International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 19s, 2025
https://theaspd.com/index.php

"Bridging Accuracy AND Transparency: Explainable Ai IN Healthcare -A Review"

Anindita Chakraborty[1*], Suvojit Mukhopadhyay[2], Piyali De[3], Sreelekha Paul[4], Sayak Banerjee[5], Nirmallya Roy[6]

[1]Department of CSE-AI, Brainware University, Kolkata, West Bengal, India, Email: ani.9012@gmail.com*
[2]Department of CSE, Indian Institute of Information Technology, Kalyani, West Bengal, India, Email: suvojitmukhopadhyayr@gmail.com
[3]Department of CSE, Brainware University, Kolkata, West Bengal, India, Email: me.piyalide@gmail.com
[4]Department of CSE, Brainware University, Kolkata, West Bengal, India, Email: paulsree350@gmail.com
[5]Department of CSE-AI, Brainware University, Kolkata, West Bengal, India, Email: banerjeesayak81@gmail.com
[6]Department of CSE-AI, Brainware University, Kolkata, West Bengal, India, Email: work.nirmallya@gmail.com

ABSTRACT
The application of artificial intelligence (AI) in healthcare has demonstrated revolutionary potential in the areas of patient outcome prediction, treatment planning, and diagnosis. But many high-performing models, particularly deep learning systems, are opaque, which is problematic in high-stakes situations where accountability and interpretability are crucial. This study investigates how Explainable AI (XAI) might improve decision-making in the healthcare industry by fostering greater transparency, trust, and dependability. We explore important XAI approaches, such as interpretable-by-design architectures, model-agnostic methodologies, and visualization techniques, and look at how they are used in clinical decision support systems, electronic health record analysis, and medical imaging. In order to ensure that judgments are in line with clinical and ethical norms, XAI helps close the gap between human expertise and sophisticated computer models by empowering physicians to comprehend the reasoning behind AI-generated recommendations. We also go over issues like treating biases that may result in unfair treatment outcomes, controlling uncertainty, and striking a balance between interpretability and accuracy. In the end, safer and more efficient patient care can result from the improved collaboration between AI systems and healthcare workers, as demonstrated by case studies. Research gaps are identified in the paper's conclusion, including the need for domain-specific interpretability methodologies, uniform evaluation frameworks, and integration strategies that take clinical workflow into consideration. The significance of developing AI systems that are not just clever but also reliable and accountable is highlighted by this work's emphasis on explain ability, opening the door for their responsible implementation in high-stakes healthcare settings.

Keywords: *Explainable AI, healthcare, interpretability, decision-making, trust, transparency, clinical support systems*

INTRODUCTION

With the growing value for the healthcare sector, artificial intelligence (AI) is highly valued for improving clinical diagnostics, operational efficiency, and treatment decisions. Better predictions in clinical diagnostics reduce the chance of human errors. This has led to investment in personalized treatment planning. However, many of these techniques have created what can be seen as "black box" interventions that deliver results without clarity. In critical areas like healthcare, where outcomes can involve life-or-death situations, this lack of insight can cause distrust, reluctance to use these tools, and ethical concerns.

Explainable AI (XAI) aims to address this issue. The goal is to create models that are both clear and understandable while maintaining their effectiveness. This allows clinicians to grasp, if not fully agree with, the decisions made by AI. This review examines how the development, application, and future of XAI work within healthcare to build trust, uphold ethical standards, and encourage collaboration between humans and AI.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 19s, 2025
https://theaspd.com/index.php

HISTORICAL CONTEXT OF AI IN HEALTHCARE

Early Developments

In the 1970s, the healthcare field started looking into AI with rule-based expert systems like MYCIN. These systems aimed to diagnose infectious diseases. They operated on simple principles. They used clear rules and decision trees, making their reasoning easy to understand. Still, their simplicity had downsides. They could only deal with specific, predefined problems and had a hard time adapting to new situations. Although being transparent was a benefit, their performance relied entirely on the knowledge provided by human experts. This limitation impacted their ability to predict outcomes accurately.

The Rise of Machine Learning

Statistical machine learning models that identified patterns in huge datasets gained popularity in the 2000s. Uses expanded in areas such as disease risk assessment, medical imaging, and treatment recommendations. However, as algorithms grew more complex, they became harder to understand. This raised concerns about "black box" decision-making.

Deep Learning Era and the Interpretability Challenge

Deep learning revolutionized medical AI from the 2010s, achieving state-of-the-art outcomes in genomics, pathology, and imaging. Despite their accuracy, many of these models were often opaque, which raised practical, ethical, and regulatory concerns for clinical decision-making. This led to research on XAI methods able to break the code.

Current Trends and Innovations in Explainable AI

Some of the recent developments in healthcare's XAI are:
i. Intrinsically Interpretable Models like decision trees, rule learners, and generalized additive models (GAMs) designed specifically for clinical applications.
ii. Visualization Tools that emphasize key features in medical images employed for diagnosis.
iii. Causality-based Models incorporating domain knowledge to increase trustworthiness.

Incorporation of these methods into clinical workflows is making it possible to provide real-time, explainable recommendations in intensive care monitoring, oncology treatment planning, and predictive triage.

Applications of XAI in High-Stakes Healthcare

i. Diagnostic Decision-Making Support

XAI models aid pathologists and radiologists, among others, by pinpointing tumor regions on MRI scans and providing explanations alongside their predictions.

ii. Therapy Recommending Systems

XAI explains how patient-specific factors influence the recommendation, thus supporting reasoning for therapy decisions in critical-care and oncology.

iii. System for Risk Prediction and Early Warning Systems

In the ICU, these interpretable tools help clinicians understand why a patient is at high risk of sepsis or cardiac arrest so that emergency interventions can be made.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 19s, 2025
https://theaspd.com/index.php

Comparative Literature Review Table

Category	Period	Key XAI Techniques	Applications	Reference
Early AI in Healthcare	1976	Rule-Based Expert Systems (MYCIN)	Infectious disease diagnosis	[1] Shortliffe, 1976
Early AI in Healthcare	1990	Decision Trees, Expert Systems	Diagnostic decision support	[2] Miller, 1990
Machine Learning Era	2001	Logistic Regression, Naïve Bayes	Medical diagnosis, classification	[3] Kononenko, 2001
Machine Learning Era	2008	Predictive Data Mining	Clinical medicine, patient stratification	[4] Bellazzi & Zupan, 2008
Machine Learning Era	2015	Random Forests, SVMs	Risk prediction in medicine	[5] Deo, 2015
Deep Learning Era	2017	CNNs	Skin cancer classification	[6] Esteva et al., 2017
Deep Learning Era	2017	CNNs	Pneumonia detection from X-rays (CheXNet)	[7] Rajpurkar et al., 2017
Deep Learning Era	2017	CNNs, RNNs, LSTM	Medical image analysis survey	[8] Litjens et al., 2017
Post-hoc XAI Methods	2017	SHAP	Interpretable predictions in healthcare models	[9] Lundberg & Lee, 2017
Post-hoc XAI Methods	2016	LIME	Explaining any classifier's predictions	[10] Ribeiro et al., 2016
Post-hoc XAI Methods	2017	Grad-CAM	Visual explanations for deep models in imaging	[11] Selvaraju et al., 2017
Intrinsically Interpretable Models	2015	Generalized Additive Models (GAMs)	Pneumonia risk prediction, readmission	[12] Caruana et al., 2015
Intrinsically Interpretable Models	2012	Rule Lists, Interpretable Models	Transparent classification and regression	[13] Lou et al., 2012
Causality and Knowledge-Driven XAI	2019	Causal Graphs	Explainable inference in healthcare	[14] Pearl, 2019
Causality and Knowledge-Driven XAI	2020	Counterfactual Prediction	ICU risk, outcome forecasting	[15] Prosperi et al., 2020
Causality and Knowledge-Driven XAI	2019	Causability, Hybrid Systems	Explainability in medical decision-making	[16] Holzinger et al., 2019
Clinical Workflow & Human-Centered XAI	2019	Human-in-the-Loop ML	Clinician adoption of explainable ML	[17] Tonekaboni et al., 2019
Multidisciplinary Perspectives on XAI	2020	Ethical & Sociotechnical Frameworks	Trust, transparency, communication	[18] Amann et al., 2020
Multimodal & Imaging-Based Explainability	2020	Explainable Deep Learning Models	Medical image interpretability	[19] Singh et al., 2020
Ethical & Regulatory Frameworks	2021	Trustworthy AI Guidelines	Standards, compliance, governance	[20] EU Commission, 2021

DISCUSSION ON KEY CONTRIBUTIONS

Few of the evaluated studies stand out as having had a particularly significant influence on the development of explainable AI (XAI) in the medical field. Since it established transparent, rule-based expert systems and demonstrated how decision support tools may improve medical thinking, Shortliffe's MYCIN [1] remains a seminal work in the early AI era. Despite having a narrow focus, it prepared the way for systems that came after that prioritized interpretability as a design concept.

Deo [5] is considered a significant step in integrating machine learning and traditional healthcare practice as we enter the machine learning era. With an emphasis on both performance and real-world uses in risk assessment and diagnostics, his work offered one of the most thorough introductions to machine learning in medicine. Bellazzi & Zupan [4] also made important contributions by concentrating on predictive data mining, which created the foundation for managing patient datasets from the real world.

Results from the deep learning era were revolutionary, especially those of Rajpurkar et al. [7] and Esteva et al. [6]. These research showed that in tasks like pneumonia identification and skin cancer categorization, deep neural networks might perform on par with or even better than expert clinicians. By examining medical image applications, Litjens et al. [8] further solidified this field and established it as a fundamental reference. But these developments also brought attention to the interpretability gap, which sparked a boom in XAI research.

Lundberg & Lee [9] (SHAP) and Ribeiro et al. [10] (LIME) have had a particularly significant impact on post-hoc explain ability techniques. With immediate adoption in healthcare applications, both approaches are increasingly commonplace instruments for analysing machine learning predictions across domains. By enabling physicians to see which aspects of an image influenced an AI's choice, Selvaraju et al. [11] (Grad-CAM) further transformed medical imaging and closed a significant gap in clinical trust.

Caruana et al. [12] provided one of the most useful examples of intrinsically interpretable models in a high-stakes healthcare scenario—predicting pneumonia risk—while preserving performance in the interpretable model stream. Because it is shown that interpretability need not necessarily be sacrificed for accuracy, this study is still frequently quoted.

Pearl [14] stands out as the theoretical foundation of causal inference from the standpoint of causality and ethics. Prosperi et al. [15] and Holzinger et al. [16] have recently expanded this foundation into the healthcare industry. Together, these publications developed the idea of "causability," highlighting the need for explanations to be consistent with medical logic rather than merely statistical correlations.

Lastly, Tonekaboni et al. [17] are crucial in terms of acceptance and governance because they put the clinician's viewpoint front and centre and show that usability and reliability are just as crucial as technological innovation. Similar to this, Amann et al. [18] offered a multidisciplinary perspective that focused on ethical and patient-centered issues, and the Ethics Guidelines for Trustworthy AI [20] published by the EU Commission established regulatory criteria that are expected to influence international norms for years to come.

All things considered, the most significant works fall into one of three categories:

1. Basic clinical AI systems ([1], [5], [6], and [7]) demonstrated high performance and viability.
2. Technical interpretability was made achievable by core XAI techniques ([9], [10], [11], [12]).
3. Ethical and causal frameworks ([14], [16], [17], [20]) → guaranteed conformity with policy, clinical trust, and medical practice.

In addition to defining the direction of explainable AI in healthcare, these contributions collectively brought attention to the gaps that still exist, especially with regard to domain-specific interpretability, standardized assessment frameworks, and smooth integration into clinical processes.

CHALLENGES AND LIMITATIONS

i. Find the Balance Between Interpretability and Accuracy

Although complex models are prone to lower transparency, extreme interpretability might lose prediction performance.

ii. Safety and Privacy of Data

Embedding XAI in healthcare should be done in full compliance with applicable privacy laws such as GDPR and HIPAA.

iii. Injustices and Prejudices

Models must be able to explain not only their outcomes but also how they find and remove prejudices within healthcare data.

Future Scope

i. Human-AI Hybrid Decision Systems

Combining human knowledge with AI prediction capabilities to increase trust and accountability.

ii. Frameworks for Regulation and Ethics

Creation of worldwide XAI standards for deployment and clinical validation.

iii. Explain ability in the Patient Centre

Customizing AI justifications for various audiences, including legislators, patients, and clinicians.

CONCLUSION

Explainable AI becomes crucial to achieving ethical compliance, transparency, and credibility in healthcare decision-making. XAI is more acceptable for regulators, patients, and physicians because it blends human intuition with algorithmic forecasts, which are by nature hard to understand. The future of safe and reliable AI applications in clinical practice looks bright due to continuous improvements in XAI techniques and human-AI collaboration. It's still challenging to overcome biases, stay in compliance with regulations, and balance interpretability and performance.

REFERENCES

1. Shortliffe, E. H. (1976). Computer-Based Medical Consultations: MYCIN. Elsevier.
2. Miller, R. A. (1990). "Medical diagnostic decision support systems—past, present, and future." JAMA.
3. Kononenko, I. (2001). "Machine learning for medical diagnosis: history, state of the art and perspective." Artificial Intelligence in Medicine.
4. Bellazzi, R., & Zupan, B. (2008). "Predictive data mining in clinical medicine." Current Opinion in Critical Care.
5. Deo, R. C. (2015). "Machine learning in medicine." Circulation.
6. Esteva, A., et al. (2017). "Dermatologist-level classification of skin cancer with deep neural networks." Nature.
7. Rajpurkar, P., et al. (2017). "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning." arXiv preprint.
8. Litjens, G., et al. (2017). "A survey on deep learning in medical image analysis." Medical Image Analysis.
9. Lundberg, S. M., & Lee, S.-I. (2017). "A unified approach to interpreting model predictions." NeurIPS.
10. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you? Explaining the predictions of any classifier." KDD.
11. Selvaraju, R. R., et al. (2017). "Grad-CAM: Visual explanations from deep networks via gradient-based localization." ICCV.
12. Caruana, R., et al. (2015). "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." KDD.
13. Lou, Y., Caruana, R., & Gehrke, J. (2012). "Intelligible models for classification and regression." KDD.
14. Pearl, J. (2019). The Book of Why: The New Science of Cause and Effect. Basic Books.
15. Prosperi, M., et al. (2020). "Causal inference and counterfactual prediction in machine learning for healthcare." npj Digital Medicine.
16. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). "Causability and explainability of artificial intelligence in medicine." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312.
17. Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). "What clinicians want: contextualizing explainable machine learning for clinical end use." Proceedings of Machine Learning Research (PMLR), 106: 359–380.

18. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective." BMC Medical Informatics and Decision Making, 20(1), 310.
19. Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020). "Explainable deep learning models in medical image analysis." Journal of Imaging, 6(6), 52.
20. European Commission. (2021). "Ethics guidelines for trustworthy AI." High-Level Expert Group on Artificial Intelligence.