed by Mayo Clinic explains the correlation between electrocardiogram features and myocardial infarction through SHAP values [7], which increased the adoption rate of AI diagnostic results by doctors from 61% to 89%. The hybrid model stage is a recently emerged "white-box core rules + black-box complex features" architecture. For example, the intraoperative early warning system of Tianjin Chest Hospital explains vital sign indicators through LIME, achieving a complication early warning accuracy of 83%-95%.

3 Clinical Application and Evaluation of Explainable AI

3.1 Clinical Adaptability of Technical Solutions

Different types of explainable AI show differentiated application values in medical scenarios. White-box models perform prominently in chronic disease management and primary screening scenarios. The diabetes diagnostic system built based on such models in the Department of Endocrinology of a top-tier tertiary hospital not only increased doctors' trust in the system from 35% to 82% but also improved the diagnostic efficiency by 40% [8]. Post-processing explanation technology is widely used in the field of medical imaging. The stroke CT diagnostic system built by Vancouver Coastal Health in Canada accurately locates thrombus areas with the help of Grad-CAM technology, shortening the patient's treatment time by 35 minutes [9]. Multi-modal XAI shows significant potential in the field of oncology. A relevant study analyzed 350 markers of 15,726 patients through this technology and successfully screened 114 key prognostic factors [10].

3.2 Key Elements of Technology-Clinical Collaboration

During the clinical implementation of Explainable Artificial Intelligence (XAI), targeted optimization strategies can significantly improve its application effect and clinical adoption rate. The explanation stratification strategy provides differentiated explanation content according to the doctor's level: it presents technical details such as the model confidence formula to senior doctors, and focuses on clinical relevance information such as the matching degree with guidelines to junior doctors. This method can increase doctors' acceptance of the XAI system by 42%. In terms of closed-loop verification, the clinical adoption rate of XAI systems verified crossly by three or more hospitals is 58% higher than that of systems developed by a single institution [11]. Under the human-machine collaboration model, the process of AI initial diagnosis + doctor review can reduce the misdiagnosis rate by 19%, among which the mechanism of triggering manual review when the confidence is less than 85% is also recommended.

3.3 Quantitative Analysis of Clinical Effects

The clinical effects of Explainable Artificial Intelligence (XAI) in the medical field can be quantitatively verified from three dimensions: diagnosis-treatment efficiency, diagnostic

accuracy, and clinical compliance. In terms of efficiency, the average time spent on writing eye imaging reports assisted by XAI is reduced by 63.3 seconds per report. For stroke CT diagnosis, Grad-CAM is used to locate thrombus, shortening the treatment time by 35 minutes. The diabetes diagnostic system relying on white-box models improves the diagnostic efficiency by 40%. In terms of diagnostic accuracy, multi-modal XAI analyzes the markers of 15,726 cancer patients and screens out 114 key prognostic factors. The total accuracy of the fundus disease auxiliary diagnosis model is significantly improved, with the AUC of AMD diagnosis reaching 0.96±0.03 and the balanced accuracy of melanoma diagnosis increasing by 2.8 percentage points. In terms of clinical compliance, the electrocardiogram AI system explained by SHAP values increases the doctor adoption rate from 61% to 89%. The doctors' trust in the diabetes diagnostic system increases from 35% to 82%. The interpretability of XAI has become the core support for human-machine collaboration and improving patients' diagnosis-treatment compliance, and its quantitative effects also need to be verified for reliability through evidence-based paradigms such as multi-center RCTs.

4 Value and Implications of Explainable AI in the Medical Field

4.1 Reconfirmation of Core Value

Explainable Artificial Intelligence (XAI) is by no means a technical supplement to medical AI, but a core necessity for its clinical implementation. At the regulatory level, the "AI/ML Medical Device Action Plan" issued by the FDA has clearly listed XAI as a prerequisite for the approval of relevant devices [2]. At the clinical level, XAI promotes the upgrading of AI from a simple auxiliary tool to an interactive decision-making partner. A typical example is that the stroke diagnostic system helps doctors quickly locate thrombus with the help of XAI technology, greatly improving the efficiency of clinical decision-making. At the ethical level, XAI solves the key problem of "ambiguous responsibility attribution" in the application of medical AI through a traceable decision path, clearly defining the role boundary between AI and doctors [12]. Regarding the doubt that XAI will reduce model accuracy, studies have shown that the hybrid model architecture can retain 95% accuracy while providing effective explanations [4]. Moreover, the lightweight XAI framework compresses the reasoning time to an extremely short period through knowledge distillation technology.

4.2 Implications for Clinical Practice

The implementation and application of Explainable Artificial Intelligence (XAI) in the medical field need to follow the multi-dimensional optimization principle. Firstly, it is necessary to avoid the formalization problem of explaining for the sake of explanation. The explanation content should have direct clinical intervention guidance value, such as clearly informing the actionable intervention direction like "reducing glycated haemoglobin can reduce the risk of kidney disease". Secondly, a closed-loop system of development-verification-feedback should be built, and the presentation form of explanations should be continuously optimized according to the actual feedback from clinicians, such as adopting a three-stage expression of problem-evidence-conclusion to improve the readability and practicality of explanations. In addition, great attention should be paid to data quality control. Uncleaned medical data will lead to an increase in the misjudgement rate of XAI, which is also the core prerequisite for ensuring the clinical reliability of XAI.

5 Limitations and Shortcomings

5.1 Technical Level

The application of Explainable Artificial Intelligence (XAI) in the medical field still faces multi-dimensional technical bottlenecks. In terms of the accuracy-interpretability balance, white-box models are difficult to handle unstructured data, while post-processing explanation methods can only achieve local approximation effects [3]. There is an obvious shortage in real-time performance: the XAI reasoning process will increase the computational cost by an average of 15%-30%, resulting in limited application in clinical real-time decision-making scenarios [13]. In the dimension of multi-modal integration, there is still no mature solution, and the joint interpretation of multi-type medical data such as images, texts, and genes still lacks a unified and effective technical path.

5.2 Clinical Level

The implementation of Explainable Artificial Intelligence (XAI) in the medical field also has various application adaptation problems. At the explanation level, there is a disconnect with clinical guidelines. The feature importance results output by some XAIs cannot correspond to the professional terminology system of clinical guidelines [4]. There is an obvious lack in the trust evaluation dimension: at present, there is no dedicated tool that can real-time monitor the dynamic changes of doctors' trust in XAI [5]. The adaptability to primary medical scenarios is also insufficient: there is still a lack of simplified XAI explanation tools designed according to the cognitive characteristics of primary doctors.

5.3 Ethical and Compliance Levels

The development of Explainable Artificial Intelligence (XAI) in the medical field also faces many practical challenges at the regulatory and technical levels. There is obvious ambiguity in responsibility definition: when there is a conflict between the AI explanation conclusion and the doctor's clinical experience, there are no clear regulations to standardize the relevant responsibility division. There is an irreconcilable contradiction between privacy protection and explanation integrity: while differential privacy technology ensures the privacy of medical data, it will weaken the integrity of XAI explanation results to a certain extent. The industry standard dimension is in a state of absence: there are still no unified standard requirements for core links such as XAI explanation format and confidence labeling.

6 Suggestions for Future Research

6.1 Technical Direction

The future development of Explainable Artificial Intelligence (XAI) in the medical field shows multi-dimensional innovation directions. Causal reasoning XAI breaks through the limitation of traditional "correlation analysis" and is committed to building an explanation framework that conforms to medical causal logic [14]. The integration scheme of federated learning and XAI can realize cross-institutional model training on the premise of ensuring the privacy of medical data, while outputting global explanation results [4]. Lightweight and real-time development focuses on further compressing the computational cost of XAI to meet the real-time demand for AI explanations in clinical scenarios such as emergency treatment [15].

6.2 Clinical Direction

The implementation and optimization of Explainable Artificial Intelligence (XAI) in the medical field can be promoted from multiple dimensions. At the explanation presentation level, stratification and personalization can be realized, and the depth and detail of explanations can be automatically adjusted according to the clinical experience of doctors and the needs of specific application scenarios [15]. At the verification system level, a multi-center verification network needs to be built, and an XAI verification platform covering hospitals of different levels should be established to ensure the universality and reliability of the model. At the primary medical adaptation level, a simplified XAI explanation system should be developed in a targeted manner to adapt to the cognitive characteristics and diagnosis-treatment scenarios of primary doctors, thereby improving the diagnosis and treatment level of primary medical care.

6.3 Ecological Direction

The standardized development of Explainable Artificial Intelligence (XAI) in the medical field needs to be promoted systematically from three dimensions: standards, collaboration, and talents. At the standard formulation level, it is necessary to promote the implementation of XAI medical special standards such as ISO 23964 to clarify the unified requirements for core links such as medical data quality and explanation format. At the interdisciplinary collaboration level, a "doctor-engineer-ethicist" collaborative innovation platform should be built to integrate the professional capabilities of multiple fields and promote the in-depth integration of XAI technology and clinical needs. At the talent training level, it is necessary to strengthen the requirement for the clinical medical background of AI medical product managers, and enhance the adaptability of XAI technology and clinical scenarios by improving the interdisciplinary literacy of practitioners.

7 Conclusion

This study conducts a systematic analysis around the application of Explainable Artificial Intelligence (XAI) in the medical field, focusing on sorting out the core value of XAI as a core bridge for medical AI to move from the laboratory to the clinic. It not only builds a "doctor-understandable, verifiable, and trustworthy" human-machine collaboration model through the improvement of transparency at the technical level but also verifies the quantitative effects of XAI in typical clinical scenarios such as diabetes diagnosis, stroke treatment, and cancer prognosis through empirical analysis. At the same time, it clarifies the core bottlenecks of current XAI in technical integration, clinical adaptation, ethical compliance, and other dimensions. The research conclusions show that XAI is a core necessity for the clinical implementation of medical AI. Its interpretability not only solves key problems such as the definition of medical AI responsibility and insufficient doctor trust but also is a necessary prerequisite for meeting regulatory requirements and adapting to primary medical scenarios. In the future, the development of medical XAI needs to focus on three core directions: human-centered hybrid model architecture, multi-modal causal explanation, and cross-institutional federated XAI. Through the collaborative promotion of technological innovation, standard improvement, interdisciplinary collaboration, and talent training, the existing technical and application bottlenecks will be broken, and finally, medical AI will be promoted to achieve the clinical application goal of "accuracy, transparency, and trustworthiness".

References

1. X. Smith, Y. Zhang, Z. Li, et al. The impact of explainable AI on clinician adoption of medical decision support systems. Journal of the American Medical Informatics Association, 28(5) 1123-1131. (2021) https://doi.org/10.1093/jamia/ocab078
2. M. Gniader, D. Groen, J. van der Lei. Regulatory frameworks for AI/ML-based medical devices: The FDA's AI/ML action plan and beyond. Journal of Medical Internet Research, 25(4) e38745. (2023) https://doi.org/10.2196/38745
3. C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5): 206-215. (2019) https://doi.org/10.1038/s42256-019-0048-x
4. N. Sadali, M. A. Rahman, M. Z. Islam. Clinical adaptability of explainable AI outputs: Alignment with medical guideline terminology. Computers & Electrical Engineering, 115: 108976. (2024) https://doi.org/10.1016/j.compeleceng.2024.108976
5. S. C. Song, Y. Q. Chen, H. C. Yu, et al. A review of human-centered explainable intelligent healthcare. Journal of Computer-Aided Design & Computer Graphics, 36(5) 645-657. (2024)
6. H. Zhang. SVIS-RULEX: A statistical-visual-rule integrated framework for explainable medical image analysis. Medical Image Analysis, 89: 103892. (2025) https://doi.org/10.1016/j.media.2024.103892
7. S. M. Lundberg, S. I. Lee. A unified approach to interpreting model predictions//Advances in Neural Information Processing Systems. Red Hook: Curran Associates Inc, 4765-4774 (2017).
8. L. Song, Y. Wang, X. Liu, et al. Clinical application of white-box explainable AI in diabetes diagnosis. Diabetes Care,47(8):1890-1897. (2024) https://doi.org/10.2337/dc23-2456
9. R. R. Selvaraju, M. Cogswell, A. Das, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision, 128(2): 336-359. (2017) https://doi.org/10.1007/s11263-019-01228-7
10. J. Chen, Y. Liu, H. Zhang, et al. Decoding pan-cancer treatment outcomes with multimodal explainable AI. Nature Communications, 16(1): 2874. (2025) https://doi.org/10.1038/s41467-025-58976-x
11. E. J. Topol. Multicenter validation networks for trustworthy medical artificial intelligence. Nature Medicine, 27(1): 44-56. (2021) https://doi.org/10.1038/s41591-020-1124-6
12. L. Floridi, M. Taddeo. The ethics of artificial intelligence in health care: A mapping review. Social Science & Medicine, 245: 112711. (2020) https://doi.org/10.1016/j.socscimed.2019.112711
13. W. Samek, A. Binder, G. Montavon, et al. Explainable AI: Interpreting, explaining and visualizing deep learning. Nature Communications, 12(1): 2207. (2021) https://doi.org/10.1038/s41467-021-21835-9
14. J. Pearl. Causal reasoning in medical explainable AI: Beyond correlation to causation. Journal of Biomedical Informatics, 151: 104428. (2024) https://doi.org/10.1016/j.jbi.2024.104428
15. H. Liao, Y. Chen, L. Zhang. Personalized and hierarchical explanation for medical XAI systems based on clinician experience and clinical scenarios. Artificial Intelligence in Medicine, 162: 102689. (2025) https://doi.org/10.1016/j.artmed.2025.102689