Various Techniques for Explainable AI Evaluation Metrics and Benchmarks for Enhancing Predictive Analytics in Healthcare

Dr. Vikas Goel[1], Dr. Nitesh Pathak[2]
[1] Professor Dept. of CSE(AIML), IMSEC, Ghaziabad, India; [2] Associate Prof. Dept of CSE/IT, BPIT, GGSIPU, New Delhi, India
rvikasgoel@gmail.com, nitishpathak@bpitindia.com

Abstract: The integration of Artificial Intelligence (AI) in healthcare has significantly improved predictive analytics. It is used for disease diagnosis, risk assessment, and clinical decision-making. However, many AI models operate as black-box systems. These are creating challenges related to interpretability, transparency, trust, accountability, and regulatory acceptance. Explainable Artificial Intelligence (XAI) has emerged as a promising approach to address these concerns. XAI provides understandable explanations of model behavior and predictions. Despite growing adoption of XAI techniques in healthcare. The evaluation of explainability remains fragmented. It is due to the absence of standardized metrics, benchmark datasets, and unified assessment frameworks. Existing studies primarily focus on isolated technical metrics. It is neglecting clinical relevance, usability, fairness, and safety considerations. This study presents a systematic literature review. The research analyzes existing parameters of concern like quantitative and qualitative evaluation metrics, benchmark protocols, and clinical validation approaches. They are used in recent healthcare AI studies. Furthermore, the study identifies critical gaps between research-level XAI implementations and real-world clinical deployment requirements.

Keywords: Explainable Artificial Intelligence (XAI), Healthcare Predictive Analytics, Evaluation Metrics, Benchmark Frameworks, Clinical Decision Support Systems

---

Introduction

Healthcare predictive analytics has been revolutionized through the use of Artificial Intelligence (AI). AI supports the automatic diagnosis, risk stratification, and relief planning. There are two models: Machine learning (ML) and deep learning (DL). These models assist a variety of clinical tasks, including risk of disease prediction to intensive care outcome prediction. Opacity and absence of transparency of these AI models limits the clinical practice. The research works indicate that such black-box models tend to be highly predictive. It only goes so far regarding how such predictions are made. It hinders the trustworthiness of clinicians, regulatory obligations, and ethical responsibility. The Explainable Artificial Intelligence (XAI) is now an important research field. It makes sure that the decisions made by AI can be interpreted and justified especially in the high stakes medical setting where human life is directly being affected. [1]

Explainability includes technical transparency (desiring knowledge on how the model operates) and user-friendly interpretability (assuring that the users of the model can put the outputs of the model to a

purposeful use). Researchers have found that lack of explainability can lead to a low adoption of AI recommendations by clinicians even when the predictive precision of the recommendations improves. It is particularly true in such fields as imaging, electronic health records (EHR) analytics, and clinical decision support systems (CDSS), since the understanding of how input features impact risk scores is significant to clinical validation and smart decision-making. [2]

Despite the progress of the XAI methods that include SHapley Additive exPlainations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME) and attention methods, benchmarking structures and standardized measurement metrics that are healthcare-use-specific have not been developed yet. These measurements which consist of fidelity (degree to which explanations are representative of the underlying model), stability (when explanations remain consistent with small data changes) and clinical relevance (whether explanations are useful to the end user) are crucial in maintaining consistent, comparative, and clinically relevant explainability claims. [3]

Explainability is often assessed either qualitatively or through heterogenous metrics, which negatively affects the cross-study comparison and clinical integration, as reported by recent systematic reviews. Measures such as interpretability, task performance, trust calibration, and user satisfaction such as have been reported to be significant but have not been uniformly applied evaluation criteria in healthcare AI research. [4]. Also, the studies in cardiology and other specialties indicate that in most cases, XAI explanations are not systematically evaluated at all, and thus, the need to use structured approaches to benchmarking. [5]

To address these gaps, this thesis examines evaluation metrics and benchmark strategies of explainable AI in healthcare predictive analytics. The current research will attempt to create the paradigm that will facilitate methodological rigor, clinical relevance, and standardization of the XAI assessment by combining both technical and human-related evaluation domains. Hopefully, this contribution may be used to come up with straightforward and reliable AI systems which can be easily used in the implementation of the Healthcare 4.0 environments, thereby improving the application of predictive performance within the actual clinical practice.

Literature Survey

Table 1 presents a systematic literature survey of existing research studies related to Explainable Artificial Intelligence (XAI) in healthcare predictive analytics. The selected studies were analyzed based on the XAI methods used, application domains, evaluation approaches, key findings, and identified limitations. The review includes widely adopted explainability techniques such as SHAP, LIME, and Grad-CAM, along with conceptual and human-centered evaluation frameworks proposed by various researchers. This comparative analysis helps in understanding current evaluation practices, identifying commonly used metrics, and recognizing the existing research gaps in the assessment of XAI systems for healthcare applications.

Table 1: Systematic literature survey of existing research studies related to Explainable Artificial Intelligence (XAI) in healthcare predictive analytics

S.No	Author(s) & Year	XAI Method Used	Application Area	Evaluation Approach	Key Findings	Limitations
1	Lundberg & Lee (2017)	SHAP	General ML / Healthcare (adapted)	Fidelity, feature importance	Provides consistent feature attribution	High computational cost
2	Ribeiro et al. (2016)	LIME	Clinical prediction models	Local interpretability	Explains individual predictions	Instability across runs
3	Selvaraju et al. (2017)	Grad-CAM	Medical imaging	Visual explanation (heatmaps)	Useful for CNN-based diagnosis	Limited to image data
4	Doshi-Velez & Kim (2017)	Conceptual framework	General healthcare AI	Human-centered evaluation	Emphasized need for evaluation standards	No concrete metrics
5	Holzinger et al. (2019)	Interactive XAI	Medical decision systems	Usability, human interaction	Importance of human-in-the-loop	Lack of quantitative metrics
6	Tjoa & Guan (2020)	Survey (multiple XAI)	Healthcare AI	Comparative evaluation	Identified lack of standardization	No unified framework
7	Arrieta et al. (2020)	Explainability taxonomy	General AI	Multi-dimensional metrics	Defined explainability categories	Not healthcare-specific
8	Ahmad et al. (2021)	SHAP, LIME	Disease prediction	Fidelity, interpretability	Improved trust in predictions	Limited clinical validation
9	Stiglic et al. (2020)	XAI tools	Clinical decision support	Usability, trust	Enhances physician understanding	Subjective evaluation
10	Bhatt et al. (2020)	Human-centered XAI	Healthcare	Human evaluation studies	Trust and usability critical	Small sample size
11	Ghassemi et al. (2021)	Critical analysis	Clinical ML	Reliability & bias	Highlighted risks of misleading explanations	Lack of benchmarks
12	Saraswat et al. (2022)	XAI for healthcare 5.0	ICU prediction	Performance + explainability	Improves clinical insight	No standard metrics
13	Nagendran et al. (2020)	AI in healthcare	Clinical deployment	Safety & regulation	Need for regulatory compliance	Limited explainability focus
14	Samek et al. (2021)	Explainability methods	Medical imaging	Robustness, stability	Importance of stable explanations	Hard to generalize

15	Alkhanbouli (2025)	Multiple XAI methods	Healthcare systems	Multi-dimensional evaluation	Identified need for benchmarking frameworks	Lack of unified standards

The literature survey presented in Table demonstrates that a wide range of Explainable Artificial Intelligence (XAI) techniques, such as SHAP, LIME, and Grad-CAM, have been extensively applied in healthcare predictive analytics to improve transparency and interpretability of machine learning models. The reviewed studies indicate that XAI methods are increasingly being utilized in clinical decision support systems, disease prediction, medical imaging, and ICU monitoring applications. Several researchers have emphasized the importance of evaluation metrics such as fidelity, interpretability, consistency, robustness, usability, and trust for assessing the quality of explanations generated by AI systems. Furthermore, human-centered evaluation approaches involving clinicians and domain experts are gaining importance to ensure practical relevance and clinical acceptance of XAI models.

Despite significant progress, the survey reveals several critical limitations in existing research. Most studies focus on isolated evaluation dimensions, primarily technical metrics, while neglecting integrated assessment involving clinical utility, fairness, safety, and regulatory compliance. In addition, there is a lack of standardized benchmarking datasets, unified evaluation frameworks, and reproducible validation protocols across healthcare applications. Many approaches also suffer from limited real-world clinical validation and over-reliance on subjective interpretation methods. These limitations highlight the necessity for a comprehensive multi-dimensional evaluation framework capable of systematically assessing XAI methods from technical, human-centered, clinical, and ethical perspectives to support trustworthy and deployable healthcare AI systems.

Research Gap Identified

We have defined clear inclusion and exclusion criteria to ensure only relevant and high-quality studies were selected. We included research focusing on explainable AI in healthcare predictive analytics with proper evaluation methods. Some studies have been excluded due to lack of explainability, clinical relevance, or methodological rigor. The following gaps are identified from the reviewed literature:

Gap 1: No Standard Evaluation Framework
- Metrics are fragmented
- No unified evaluation pipeline

Gap 2: Lack of Benchmarking Standards
- No common datasets or protocols

Gap 3: Limited Clinical Validation
- Few real-world hospital deployments

Gap 4: Missing Multi-Dimensional Integration
- Studies focus on either technical OR human factors, not both

Gap 5: Weak Regulatory Alignment

- Evaluation rarely considers compliance requirements

Conclusion

The systematic literature review reveals that although Explainable AI has been widely applied in healthcare, there is a lack of standardized, multi-dimensional evaluation frameworks. Existing studies focus on isolated metrics such as fidelity or user trust, while neglecting integrated evaluation across technical, clinical, and regulatory dimensions. Furthermore, the absence of benchmark datasets and limited real-world validation hinder the adoption of XAI systems. This highlights the need for a unified evaluation model to ensure reliable, safe, and clinically meaningful explainability.

References

1. T. Hulsen, Explainable artificial intelligence (XAI): Concepts and challenges in healthcare, AI, vol. 4, no. 3, pp. 652–666, 2023.
2. Roychowdhury S, Lanfranchi V, Mazumdar S. Evaluating explanation performance for clinical decision support systems for non-imaging data: A systematic literature review. Computer Biol Med. 2025 Oct;197(Pt A):110944. doi: 10.1016/j.compbiomed.2025.110944. Epub 2025 Aug 26. PMID: 40857813.
3. Mienye ID, Obaido G, Jere N, Mienye E, Aruleba K, Emmanuel ID, Ogbuokiri B. A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges. Informatics in Medicine Unlocked. 2024 Jan 1;51:101587.
4. Jung J, Lee H, Jung H, Kim H. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. Heliyon. 2023 May 8;9(5):e16110. doi: 10.1016/j.heliyon.2023.e16110. PMID: 37234618; PMCID: PMC10205582.
5. Salih, A.M., Galazzo, I.B., Gkontra, P. et al. A review of evaluation approaches for explainable AI with applications in cardiology. Artificial Intelligence Rev 57, 240, 2024. https://doi.org/10.1007/s10462-024-10852-w.
6. Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems, vol. 30, 2017.
7. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier." In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135-1144, 2016.
8. Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization." In Proceedings of the IEEE international conference on computer vision, pp. 618-626, 2017.
9. Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608, 2017.
10. Holzinger, Andreas, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. "Causability and explainability of artificial intelligence in medicine." Wiley interdisciplinary reviews: data mining and knowledge discovery 9, no. 4 (2019): e1312, 2019.

11. Tjoa, Erico, and Cuntai Guan. "A survey on explainable artificial intelligence (xai): Toward medical xai." IEEE transactions on neural networks and learning systems 32, no. 11, pp. 4793-4813, 2020.
12. Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." Information fusion vol. 58, pp. 82-115, 2020.
13. Ahmad, Muhammad Aurangzeb, Carly Eckert, and Ankur Teredesai. "Interpretable machine learning in healthcare." In Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics, pp. 559-560. 2018.
14. Sarwar, Tabinda, Sattar Seifollahi, Jeffrey Chan, Xiuzhen Zhang, Vural Aksakalli, Irene Hudson, Karin Verspoor, and Lawrence Cavedon. "The secondary use of electronic health records for data mining: data characteristics and challenges." ACM Computing Surveys (CSUR) 55, no. 2, pp. 1-40, 2022.
15. Gu, Jindong, and Daniela Oelke. "Understanding bias in machine learning." arXiv preprint arXiv:1909.01866, 2019.
16. Ghassemi, Marzyeh, Luke Oakden-Rayner, and Andrew L. Beam. "The false hope of current approaches to explainable artificial intelligence in health care." The lancet digital health 3, no. 11, pp. 745-750, 2021.
17. Saraswat, Deepti, Pronaya Bhattacharya, Ashwin Verma, Vivek Kumar Prasad, Sudeep Tanwar, Gulshan Sharma, Pitshou N. Bokoro, and Ravi Sharma. "Explainable AI for healthcare 5.0: opportunities and challenges." IEEE Access vol. 10 , pp. 84486-84517, 2022.
18. Nagendran, Myura, Yang Chen, Christopher A. Lovejoy, Anthony C. Gordon, Matthieu Komorowski, Hugh Harvey, Eric J. Topol, John PA Ioannidis, Gary S. Collins, and Mahiben Maruthappu. "Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies." bmj vol. 368, 2020.
19. Samek, Wojciech, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, eds. Explainable AI: interpreting, explaining and visualizing deep learning. Springer Nature, 2019.
20. Alkhanbouli, Razan, Hour Matar Abdulla Almadhaani, Farah Alhosani, and Mecit Can Emre Simsekler. "The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions." BMC medical informatics and decision making 25, no. 1, pp. 110, 2025.
21. Vani, M. Sree, Rayapati Venkata Sudhakar, A. Mahendar, Sukanya Ledalla, Marepalli Radha, and M. Sunitha. "Personalized health monitoring using explainable AI: bridging trust in predictive healthcare." Scientific Reports 15, no. 1, pp. 31892, 2025.
22. Hossain, Md Imran, Ghada Zamzmi, Peter R. Mouton, Md Sirajus Salekin, Yu Sun, and Dmitry Goldgof. "Explainable AI for medical data: current methods, limitations, and future directions." ACM Computing Surveys 57, no. 6, pp. 1-46, 2025.
23. Agrawal, Renuka, Tawishi Gupta, Shaurya Gupta, Sakshi Chauhan, Prisha Patel, and Safa Hamdare. "Fostering trust and interpretability: integrating explainable AI (XAI) with machine learning for enhanced disease prediction and decision transparency." Diagnostic Pathology 20, no. 1, pp. 105, 2025.