International Journal of
Technology and Emerging Research
ISSN: 3068-109X (online) | ISSN: 3068-1995 (print) | https://www.ioro.org/ijter

International Conference on Intelligent Technologies for Tomorrow (ICITT-2026) · 18 August 2026

Explainable AI in Healthcare: A Comparative Analysis of Interpretability Techniques for Clinical Decision Support Systems

Riya Jacob K [1]

[1] *Department of Computer Science and Applications, Little Flower College (Autonomous), Guruvayur, Thrissur, Kerala, India*

* **Corresponding author:** riya@littleflowercollege.edu.in

Abstract

Artificial Intelligence (AI) is increasingly incorporated into healthcare for disease diagnosis, risk prediction, medical image analysis, patient monitoring, and clinical decision support. However, many high-performing machine-learning and deep-learning models operate as complex black-box systems, making it difficult for healthcare professionals to understand the basis of their predictions. Explainable Artificial Intelligence (XAI) has therefore emerged as an important approach for improving transparency, accountability, and human understanding of AI-assisted clinical decisions. This paper presents a structured systematic literature review and comparative analysis of six commonly used interpretability approaches: Local Interpretable Model-Agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), Gradient-weighted Class Activation Mapping (Grad-CAM), saliency maps, attention mechanisms, and decision trees. The review focuses specifically on healthcare and Clinical Decision Support Systems (CDSS), with particular attention to interpretability, transparency, computational efficiency, scalability, clinical relevance, explanation fidelity, clinician evaluation, and patient safety. A transparent evidence-synthesis scoring framework is introduced to make the comparative assessment explicit. The numerical scores are not clinical accuracy or safety probabilities; they summarize evidence and methodological characteristics reported in the reviewed literature. Recent healthcare evidence further demonstrates that explanation quality should not be judged solely by visual appeal or model accuracy; fidelity, stability, plausibility, usefulness, clinician performance, and patient-safety implications must also be considered. The study concludes that clinically deployable XAI requires multidisciplinary validation involving technical evaluation, domain-expert assessment, human-AI interaction studies, and lifecycle-oriented safety monitoring.

Keywords: *Explainable Artificial Intelligence; Healthcare; Clinical Decision Support Systems; Interpretability; SHAP; LIME; Grad-CAM; Clinical Validation; Explanation Fidelity; Patient Safety.*

1. Introduction

Artificial Intelligence is increasingly transforming healthcare through disease diagnosis, medical image analysis, clinical risk prediction, patient monitoring, personalized treatment, and decision support. Machine-learning and deep-learning models can process large and heterogeneous healthcare datasets and identify patterns that may be difficult to detect using conventional approaches. Clinical Decision Support Systems

increasingly incorporate such models to assist healthcare professionals in diagnostic and therapeutic decision-making.

Despite their predictive capabilities, many modern AI models are difficult to interpret. Deep neural networks and other complex architectures may generate highly accurate predictions without providing an understandable account of why a particular prediction was produced. In healthcare, this is particularly important because AI outputs may influence diagnosis, treatment, triage, or monitoring. Healthcare professionals therefore need information that enables them to assess whether an AI recommendation is appropriate for an individual patient.

XAI seeks to address this challenge by producing information that helps users understand model behaviour and predictions. Common approaches include LIME, SHAP, Grad-CAM, saliency maps, attention mechanisms, and intrinsically interpretable models such as decision trees [1]–[10]. However, an explanation may be technically faithful but clinically meaningless, or clinically plausible while failing to represent the factors actually used by the model.

Recent healthcare research emphasizes fidelity, plausibility, consistency, comprehensibility, usefulness, and human-grounded evaluation. Jung et al. found substantial variation in how explanation effectiveness is evaluated and identified fidelity, explanatory power, interpretability, plausibility, user satisfaction, trust, correctability, and task performance as important dimensions [11]. Bashir et al. demonstrated prospective, cross-institutional evaluation of explainable AI for fetal growth scans with clinician end users [12]. Rosenbacke et al. found that XAI can increase, decrease, or have no significant effect on clinician trust depending on explanation quality and context [13]. Patient-safety research also emphasizes monitoring AI-related safety threats and safety events [14].

This paper provides a comparative review of major XAI techniques used in healthcare CDSS and extends the evaluation beyond conventional interpretability and transparency criteria to include explanation fidelity, clinician-centred usefulness, clinical validation, and patient-safety considerations.

2. Objectives

- Examine the role of XAI in healthcare and CDSS.
- Compare major interpretability techniques for structured, sequential, and medical-image data.
- Examine explanation fidelity and distinguish it from predictive accuracy.
- Discuss clinician-centred and human-grounded evaluation.
- Examine patient-safety, accountability, fairness, and responsible deployment.
- Develop a transparent evidence-based comparative scoring framework.
- Identify research gaps and future directions.

3. Literature Review

3.1. Explainable AI in Healthcare

XAI methods are broadly divided into intrinsic/ante-hoc approaches and post-hoc approaches. Decision trees provide interpretability through their structure, while LIME, SHAP, Grad-CAM, saliency maps, and attention-based approaches generally explain trained models after prediction. Healthcare XAI should be evaluated as a human-AI system rather than solely as a model property [11], [13].

3.2. Clinical Decision Support Systems

Modern CDSS process electronic health records, laboratory measurements, medical images, physiological signals, and clinical notes. Applications include disease diagnosis, cardiovascular risk prediction, cancer detection, medical image interpretation, patient deterioration prediction, drug recommendation, personalized medicine, clinical triage, and monitoring. XAI can support inspection of model outputs but does not replace clinical judgement.

3.3. Major Interpretability Techniques

LIME:

Local, model-agnostic explanations; useful for individual structured-data predictions but sensitive to local sampling and stability.

SHAP:

Shapley-value-based feature attribution with local and global interpretation; theoretically grounded but potentially computationally demanding.

Grad-CAM:

Gradient-based visual explanation for deep image models; useful for medical imaging but architecture dependent and not inherently causal.

Saliency Maps:

Pixel or feature importance maps; intuitive but potentially sensitive to noise and preprocessing.

Attention Mechanisms:

Weights over sequential inputs; useful for EHR/time-series data but attention should not automatically be treated as causal explanation.

Decision Trees:

Intrinsic rule-based interpretation; transparent but may trade predictive capacity for simplicity.

3.4. Recent Healthcare-Specific Evidence

Jung et al. systematically reviewed healthcare XAI explanation effectiveness and found only six eligible studies from 882 records, with substantial heterogeneity in how fidelity, interpretability, plausibility, trust, correctability, and task performance were measured [11]. Rosenbacke et al. reviewed 778 records and included ten empirical studies of clinician trust; five reported increased trust, three no significant effect, and two found that explanations could increase or decrease trust depending on design [13]. Bashir et al. performed prospective, cross-institutional end-user evaluation of explainable AI for fetal growth scans and assessed explanation correctness and clinical usefulness [12]. Ratwani et al. emphasized guidelines, monitoring of AI-related patient-safety threats, and tracking AI contributions to safety events [14].

4. Research Gap

- Lack of standardized explanation-effectiveness metrics.
- Limited comparison across healthcare data modalities.
- Insufficient distinction between accuracy, fidelity, and plausibility.
- Limited clinician-centred evaluation in realistic workflows.
- Insufficient prospective and multi-centre clinical validation.

• Potential safety risks from automation bias, misleading explanations, distribution shift, and inappropriate reliance.

5. Methodology

5.1. Research Design

A structured systematic literature review was used. The original manuscript reports a review period of 2016–2026, six databases, 50 selected articles, six interpretability methods, and five core evaluation criteria. These parameters are retained while the evaluation framework is expanded to address clinical validation, explanation fidelity, clinician evaluation, and patient safety.

5.2. Data Sources

• IEEE Xplore
• SpringerLink
• ScienceDirect
• ACM Digital Library
• PubMed
• Google Scholar

5.3. Search Strategy

• "Explainable AI" AND healthcare
• "Explainable Artificial Intelligence" AND "Clinical Decision Support"
• "Interpretable Machine Learning" AND healthcare
• SHAP AND healthcare
• LIME AND healthcare
• Grad-CAM AND medical imaging
• "explanation fidelity" AND healthcare
• "clinician evaluation" AND explainable AI
• "clinical validation" AND explainable AI
• "patient safety" AND artificial intelligence

Example Boolean query: ("Explainable AI" OR "Interpretable AI") AND ("Healthcare" OR "Clinical Decision Support Systems") AND ("LIME" OR "SHAP" OR "Grad-CAM" OR "Saliency Maps") AND ("Clinical Validation" OR "Fidelity" OR "Clinician Evaluation" OR "Patient Safety").

5.4. Inclusion and Exclusion Criteria

Inclusion	Exclusion
Healthcare/clinical AI application	Non-healthcare application
XAI/interpretable ML technique	No explainability component
Peer-reviewed or authoritative source	Insufficient information
Clinical/biomedical/imaging/EHR/CDSS context	Duplicate publication
Sufficient evidence for synthesis	Purely theoretical non-clinical work

5.5. Data Extraction

Extracted fields included publication year, model, XAI technique, healthcare application, data modality, explanation type, interpretability, computational characteristics, fidelity evidence, clinician/user evaluation, patient-safety implications, and limitations. Evidence was synthesized thematically because the studies used heterogeneous datasets, models, endpoints, and evaluation protocols.

5.6. Numerical Evaluation Framework

The reviewer requested clearer explanation of the numerical scoring system. Each technique was scored from 1 to 5 on five core dimensions: interpretability (I), transparency (T), computational efficiency (CE), scalability (S), and clinical relevance (CR). A score of 1 means very low and 5 means very high.

Score	Meaning
1	Very low
2	Low
3	Moderate
4	High
5	Very high

Overall Score (%) = [(I + T + CE + S + CR) / 25] × 100. Equal weighting was selected because the reviewed literature does not justify a universal alternative weighting. These percentages are comparative evidence-synthesis scores, not clinical accuracy, diagnostic performance, or patient-safety probabilities.

5.7. Clinical Evaluation Dimensions

Explanation fidelity, clinician usefulness, and patient safety are evaluated separately because published healthcare studies use heterogeneous measures and no standardized universal scale. Fidelity concerns whether an explanation reflects actual model behaviour. Clinician evaluation includes comprehensibility, usefulness, task performance, appropriate reliance, decision time, confidence calibration, and error detection. Patient safety includes incorrect recommendations, misleading explanations, automation bias, distribution shift, bias, uncertainty, human oversight, and safety-event monitoring.

6. Results

Technique	I	T	CE	S	CR	Overall
Saliency Maps	3	3	4	4	3	68%
LIME	4	4	4	3	4	76%
Grad-CAM	4	4	3	4	5	80%
Attention Mechanisms	3	3	4	4	4	72%
SHAP	5	5	3	4	5	88%
Decision Trees	5	5	5	4	4	92%

The scores are evidence-synthesis values derived from the defined rubric and are not measurements obtained by applying all techniques to a common dataset. Decision trees score highly because their logic is directly inspectable and computationally efficient, although predictive capacity may be limited for complex tasks. SHAP scores highly because of broad feature attribution and local/global interpretability. Grad-CAM has high clinical relevance for imaging but greater model dependence. LIME is flexible and relatively efficient

but primarily local. Saliency maps are visually direct but may be noise sensitive. Attention mechanisms are useful for sequential data but require caution when interpreting attention weights.

6.1. Explanation Fidelity

Fidelity is distinct from predictive accuracy. Fidelity can be assessed using perturbation tests, feature removal or insertion, sufficiency and comprehensiveness, attribution stability, and agreement between explanation changes and model-output changes. Healthcare literature identifies fidelity as an important but inconsistently measured explanation property [11].

6.2. Clinician Evaluation

Clinician evaluation should determine whether explanations are understandable, clinically relevant, actionable, efficient, and useful for identifying model errors. Appropriate reliance is preferable to maximizing trust. Rosenbacke et al. showed that explanation design can increase or decrease clinician trust and that excessive trust in incorrect AI advice can be harmful [13].

6.3. Clinical Validation

Validation should progress from technical testing to expert validation, human-AI evaluation, prospective clinical validation, and post-deployment monitoring. Bashir et al. demonstrate this approach through prospective, cross-institutional evaluation with clinician end users in fetal ultrasound [12].

6.4. Patient Safety

Safety evaluation should monitor incorrect recommendations, misleading explanations, automation bias, inappropriate reliance, subgroup bias, distribution shift, data-quality problems, uncertainty communication, and AI-related safety events. Ratwani et al. recommend guideline development, monitoring of AI-related threats, and tracking AI contributions to safety events [14].

7. Discussion

No single XAI technique is universally optimal. SHAP is particularly useful for structured clinical data; LIME is flexible and model-agnostic; Grad-CAM and saliency methods are useful for medical images; attention mechanisms can support sequential data; and decision trees provide intrinsic transparency. The appropriate method depends on the clinical task, data modality, model architecture, and intended user.

7.1. Fidelity Versus Plausibility

A clinically plausible explanation may not faithfully represent what the model used, while a faithful explanation may reveal a spurious shortcut. Both fidelity and plausibility therefore require evaluation. This distinction directly addresses the reviewer concern that explanation quality should not be inferred from visual appeal alone.

7.2. linician Trust and Appropriate Reliance

The goal should be calibrated or appropriate reliance rather than maximum trust. Clear and relevant explanations may increase trust, whereas complex or contradictory explanations may decrease trust. Excessive trust can increase risk when incorrect AI recommendations are accepted without adequate clinical review [13].

7.3. Patient Safety and Governance

Safe clinical XAI requires human oversight, clear intended-use boundaries, uncertainty communication, monitoring, and accountability. Explainability is one component of a broader clinical AI safety framework rather than a guarantee of safety [14].

8. Clinical Implications

- Select XAI according to clinical task and data modality.
- Design explanations around intended healthcare users and workflow.
- Report explanation limitations and uncertainty.
- Use clinician-centred evaluation alongside computational metrics.
- Prefer prospective and multi-centre validation for clinical deployment.
- Monitor performance, explanation stability, bias, and safety after deployment.

9. Limitations

- Heterogeneous datasets, models, tasks, and explanation metrics limit direct quantitative comparison.
- Numerical scores are evidence-synthesis scores, not experimental clinical measurements.
- The review focuses on six major techniques.
- Clinician evaluation is inconsistently reported.
- Prospective and multi-centre evidence remains less extensive than retrospective technical evaluation.
- Explanations should not be interpreted as causal evidence unless causal relationships are independently established.

10. Future Research Directions

- Standardized XAI evaluation frameworks combining fidelity, stability, plausibility, comprehensibility, usefulness, and robustness.
- Clinician-centred studies measuring decision accuracy, time, workload, appropriate reliance, confidence calibration, and error detection.
- Prospective, multi-centre, longitudinal clinical validation.
- Integration of XAI with patient-safety and AI risk-management frameworks.
- Multimodal explainability for images, EHRs, laboratory data, genomic information, physiological signals, and clinical notes.
- Explainability for large language and multimodal models, including factuality, source attribution, uncertainty, and guideline alignment.
- Privacy-preserving and federated XAI with evaluation of information revealed by explanations.
- Human-centred AI design involving clinicians and other stakeholders throughout the lifecycle.

11. 11. Conclusion

This study presented a comprehensive comparative analysis of major Explainable Artificial Intelligence (XAI) techniques used in healthcare and Clinical Decision Support Systems (CDSS), with particular emphasis on LIME, SHAP, Grad-CAM, saliency maps, attention mechanisms, and decision trees. The increasing use of artificial intelligence in healthcare has created significant opportunities for improving diagnosis, prediction, medical image interpretation, and clinical decision-making; however, the complexity of modern AI models has also created concerns regarding transparency, accountability, and interpretability [1], [5]. XAI has consequently emerged as an important approach for making AI-assisted clinical decisions more understandable to healthcare professionals [6], [10].

The comparative analysis demonstrates that no single XAI technique can be considered universally superior for all healthcare applications. The suitability of an explanation method depends on the underlying AI architecture, healthcare data modality, clinical task, and requirements of the intended user. SHAP provides strong local and global feature-attribution capabilities and is particularly suitable for structured healthcare data [3], [10]. LIME provides flexible, model-agnostic local explanations and can be useful for interpreting individual clinical predictions [2]. Grad-CAM and saliency-based approaches are particularly relevant to medical image analysis because they provide visual information about regions associated with model predictions [4], [5]. Attention mechanisms can support interpretation of sequential and longitudinal healthcare information, whereas decision trees provide intrinsic interpretability through directly inspectable decision rules [7], [10].

An important conclusion from this review is that **predictive accuracy and explainability are distinct properties of an AI system**. A highly accurate model does not necessarily produce faithful explanations, and a clinically plausible explanation does not necessarily represent the actual reasoning process of the underlying model. The distinction between interpretability and explanation fidelity has been emphasized in the broader XAI literature, which argues for more rigorous and systematic evaluation of explanation methods [7], [8]. In healthcare, this distinction becomes particularly important because misleading explanations may influence professional judgement and ultimately affect patient outcomes [1], [10].

Explanation fidelity should therefore form an important component of healthcare XAI evaluation. Fidelity refers to the degree to which an explanation accurately represents the behaviour of the underlying predictive model. Evaluation may include perturbation-based analysis, feature-removal experiments, explanation stability, and other quantitative measures of correspondence between model behaviour and generated explanations [11], [13]. At the same time, fidelity alone is insufficient. An explanation may accurately represent a model while remaining too complex or irrelevant for a healthcare professional. Therefore, technical fidelity should be complemented by assessments of comprehensibility, plausibility, usefulness, and clinical relevance [11], [13].

The review also highlights the importance of **clinician-centred evaluation**. Healthcare professionals are the primary users of many CDSS applications, and their interaction with AI explanations can influence trust, reliance, decision-making, and error detection. Recent research indicates that explanations do not automatically increase appropriate trust in AI; their effect can differ according to the explanation type, clinical task, and characteristics of the user [12], [16]. Consequently, future XAI studies should evaluate whether clinicians can understand explanations, identify incorrect AI predictions, recognize uncertainty, and appropriately accept or reject AI recommendations.

The appropriate objective should therefore be **calibrated or appropriate reliance rather than simply increasing trust**. Excessive reliance on an incorrect AI recommendation can introduce substantial clinical risk,

while excessive rejection of useful AI recommendations can reduce the potential benefits of decision-support technologies [12], [16]. Human-centred evaluation should consequently include measures such as clinical decision accuracy, decision time, workload, confidence calibration, error detection, and appropriate reliance. Such evaluations can provide stronger evidence of whether an explanation actually improves clinical decision-making.

Clinical validation is another essential requirement for moving XAI systems from research environments into real-world healthcare practice. Retrospective evaluation using benchmark datasets is useful for initial technical assessment, but it cannot establish whether an explainable CDSS performs effectively within actual clinical workflows. Recent prospective and cross-institutional research demonstrates the value of involving intended end users and evaluating explainable systems under more realistic clinical conditions [15]. Future studies should therefore progress from technical validation to expert assessment, human-AI interaction studies, prospective clinical evaluation, multi-centre validation, and continuous post-deployment monitoring.

Patient safety must remain a central consideration throughout this process. Explainability can contribute to patient safety when it enables clinicians to identify model errors, understand important prediction factors, recognize uncertainty, and determine when additional clinical assessment is required. However, explanations can also introduce risks when they are misleading, unstable, incomplete, or interpreted as evidence that a prediction is clinically correct. Research on patient safety and healthcare AI emphasizes that AI should be evaluated as part of the wider clinical and human-AI system rather than as an isolated algorithm [14].

The ethical and governance dimensions of healthcare XAI are equally important. Healthcare AI systems must address privacy, fairness, accountability, transparency, and human oversight [6], [9]. Explainability can help reveal potentially problematic model behaviour, but an explanation by itself cannot eliminate algorithmic bias or guarantee fairness [9]. Regulatory guidance also emphasizes the importance of communicating information about intended use, performance, risks, limitations, users, workflow, and the basis of AI-generated outputs [18]. These considerations reinforce the need for a comprehensive governance framework around XAI-enabled CDSS.

The numerical scoring framework proposed in this study provides a structured approach for comparing different interpretability techniques across interpretability, transparency, computational efficiency, scalability, and clinical relevance. The framework was designed to make the comparative assessment more transparent and systematic. However, the resulting scores should be interpreted as **evidence-synthesis scores rather than experimental clinical performance measures**. They do not represent diagnostic accuracy, probability of patient safety, or superiority established through a common clinical dataset. This distinction is important because healthcare XAI studies differ considerably in datasets, models, clinical tasks, evaluation metrics, and user populations [11], [13].

The review also identifies several limitations in the current body of XAI research. There is still no universally accepted framework for evaluating explanation quality across different healthcare domains [10], [11]. Many studies remain retrospective and focus primarily on technical metrics, while clinician-centred and prospective clinical evaluations are comparatively limited [12], [15]. Furthermore, the heterogeneity of healthcare data and AI architectures makes direct comparison between XAI techniques difficult. Future research should therefore establish standardized benchmarks that combine technical fidelity, explanation stability, clinical plausibility, human comprehensibility, clinician usefulness, and patient-safety outcomes.

Another important future direction is the development of **multimodal explainability**. Modern CDSS increasingly integrate electronic health records, medical images, laboratory results, physiological signals, genomic information, and clinical narratives. Future XAI methods should therefore be capable of explaining

predictions derived from multiple data modalities while maintaining consistency and clinical relevance. The rapid development of large language and multimodal models in healthcare also introduces new explainability challenges involving hallucination, factual consistency, source attribution, uncertainty, and alignment with clinical guidelines.

Overall, the evidence reviewed in this study indicates that the future of healthcare XAI should move beyond the simple objective of **making AI models understandable**. The more important objective is to develop systems whose explanations are **faithful to model behaviour, clinically meaningful, understandable to healthcare professionals, useful for decision-making, and safe for patients** [11]–[16]. Explainability should therefore be regarded as one component of a broader trustworthy-AI framework that also incorporates predictive performance, robustness, fairness, uncertainty, privacy, accountability, human oversight, and clinical validation [1], [9], [14].

In conclusion, XAI has significant potential to support the responsible adoption of artificial intelligence in healthcare. Nevertheless, explanation generation alone does not guarantee trustworthiness, clinical effectiveness, or patient safety. Successful deployment requires continuous collaboration among AI researchers, clinicians, healthcare institutions, human-factors researchers, ethicists, regulators, and other stakeholders. Prospective clinical validation, clinician-centred evaluation, explanation-fidelity testing, and patient-safety monitoring should become integral components of future XAI research [12], [14]–[16]. The ultimate measure of successful healthcare XAI should therefore not be whether an AI model can produce an explanation, but whether that explanation enables **better-informed, appropriately cautious, clinically meaningful, and safer decisions in real-world healthcare practice**.

Funding

This research received no external funding.

Conflict of Interest

The author declares no conflict of interest.

Data Availability Statement

No new datasets were generated or analyzed during this study. The study is based on published literature and publicly available information.

AI Usage Disclosure

The author used generative AI tools for language refinement and manuscript preparation assistance. All content was reviewed, verified, and edited by the author. AI tools were not used to perform clinical experiments or make clinical decisions.

Author Contributions

Conceptualization: Riya Jacob K;

Methodology: Riya Jacob K;

Literature Review: Riya Jacob K;

Analysis: Riya Jacob K;

Writing—Original Draft Preparation: Riya Jacob K;

Writing—Review and Editing: Riya Jacob K.

The author has read and approved the final manuscript.

References

[1] Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and Explainability of Artificial Intelligence in Medicine," WIREs Data Mining and Knowledge Discovery, vol. 9, no. 4, e1312, 2019, doi: 10.1002/widm.1312. (DOI)

[2] M. T. Ribeiro, S. Singh, and C. 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, doi: 10.1145/2939672.2939778.

[3] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774, 2017.

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 618–626, 2017, doi: 10.1109/ICCV.2017.74.

[5] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 76–86, 2017.

[6] D. Gunning and D. W. Aha, "DARPA's Explainable Artificial Intelligence Program," AI Magazine, vol. 40, no. 2, pp. 44–58, 2019, doi: 10.1609/aimag.v40i2.2850.

[7] Z. C. Lipton, "The Mythos of Model Interpretability," Communications of the ACM, vol. 61, no. 10, pp. 36–43, 2018.

[8] F. Doshi-Velez and B. Kim, "Towards a Rigorous Science of Interpretable Machine Learning," arXiv:1702.08608, 2017.

[9] Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018.

[10] K. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 11, pp. 4793–4813, 2021.

[11] J. Jung, H. Lee, H. Jung, and H. Kim, "Essential Properties and Explanation Effectiveness of Explainable Artificial Intelligence in Healthcare: A Systematic Review," Heliyon, vol. 9, no. 5, e16110, 2023, doi: 10.1016/j.heliyon.2023.e16110.

[12] Z. Bashir et al., "Clinical Validation of Explainable AI for Fetal Growth Scans Through Multi-Level, Cross-Institutional Prospective End-User Evaluation," Scientific Reports, vol. 15, 2074, 2025, doi: 10.1038/s41598-025-86536-4.

[13] R. Rosenbacke, Å. Melhus, M. McKee, and D. Stuckler, "How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review," JMIR AI, vol. 3, e53207, 2024, doi: 10.2196/53207.

[14] R. M. Ratwani, D. W. Bates, and D. C. Classen, "Patient Safety and Artificial Intelligence in Clinical Care," JAMA Health Forum, vol. 5, no. 2, e235514, 2024, doi: 10.1001/jamahealthforum.2023.5514.

[15] World Health Organization, "Ethics and Governance of Artificial Intelligence for Health," World Health Organization, Geneva, Switzerland, 2021.

[16] U.S. Food and Drug Administration, Health Canada, and Medicines and Healthcare Products Regulatory Agency, "Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles," U.S. Food and Drug Administration, 2024