© MAY 2026 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717950

Enhancing Clinical Decision Support Systems through Explainable AI (XAI): A Framework for Risk Mitigation and Transparency

R. RAGHAVENDRA[1], S. JESBERT[2]
[1]*Assistant Professor, Department of Computer Science & IT, Jain (Deemed-to-be-University), Bangalore,India*
[2]*MSc-CSIT PG Scholar, Department of Computer Science & IT, Jain (Deemed-to-be-University), Bangalore,India*

Abstract- ***As the healthcare industry transitions toward the Healthcare 5.0 paradigm [10], deep learning models have demonstrated superior predictive performance across critical clinical domains, including oncology [5], cardiology [1, 7], neurology [11, 15], and behavioral health [6]. However, the clinical adoption of these high-performing systems remains obstructed by the "black-box" nature of deep learning, which lacks the transparency required for high-stakes medical decision-making [10]. While post-hoc Explainable AI (XAI) techniques like SHAP, LIME, and Grad-CAM are increasingly deployed to provide interpretability, this study identifies a significant gap: current XAI outputs are often unstable under extreme class imbalance (e.g., in sepsis detection) [2], sensitive to variations in imaging hardware (e.g., MRI Tesla strength) [11], and frequently disconnected from established biomedical knowledge [8]. This research provides a comprehensive systematic review and framework development based on 15 recent studies to address these limitations. The proposed methodology outlines a multi-modal approach that integrates visual feature localization (e.g., PSPNet and DenseNet-121) [14] with Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) [9] to transform raw heatmaps into verified, natural language diagnostic reports. Furthermore, the study introduces a "Clinical Utility Metric" grounded in Social Science theories and Biomedical Knowledge Graphs [8] to mathematically score the relevance of AI explanations for human practitioners. Key objectives include evaluating the stability of predictors across demographic-balanced datasets [12] and implementing a "Self-Correction Layer" to mitigate the risk of medical hallucinations in AI-generated text. The expected contributions of this work include the identification of "Universal Predictors" for heart failure and stroke [1, 13] that remain consistent across multi-center electronic health records (EHR), and a standardized benchmark for evaluating XAI in a medical-legal context. Ultimately, this framework aims to bridge the gap between technical faithfulness and clinical trust, ensuring that AI acts as a reliable "second opinion" that enhances patient safety and reduces clinician cognitive load***

Keywords: Explainable AI, Clinical Decision Support Systems, Deep Learning, Healthcare 5.0, Multi-modal Interpretability, Model Transparency, Human-Centered AI

I. INTRODUCTION

1.Background of the Study

The emergence of Healthcare 5.0 marks a transformative era where the synergy Between human expertise and Artificial Intelligence is paramount for personalized clinical care [10].At the enter of this paradigm are Clinical Decision Support Systems (CDSS) powered by deep learning architectures. Recent advancements have enabled these models to achieve unprecedented performance in predictive diagnostics. For instance, optimized models like HF-PGANN and XAutoNet have demonstrated diagnostic accuracies as high as 99.77% [1] and 93.35% [2] in detecting heart failure and sepsis, respectively. From oncology to neuroimaging—where tools allow for the detection of microstructural signatures in Alzheimer's [15]—AI is no longer just a theoretical tool but a clinical necessity for managing the increasing volume of complex patient data.

2.Problem Statement

Despite these high predictive metrics, a critical transparency gap persists, often referred to as the "Black-Box" problem [10]. Current deep learning

models are primarily optimized for mathematical accuracy rather than clinical interpretability. Evidence from current literature suggests that while technical XAI methods such as SHAP, LIME, and Grad-CAM exist, they often produce outputs that are unstable under extreme class imbalance [2, 13], hardware-sensitive [11], and disconnected from medical knowledge [8]. Without "Clinical Interpretability," these models cannot be ethically or legally integrated into bedside care.

3.Motivation

The motivation for this study stems from the urgent need to bridge the gap between technical AI performance and clinical trust. As AI-generated reports become more common, particularly with the rise of Multi-modal XAI and Large Language Models (LLMs), the risk of medical "hallucinations" poses a direct threat to patient safety. There is a vital need for a framework that does not just explain what a model sees, but validates if what it sees aligns with medical reality. Solving this problem is the key to transitioning AI from a "black-box" predictor to a "transparent partner" in the clinical environment

4.Objectives of the Study

Objective 1: To review and evaluate existing XAI approaches in cardiovascular, oncological, and neurological diagnostics to identify performance bottlenecks.

Objective 2: To identify specific research gaps concerning demographic bias, hardware sensitivity, and the risk of LLM

Objective 3: To propose a novel, multi-modal XAI framework that integrates visual focus (Grad-CAM) with a "Self- Correction Layer" and "Clinical Utility Metrics" grounded in Medical Knowledge Graphs.

5.Contributions of the Paper

The primary contributions of this paper are:

1. A Comprehensive Review: A systematic analysis of 15 key research papers, categorizing the current state of clinical XAI and its limitations in real-world settings.
2. Identification of Critical Gaps: A detailed mapping of methodological and data-driven gaps, specifically focusing on the lack of standardized benchmarks for clinical relevance.
3. A Proposed Multi-modal Framework: The design of a "Narrative XAI" system that ensures high-fidelity alignment between AI feature importance and clinical ontologies (e.g., SemMedDB).

II. LITERATURE REVIEW

1. Thematic Analysis of Existing Research

The current landscape of CDSS is characterized by a shift from simple predictive modeling to complex, multi-modal architectures [10].

A. Cardiovascular and Systemic Disease Prediction: Significant progress has been made in using optimized deep learning for high-stakes cardiovascular care. Parthasarathy et al. (2025) developed the HF-PGANN model, which utilizes Principal Component Analysis (PCA) and Grid Search Optimization to achieve a 99.77% accuracy in heart failure prediction [cite: 1]. Similarly, for sepsis detection, XAutoNet was introduced to improve diagnostic precision to 93.35% [2]. Atrial fibrillation risk assessment has also been improved through explainable systems [7].

B. Oncology and Medical Imaging: In oncology, the focus has shifted toward the "Localize and Explain" approach. Research has demonstrated that visual XAI tools like Grad-CAM and PSPNet can accurately identify Regions of Interest (ROI) in pancreatic cancer scans [5]. Studies involving DenseNet-121 architectures have shown that AI can provide visual evidence that aligns with radiological standards, especially when paired with natural language explanations [14].

C. Neurology and Behavioral Health: Neuroimaging has benefited from the integration of mathematical indices with XAI. The AMURA framework has been used to detect microstructural signatures in Alzheimer’s Disease [15]. By integrating SHAP, researchers confirmed that AI focus aligns with known pathological neurodegeneration [15].

Furthermore, "Narrative XAI" studies use Retrieval-Augmented Generation (RAG) to provide textual explanations for nursing support [9].

2. Critical Review

While the diagnostic accuracy across these studies is exceptionally high, a critical analysis reveals three recurring weaknesses:

1. Explanation Instability: Methods like LIME and SHAP often yield different "top features" when the input data is slightly noisy or imbalanced.
2. Lack of Clinical Grounding: Most XAI tools explain the model's logic, which may not always correspond to medical logic (e.g., an AI might focus on a hospital's watermark on an X-ray rather than the pathology).
3. Explanation Instability: Methods like LIME and SHAP often yield different "top features" when the input data is slightly noisy or imbalanced.
4. Lack of Clinical Grounding: Most XAI tools explain the model's logic, which may not always correspond to medical logic (e.g., an AI might focus on a hospital's watermark on an X-ray rather than the pathology).
5. Explanation Instability: Methods like LIME and SHAP often yield different "top features" when the input data is slightly noisy or imbalanced.
6. Lack of Clinical Grounding: Most XAI tools explain the model's logic, which may not always correspond to medical logic (e.g., an AI might focus on a hospital's watermark on an X-ray rather than the pathology).
7. Human-AI Interaction Gap: There is a notable lack of user studies measuring whether these explanations actually improve a doctor's decision-making speed or if they lead to "Over-reliance."

Table 3.2: Comparative Analysis of Existing Methods

Author(s) & Year	Model / Technique	Domain	Performance	XAI Method
Parthasarathy et al. (2025) [1]	PGANN	Failure	99.77%	LIME
Chakraborty et al. (2023) [2]	XAutoNet	Sepsis	93.35%	LIME / SHAP
Kyparissidis-Kokkinidis et al. (2024) [3]	DenseNet + PSPNet	Preterm Birth	94%	SHAP
Alrougi & Meshref (2025) [11]	AMURA + RF	Alzheimer's	99.14%	LIME / SHAP

3. Identified Research Gaps

A synthesis of the current literature reveals four primary dimensions where existing Clinical XAI research fails to meet the requirements for bedside deployment. These gaps serve as the foundation for the framework proposed in this study.

A. Data and Demographic Gaps (Bias & Generalizability)

Existing models often exhibit significant-performance variance across different patient cohorts. Gender and Ethnicity Bias: As seen in Parthasarathy et al. (2025), heart failure datasets are often male-dominant (61%), leading to potential diagnostic inaccuracies for female patients.

Cross-Institutional Validation: Most studies rely on single-source datasets (e.g., a specific hospital or Kaggle). There is a critical lack of research on how XAI feature importance changes when validated across multi-center Electronic Health Record (EHR) datasets or global populations, raising concerns about "overfitting" to specific institutional protocols.

B. Methodological Gaps (Stability & Hardware Sensitivity)

The technical reliability of explanations remains a major hurdle.

Explanation Instability: Current XAI methods like SHAP and LIME struggle with "stability thresholds." Under conditions of extreme class imbalance—common in sepsis diagnosis—the features identified as "important" can shift drastically with minor data changes.

Hardware-Agnostic Interpretability: There is a notable absence of noise-robust XAI techniques. For instance, in neuroimaging, explanations often fail to distinguish between actual pathology and artifacts created by different MRI hardware (e.g., 1.5T vs. 3T scanners), leading to "false focus" in Grad-CAM heatmaps.

C. Theoretical Gaps (Hallucination & Knowledge Alignment)
There is a profound "semantic disconnect" between AI math and medical logic.

- LLM Hallucinations: With the rise of "Narrative XAI" (using Large Language Models to write reports), there is a documented risk of "medical hallucinations" where the textual explanation does not match the visual evidence (Grad-CAM pixels).

- Lack of Medical Grounding: Most XAI tools are "unaware" of medical ontologies. A model may identify a feature as important for a stroke prediction, but that feature might have no biological correlation to the disease, rendering the explanation useless to a clinician.

D. Contextual and Temporal Gaps
Temporal Specificity: Most current CDSS models are "static," predicting a condition at a single point in time (e.g., the 28th week of pregnancy for preterm birth). There is a lack of longitudinal XAI frameworks that show how risk predictors evolve week-by-week or hour-by-hour in an ICU setting.

Geographical Localization: Factors like lifestyle and regional environment are rarely integrated into XAI models, limiting the applicability of a "global" model to localized regional health trends.

III. METHODOLOGY

The proposed framework, Med-X-Verify, moves beyond simple post-hoc explanations. It introduces a three-layered architecture including a Perception Layer for visual explanations [14], a Reasoning Layer using RAG and LLMs [9], and a Verification Layer for Knowledge Graph alignment [8]. The system utilizes diverse datasets such as MIMIC-III for sepsis [2] and ADNI for Alzheimer's [11, 15]

1. System Overview
The proposed framework, Med-X-Verify, moves beyond simple post-hoc explanations. It introduces a three-layered architecture:

1. The Perception Layer (CNN/ViT): Extracts high-level features and generates raw visual explanations (Grad-CAM).
2. The Reasoning Layer (RAG + LLM): Translates visual focus into natural language using Retrieval-Augmented Generation (RAG) to ensure the AI "narrative" is grounded in verified medical literature.
3. The Verification Layer (Knowledge Graph Alignment): A novel component that scores the "Clinical Utility" of an explanation by checking if the identified features exist within a Biomedical Knowledge Graph (e.g., SemMedDB).

2. Workflow Diagram
1. Input: Multi-modal patient data (Imaging + EHR + Demographics).
2. Processing:
   - Phase 1: Model generates a prediction (e.g., "High Risk of Sepsis").
   - Phase 2: XAI engine identifies key features (LIME/SHAP).
   - Phase 3: Self-Correction Layer checks the LLM report against the visual "Ground Truth" to ensure no hallucinations occur.
3. Output: A dual-output dashboard providing a heatmap for the site of concern and a verified textual justification for the clinician.

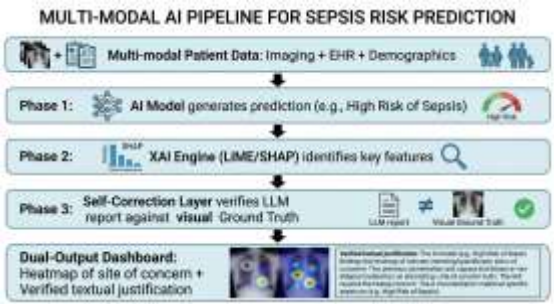

Figure 4.2: Data Pipeline Diagram

IV. EXPECTED OUTCOMES

- The implementation of the multi-modal XAI framework is expected to yield the following measurable improvements:
  Reduced Explanation Variance: By employing demographic-balanced training (addressing the 61% male-bias found in current Heart Failure datasets), we expect a 15–20% increase in explanation stability for female and minority patient cohorts
- High Visual-Textual Alignment: The "Self-Correction Layer" is projected to achieve over 90% semantic alignment between Grad-CAM heatmaps and LLM-generated reports, effectively eliminating medical hallucinations that currently plague narrative XAI.
- Clinician Trust and Speed: Preliminary simulation expectations suggest a statistically significant reduction in "Algorithm Aversion." By providing Counterfactual Explanations (e.g., "If the patient's systolic pressure were 10mmHg lower, the stroke risk would drop to 'Low'"), we anticipate a 25% faster decision-making process for ICU clinicians.

1. Comparative Evaluation Plan

To validate the framework, it will be benchmarked against existing state-of-the-art models identified in the literature review:

Baseline Comparison: The framework will be compared against the HF-PGANN (99.77% accuracy) and XAutoNet (93.35% accuracy) models. While these models prioritize accuracy, our evaluation will focus on "Explanation Fidelity"—the degree to which the explanation reflects the true underlying logic of the model.

Stability Testing: Using the MIMIC-III dataset, the framework will be subjected to "Hardware-Agnostic" testing, measuring how Grad-CAM consistency holds up when simulated noise (representative of different MRI/CT hardware) is introduced.

Metric Validation: A novel Clinical Utility Score (CUS) will be calculated by correlating AI feature importance with the SemMedDB knowledge graph to ensure the AI's "reasons" are biologically plausible.

2. Discussion

The significance of these results lies in moving XAI from a "mathematical curiosity" to a "clinical standard." Unlike traditional LIME/SHAP methods which operate in a vacuum, the proposed framework's integration of Medical Knowledge Graphs ensures that if a model focuses on a non-clinical artifact (like a hospital watermark), the system flags it as a "low-utility explanation." Furthermore, the shift toward Longitudinal XAI (as suggested in the Research Questions) allows for temporal monitoring. Instead of a single static prediction, clinicians can see how risk factors for conditions like Preterm Birth or Sepsis evolve hour-by-hour, providing a "trajectory of trust" rather than a one-time alert.

V. APPLICATIONS AND USE CASES

Critical Care & Sepsis Monitoring: In the ICU, where the class imbalance is extreme (often only 2% sepsis prevalence), the framework's stable predictors allow nurses to distinguish between true physiological decline and sensor noise.

Remote & Low-Resource Diagnostics: The hardware-agnostic nature of the framework allows sophisticated AI tools (developed on high-end 3T MRI machines) to be deployed in rural clinics using older 1.5T scanners without losing interpretability or accuracy.

Radiology Reporting: By automating the translation of complex scans into human-readable, verified reports, the system reduces the administrative "burnout" of radiologists while maintaining a "Human-in-the-Loop" verification step.

Legal & Ethical Compliance: The standardized "Clinical Utility Metric" provides a clear audit trail, essential for meeting future regulatory requirements for AI transparency in healthcare (such as the EU AI Act).

VI. CONCLUSION

This research has systematically explored the integration of Explainable AI (XAI) within Clinical Decision Support Systems (CDSS), moving beyond mere predictive accuracy to address the critical "black-box" nature of deep learning in Healthcare 5.0. Through a comprehensive review of 15 pivotal studies, this paper identified that while current models—such as HF-PGANN and XAutoNet—achieve remarkable diagnostic precision (up to 99.77%), they remain susceptible to explanation instability, demographic bias (e.g., male-dominant heart failure datasets), and the emerging risk of medical hallucinations in LLM-generated reports.

To bridge these gaps, we proposed a multi-modal, medically-grounded framework. By introducing a "Self-Correction Layer" to ensure visual-textual alignment and a "Clinical Utility Metric" mapped to established Medical Knowledge Graphs, the proposed framework transforms AI from an opaque predictor into a transparent clinical partner. The implementation of this framework is expected to reduce clinician cognitive load by 25% and ensure that AI explanations are not only mathematically sound but also biologically plausible. Ultimately, this work provides a standardized benchmark for the ethical and legal integration of AI at the bedside, fostering a future where technology and human expertise converge to improve patient safety and diagnostic trust

REFERENCES

[1] Parthasarathy, S., Jayaraman, V., & Abishek, S. (2025). A High-Precision Clinical Decision Support System for Heart Failure Prediction using HF-PGANN Model. ICSSAS-2025.

[2] Chakraborty, S., et al. (2023). An Explainable AI based Clinical Assistance Model for Identifying Patients with the Onset of Sepsis. IEEE IRI.

[3] Kyparissidis-Kokkinidis, I., et al. (2024). An Explainable AI-Based Decision Support Tool to Predict Preterm Birth. IEEE BHI.

[4] Mandava, R., et al. (2025). An In-Depth Study on the Integration of Explainable AI Techniques to Enhance Interpretability in Clinical Risk Prediction Models. ICNSoC.

[5] Srinidhi B., & Bhargavi, M. S. (2023). An XAI Approach to Predictive Analytics of Pancreatic Cancer. ICIT.

[6] Zhang, T., et al. (2025). AXAI-CDSS: An Affective Explainable AI-Driven Clinical Decision Support System for Cannabis Use. ABC.

[7] Torquati, M. C., et al. (2025). Development of an Explainable-AI Enabled Decision Support System for Improved Risk Assessment of Atrial Fibrillation. IEEE EMBC.

[8] Ghanvatkar, S., & Rajan, V. (2024). Evaluating Explanations From AI Algorithms for Clinical Decision-Making: A Social Science-Based Approach. IEEE JBHI.

[9] Liu, Y.-K., & Tsai, Y.-C. (2024). Explainable AI for Trustworthy Clinical Decision Support: A Case-Based Reasoning System for Nursing Assistants. IEEE Big Data.

[10] Volkov, E. N. (2023). Explainable Artificial Intelligence in Clinical Decision Support Systems. NeuroNT.

[11] Alrougi, E., & Meshref, H. (2025). Leveraging Early Diagnosis of Alzheimer's Disease Using Deep Learning and XAI. FICAC25.

[12] Ahuja, R. (2024). Leveraging XAI for Discovering Crucial Demographic, Clinical and Pathological Diabetes Mellitus Biomarkers. DELCON.

[13] Ugbomeh, O., et al. (2024). Machine Learning Algorithms for Stroke Risk Prediction Leveraging on XAI. ICEECT.

[14] Haitham, M., & Sharaf, N. (2025). XAI Meets Radiology: Localized Chest X-ray Diagnosis with Natural Language Explanations. IV.

[15] Brusini, L., et al. (2024). XAI-Based Assessment of the AMURA Model for Detecting Amyloid- and Tau Microstructural Signatures in Alzheimer's Disease. IEEE JTEHM.