Explainable AI in Healthcare

Prof.Gandhi.R.S[1], Prof.Barvkar.B.Y[2], Srushti B.Lembhe[3], Jyoti D.Gadadare[4]
[1,2,3,4]Dattakala Group of Institution Faculty of Engineering Department of Information Technology Savitribai Phule Pune University

Abstract-Artificial intelligence (AI) with deep learning models has been widely applied in numerous domains, including medical imaging and healthcare tasks. In the medical field, any judgment or decision is fraught with risk. A doctor will carefully judge whether a patient is sick before forming a reasonable explanation based on the patient's symptoms and/or an examination. Therefore, to be a viable and accepted tool, AI needs to mimic human judgment and interpretation skills. Specifically, explainable AI (XAI) aims to explain the information behind the black-box model of deep learning that reveals how the decisions are made. This paper provides a survey of the most recent XAI techniques used in healthcare and related medical imaging applications. We summarize and categorize the XAI types, and highlight the algorithms used to increase interpretability in medical imaging topics. In addition, we focus on the challenging XAI problems in medical applications and provide guidelines to develop better interpretations of deep learning models using XAI concepts in medical image and text analysis. Furthermore, this survey provides future directions to guide developers and researchers for future prospective investigations on clinical topics, particularly on applications with medical imaging.

Keywords- explainable AI; medical imaging; deep learning; radiomics

I.INTRODUCTION

Currently, artificial intelligence, which is widely applied in several domains, can perform well and quickly. This is the result of the continuous development and optimization of machine learning algorithms to solve many problems, including in the healthcare field, making the use of AI in medical imaging one of the most important scientific interests. However, AI based on deep learning algorithms is not transparent, making clinicians uncertain about the signs of diagnosis. The key question then is how one can provide convincing evidence of the responses. However, there exists a gap between AI models and human understanding, currently known as "black-box" transparency. For this reason, many research works focus on simplifying the AI models for better understanding by clinicians, in order to improve confidence in the use of AI models. For example, the Defense Advanced Research Projects Agency (DARPA) of the United States developed the explainable AI (XAI) model in 2015. Later, in 2021, a trust AI project showed that the XAI can be used in interdisciplinary types of application problems, including psychology, statistics, and computer science, and may provide explanations that increase the trust of users.

i. Importance of AI in modern healthcare:
Artificial Intelligence (AI) has emerged as a transformative force in modern healthcare, enabling more accurate, efficient, and personalized medical services. By leveraging vast datasets—from electronic health records (EHRs) to medical imaging and genomics—AI systems can support clinical decision-making, automate routine tasks, and detect patterns that may elude human clinicians. Applications such as early disease detection, predictive analytics, drug discovery, and robotic- assisted surgery are reshaping how care is delivered. Especially during global health crises like the COVID-19 pandemic, AI played a critical role in tracking outbreaks, diagnosing infections, and managing healthcare resources.

II.LITERATURE REVIEW

ii. Systematic Review and Survey:
a. AI-yateem and Li(2026) conducted a systematic review analyzing 112 peer-reviewed articles from PubMed, Scopus, IEEE Xplore, and Web of Science. They categorized studies by application domain (e.g., radiology, pathology), AI model type (e.g., decision trees, deep neural networks), and explanation technique (e.g., SHAP, LIME, attention mechanisms).

The review highlighted SHAP and attention-based models as widely applicable due to their balance between fidelity and usability. The authors proposed a maturity model for human-in-the-loop XAI and emphasized the need for domain-specific interpretability benchmarks and regulatory-compliant XAI system.

b. Aziz et AI(2024) examined the evolution of XAI in Clinical Decision Support Systems (CDSS) by analyzing 68 articles published between 2000 and 2024. They focused on datasets, application areas, machine learning models, explainable AI methods, and evaluation strategies.

iii. Conceptual Frame Works and Terminology

Markus et AI (2020) surveyed the terminology, design choices, and evaluation strategies in creating trustworthy AI for healthcare. They proposed a framework to guide the choice between classes of explainable AI methods, such as explainable modeling versus post-hoc explanation, and model-based, attribution-based, or example-based explanations.

iv. Application and Case Studies

a. Hong et AI (2025) developed an XAI framework to combat medical misinformation and enhance evidence-based healthcare delivery. Their systematic review of 17 studies revealed the urgent need for transparent AI systems in healthcare. The proposed solution demonstrated 95% recall in clinical evidence retrieval and integrated novel trustworthiness classifiers, achieving a 76% F1 score in detecting biomedical misinformation.

b. Lai (2023) Reviewed the use of Vision Transformers (ViT) in medical imagery diagnosis, focusing on their interpretability.

v. Challenges and Limitations

Trade-off Between Accuracy and Interpretability: Many deep learning models offer high accuracy but lack transparency, making it challenging to understand their decision-making process.

Lack of Standardized Evaluation Metrics: There is a need for standardized metrics to evaluate the effectiveness and reliability of XAI methods in healthcare.

Data Bias: Bias in training data can lead to biased AI models, which may affect the fairness and equity of healthcare decisions.

Integration into Clinical Workflows: Incorporating XAI into existing clinical workflows requires addressing technical, organizational, and regulatory challenges.

IV. PROBLEM STATEMENT

Artificial Intelligence (AI) is increasingly being adopted in healthcare for tasks such as diagnosis, prognosis, and treatment planning. However, many of these AI models—especially deep learning systems—operate as "black boxes," offering little to no insight into how they arrive at their decisions. This lack of transparency poses a serious challenge in clinical settings, where understanding the reasoning behind a decision is critical for patient safety, ethical accountability, and regulatory compliance. The absence of explainability not only limits the trust and adoption of AI by healthcare professionals but also increases the risk of biased or erroneous outcomes. Therefore, there is an urgent need to explore and evaluate Explainable AI (XAI) methods that can make AI models more interpretable and trustworthy, without significantly compromising their performance.

V. PROPOSED SYSTEM

This review aims to systematically explore and evaluate current Explainable Artificial Intelligence (XAI) techniques applied in healthcare. The methodology consists of the following steps:

1. literature search strategy

A comprehensive search was conducted across major academic databases including PubMed, IEEE Xplore, ACM Digital Library, Scopus, and Google Scholar. The keywords used included:

- "Explainable AI in healthcare"
- "Interpretable machine learning medical"
- "XAI clinical decision support"
- "SHAP, LIME, Grad-CAM in medicine"
- "Trustworthy AI in health"

2. Inclusion and Exclusion Criteria

- Inclusion: Peer-reviewed journal articles, conference papers, and preprints published between 2016 and 2025 that focus on the application of explainable AI in healthcare settings.

- Exclusion: Non-English articles, papers without a clear focus on explainability, or those unrelated to clinical or biomedical applications.

3. Data Extraction and Categorization

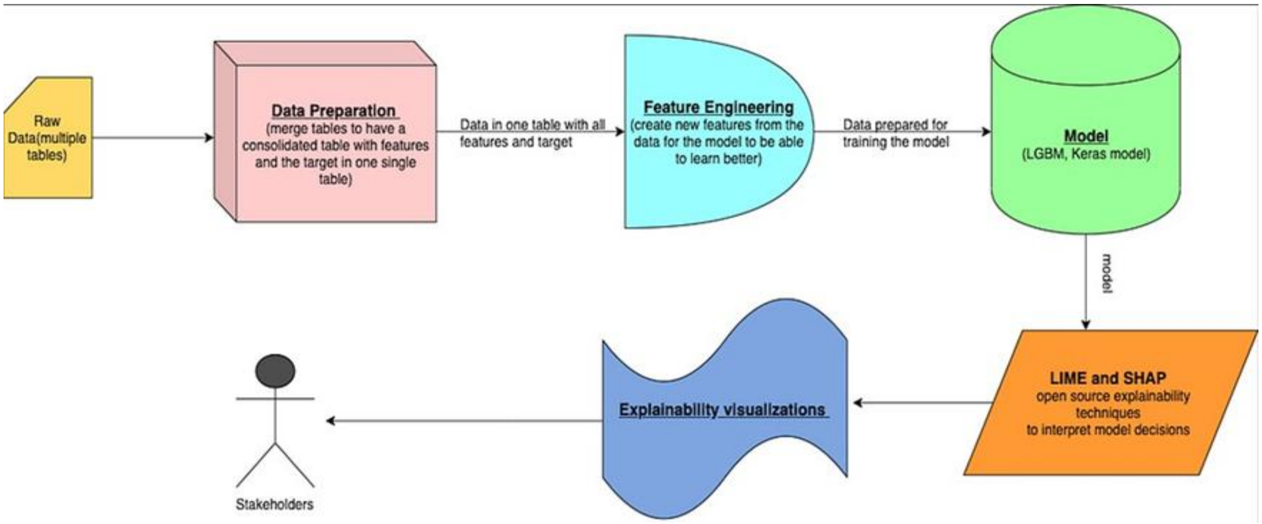
Relevant papers were categorized based on:

- Type of XAI Method (e.g., SHAP, LIME, attention mechanisms)
- AI Model Used (e.g., decision trees, neural networks, ensemble models)
- Healthcare Application (e.g., radiology, cardiology, oncology, EHR-based prediction)
- Type of Explanation (post-hoc vs. intrinsic, visual vs. textual)
- Evaluation Metrics (fidelity, human usability, clinical trustworthiness)

4. Analysis Framework

The selected studies were analyzed using both qualitative and quantitative approaches to identify:

- Most commonly used XAI techniques in healthcare
- Trends in explainability research
- Gaps and limitations in current approaches
- Real-world adoption barriers
- Suggestions for future research directions.


Raw Data(multiple tables)
Data Preparation
(merge tables to have a consolidated table with features and the target in one single table)
Data in one table with all features and target
Feature Engineering
(create new features from the data for the model to be able to learn better)
Data prepared for training the model
Model
(LGBM, Keras model)
model
LIME and SHAP
open source explainability techniques to interpret model decisions
Explainability visualizations
Stakeholders

VI. IMPLEMENTATION

To understand how Explainable AI (XAI) can be implemented in healthcare, consider a typical use case like predicting patient risk using Electronic Health Records (EHR). The following steps outline a simple, general process:

1. Data Collection

Patient data is collected from electronic health records, including lab results, diagnoses, medications, age, gender, and medical history.

2. AI Model Development

A machine learning model (e.g., Random Forest, XGBoost, or Deep Neural Network) is trained on the patient data to predict clinical outcomes—such as the risk of developing diabetes or heart disease.

3. Explainability Layer (XAI)

To make the model's decisions understandable, an XAI technique is applied:

- SHAP (SHapley Additive exPlanations): Shows how each feature (like blood pressure or glucose level) contributes to the prediction.
- LIME (Local Interpretable Model-Agnostic Explanations): Explains a single prediction by creating a simplified, local model.
- Grad-CAM (for medical imaging): Highlights areas in an image that the model focused on when making a decision (e.g., identifying a tumor in an MRI scan).

4. User Interface for Clinicians

The explanations are presented in a simple visual format (e.g., bar graphs, heatmaps, or highlighted images), so doctors can easily interpret why the AI made a certain recommendation or diagnosis.

5. Clinical Use

Doctors use the model's prediction and explanation to make more informed decisions, improve patient trust, and ensure that the AI aligns with medical guidelines and ethical standards.

VII. CONCLUSION

As artificial intelligence continues to transform healthcare, the need for transparency and trust in AI-driven decisions becomes increasingly important. Explainable AI (XAI) addresses this need by making complex models more understandable to clinicians, patients, and stakeholders. This review has highlighted key XAI methods, their applications across various healthcare domains, and the challenges associated with their implementation. While current techniques like SHAP, LIME, and attention-based models have shown promise, limitations such as performance trade- offs, lack of standardized evaluation, and integration into clinical workflows remain significant barriers. For AI to be safely and effectively adopted in healthcare, future research must focus on developing user-friendly, clinically validated, and ethically sound XAI systems.

VIII. FUTURE SCOPE

The future of Explainable AI (XAI) in healthcare is promising and full of potential. As AI systems become more integrated into clinical workflows, there will be a growing demand for tools that are not only accurate but also interpretable and trustworthy. Future research should focus on:

- Developing standardized evaluation metrics to measure the quality and usefulness of AI explanations.
- Designing user-friendly interfaces that present explanations clearly to clinicians and patients.
- Creating domain-specific XAI models tailored for specialties like cardiology, oncology, or radiology.
- Combining multiple explanation methods to improve reliability and completeness.
- Integrating XAI into clinical decision support systems (CDSS) in real time.
- Ensuring ethical and regulatory compliance by aligning XAI systems with healthcare laws and patient privacy requirements.

REFERENCE

[1] 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. https://doi.org/10.1186/s12911-020-01332- 6

[2] Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2020). The role of explainability in creating trustworthy artificial intelligence for healthcare: A comprehensive survey. *arXiv preprint* arXiv:2007.15911. https://arxiv.org/abs/2007.15911

[3] Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). Survey of explainable AI techniques in healthcare. *Sensors, 23*(2), 634. https://doi.org/10.3390/s23020634

[4] Sheu, R.-K., & Pardeshi, M. S. (2022). A survey on medical explainable AI (XAI): Recent progress, explainability approach, human interaction and scoring system. *Sensors, 22*(20), 8068. https://doi.org/10.3390/s22208068

[5] Bhati, D., Neha, F., & Amiruzzaman, M. (2024). A survey on explainable AI (XAI) techniques for visualizing deep learning models in medical imaging. *Journal of Imaging, 10*(10), 239. https://doi.org/10.3390/jimaging10100239

[6] Tulsani, V., Sahatiya, P., Parmar, J., & Parmar, J. (2023). XAI applications in medical imaging: A survey of methods and challenges. *International Journal on Recent and Innovation Trends in Computing and Communication, 11*(9), 181–186. https://ijritcc.org/index.php/ijritcc/article/view/8332

[7] Lai, V. (2023). Interpretability in vision transformers for medical imaging: A systematic

review. *arXiv preprint* arXiv:2309.00252. https://arxiv.org/abs/2309.00252

[8] Hong, S., et al. (2025). Safeguarding patient trust in the age of AI: Tackling health misinformation with explainable AI. *arXiv preprint* arXiv:2509.04052. https://arxiv.org/abs/2509.04052

[9] Khamparia, A., & Gupta, D. (2025). *Explainable artificial intelligence for biomedical and healthcare applications*. CRC Press.

[10] Kataria, A., & Rani, S. (2026). *Explainable AI for healthcare: Real-life applications and use-cases for practitioners*. Routledge.

[11] Al-Yateem, N., & Li, Q. H. (2026). Explainable artificial intelligence (XAI) in healthcare: A systematic review of algorithms, interpretability techniques, and clinical integration strategies. *Innovative Reviews in Engineering and Science, 3*(2), 145–153.

[12] Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (XAI): Towards medical XAI. *IEEE Transactions on Neural Networks and Learning Systems, 32*(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

[13] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD* (pp. 1135– 1144). https://doi.org/10.1145/2939672.2939778

[14] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 30. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28 b67767-Paper.pdf

[15] Rajkomar, A., et al. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine, 1*(18). https://doi.org/10.1038/s41746-018-0029-1