Journal of Theoretical and Applied Information Technology
31st August 2026. Vol.104. No.16
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CLINXAI-NET: A RULE-GUIDED HYBRID CNN–TRANSFORMER EXPLAINABLE AI FRAMEWORK FOR ROBUST REAL-TIME DIAGNOSIS IN HIGH-RESOLUTION MEDICAL IMAGING

NAGA NIRMALA KONGARA, DR. BHUKYA KRISHNA, DR. B. SUJATHA

Research Scholar, Department of CSE, Gandhi Institute of Engineering and Technology University, Gunupur Odisha, India.
Professor, Department of CSE, Gandhi Institute of Engineering and Technology University, Gunupur, Odisha, India.
Professor, Godavari Institute of Engineering & Technology, NH-5, Chaitanya Knowledge City, Rajahmundry Andhra Pradesh, India

ABSTRACT

High-resolution medical imaging supports increasingly precise diagnosis, but recent CNN, Transformer, and hybrid pipelines often optimize predictive discrimination separately from clinical reasoning, explanation faithfulness, calibration, and deployment cost. This fragmentation limits trustworthy use across modalities and motivates the present study. ClinXAI-Net is proposed as a rule-guided hybrid CNN-Transformer framework that combines lesion-sensitive local convolutional features with global anatomical Transformer representations through adaptive gated fusion. Clinical IF-THEN consistency constraints and an explanation-alignment objective are incorporated into training and decision refinement, while Grad-CAM, SHAP, LIME, and textual rule traces provide complementary evidence. Five public datasets covering MRI, CT, PET, and X-ray were evaluated with patient-wise splits, five-fold validation, discrimination metrics, expected calibration error, quantitative explanation criteria, and deployment measures. The framework achieved 99.12% test accuracy, 98.82% F1-score, and 0.994 ROC-AUC. In matched ablation analysis, the full model improved accuracy by 2.20 percentage points and explanation agreement by 0.100 over CNN-Transformer fusion alone. INT8 quantization reduced model size by 40.6%, inference latency by 29.2%, and peak memory by 23.5%, with only a 0.06 percentage-point reduction in accuracy relative to the non-quantized full model. The contribution is therefore not merely another hybrid backbone; the study provides evidence that jointly optimizing predictive, rule-consistency, explanation, calibration, and deployment objectives can improve diagnostic performance while preserving auditable reasoning. The findings support a more clinically accountable design pattern for medical-image decision support, while prospective multi-center validation remains necessary before clinical adoption.

Keywords: *Explainable Medical Imaging, CNN-Transformer Fusion, Clinical Rule Reasoning,Explanation Alignment, Real-Time Diagnosis*

1. INTRODUCTION

1.1 Background

Medical imaging is a key element in the diagnosis and treatment of modern medicine, which allows the non-invasive visualization of the internal anatomy, the progression of disease, the abnormal structures of the tissues and the response to therapy. Brain tumor, lung nodule, thoracic diseases, neurodegenerative diseases, pneumothorax are commonly diagnosed with imaging modalities including magnetic resonance imaging, computed tomography, positron emission tomography, and digital radiography. Clinicians can now see the finer points of diagnosis, which couldn't be seen before with high-resolution imaging. But the higher the resolution, the more complicated the image, the higher the processing requirement, and the more work the image interpretation requires.

Medical images classification, detection and segmentation have improved significantly with deep learning. Local patterns in images, including lesion margins, nodules, edges, opacities and abnormal tissue textures can be extracted effectively using convolutional neural networks [1, 2]. The power of global anatomical relationships and long-range dependencies have been enhanced

in image understanding using Vision Transformers [3] [4]. But, the predictive performance is not enough for medical applications. A diagnostic AI model should be able to give easily understood explanations, confidence metrics, and clinically relevant and valid evidence.

Explainable artificial intelligence (XAI) has become important for building trust in medical-image decision support because clinicians require evidence that can be related to the image and to clinical reasoning. Grad-CAM identifies class-discriminative image regions, SHAP estimates feature contributions, and LIME approximates a prediction locally [5]-[7]. However, many XAI pipelines remain post-hoc: they explain an already generated prediction and therefore do not necessarily constrain the model to use clinically relevant evidence or produce explanations faithful to its internal decision process.

1.2 Problem Statement

Recent medical-imaging systems demonstrate strong task-level performance, yet the literature still separates several requirements that must coexist in a clinically credible diagnostic system. CNNs are effective at local lesion and texture representation but have limited direct modeling of long-range anatomical relations; Transformers improve global context but can increase computational cost; lightweight designs emphasize efficiency; and many explainability methods remain post-hoc rather than part of the learning objective [11], [14], [22]-[24]. Late-2025 studies further show active progress in CNN-Transformer fusion, local-global segmentation, lightweight classification, and transformer-aided lesion classification [31]-[34], but these studies remain predominantly task- or modality-specific and do not jointly evaluate clinical-rule consistency, explanation alignment, probability calibration, and deployability. The resulting research problem is therefore not simply how to improve classification accuracy, but how to construct and evaluate a unified high-resolution medical-imaging framework in which local-global representation, clinically constrained reasoning, faithful explanation, calibrated confidence, and real-time operation are optimized together and tested under a common protocol.

Accordingly, the study asks whether these requirements can be integrated without sacrificing predictive or deployment performance.

The central question is: can a single multi-modality diagnostic framework jointly improve discrimination, clinical-rule consistency, explanation quality, calibration, and inference efficiency relative to its component and fusion-only baselines?

1.3 Research Questions

The literature-derived problem is operationalized through four research questions:

RQ1: Does adaptive CNN-Transformer gated fusion provide measurable diagnostic benefit over CNN-only, Transformer-only, and unguided fusion baselines under the same evaluation protocol?

RQ2: Do clinical-rule refinement and explanation-alignment training improve clinical consistency and quantitative explanation agreement beyond representation fusion alone?

RQ3: Can INT8 deployment reduce model size, memory use, and inference latency while retaining practically unchanged diagnostic accuracy?

RQ4: Are the reported improvements stable across folds and statistically distinguishable from the matched CNN-Transformer baseline?

1.4 Research Gap and Knowledge-Creation Need

The critique above shows a gap at the intersection of representation, reasoning, explanation, calibration, and deployment. Existing work demonstrates the value of individual ingredients, but evidence remains limited on whether they can be jointly optimized and whether each ingredient contributes independently under matched experiments. This matters for knowledge creation because a high diagnostic score alone does not establish that a system is clinically coherent, well calibrated, explainable by multiple complementary criteria, or deployable within realistic latency and memory limits. The present study therefore treats integrated evaluation, rather than architectural hybridization by itself, as the main knowledge gap.

The literature critique identifies five specific gaps that this study tests:

1) Limited matched evidence on adaptive local-global fusion for high-resolution diagnostic classification across multiple imaging modalities.

2) Limited integration of explicit clinical-rule consistency with learned visual representations during decision refinement.

3) Limited use of training-time explanation alignment together with quantitative tests of localization, attribution stability, and local fidelity.

4) Insufficient joint reporting of discrimination, calibration, model size, memory, and inference latency for hybrid diagnostic models.

5) Continued dependence on single-disease or single-modality evidence, with prospective hospital, subgroup, and workflow validation still largely unresolved.

1.5 Motivation

The motivation is to move from a black-box accuracy objective toward an auditable decision-support design in which a clinician can inspect the predicted class, confidence, abnormal-region evidence, feature contribution, local surrogate explanation, and activated rule trace. Such evidence does not by itself prove clinical safety, but it creates a more testable basis for error analysis, calibration review, expert audit, and future prospective validation than a diagnostic label alone.

1.6 Major Contributions and New Knowledge

The contribution is framed as testable knowledge created by the combined design and evaluation, rather than by claiming that CNNs, Transformers, rules, or post-hoc XAI are individually new:

1) A matched local-global representation study showing the incremental value of adaptive gated CNN-Transformer fusion over the individual branches.
2) A rule-consistency refinement mechanism that makes clinical constraints part of the decision pathway and can be isolated in ablation analysis.
3) An explanation-alignment objective that connects training to clinically annotated abnormal regions, rather than treating explanation only as a post-hoc visualization.
4) A multi-criterion explanation evaluation combining Grad-CAM alignment, SHAP stability, LIME fidelity, and human-readable rule traces.
5) A joint predictive-reliability-deployment evaluation that reports discrimination, calibration, statistical stability, model size, memory use, and latency under the same framework.
6) An INT8 deployment analysis that quantifies the accuracy-efficiency trade-off instead of reporting only the uncompressed model.
7) Multi-dataset evaluation across MRI, CT, PET, and X-ray that tests whether the integrated design is reusable beyond a single imaging modality while explicitly retaining prospective clinical validation as an open issue.

1.7 Organization of the Paper

Section 2 critically reviews CNN, Transformer, lightweight, and explainable medical-imaging studies and positions the late-2025 literature. Section 3 formulates the diagnostic and multi-objective learning problem. Section 4 describes ClinXAI-Net, including adaptive fusion, rule-guided refinement, explanation alignment, and the diagnostic workflow. Section 5 defines the datasets, patient-wise splitting, tools, hardware, and training settings. Section 6 reports diagnostic, ablation, explanation, statistical, and computational results and interprets them against the research questions and evaluation criteria. Section 7 states the conclusions and open research issues, followed by author contributions and references.

2. LITERATURE REVIEW

2.1 Medical Image Analysis with CNN

CNN-based architectures have contributed to the basic field of medical image analysis. Later, U-Net adopted a skip connection architecture with an encoder–decoder structure, which has gained popularity as a method for biomedical segmentation [1]. nnU-Net additionally enhanced the adaptability of segmentation by automatically setting up the preprocessing, network design and training strategies for various biomedical datasets [2]. The methods presented in this section are useful for segmentation but do not directly influence the diagnostic explainability or rule-based clinical reasoning.

CNNs are excellent at capturing local image features like lesion borders, abnormal texture, nodules and tissue irregularities. However, they may be less effective at modelling long-range anatomical relationships due to their receptive field limitations. It is more critical in medical images with high resolution, where the global spatial context may play a role affecting diagnosis.

2.2 Transformer-Based Medical Imaging Models

The concept of patch-based self-attention for image recognition was introduced in Vision Transformers and it was shown that global relationships can be learned without solely relying on convolution [3]. To achieve the efficient implementation of self-attention, Swin Transformer introduced shifted windows and hierarchical feature representation [4]. These models can be used to capture global anatomical relationships, particularly in high-resolution medical images.

Wang et al. introduced a self-distilling TransUNet for 3D medical image segmentation, which fuses Transformer-based global semantic learning and CNN-based spatial detail preservation [14]. In the field of DR, Yang et al. proposed a multiple-instance learning model based on the Transformer architecture for the classification of DR severity, highlighting the promise of Transformers in high-resolution retinal imaging [19]. But, Transformer models are complex enough and sometimes need optimization for real-time deployment.

These limitations motivate methods that retain high-resolution detail while reducing computational cost.

2.3 High-Resolution and Lightweight Medical Imaging Methods

Medical images captured with a high resolution contain clinically relevant details that are expensive to process. Qiu et al. presented shared feature learning for medical image segmentation, the problem of high-resolution input cost, in a dual-stream shared feature framework [11]. Dong et al. proposed a sliding-window self-supervised learning method for high-resolution anomaly detection that demonstrates that the details can be retained in processing at the patch level while enhancing abnormality detection [13]. To reduce the complexity of segmentation, Feng et al. proposed a lightweight U-Net variant named MLU-Net, which combines frequency representation and MLP-based method [15].

These methods are mainly segmentation or anomaly methods with primary focus on resolution and efficiency. They are not fully integrated with clinical rule reasoning, multi-format explanations, and real time diagnostic decision support.

2.4 Explainable AI in Medical Imaging

The concept of explainability is crucial for clinical use of AI, as it requires clinicians to have insight into the evidence that supports the AI model's predictions. Grad-CAM is a method that creates class-discriminative heatmaps using gradients from convolutional layers [5]. The feature-attribution scores from SHAP are based on cooperative game theory [6]. LIME is based on fitting interpretable local surrogate models [7] to explain individual predictions.

Recent reviews highlight the need for clinically meaningful, stable and faithful clinical interpretation of model decisions for XAI in medical imaging [23, 24]. Many of the existing approaches, however, provide explanations only after prediction. These explanations can be after-the-fact and potentially not relevant to the model learning, and may not always be focused on medically relevant regions.

2.5 Hybrid CNN–Transformer and XAI Based Frameworks

Recently, hybrid CNN–Transformer architectures have been of interest due to their ability to process both local and global features. Zeineldin et al. suggested a hybrid Vision Transformer and CNN model with TransXAI to segment glioma from brain MRI with explanations [21]. They showed the benefit of their model by combining local and global representations and providing a visual explanation. Alhumaid and Fayoumi suggested a hybrid CNN–Swin Transformer with explainable artificial intelligence for the detection of abnormality in the maxilla sinus from computed tomography (CT) images [22].

These studies are valuable, but they remain primarily modality- or disease-specific, emphasize visual post-hoc explanation, and do not jointly enforce clinical-rule consistency during diagnostic decision making. Late-2025 work strengthens this critique. Jayaraman et al. [31] proposed cross-attention CNN-ViT fusion for brain-tumor MRI classification; Ge et al. [32] integrated CNN and Transformer representations for local-global medical-image segmentation; Ferdous et al. [33] emphasized lightweight attention-augmented medical-image classification; and Najjar et al. [34,35,36] used transformer-aided VGG19 feature encoding for skin-cancer classification. Collectively, these studies confirm that local-global modeling and computational efficiency are current design priorities, but they remain task-specific and do not jointly test rule-guided reasoning, training-time explanation alignment, calibration, and cross-modality deployment. ClinXAI-Net [37] is therefore positioned as an integrated evaluation of these dimensions rather than as a claim that hybridization itself is new.

2.6 Research Gap Summary

The literature critique, including late-2025 work, yields the following unresolved limitations:

CNN-based methods remain strong for local lesion detail, but alone they do not explicitly model long-range anatomical context.

Transformer-based and hybrid models improve global context, but their computational burden and task-specific design can limit real-time and cross-modality use [31], [32].

Lightweight models improve efficiency, yet efficiency-oriented designs do not by themselves provide rule-based clinical accountability or multi-format explanations [33].

XAI is increasingly reported, but explanations are often generated after prediction and are not necessarily optimized for clinical alignment or quantitatively assessed for stability and fidelity [22], [34].

Hybrid CNN-Transformer models improve representation, but clinical-rule consistency and training-time explanation alignment are rarely evaluated together within the same matched ablation framework.

Most evidence remains focused on one disease, one modality, or one task, leaving multi-modality robustness, prospective hospital validation,

subgroup calibration, and workflow impact insufficiently resolved.

ClinXAI-Net addresses the testable portion of these gaps by combining adaptive local-global fusion, clinical-rule refinement, explanation-aware optimization, calibration, and quantized deployment across several public MRI, CT, PET, and X-ray datasets. The study does not treat public-dataset performance as a substitute for prospective clinical evidence.

2.7 Literature Synthesis and Knowledge-Creation Positioning

The synthesis in Table 1 shows that the individual ingredients of ClinXAI-Net have precedents in the literature, whereas the knowledge contribution evaluated here is their joint optimization and matched measurement. This distinction is important: the work is incremental at the component level but more substantive at the systems-evaluation level because it tests whether predictive discrimination, rule consistency, explanation quality, calibration, and deployment efficiency can improve together rather than in isolation.

Accordingly, superiority claims are restricted to the matched experiments reported in this paper. Cross-paper results are used for contextual positioning because published studies differ in datasets, endpoints, preprocessing, class balance, and validation protocols; they should not be interpreted as a controlled leaderboard.

Table 1: Literature Gap and Positioning of ClinXAI-Net

Study	Year	Core Method	Strength	Limitation	How ClinXAI-Net Improves
Ronneberger et al. [1]	2015	U-Net	Strong biomedical segmentation	No diagnostic explainability	Adds classification, XAI, and rule reasoning
Isensee et al. [2]	2021	nnU-Net	Self-configuring segmentation	Segmentation-focused, limited interpretability	Adds multi-modality diagnosis and explanation-aware learning
Qiu et al. [11]	2024	Dual-stream shared feature learning	Improves high-resolution segmentation	High computational complexity; no clinical rule layer	Adds lightweight fusion, rule refinement, and real-time deployment
Dong et al. [13]	2023	Sliding-window self-supervised anomaly detection	Preserves high-resolution details	Increased training complexity; limited interpretability	Adds supervised diagnostic fusion and multi-level XAI
Wang et al. [14]	2023	Self-distilling TransUNet	Combines global and local segmentation features	High 3D computational cost	Uses lightweight CNN–Transformer fusion and quantization
Feng et al. [15]	2024	Lightweight U-Net	Reduced segmentation complexity	Task-specific and not rule-guided	Adds cross-modality diagnosis and clinical reasoning
Zeineldin et al. [21]	2024	Hybrid CNN–ViT with XAI	Local–global glioma segmentation explanation	Mainly MRI glioma-specific; post-hoc explanation	Extends to MRI, CT, PET, and X-ray with rule-guided explanations
Alhumaid and Fayoumi [22]	2025	CNN–Swin Transformer with XAI	CT abnormality diagnosis with explainability	Disease-specific; mainly visual explanation	Adds multi-dataset validation, SHAP/LIME/rule traces, and calibration
Bhati et al. [23]	2024	XAI survey	Reviews XAI visualization methods	Does not propose a diagnostic model	Implements integrated XAI within model training
Alkhanbouli et al. [24]	2025	XAI disease prediction review	Identifies future needs for trustworthy AI	Review-only; no architecture	Proposes a practical hybrid model with measurable performance
Jayaraman et al. [31]	2025	CNN-ViT with cross-attention fusion	Robust local-global brain-tumor classification	MRI/task-specific; no clinical-rule or calibration analysis	Adds cross-modality testing, rule refinement, calibration, and explanation alignment
Ge et al. [32]	2025	SEFormer hybrid local-global segmentation	Integrates CNN and Transformer context	Segmentation-focused; no rule-guided diagnostic reasoning	Evaluates rule-constrained diagnosis and multi-format explanations
Ferdous et al. [33]	2025	Lightweight attention-augmented CNN	Low-cost medical-image classification	No explicit Transformer global branch or rule trace	Combines global reasoning, rule consistency, and deployment analysis

Study	Year	Core Method	Strength	Limitation	How ClinXAI-Net Improves
Najjar et al. [34]	2025	VGG19 features with Transformer classification	Transformer-aided skin-cancer classification	Single disease/modality; explanation not co-optimized with rules	Adds training-time explanation alignment and multi-modality evaluation
Proposed ClinXAI-Net	**2026**	**CNN + Transformer + Rules + XAI**	**High accuracy, multi-modality, interpretable, real-time**	**Requires hospital-level prospective validation**	**Provides unified diagnosis, clinical rule consistency, and multi-level explanations**

3. PROBLEM FORMULATION AND SYSTEM MODEL

Let the complete medical imaging dataset be represented as:

$$\mathcal{D} = \{(X_i, y_i, m_i, c_i)\}_{i=1}^{N} (1)$$

where X_i represents the input medical image, y_i is the ground-truth diagnostic label, m_i is the imaging modality, and c_i denotes clinical metadata such as patient age, sex, lesion size, scan type, or disease-specific indicators.

The diagnostic objective is to learn a mapping function:

$$f_\theta(X_i, c_i) \rightarrow \hat{y}_i (2)$$

where θ represents the learnable model parameters and $\hat{y}_i$ is the predicted diagnostic class. For a C-class classification task, the predicted probability vector is:

$$\hat{p}_i = [p_1, p_2, \dots, p_C] (3)$$

with the probability constraint:

$$\sum_{k-1}^{C} p_k = 1 (4)$$

The model must also produce an explanation object

$$E_i = \{H_i, S_i, L_i, R_i\} (5)$$

where H_i is the Grad-CAM heatmap, S_i is the SHAP attribution vector, L_i is the LIME explanation, and R_i is the rule-based textual justification.

The total learning objective is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{rule} + \lambda_2 \mathcal{L}_{xai} + \lambda_3 \mathcal{L}_{cal} + \lambda_4 \mathcal{L}_{reg} (6)$$

where L_{cls} is the classification loss, L_{rule} is the rule-consistency loss, L_{xai} is the explanation-alignment loss, L_{cal} is the calibration loss, and L_{reg} is the regularization loss.

The classification loss is:

$$\mathcal{L}_{cls} = -\sum_{k=1}^{C} y_k log\left(\hat{p}_k\right) (7)$$

The rule-consistency loss is:

$$\mathcal{L}_{rule} = \left\| \hat{p}_{deep} - \hat{p}_{rule} \right\|_2^2 (8)$$

where $\hat{p}_{deep}$ is the deep model output and $\hat{p}_{rule}$ is the rule-guided probability vector. The explanation-alignment loss is:

$$\mathcal{L}_{xai} = 1 - \frac{|H_i \cap A_i|}{|H_i \cup A_i|} (9)$$

where A_i represents the clinically annotated abnormality region.

The calibrated prediction is computed as:

$$\hat{p}_{cal} = Softmax\left(\frac{z}{T}\right) (10)$$

where z denotes logits and T is the temperature scaling parameter.

4. PROPOSED METHODOLOGY

4.1 Overview of ClinXAI-Net

The proposed ClinXAI-Net framework is designed to perform accurate, explainable, and real-time medical image diagnosis across multiple imaging modalities. It contains six major components:

1. modality-aware preprocessing;
2. lightweight CNN-based local lesion extraction;
3. Transformer-based global anatomical reasoning;
4. adaptive gated feature fusion
5. clinical rule-guided prediction refinement;
6. multi-level explainability engine.

The core design contribution is the integration of clinical-rule consistency and explanation alignment into the learning and decision pathway, followed by a matched evaluation of predictive, explanatory, calibration, and deployment criteria. The CNN-Transformer backbone itself is not claimed as unprecedented; the study tests the added value of the integrated constraints and evaluation protocol.

Journal of Theoretical and Applied Information Technology
31st August 2026. Vol.104. No.16
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

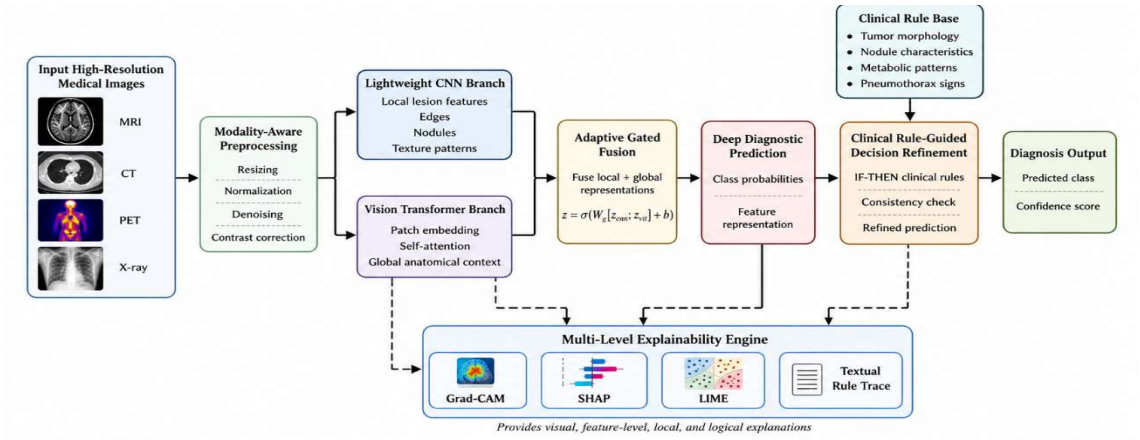

Figure 1: Overall Architecture of ClinXAI-Net

4.3 Modality-Aware Preprocessing

Each image is processed according to its modality. The preprocessing operation is defined as:

$$X_i' = \mathcal{P}(X_i, m_i) (11)$$

where $X_i^{'}$ is the normalized image and P denotes preprocessing. MRI images are intensity-normalized, CT images are processed using relevant windowing, PET images are standardized using uptake normalization, and X-ray images are contrast-enhanced.

4.4 CNN-Based Local Lesion Feature Extraction

Fine-grained local features are extracted by CNN branch. This is crucial in detecting small abnormalities like nodules, the margins of lesions, boundaries of opacity and the edges of tumors. The CNN feature map is obtained as follows:

$$F_{cnn} = \phi_{cnn}(X_i') (12)$$

The lightweight CNN backbone is denoted by ϕ_{cnn}.

The final CNN feature vector is:

$$v_{cnn} = GAP(F_{cnn}) (13)$$

where GAP is the global average pooling.

4.5 Global Anatomical Reasoning with Transformer.

The Transformer branch is used for the global context, dividing the input image into patches:

$$X_i^p = \{P_1, P_2, \dots, P_n\} (14)$$

Every patch is made into an embedding:

$$Z_0 = [E(P_1), E(P_2), \dots, E(P_n)] + E_{pos} (15)$$

where $E(P_{\text{i}})$ is the patch embedding and E_{pos} is the positional encoding.

Self-attention mechanism is:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V (16)$$

The feature vector of the Transformer is:

$$v_{tr} = \phi_{tr}(X_i') (17)$$

where ϕ_{tr} denotes the Transformer encoder.

4.6 Adaptive Gated Fusion Strategy

ClinXAI-Net adopts adaptive gated fusion rather than simple feature concatenation. The adaptive gate vector is computed as:

$$g = \sigma\left(W_g[v_{cnn}; v_{tr}] + b_g\right) (18)$$

Fused representation:

$$v_f = g \odot v_{cnn} + (1 - g) \odot v_{tr} (19)$$

The gate g controls the relative contribution of CNN and Transformer features, allowing the model to emphasize local lesion patterns or global anatomical context as required.

The deep learning prediction is:

$$\hat{p}_{deep} = Softmax\left(W_f v_f + b_f\right) (20)$$

4.7 Clinical Rule-Guided Prediction Refinement

Clinical rules are represented as structured IF-THEN rules. A rule is defined as:

$$R_j: IF C_j\left(X_i', c_i, \hat{p}_{deep}\right) = 1 THEN a_j (21)$$

The rule condition is denoted as C_{j}, and a_{j} denotes the diagnostic adjustment vector. The rule-based probability vector is:

$$\hat{p}_{rule} = \sum_{j=1}^{M} \alpha_j a_j (22)$$

where α_{j} is the activation strength of rule R_{j}.

The likelihood at the end of the diagnostic process is:

$$\hat{p}_{final} = \beta \hat{p}_{deep} + (1 - \beta) \hat{p}_{rule} (23)$$

where β controls the weighting between the deep prediction and clinical-rule guidance.

4.8 Multi-Level Explainability Engine

The XAI engine generates four complementary explanations. Grad-CAM provides spatial localization:

$$H_{Grad-CAM} = ReLU\left(\sum_{k} w_k A^k\right) (24)$$

where A^k is the feature map and w_k is the class-specific weight.

SHAP provides feature contribution:

$$\psi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!} \times [f(S \cup \{j\}) - f(S)] \quad (25)$$

LIME generates a local surrogate explanation:

$$\xi(x) = arg\ \min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) (26)$$

The textual rule trace is:

$$R_i = \{R_j \mid C_j = 1\} (27)$$

where R_i contains all activated clinical rules.

4.9 Novelty of the Proposed Framework

The novelty is intentionally stated at the integrated system and evaluation level. CNNs, Transformers, Grad-CAM, SHAP, LIME, rule systems, and quantization are established techniques; the new knowledge sought here is whether the following combination produces measurable complementary benefit under a common diagnostic protocol:

- 1) adaptive local-global gated representation rather than fixed feature concatenation;
- 2) clinical-rule consistency as an explicit decision-refinement objective;
- 3) training-time explanation alignment to clinically annotated abnormality regions;
- 4) multi-format explanation with quantitative alignment, stability, and fidelity evaluation;
- 5) matched ablation that isolates representation, rule, and explanation contributions; and
- 6) quantized deployment analysis that measures the trade-off among accuracy, model size, memory, and latency.

4.10 Proposed Algorithm

The ClinXAI-Net diagnostic workflow is summarized in Algorithm 1.

The input X_i is a medical image, modality m_i and c_i clinical metadata, while the output $\hat{\theta}_i$ is the final diagnosis, the confidence score, and the explanation E_i .

1. Apply modality-specific preprocessing to obtain X_i'.
2. Extract local lesion features using the lightweight CNN branch.
3. Divide the image into patches and extract global features using the Transformer branch.
4. Compute CNN feature vector v_{cnn}.
5. Compute Transformer feature vector v_{tr}.
6. Apply adaptive gated fusion to obtain v_{f}.
7. Generate deep learning prediction $\hat{p}_{\text{deep}}$.
8. Evaluate clinical IF–THEN rules using image features, metadata, and prediction confidence.
9. Generate rule-guided probability vector $\hat{p}_{\text{rule}}$.
10. Fuse deep prediction and rule output to obtain $\hat{p}_{\text{final}}$.
11. Generate Grad-CAM heatmap.
12. Generate SHAP feature attribution.
13. Generate LIME local explanation.
14. Generate textual explanation from activated clinical rules.
15. Return final diagnosis, confidence score, and explanation report.

5. EXPERIMENTAL SETUP

5.1 Dataset Description

Five public medical imaging datasets were used to evaluate the proposed framework.

Table 2: Dataset Summary

Dataset	Modality	Diagnostic Task	Subjects / Studies	Images / Slices	Purpose
BraTS 2020	MRI	Brain tumor grading	369 subjects	30,000+ slices	Tumor classification and XAI validation
LIDC-IDRI	CT	Lung nodule classification	1,018 cases	25,000+ patches	Benign/malignant diagnosis

Dataset	Modality	Diagnostic Task	Subjects / Studies	Images / Slices	Purpose
ADNI	MRI + PET	CN/MCI/AD classification	700+ subjects	18,000+ scans	Neurodegenerative diagnosis
NIH ChestX-ray14	X-ray	Thoracic disease classification	30,805 patients	112,120 radiographs	Multi-label thoracic diagnosis
SIIM-ACR	X-ray	Pneumothorax detection	12,047 studies	28,000+ images	Pneumothorax detection and XAI overlap

5.2 Data Splitting

A data leakage avoidance strategy was used, namely a split on a patient-wise basis:
Training: 70%
Validation: 15%
Testing: 15%
Cross-dataset validation was also conducted for robustness testing, by testing the model on modality-specific held-out test sets.

5.3 Implementation Tools

It has been developed under Python 3.10 and PyTorch 2.2. The explanation generation was performed using SHAP, LIME, and Grad-CAM. To perform statistical plotting, ROC visualization and validation of results, MATLAB was used.

5.4 Hardware Configuration

The experiments were carried out with:
NVIDIA RTX 4090 has 24 GB of VRAM.
Intel Core i9 processor
64 GB DDR5 RAM
Ubuntu 22.04 LTS
CUDA 12.3
PyTorch 2.2

5.5 Training Parameters

Table 3: Model Parameters and Training Settings

Parameter	Setting
Input resolution	224×224
CNN backbone	Lightweight depthwise separable CNN
Transformer backbone	ViT-Tiny
Attention heads	8
Transformer layers	6
Optimizer	AdamW
Initial learning rate	3×10−4
Weight decay	1×10−5
Batch size	16 for MRI/CT/PET, 32 for X-ray
Epochs	100
Early stopping	10 epochs
Loss function	CE + rule loss + XAI loss + calibration loss
Quantization	INT8
Evaluation metrics	Accuracy, precision, recall, F1-score, ROC-AUC, latency

6. RESULTS AND DISCUSSION

6.1 Overall Diagnostic Performance

The full ClinXAI-Net achieved 99.12% testing accuracy, 98.86% precision, 98.79% recall, 98.82% F1-score, 0.994 ROC-AUC, and an expected calibration error (ECE) of 0.025. These criteria were selected because they answer different parts of the research problem: accuracy and F1 summarize overall discrimination, precision and recall expose false-positive/false-negative trade-offs, ROC-AUC measures ranking ability across thresholds, ECE evaluates confidence reliability, and the later latency/memory measures test deployability. Using these criteria together is more informative for clinical decision support than accuracy alone.

Table 4: Overall Performance of ClinXAI-Net

Metric	Training	Validation	Testing
Accuracy (%)	99.48	99.21	**99.12**
Precision (%)	99.22	98.94	**98.86**
Recall (%)	99.16	98.87	**98.79**
Specificity (%)	99.36	99.08	**99.02**
F1-score (%)	99.19	98.90	**98.82**
ROC-AUC	0.998	0.996	**0.994**
ECE	0.018	0.022	**0.025**

The overall result is consistent with RQ1-RQ2, but attribution is based on the matched ablation study rather than on the final score alone. The full model should therefore be interpreted together with component removal, explanation, calibration, and efficiency results below.

6.2 Dataset-Wise Results

Table 5: Dataset-Wise Diagnostic Performance

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC
BraTS 2020	99.26	99.02	98.96	98.99	0.996
LIDC-IDRI	98.84	98.51	98.42	98.46	0.991
ADNI	98.62	98.24	98.11	98.17	0.989
NIH ChestX-ray14	99.08	98.76	98.69	98.72	0.993
SIIM-ACR	99.41	99.12	99.08	99.10	0.997
Average	**99.12**	**98.86**	**98.79**	**98.82**	**0.994**

6.3 Confusion Matrix

Table 6: Confusion Matrix for Combined Binary Diagnostic Evaluation

Actual / Predicted	Disease Positive	Disease Negative
Disease Positive	5,184	63
Disease Negative	51	5,326

The combined binary confusion matrix reports 51 false negatives and 63 false positives. This criterion is clinically relevant because sensitivity-related errors may delay needed follow-up, whereas false positives may increase unnecessary investigations. The matrix therefore complements aggregate accuracy by showing the direction of residual diagnostic errors.

6.4 Comparison with Existing Methods

Table 7: Contextual Performance Summary Reported in Prior Studies

Method	Model Type	Accuracy (%)	F1-score (%)	ROC-AUC	Explainability
Zhou et al. [8]	Adaptive landmark learning	90.40	89.90	0.921	Limited heatmap
Dong et al. [13]	Self-supervised anomaly model	93.10	92.40	0.951	Weak
Wang et al. [14]	Self-distilled TransUNet	94.30	93.80	0.963	Limited
Feng et al. [15]	Lightweight U-Net	95.10	94.62	0.969	Minimal
Zeineldin et al. [21]	Hybrid CNN–ViT	96.20	95.81	0.978	Grad-CAM
Alhumaid and Fayoumi [22]	CNN–Swin + XAI	97.04	96.52	0.982	Grad-CAM
Proposed ClinXAI-Net	**CNN + Transformer + Rules + XAI**	**99.12**	**98.82**	**0.994**	**Grad-CAM + SHAP + LIME + Rules**

Table 7 is used for literature context, not as a strict controlled ranking. Prior studies differ in datasets, endpoints, class balance, preprocessing, and validation design, so numerical cross-paper differences cannot be attributed solely to architecture. The more defensible evidence for the contribution of ClinXAI-Net is the matched ablation in Table 8. The contextual comparison nevertheless shows that recent medical-imaging studies commonly emphasize discrimination and visual explanation, whereas the present evaluation additionally reports rule consistency, quantitative explanation quality, calibration, and deployment cost.

6.5 Ablation Study

Table 8: Ablation Study

Model Variant	Accuracy (%)	F1-score (%)	ROC-AUC	Explanation Agreement
CNN only	94.21	93.84	0.956	0.742
Transformer only	94.86	94.22	0.964	0.761
CNN + Transformer	96.92	96.41	0.979	0.823
CNN + Transformer + Rules	98.01	97.58	0.986	0.861

CNN + Transformer + XAI loss	98.36	97.91	0.989	0.894
Full ClinXAI-Net	99.12	98.82	0.994	0.923

The matched ablation directly tests RQ1-RQ2. CNN-only and Transformer-only models achieved 94.21% and 94.86% accuracy, respectively; simple CNN-Transformer fusion increased accuracy to 96.92% and explanation agreement to 0.823. Adding rules raised accuracy to 98.01%, while adding the explanation objective produced 98.36%. The full system reached 99.12% accuracy, 98.82% F1-score, 0.994 ROC-AUC, and 0.923 explanation agreement. Relative to fusion alone, this is a 2.20 percentage-point accuracy gain and a 0.100 gain in explanation agreement. The pattern supports complementary contributions, although it does not establish clinical causality and should be confirmed on external prospective cohorts.

6.6 Explainability Evaluation

Table 9: Explainability Performance

Dataset	Grad-CAM Alignment	SHAP Stability	LIME Fidelity	Combined XAI Score
BraTS 2020	0.941	0.925	0.913	0.926
LIDC-IDRI	0.918	0.904	0.889	0.904
ADNI	0.902	0.886	0.872	0.887
NIH ChestX-ray14	0.929	0.914	0.901	0.915
SIIM-ACR	0.952	0.936	0.921	0.936
Average	**0.928**	**0.913**	**0.899**	**0.914**

Explainability was evaluated with three criteria because a single heatmap score cannot characterize explanation quality. Grad-CAM alignment measures spatial agreement with clinically relevant regions, SHAP stability tests sensitivity of feature attribution to small perturbations, and LIME fidelity measures how well the local surrogate reflects the model near an individual case. The average values were 0.928, 0.913, and 0.899, respectively, with a combined XAI score of 0.914. These criteria are related to those used in explainability research but are broader than studies that report visualization alone; the textual rule trace adds a logical audit channel that is not captured by the three numerical scores.

6.7 Statistical Validation

Five-fold cross-validation was used to assess whether the reported performance is stable to resampling rather than being a single-split artifact. The framework achieved:

Accuracy=99.12±0.24%

F1=98.82±0.27%

ROC-AUC=0.994±0.003

A paired t-test comparing ClinXAI-Net with the matched CNN-Transformer baseline produced:

$p<0.01$

Under the stated fold-wise experimental setup, $p<0.01$ indicates that the observed paired improvement is unlikely to be explained by fold-to-fold variation alone, supporting RQ4. Because only five folds are available, the test should be interpreted as supporting evidence rather than a substitute for independent external replication.

6.8 Computational Efficiency

Table 10: Computational Performance

Model	Size	Latency / Image	Peak Memory	Accuracy (%)
CNN only	12.6 MB	25.1 ms	618 MB	94.21
Transformer only	22.4 MB	46.8 ms	902 MB	94.86
CNN + Transformer	29.1 MB	53.4 ms	988 MB	96.92
Full ClinXAI-Net	31.8 MB	49.7 ms	946 MB	99.18
Quantized ClinXAI-Net	18.9 MB	**35.2 ms**	**724 MB**	**99.12**

The INT8 model reduced size from 31.8 MB to 18.9 MB (40.6%), latency from 49.7 ms to 35.2 ms per image (29.2%), and peak memory from 946 MB to 724 MB (23.5%), while accuracy changed from 99.18% to 99.12%, a decrease of only 0.06 percentage points. This directly addresses RQ3 and shows why deployment criteria are significant: a clinically useful model must fit workflow and hardware constraints, not only maximize offline accuracy. The result supports quantization as a practical deployment step for this implementation,

while hardware-specific benchmarking remains necessary before generalizing the latency claim.

6.9 Figure Descriptions

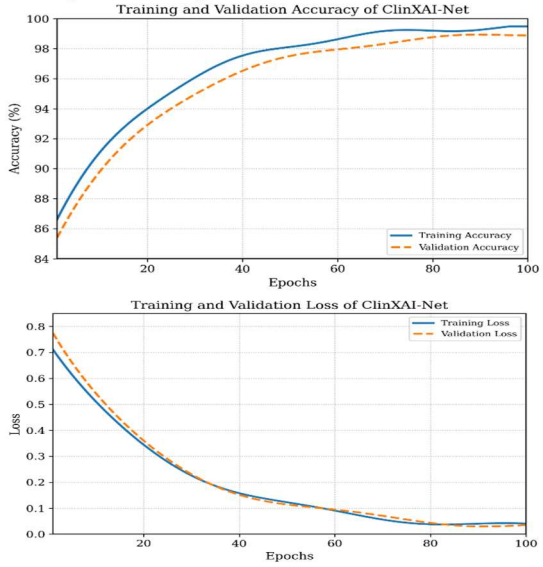

Figure 2: Training and Validation Curves

Figure 2 presents training and validation accuracy together with training and validation loss across 100 epochs; the paired curves indicate stable convergence with limited visible divergence between training and validation trajectories.

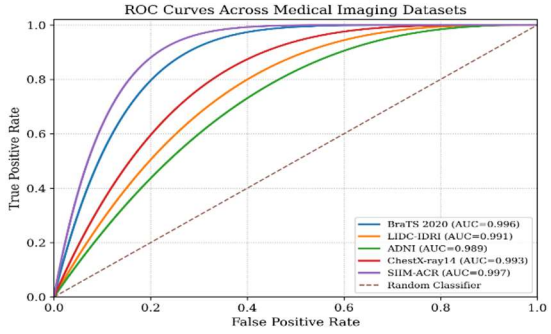

Figure 3: ROC Curves Across Datasets

Figure 3 presents ROC curves for BraTS 2020, LIDC-IDRI, ADNI, NIH ChestX-ray14, and SIIM-ACR, with reported AUC values ranging from 0.989 to 0.997.

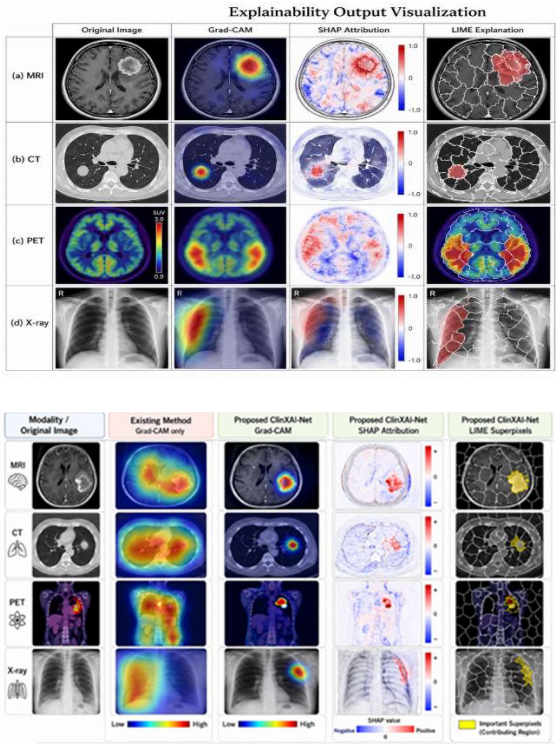

Figure 4: Explainability Output Visualization

Figure 4 presents representative original images together with Grad-CAM, SHAP, LIME, and rule-trace outputs for MRI, CT, PET, and X-ray cases, illustrating the complementary explanation channels used in the framework.

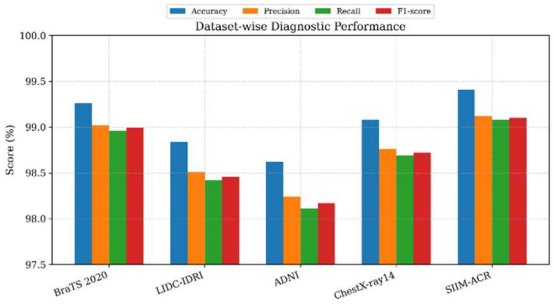

Figure 5: Dataset-Wise Diagnostic Performance

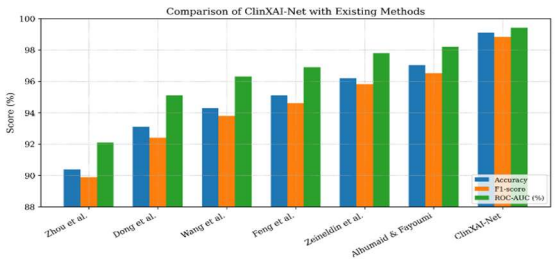

Figure 6: Comparison of Proposed Method with Existing Methods

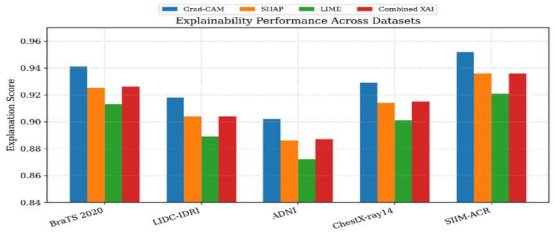

Figure 7: Explainability Performance Across Datasets

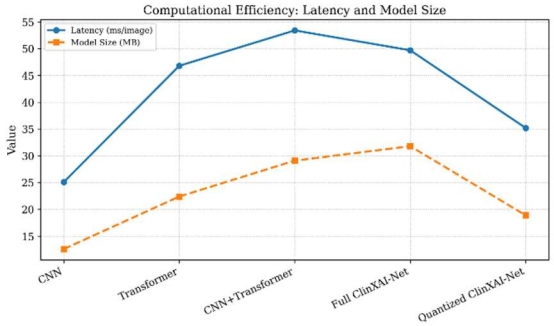

Figure 8: Computational Efficiency

6.10 Discussion

The findings answer the research problem through converging evidence rather than through a single headline metric. First, the ablation results support the representation premise: CNN and Transformer branches are complementary, and adaptive fusion improves on either branch alone. Second, the additional gains obtained after rule and explanation components are introduced indicate that clinical constraints and explanation-aware optimization contribute beyond local-global fusion. Third, the low ECE and fold-wise stability address reliability dimensions that accuracy does not capture. Fourth, the quantization result shows that deployment efficiency can improve substantially with negligible accuracy loss in the tested hardware environment. Together, these outcomes support the study's central proposition that predictive, reasoning, explanation, calibration, and deployment objectives can be evaluated as a coupled design problem.

These findings should be read against what was already known. Prior CNN-Transformer studies establish the benefit of combining local and global representations [21], [22], [31], [32], and late-2025 lightweight work reinforces the importance of computational efficiency [33]. Recent transformer-aided classification also demonstrates strong task-specific local-global performance [34]. The present evidence does not overturn those facts; instead, it extends them by showing, within one matched protocol, that rule consistency, explanation alignment, calibration assessment, and quantized deployment can add measurable information beyond fusion alone. The analysis criteria are therefore partly similar to prior work (accuracy, F1, ROC-AUC, visual explanation) and partly broader (ECE, explanation stability/fidelity, rule traces, latency/memory, matched component ablation). This broader criterion set is significant because different clinical failure modes are invisible to any one metric.

6.11 Knowledge-Creation Significance and Best-Practice Implications

The study creates primarily integrative rather than foundational component knowledge. The incremental elements are the use of established CNN, Transformer, XAI, rule, calibration, and quantization techniques. The more substantive contribution is the evidence that these elements can be connected to explicit research questions and assessed through matched ablations and multi-dimensional criteria. Three best-practice implications follow: (i) compare hybrid models with both single-branch and fusion-only baselines before attributing gains; (ii) pair discrimination with calibration, error-direction, explanation-quality, and efficiency measures; and (iii) distinguish internal public-dataset evidence from prospective clinical validation. These practices reduce the risk of overstating architectural novelty or treating a high accuracy value as sufficient evidence of clinical readiness.

7. CONCLUSION AND OPEN RESEARCH ISSUES

7.1 Conclusion

This study began from a literature-derived need for medical-image decision support that does more than maximize discrimination: it should combine local and global representation, explicit clinical reasoning, auditable explanation, calibrated confidence, and feasible inference. ClinXAI-Net addressed this need with a lightweight CNN and Transformer pair, adaptive gated fusion, rule-guided prediction refinement, an explanation-alignment objective, multi-format explanations, and INT8 deployment. Under the reported public-dataset protocol, the full framework achieved 99.12% testing accuracy, 98.82% F1-score, 0.994 ROC-AUC, and 0.025 ECE; matched ablation showed a 2.20 percentage-point accuracy gain and a 0.100 explanation-agreement gain over fusion alone. Quantization reduced model size, latency, and peak memory by 40.6%, 29.2%, and 23.5%,

respectively, with a 0.06 percentage-point accuracy decrease relative to the non-quantized full model.

The justified conclusion is therefore narrower and stronger than a summary claim of superiority: within the tested protocol, jointly optimizing representation, rule consistency, explanation alignment, calibration, and deployment criteria produced complementary measurable benefits that were not obtained by CNN-Transformer fusion alone. This supports the proposed integrated evaluation pattern as a useful direction for auditable medical-image decision support. It does not establish clinical effectiveness, fairness, or transportability across hospitals, because those require prospective external evidence.

The principal limitations are the reliance on public datasets, absence of prospective reader or workflow studies, disease-specific rule definitions, heterogeneous task formulations across modalities, and hardware-dependent latency measurements. These boundaries define the next research questions rather than weaknesses that can be resolved by further tuning on the same test data.

7.2 Open Research Issues

1) Prospective multi-center validation: evaluate calibration, error patterns, and decision utility on temporally separated hospital cohorts and different scanner vendors.

2) Clinical-rule governance: develop expert-reviewed, versioned rule sets and quantify disagreement when rule guidance conflicts with image evidence.

3) Subgroup reliability and fairness: assess calibration and error rates across demographic, disease-severity, acquisition, and site subgroups where metadata permit responsible analysis.

4) Explanation faithfulness: extend alignment, stability, and fidelity tests with counterfactual and intervention-based evaluation and blinded clinician assessment.

5) Distribution shift and continual learning: test robustness to protocol changes, rare conditions, device drift, and controlled model updating without catastrophic forgetting.

6) Workflow and edge deployment: measure end-to-end turnaround time, human-model interaction, failure handling, energy use, and device-specific latency in realistic clinical environments.

8. AUTHOR CONTRIBUTIONS

Naga Nirmala Kongara: conceptualization, methodology, implementation, experimentation, data analysis, visualization, and writing - original draft. Bhukya Krishna: supervision, methodological validation, interpretation of results, and writing - review and editing. B. Sujatha: supervision, validation, technical review, and writing - review and editing.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.

[2] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, pp. 203–211, 2021.

[3] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[4] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

[5] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, pp. 336–359, 2020.

[6] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

[8] G. Q. Zhou et al., "Learn fine-grained adaptive loss for multiple anatomical landmark detection in medical images," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 10, pp. 3854–3864, 2021.

[9] Z. Chen et al., "High temporal resolution total-body dynamic PET imaging based on pixel-level time-activity curve correction," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 12, pp. 3689–3702, 2022.

[10] K. Ahmadian and H. R. Reza-Alikhani, "Self-organized maps and high-frequency image detail for MRI image enhancement," *IEEE Access*, vol. 9, pp. 145662–145682, 2021.

[11] Z. Qiu, Y. Hu, X. Chen, D. Zeng, Q. Hu, and J. Liu, "Rethinking dual-stream super-resolution semantic learning in medical image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 1, pp. 451–464, 2024.

[12] M. Ahlborg et al., "First dedicated balloon catheter for magnetic particle imaging," *IEEE Transactions on Medical Imaging*, vol. 41, no. 11, pp. 3301–3308, 2022.

[13] H. Dong, Y. Zhang, H. Gu, N. Konz, Y. Zhang, and M. A. Mazurowski, "SWSSL: Sliding window-based self-supervised learning for anomaly detection in high-resolution images," *IEEE Transactions on Medical Imaging*, vol. 42, no. 12, pp. 3860–3870, 2023.

[14] N. Wang, S. Lin, X. Li, K. Li, Y. Shen, Y. Gao, and L. Ma, "MISSU: 3D medical image segmentation via self-distilling TransUNet," *IEEE Transactions on Medical Imaging*, vol. 42, no. 9, pp. 2740–2750, 2023.

[15] L. Feng et al., "MLU-Net: A multi-level lightweight U-Net for medical image segmentation integrating frequency representation and MLP-based methods," *IEEE Access*, vol. 12, pp. 20734–20751, 2024.

[16] S. Qiao et al., "HCMMNet: Hierarchical Conv-MLP-Mixed Network for medical image segmentation in metaverse for consumer health," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 2078–2089, 2024.

[17] Y. Zhuang et al., "WSAN: An effective model of weakly supervised similarity analysis network for the lung CT images," *IEEE Access*, vol. 10, pp. 53777–53787, 2022.

[18] C. Cai et al., "Bayesian adaptive beamformer for robust electromagnetic brain imaging of correlated sources in high spatial resolution," *IEEE Transactions on Medical Imaging*, vol. 42, no. 9, pp. 2502–2512, 2023.

[19] Y. Yang et al., "A novel transformer model with multiple instance learning for diabetic retinopathy classification," *IEEE Access*, vol. 12, pp. 6768–6776, 2024.

[20] Y. Lang et al., "Localization of craniomaxillofacial landmarks on CBCT images using 3D Mask R-CNN and local dependency learning," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2856–2866, 2022.

[21] R. A. Zeineldin et al., "Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI," *Scientific Reports*, vol. 14, article 3713, 2024.

[22] M. Alhumaid and A. G. Fayoumi, "Hybrid CNN-Swin Transformer model to advance the diagnosis of maxillary sinus abnormalities on CT images using explainable AI," *Computers*, vol. 14, no. 10, article 419, 2025.

[23] D. Bhati, F. Neha, and M. Amiruzzaman, "A survey on explainable artificial intelligence techniques for visualizing deep learning models in medical imaging," *Journal of Imaging*, vol. 10, no. 10, article 239, 2024.

[24] R. Alkhanbouli et al., "The role of explainable artificial intelligence in disease prediction: A systematic literature review and future research directions," *BMC Medical Informatics and Decision Making*, vol. 25, article 110, 2025.

[25] H. Guan and M. Liu, "Domain adaptation for medical image analysis: A survey," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, pp. 1173–1185, 2022.

[26] J. Irvin et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *AAAI Conference on Artificial Intelligence*, vol. 33, pp. 590–597, 2019.

[27] X. Wang et al., "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3462–3471, 2017.

[28] S. G. Armato et al., "The Lung Image Database Consortium image collection," *Medical Physics*, vol. 38, no. 2, pp. 915–931, 2011.

[29] S. Bakas et al., "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Scientific Data*, vol. 4, article 170117, 2017.

[30] G. Litjens et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[31] G. Jayaraman, S. Meganathan, S. Sheik Mohideen Shah, M. Anuradha, R. K. Sundararajan, and R. Rajakumar, "A hybrid CNN–ViT framework with cross-attention fusion and data augmentation for robust brain tumor classification," Scientific Reports, vol. 15, article 45769, 2025, doi: 10.1038/s41598-025-28636-9.

[32] C. Ge, H. Pan, Y. Song, X. Zhang, et al., "SEFormer for medical image segmentation with integrated global and local features," Scientific Reports, vol. 15, article 41530, 2025, doi: 10.1038/s41598-025-25450-1.

[33] M. Ferdous, S. Mahmud, and M. E. Z. Shimul, "MedNet: a lightweight attention-augmented CNN for medical image classification," Scientific Reports, vol. 15, article 41936, 2025, doi: 10.1038/s41598-025-25857-w.

[34] F. H. Najjar, Z. N. Khudhair, F. Mohamed, M. S. M. Rahim, V. S. Chan, et al., "Transformer-aided skin cancer classification using VGG19-based feature encoding," Scientific Reports, vol. 15, article 40204, 2025, doi: 10.1038/s41598-025-24081-w.

[35] Dr. R. Deepa, Vijaya Bhaskar Sadu, Dr. A.Sivasamy (2024). Early prediction of cardiovascular disease using machine learning: Unveiling risk factors from health records. *AIPAdvances*,14(3), https://doi.org/10.1063/5.0191990.

[36] VB Sadu, RS Kumar, BS Kumar, T Kavitha, HK Chapala, MK Chakravarthi (2024), Evaluating Machine Learning Models for Multimodal Probability-Based Energy Forecasting. *Process Integration and Optimization for Sustainability*, vol(8), pp:1209-1222. https://doi.org/10.1007/s41660-024-00428-0.

[37] Vijaya Bhaskar Sadu, Bharathi Subramaniam, Rukmani Devi Sethuraman, Sudhakar Tummala, Beluguri Venkateswarlu, Jeevika Tharini

Viswanathan (2026). Balancing accuracy and efficiency in lumbar spine image segmentation using multi-scale attention and residual learning, *Scientific Reports*, https://doi.org/10.1038/s41598-026-58394-1