www.jcmcri.org

JCMCRI OPEN ACCESS

Journal Of
Clinical, Medical Case Reports And Images

ISSN ---- - ----

Narrative Review

Open Access, Volume 1

The role of explainable artificial intelligence (XAI) in clinical decision-making: a narrative review

Mehrdad Motamed Shariati*

Eye Research Centre, Mashhad University of Medical Sciences, Mashhad, Iran.

***Corresponding Author:**

Mehrdad Motamed Shariati MD,

Eye Research Center, Khatam Al-Anbia Eye Hospital, Gharani Boulevard, Mashhad, Iran.

Tel: +989377388690.

Email: Mehrdad_shariati2005@yahoo.com.

Received: July 16, 2026

Accepted: July 22, 2026

Published: July 23, 2026

Archived: www.jcmcri.org

Abstract

Artificial Intelligence (AI) is increasingly integrated into clinical workflows, supporting diagnostics, risk prediction, and treatment planning. However, traditional "black-box" models—particularly in deep learning—often lack transparency, hindering trust, accountability, and adoption in healthcare settings. This narrative review aims to explore the concept and significance of Explainable AI (XAI) in clinical decision-making. It critically examines the current landscape of XAI techniques, their practical applications in various medical domains, and the challenges that limit their clinical translation. A literature review was conducted using major databases (PubMed, Scopus, IEEE Xplore) to identify articles published between 2015 and 2025 related to XAI in healthcare. Emphasis was placed on studies describing the implementation and evaluation of XAI tools in real or simulated clinical settings. XAI techniques such as SHAP, LIME, Grad-CAM, and attention mechanisms have shown promise in enhancing interpretability across diverse applications, including radiology, ophthalmology, pathology, and electronic health records. These methods help clinicians understand AI outputs, facilitate trust, support shared decision-making, and fulfill regulatory requirements. Nonetheless, key challenges remain—particularly in balancing model accuracy with interpretability, standardizing explanations, and validating their utility from a clinician's perspective. XAI represents a crucial step toward trustworthy AI in healthcare. Future research should focus on developing domain-specific explanations, incorporating user feedback, and establishing evaluation frameworks to measure explanation quality and clinical impact.

Keywords: Explainable Artificial Intelligence (XAI), Clinical Decision-Making, Medical AI, Model Interpretability, Trust in Healthcare AI

Introduction

By providing tools that assist physicians with diagnosis, risk assessment, and treatment planning, artificial intelligence (AI) has become a vital part of modern healthcare [1, 2]. Despite these advances, caution remains regarding the use of AI in real healthcare settings. One major barrier is the lack of transparency in many AI systems, especially deep learning models, which often act as "black boxes"—offering little to no explanation for their predictions [3]. Accountability, interpretability, and trust are crucial in a high-stakes field like healthcare. When patient lives are at risk, clinicians need to understand why an AI model reached a particular conclusion [4, 5]. The aim of Explainable Artificial Intelligence (XAI) is to address this need by developing tools and methods that enable humans to understand AI outputs. XAI is not only desired but also essential in healthcare for ethical accountability, regulatory compliance, and informed shared decision-making [6-8]. This paper explores the role of XAI in clinical decision-making by examining its foundational concepts, current methods, real-world applications, and the challenges that must be addressed to fully integrate AI into medicine.

Methods

To create this narrative review, a comprehensive literature search was conducted across the PubMed, Scopus, and IEEE Xplore databases, encompassing publications from 2015 to 2025. The search included combinations of terms such as "explainable AI," "XAI," "interpretability," "clinical decision-making," "medical artificial intelligence," and "trust in AI." Priority was given to peer-reviewed original research articles, review papers, and position statements that describe the development, evaluation, or implementation of XAI techniques in healthcare. Studies were selected based on their relevance to clinical decision support, human-AI interaction, model transparency, or real-world applications in various medical fields such as radiology, ophthalmology, pathology, and electronic health records. There was no restriction on study design. Reference lists of key articles were also examined to find additional relevant literature. Articles were reviewed and analyzed to identify common themes, application areas, and unresolved challenges in implementing XAI in clinical settings. This review aims not to provide a systematic or quantitative synthesis of evidence but to describe the current landscape and emphasize conceptual and practical issues of XAI in medicine.

Overview of Explainable Artificial Intelligence (XAI)

Making AI systems' decision-making processes more visible and intelligible to human users is the goal of the large and developing area of explainable AI [9]. Such explainability is particularly important in the healthcare industry, where professionals are held to ethical and legal standards and choices can have a profound impact on patients' lives [10]. XAI methods may be divided into two categories: those that offer local explanations, which concentrate on the reasons behind a model's particular prediction in a given scenario, and those that offer global explanations, which give a broad grasp of how the model functions generally [11, 12]. Additionally, methods may be model-agnostic, meaning they can be applied to any machine learning model, or model-specific, meaning they are tailored to particular architectures like convolutional neural networks or decision trees [13-15]. Local Interpretable Model-agnostic Explanations (LIME), which uses simpler interpretable models to approximate the model locally, gradient-based techniques like Grad-CAM, which highlight key regions in image classification tasks, and SHapley Additive exPlanations (SHAP), which use game theory to assign feature importance scores, are some of the most popular XAI techniques [16, 17]. Other emerging tools include counterfactual explanations, which suggest what changes would have resulted in a different outcome, and attention mechanisms, particularly in transformer models, that identify which parts of the input the model focused on. These tools collectively aim to foster interpretability without sacrificing too much performance, although achieving this balance remains a challenge.

Applications of XAI in Clinical Practice

Explainable AI has been used in several medical professions, bridging the gap between clinical reasoning and complex algorithms. In radiology, interpretability tools like Grad-CAM allow physicians to see which areas of an image contributed most to a diagnosis, such as pneumonia or cancer. AI is widely used in radiology to identify abnormalities in imaging modalities, including X-rays, CT scans, and MRIs. Research has indicated that there is a considerable boost in diagnostic confidence and clinician trust when model attention maps correspond with areas radiologists find relevant [18-20].

AI has also been quickly adopted in ophthalmology, especially in the interpretation of fundus photos and OCT scans for conditions including glaucoma, age-related macular degeneration, and diabetic retinopathy [21, 22]. XAI techniques such as SHAP and attention-based visualization allow ophthalmologists to determine if the model is responding to clinically important characteristics like retinal layer disturbances or vascular abnormalities in these situations [23, 24].

Similarly, predictive models based on electronic health records are increasingly being used for early warning systems, such as sepsis prediction or ICU triage. Here, XAI enables real-time treatments and promotes alignment with clinical intuition by enabling physicians to observe how vital signs, laboratory data, and historical factors contribute to a risk prediction [25, 26].

In digital pathology, XAI helps interpret high-resolution histopathological slides by highlighting relevant cellular features such as nuclear atypia or mitotic figures, which may support or contradict human evaluation [27, 28].

Across these domains, XAI not only enhances trust and usability but also helps identify potential biases or errors in model development.

Benefits of XAI in Clinical Decision-Making

Increasing clinician confidence in AI-assisted judgments is one of the main advantages of XAI in clinical settings. Physicians are more likely to incorporate a model into their clinical workflow when they can comprehend and verify the reasoning behind the advice [29]. Explainability also supports human-AI collaboration, allowing for a feedback loop in which clinicians refine their decisions with AI input and, in turn, improve the performance of AI models through supervised learning [30]. Regulatory compliance is another critical advantage. Legal frameworks such as the European Union's General Data Protection Regulation (GDPR) increasingly require a "right to explanation" for automated decision-making, and XAI provides the means to meet such standards [31]. In addition to supporting professional accountability, explainability promotes patient-centered care. Patients are more likely to consent to AI-influenced diagnoses or treatment plans when clinicians can explain the rationale behind them in accessible language. This transparency fosters informed consent and shared decision-making, key elements of modern medical ethics [32, 33]. Moreover, XAI can improve training and education by helping less experienced clinicians understand the clinical reasoning behind certain diagnoses or predictions, essentially serving as a digital second opinion [34]. Ultimately, XAI strengthens the credibility, safety, and utility of AI in medicine, paving the way for its more widespread and ethical integration [35].

Limitations and Challenges

Despite its promise, explainable AI in medicine faces several substantial challenges. One of the most persistent is the trade-off between model accuracy and interpretability. Simple models such as logistic regression or decision trees are inherently more transparent but often underperform compared to deep neural networks, which are more accurate yet opaque. Another issue is the lack of standardization in evaluating explanation quality. Different XAI methods can yield contradictory explanations for the same input, leading to confusion or even mistrust among clinicians. This problem is exacerbated by the absence of widely accepted metrics for measuring the effectiveness, clarity, or faithfulness of an explanation. Cognitive overload is another practical concern. In busy clinical environments, overly detailed or technical explanations may burden rather than assist decision-makers, especially if the explanation interfaces are not well-integrated into electronic health records or imaging platforms. Moreover, there is a fundamental philosophical problem: explainability does not always equate to correctness. A model can produce a plausible explanation for a flawed prediction, creating a false sense of security. Finally, biases in training data, lack of generalizability, and legal uncertainty regarding liability when using AI tools add to the complexity of deploying XAI in clinical practice. These limitations highlight the need for caution, continuous validation, and human oversight in the deployment of AI systems, no matter how interpretable they claim to be.

Emerging Trends and Opportunities

The next frontier for XAI in healthcare involves the development of personalized explanations that can adapt to the expertise and preferences of different users. For instance, a radiologist may benefit from visual heatmaps, whereas a general physician might prefer a concise list of contributing variables. Human-in-the-loop systems represent another promising area, where clinicians can interact with AI tools in real time, query model behavior, or provide corrective feedback. Incorporating clinician feedback into the training process could significantly improve both performance and usability. Education and training will also be essential. As AI becomes more integrated into healthcare, medical schools and continuing education programs must incorporate basic AI and XAI literacy to prepare clinicians to work alongside intelligent systems. There is also a critical need to establish robust evaluation frameworks and benchmarks for XAI tools. These frameworks should consider not only technical metrics but also clinical relevance, user satisfaction, and impact on patient outcomes. Finally, interdisciplinary collaboration between computer scientists, clinicians, ethicists, and regulatory bodies will be vital to ensure that XAI technologies are developed and implemented in ways that align with the values and complexities of clinical medicine.

Conclusion

Explainable artificial intelligence represents a vital bridge between complex computational models and the human-centered world of clinical decision-making. By making AI systems more interpretable and transparent, XAI enhances trust, supports collaboration, and facilitates ethical and regulatory compliance in healthcare. Real-world applications across radiology, ophthalmology, pathology, and predictive analytics show that XAI can improve not only model usability but also clinical confidence and patient engagement. Nonetheless, significant challenges remain in balancing accuracy with interpretability, standardizing explanation quality, and integrating these tools seamlessly into clinical workflows. As the healthcare system continues to adopt AI technologies, it must do so with an emphasis on clarity, accountability, and partnership between machines and humans. The continued evolution of XAI holds the potential to make AI not only more powerful but also more acceptable, equitable, and human-centered.

Statements and Declarations

Competing interests: The author declares no competing interests.

Funding: None.

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and material: Not applicable.

Authors' contributions: The author confirms sole responsibility for all aspects of the manuscript. This includes the conception and design of the work; acquisition, analysis, and interpretation of the literature; drafting and critical revision of the manuscript; final approval of the version to be published; and agreement to be accountable for all aspects of the work in ensuring accuracy and integrity.

Acknowledgements: None.

Declarations

Competing interests: The author declares no competing interest.

References

1. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. Journal of the American medical informatics association. 2020;27(3):491-7.
2. Babic B, Gerke S, Evgeniou T, Cohen IG. Beware explanations from AI in health care. Science. 2021;373(6552):284-6.
3. Park C-W, Seo SW, Kang N, Ko B, Choi BW, Park CM, et al. Artificial intelligence in health care: current applications and issues. Journal of Korean medical science. 2020;35(42).
4. Maddox TM, Rumsfeld JS, Payne PR. Questions for artificial intelligence in health care. Jama. 2019;321(1):31-2.
5. Väänänen A, Haataja K, Vehviläinen-Julkunen K, Toivanen P. AI in healthcare: A narrative review. F1000Research. 2021;10:6.
6. Howell MD, Corrado GS, DeSalvo KB. Three epochs of artificial intelligence in health care. Jama. 2024;331(3):242-4.
7. Saraswat D, Bhattacharya P, Verma A, Prasad VK, Tanwar S, Sharma G, et al. Explainable AI for healthcare 5.0: opportunities and challenges. IEEe Access. 2022;10:84486-517.
8. Srinivasu PN, Sandhya N, Jhaveri RH, Raut R. From blackbox to explainable AI in healthcare: existing tools and case studies. Mobile Information Systems. 2022;2022(1):8167821.
9. Pawar U, O'shea D, Rea S, O'reilly R, editors. Explainable AI in

healthcare. 2020 international conference on cyber situational awareness, data analytics and assessment (CyberSA); 2020: IEEE.

10. Chaddad A, Peng J, Xu J, Bouridane A. Survey of explainable AI techniques in healthcare. Sensors. 2023;23(2):634.

11. Sindiramutty SR, Tee WJ, Balakrishnan S, Kaur S, Thangaveloo R, Jazri H, et al. Explainable AI in healthcare application. Advances in Explainable AI Applications for Smart Cities: IGI Global Scientific Publishing; 2024. p. 123-76.

12. Cinà G, Röber T, Goedhart R, Birbil I. Why we do need explainable ai for healthcare. arXiv preprint arXiv:220615363. 2022.

13. Gerlings J, Jensen MS, Shollo A. Explainable AI, but explainable to whom? An exploratory case study of xAI in healthcare. Handbook of Artificial Intelligence in Healthcare: Vol 2: Practicalities and Prospects: Springer; 2021. p. 169-98.

14. Bharati S, Mondal MRH, Podder P. A review on explainable artificial intelligence for healthcare: Why, how, and when? IEEE Transactions on Artificial Intelligence. 2023;5(4):1429-42.

15. Ladbury C, Zarinshenas R, Semwal H, Tam A, Vaidehi N, Rodin AS, et al. Utilization of model-agnostic explainable artificial intelligence frameworks in oncology: a narrative review. Translational cancer research. 2022;11(10):3853.

16. Wang Y. A comparative analysis of model agnostic techniques for explainable artificial intelligence. Research Reports on Computer Science. 2024:25–33-25–33.

17. Devireddy K. A comparative study of explainable AI methods: Model-agnostic vs. model-specific approaches. arXiv preprint arXiv:250404276. 2025.

18. De Vries BM, Zwezerijnen GJ, Burchell GL, van Velden FH, Menke-van der Houven van Oordt CW, Boellaard R. Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review. Frontiers in medicine. 2023;10:1180773.

19. Neri E, Aghakhanyan G, Zerunian M, Gandolfo N, Grassi R, Miele V, et al. Explainable AI in radiology: a white paper of the Italian Society of Medical and Interventional Radiology. La radiologia medica. 2023;128(6):755-64.

20. Groen AM, Kraan R, Amirkhan SF, Daams JG, Maas M. A systematic review on the use of explainability in deep learning systems for computer aided diagnosis in radiology: Limited use of explainable AI? European journal of radiology. 2022;157:110592.

21. Tan TF, Dai P, Zhang X, Jin L, Poh S, Hong D, et al. Explainable artificial intelligence in ophthalmology. Current Opinion in Ophthalmology. 2023;34(5):422-30.

22. Singh A, Jothi Balaji J, Rasheed MA, Jayakumar V, Raman R, Lakshminarayanan V. Evaluation of explainable deep learning methods for ophthalmic diagnosis. Clinical Ophthalmology. 2021:2573-81.

23. Shipra EH, Rahman MS, editors. An explainable artificial intelligence strategy for transparent deep learning in the classification of eye diseases. 2024 IEEE International Conference on Computing, Applications and Systems (COMPAS); 2024: IEEE.

24. Obayya M, Nemri N, Nour MK, Al Duhayyim M, Mohsen H, Rizwanullah M, et al. Explainable artificial intelligence enabled TeleOphthalmology for diabetic retinopathy grading and classification. Applied Sciences. 2022;12(17):8749.

25. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, Miller T, Bian J, Chen JH, et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. Journal of the American Medical Informatics Association. 2020;27(7):1173-85.

26. Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jørgensen MJ, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nature communications. 2020;11(1):3852.

27. Klauschen F, Dippel J, Keyl P, Jurmeister P, Bockmayr M, Mock A, et al. Toward explainable artificial intelligence for precision pathology. Annual Review of Pathology: Mechanisms of Disease. 2024;19(1):541-70.

28. Holzinger A, Malle B, Kieseberg P, Roth PM, Müller H, Reihs R, et al. Towards the augmented pathologist: Challenges of explainable-ai in digital pathology. arXiv preprint arXiv:171206657. 2017.

29. Phillips V. A counterintuitive approach to explainable AI in healthcare: balancing transparency, efficiency, and cost. AI & SOCIETY. 2025:1-7.

30. Talukder N. Clinical decision support system: an explainable AI approach: N. Talukder; 2024.

31. Freyer N, Groß D, Lipprandt M. The ethical requirement of explainability for AI-DSS in healthcare: a systematic review of reasons. BMC Medical Ethics. 2024;25(1):104.

32. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Consortium PQ. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC medical informatics and decision making. 2020;20(1):310.

33. Vainio-Pekka H, Agbese MO-O, Jantunen M, Vakkuri V, Mikkonen T, Rousi R, et al. The role of explainable AI in the research field of AI ethics. ACM Transactions on Interactive Intelligent Systems. 2023;13(4):1-39.

34. Fiok K, Farahani FV, Karwowski W, Ahram T. Explainable artificial intelligence for education and training. The Journal of Defense Modeling and Simulation. 2022;19(2):133-44.

35. Hossain MI, Zamzmi G, Mouton PR, Salekin MS, Sun Y, Goldgof D. Explainable AI for medical data: Current methods, limitations, and future directions. ACM Computing Surveys. 2025;57(6):1-46.