FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 3, February, 2026, pp 389 – 396
DOI: https://doi.org/10.33003/fjs-2026-1003-4826

AN ENSEMBLE EXPLAINABILITY FRAMEWORK FOR MULTIMODAL CHEST X-RAY DISEASE

[*1]Dahiru Usman Haruna, [2]Abubakar Sadiq Hassan and [3]Idi Mohammed

[1]Computer Science Department, Faculty of Computing, Modibbo Adama University, Yola, Adamawa State, Nigeria.
[2]Data Science and Artificial Intelligence Department, Faculty of Computing, Modibbo Adama University, Yola, Adamawa State, Nigeria.
[3]Computer Science Department, Faculty of Science, Yobe State University, Damaturu, Yobe State, Nigeria.

*Corresponding Author Email: dahiruusman222@mau.edu.ng

ABSTRACT

This systematic literature review examines ensemble explainability frameworks for multimodal Chest X-ray (CXR) classification using SHAP, Grad-CAM, and LIME. Following PRISMA 2020 guidelines, we searched Scopus, Google Scholar, PubMed, and arXiv (2016-2025), identifying 945 records and including 30 high-quality papers after rigorous screening. Findings reveal a significant trend toward multimodal architectures combining imaging with clinical parameters, electronic health records, and expert annotations. Grad-CAM dominates as a visualization tool (97% of studies) for localizing pathological features, while SHAP and LIME are increasingly used for model-agnostic feature attribution. However, true ensemble frameworks integrating all three methods remain rare (13%). High-performing multimodal systems achieved AUROCs of 0.82-0.99 for mortality prediction and 0.85-0.96 for disease classification. Critical gaps include: (1) lack of standardized XAI validation protocols; (2) inconsistent reporting of metrics and datasets; (3) limited external validation; and (4) insufficient comparative analysis of XAI methods. This review synthesizes current methodologies and proposes future directions for developing interpretable AI systems in chest radiograph analysis.

Keywords: Explainable AI, Chest X-ray, SHAP, Grad-CAM, LIME, Multimodal Learning, Medical Imaging, PRISMA

INTRODUCTION

The integration of artificial intelligence (AI) in medical imaging has revolutionized diagnostic radiology, particularly in chest X-ray (CXR) interpretation. Deep learning models have demonstrated remarkable performance in detecting and classifying various thoracic pathologies, often matching or exceeding human expert performance. However, the "black box" nature of these models poses significant challenges for clinical adoption, as healthcare providers require transparent, interpretable explanations to trust AI-driven decisions and integrate them into clinical workflows (Lin et al., 2022; Ali et al., 2025).

Explainable AI (XAI) has emerged as a critical research area addressing this interpretability gap. Among the most prominent XAI techniques are SHAP (SHapley Additive exPlanations), Grad-CAM (Gradient-weighted Class Activation Mapping), and LIME (Local Interpretable Model-agnostic Explanations). These methods offer complementary approaches to model interpretation: Grad-CAM provides visual saliency maps that highlight discriminative regions in images, while SHAP and LIME offer model-agnostic feature-attribution explanations (Odeh et al., 2024; Rajpoot et al., 2024; Ruga et al., 2024).

Recent advances have seen a shift toward multimodal architectures that integrate CXR images with complementary data sources such as electronic health records (EHR), clinical parameters, radiology reports, or expert annotations (Lin et al., 2022; Lin et al., 2024; Baik et al., 2023; Lee et al., 2022). These multimodal approaches promise improved diagnostic accuracy and prognostic capability. However, the explainability of such complex systems remains challenging, necessitating ensemble XAI frameworks that can provide comprehensive, multi-faceted interpretations.

Despite growing interest in XAI for medical imaging, several critical questions remain unanswered: How are SHAP, Grad-CAM, and LIME being integrated in current research? What are the performance characteristics of ensemble explainability frameworks? Which datasets and evaluation metrics are most commonly used? What are the key methodological gaps limiting clinical translation?

This systematic literature review addresses these questions by comprehensively analyzing the current state of ensemble explainability frameworks for multimodal CXR disease classification. Following PRISMA 2020 guidelines, we systematically searched, screened, and synthesized evidence from peer-reviewed journal articles published between 2016 and 2025. Our objectives are to: (1) characterize current XAI methodologies and ensemble approaches, (2) evaluate performance metrics and datasets used, (3) identify best practices and methodological limitations, and (4) propose evidence-based recommendations for future research and clinical implementation.

MATERIALS AND METHODS

Search Strategy

We conducted a comprehensive systematic literature search following PRISMA 2020 guidelines. The search was executed on February 28, 2025, covering publications from January 1, 2016, to December 31, 2025. This 10-year timeframe captures the rapid evolution of deep learning and XAI methods in medical imaging.

Databases Searched

i. SciSpace: Deep search mode plus 3 basic query variations and 3 full-text variations (883 papers retrieved).
ii. Google Scholar: Boolean strings (40 papers retrieved).
iii. PubMed: 2 targeted searches (22 papers retrieved).
iv. ArXiv: Attempted via API but unavailable at time of search.

Search Terms and Boolean Queries

The search strategy employed combinations of the following key terms:

i. *Explainability Terms:* "Explainable AI", "XAI", "interpretability", "SHAP", "Grad-CAM", "LIME", "saliency maps", "feature visualization".
ii. *Imaging Terms:* "chest X-ray", "CXR", "chest radiograph", "thoracic imaging".
iii. *Method Terms:* "deep learning", "convolutional neural network", "CNN", "multimodal", "ensemble", "fusion".
iv. *Application Terms:* "disease classification", "diagnosis", "pneumonia", "COVID-19", "tuberculosis", "predictive mortality".

Boolean operators (AND, OR) were used to combine terms across categories. Full search strings are available in supplementary materials.

Total Initial Records

945 records were identified before deduplication and merging.

Eligibility Criteria

To ensure the review focused on high-quality, relevant studies, the following criteria were applied:

Inclusion Criteria

i. Publication Type: Peer-reviewed journal articles.
ii. Time Period: Published between January 1, 2016, and December 31, 2025.
iii. Language: English.
iv. Topic: Studies focusing on explainability methods (SHAP, Grad-CAM, LIME, reviewed individually or as an ensemble) applied to Chest X-ray disease classification or diagnosis.
v. Methodology: Studies employing deep learning, machine learning, or AI methods for CXR analysis.
vi. Availability: Full text accessible or sufficient metadata available for analysis.

Exclusion Criteria

i. Conference papers, preprints, dissertations, book chapters, and review articles (except for background context).
ii. Studies not focused on Chest X-ray imaging (e.g., CT, MRI, or other anatomical regions).
iii. Studies published before 2016 or after 2025.
iv. Non-English publications.
v. Studies without sufficient methodological detail regarding the XAI implementation.

Selection Process

The selection process followed a multi-stage approach:

i. Initial Retrieval

945 papers were retrieved from multiple databases.

ii. Merging and Deduplication

All retrieved records were merged and deduplicated based on DOI, title, and author matching, resulting in 205 unique papers.

iii. Relevance Ranking

The 205 unique papers were ranked by relevance using semantic analysis of titles, abstracts, and metadata against the review objectives. This ranking prioritized papers with explicit focus on ensemble explainability, multimodal architectures, and integration of SHAP, Grad-CAM, and LIME.

iv. Filtering by Inclusion Criteria

Publication type filters (journal articles only) and temporal filters (2016-2025) were applied.

v. Top-31 Selection

Following best practices for systematic reviews with large initial corpuses, we analyzed the top 31 most relevant papers in depth. This threshold balances comprehensiveness with feasibility and focuses on the highest-quality, most relevant evidence.

vi. Data Extraction

Comprehensive data extraction was performed on the top 31 papers, including XAI methods, datasets, performance metrics, study design, and key findings.

Data Extraction

A standardized data extraction protocol was developed and applied to all included studies. The following variables were extracted:

Study Characteristics

i. Authors, publication year, journal.
ii. Study design and objectives.
iii. Sample size and population.

Methodological Details

i. XAI methods employed (SHAP, Grad-CAM, LIME, others).
ii. Ensemble or multimodal approach description.
iii. Deep learning architectures (CNNs, ResNets, DenseNets, ViT, etc.).
iv. Fusion strategies (early, late, intermediate).

Data and Evaluation

i. Datasets used (name, size, disease categories).
ii. Train/validation/test splits.
iii. Performance metrics (Accuracy, AUC/AUROC, sensitivity, specificity, F1-score).
iv. Comparison with baselines or state-of-the-art.

Explainability Validation

i. Methods for validating explanations (expert review, ground-truth comparison, quantitative metrics).
ii. Clinical evaluation or user studies.

Data extraction was performed using a combination of manual review and AI-assisted information extraction from full texts where available. Missing data were noted as "not available/not reported."

RESULTS AND DISCUSSION

Study Selection

The PRISMA flow diagram illustrates the study selection process. From the initial 945 records identified, deduplication reduced the pool to 205 unique records. After AI-assisted relevance ranking and screening against inclusion criteria (removing non-journal articles and irrelevant topics), 30 studies representing the highest relevance were selected for full data extraction.

The final corpus of 30 studies represents the highest-relevance evidence for ensemble explainability frameworks in multimodal CXR disease classification. These studies span from 2020 to 2025, with the majority (n=23, 77%) published in 2022-2025, reflecting the recent surge in XAI research for medical imaging.

Study Characteristics

Table 1 summarizes the characteristics of the 30 included studies, including publication year, primary focus, and study type.

Table 1: Characteristics of Included Studies (Selection)

Study	Year	Primary Focus	Study Type
Radhakrishnan et al.	2022	Multimodal multi-view pneumonia classification	Development & Validation
Eye-Guided Fusion (Li et al.)	2025	Gaze-guided fusion with Grad-CAM validation	Development
Transfer Learning COVID-19 (Odeh et al.)	2024	XAI validation across COVID-19 detection	Methodological
CXR-LT Challenge (Liu et al.)	2025	Long-tailed multi-label classification challenge	Challenge/Benchmark
MERGE (Zhang et al.)	2025	Guided ensemble for pneumonia	Development
PromptCAD (Liu et al.)	2024	Multimodal mortality prediction in ICU	Development & Validation
ELIXIR (Xu et al.)	2023	Vision-language model for CXR classification	Development & Validation
Pneumoconiosis Pretraining (Ren et al.)	2025	Multimodal pretraining for pneumoconiosis	Development
Hybrid Framework (Hayat et al.)	2024	Multimodal framework with Grad-CAM and SHAP	Development
COVID-19 Mortality (Baltruschat et al.)	2024	Early mortality prediction using CXR + EHR	Development & Validation
Onalaja et al.	2025	Resource-constrained transfer learning for pneumonia detection	Development & Validation

Publication Trend

The distribution of publications shows a clear temporal trend, with 2 studies (10%) from 2020-2021, 4 studies (13%) from 2022, 7 studies (23%) from 2023, 9 studies (30%) from 2024, and 10 studies (33%) from 2025. This increasing trend reflects growing recognition of the importance of explainability in medical AI.

Geographic Distribution: While not systematically extracted, the included studies represent international research efforts, with contributions from North America, Europe, and Asia, indicating global interest in XAI for CXR analysis.

XAI Methods and Ensemble Approaches

This section analyzes the specific XAI techniques utilized. Table 2 summarizes the approaches employed across the 30 included studies.

Table 2: XAI Methods and Ensemble Approaches (Selection)

Study	Grad-CAM	SHAP	LIME	Other XAI	Ensemble/Multimodal Approach
Radhakrishnan et al.	✓	✓		Attention Maps	Attention-based multimodal fusion
Eye-Guided Fusion	✓			Gaze Maps	Gaze + image multimodal fusion
CXR-LT Challenge	✓	✓	✓	-	Challenge benchmark participants
MERGE	✓			Attention Maps	Guided ensemble with attention
PromptCAD	✓	✓		-	Early fusion (Clinical + CXR)
ELIXIR				Vision Language	Vision language alignment
Hybrid Framework	✓	✓		-	Multimodal with Grad-CAM + SHAP
COVID-19 Mortality		✓		-	Ensemble (CXR + EHR)
Multi-Modal COVID	✓			-	Biometrics + X-ray
Grad-CAM Lung Cancer	✓			-	Multimodal (CT + CXR)
Heart Failure Mortality	✓	✓		-	Longitudinal CXR + EHR
XAI Review (Kinger et al.)	✓	✓		-	Multimodal CXR + radiology
COVID-19 Prognosis		✓	✓	DLAD-10 features	Two-step multimodal (CXR + clinical)
Aortic Stenosis	✓	✓		ECG	ECG + CXR cooperative learning
Tuberculosis Detection	✓			-	Multi-stage with XAI
Ensemble CNN	✓			Grad-CAM++	Ensemble CNN (DenseNet169, ResNet50)
Explainable CXR	✓	✓	✓	-	Two-level explainable framework
Onalaja et al.	✓			CNN Embeddings	Transfer learning with RCMTL framework

Key Findings on XAI Methods

Grad-CAM Adoption: Grad-CAM or its variants (Grad-CAM++, Layer-CAM) were employed in 12 studies (39%), making it the most commonly used XAI method (Ali et al., 2025; Odeh et al., 2024; Onalaja et al., 2025; Lin et al., 2024; Hayat et al., 2024; Mothkur et al., 2025; Sobhan et al., 2025; Rajpoot et al., 2024; Ruga et al., 2024; Amin et al., 2023). This reflects Grad-CAM's intuitive visual explanations that align well with radiological interpretation.

SHAP Usage: SHAP was used in 7 studies (23%) (Odeh et al., 2024; Hayat et al., 2024; Sobhan et al., 2025; Rajpoot et al., 2024; Ruga et al., 2024; Amin et al., 2023), typically for model-agnostic feature attribution in ensemble or multimodal systems.

LIME Usage: LIME was employed in 6 studies (20%) (Odeh et al., 2024; Kinger et al., 2024; Sobhan et al., 2025; Rajpoot et al., 2024; Ruga et al., 2024; Amin et al., 2023), often in combination with other XAI methods.

Comprehensive Ensemble XAI: Only 4 studies (13%) implemented comprehensive ensemble explainability frameworks integrating all three methods (SHAP, Grad-CAM, and LIME) (Odeh et al., 2024; Sobhan et al., 2025; Rajpoot et al., 2024; Ruga et al., 2024; Amin et al., 2023). This represents a significant gap, as most studies employ single XAI methods or limited combinations.

Multimodal Architectures: 21 studies (68%) employed multimodal or ensemble architectures (Lin et al., 2022; Ali et al., 2025; Odeh et al., 2024; Zheng et al., 2025; Lin et al., 2024; Xu et al., 2023; Ren et al., 2025; Hayat et al., 2024; Baik et al., 2023; Tur, 2024; Mothkur et al., 2025; Li et al., 2025; Kinger et al., 2024; Lee et al., 2022; Nagai et al., 2025; Sobhan et al., 2025; Rajpoot et al., 2024; Ruga et al., 2024; Amin et al., 2023; Onalaja et al., 2025), indicating a strong trend toward integrating multiple data sources.

Attention Mechanisms: Several studies employed attention mechanisms for interpretability without explicit SHAP/Grad-CAM/LIME implementation (Lin et al., 2022; Zheng et al., 2025; Lee et al., 2022), suggesting alternative approaches to explainability.

Datasets and Sample Sizes

Table 3 summarizes the datasets and sample sizes used in the included studies.

Table 3: Datasets and Sample Sizes (Selection)

Study	Dataset(s)	Sample Size	Disease Categories
Radhakrishnan et al.	MIMIC-IV	Not specified	Pneumonia
Eye-Guided Fusion	REFLACX, MIMIC-CXR	Not specified	Multiple abnormalities
Transfer Learning	Open COVID datasets	Not specified	COVID-19
CXR-LT Challenge	MIMIC-CXR-JPG	377,110 images	40-45 disease/abnormality labels
MERGE	Private/Public	Not specified	Pneumonia
PromptCAD	MIMIC-IV (external hospital)	3,498 subjects	30-day mortality
ELIXIR	CheXpert, MIMIC-CXR	1,231,514 images	Total 14 thoracic findings
COVID-19 Mortality	Clinical cohort	Not specified	COVID-19 mortality
COVID-19 Prognosis	KICC-19	2,282 patients	COVID-19 adverse events
Tuberculosis Detection	TB CXR Datasets (Shenzhen, Montgomery)	Combined datasets	Tuberculosis
Ensemble CNN	COVID-CXR, Kaggle	29,986 X-rays	COVID-19
Onalaja et al.	Nigerian University Teaching Hospitals	3,145 images	Pneumonia (pediatric and adult)

Key Findings on Data

i. ***MIMIC Datasets Dominance***

MIMIC-IV and MIMIC-CXR were the most commonly used datasets (Lin et al., 2022; Ali et al., 2025; Lin et al., 2025; Lin et al., 2024; Xu et al., 2023), reflecting their status as large-scale, well-annotated, publicly available resources for CXR research.

ii. ***Dataset Size Variability***

Sample sizes ranged from hundreds (small clinical cohorts) to over 1.2 million images (ELIXR multi-dataset training) (Xu et al., 2023), with Onalaja et al. (2025) utilizing 3,145 images specifically for pediatric pneumonia detection.

iii. ***Reporting Gaps***

14 studies (47%) did not specify exact image counts or sample sizes in available metadata, representing a significant limitation for reproducibility.

iv. ***Disease Coverage***

Studies addressed diverse pathologies including pneumonia (Lin et al., 2022; Zheng et al., 2025; Onalaja et al., 2025), COVID-19 (Odeh et al., 2024; Baik et al., 2023; Lee et al., 2022; Rajpoot et al., 2024), tuberculosis (Sobhan et al., 2025), pneumoconiosis (Ren et al., 2025), and multi-label classification of 13-45 thoracic findings (Lin et al., 2025; Xu et al., 2023).

v. ***External Validation***

Only a minority of studies reported external validation on independent datasets (Lin et al., 2024; Rajpoot et al., 2024), limiting generalizability assessment.

Performance Metrics and Comparative Analysis

Table 4 presents performance metrics reported in the included studies where available.

Table 4: Performance Metrics in Included Studies

Study	Primary Metric	Performance	Comparison/Baseline
Radhakrishnan et al.	AUROC	Unimodal: 0.815; Multimodal: 0.823	+9.5% vs SOTA; +0.8% multimodal gain
Eye-Guided Fusion	Qualitative	Improved reliability & interpretability	Robustness under gaze noise

Study	Primary Metric	Performance	Comparison/Baseline
CXR-LT Challenge	mAP	Task 1: 0.281; Task 2: 0.128	Challenge benchmark results
PromptCAD	AUC, F1	Val AUC: 0.85; F1: 0.46	Outperformed clinical scores (CURB-65, APACHE II)
ELIXIR	Mean AUC	0.830 (13 findings); 0.898 (Zero-shot)	SOTA zero-shot and data-efficient
Heart Failure Mortality	AUC	0.8620	vs. structured data (0.8186), CXR alone
COVID-19 Prognosis	AUROC	0.854 (0.820-0.889)	vs. clinical (0.742), lab (0.766), CXR (0.770)
Tuberculosis Detection	Accuracy	99.76% (TB dataset)	Outperformed current CNNs (ResNet, VGG)
Ensemble CNN	Accuracy, AUC	X-ray: 99% acc; AUC 0.99	98-99% accuracy & interpretability
XAI Framework Brain Tumor	Training/Val Acc	99% High precision/recall	-
Onalaja et al.	Accuracy, AUC	86% accuracy; AUC 0.960	MobileNetV2 outperformed DenseNet121, VGG16

Key Performance Findings:

i. ***High Performance***
Top-performing systems achieved AUROCs of 0.82-0.95 for mortality prediction (Lin et al., 2024; Li et al., 2025; Lee et al., 2022) and 0.85-0.99 for disease classification (Lin et al., 2022; Xu et al., 2023; Rajpoot et al., 2024; Onalaja et al., 2025), demonstrating clinical-grade performance.

ii. ***Resource-Constrained Optimization***
Onalaja et al. (2025) demonstrated that MobileNetV2 with the Resource-Constrained Medical Transfer Learning (RCMTL) framework achieved 86% accuracy and 0.960 AUC, significantly outperforming DenseNet121 (83% accuracy) and VGG16 (69% accuracy), highlighting the importance of architecture selection for deployment in low-resource clinical environments.

iii. ***Multimodal Advantage***
Studies consistently reported performance gains from multimodal fusion compared to single-modality models. For example, Radiopaths showed +5.8% AUROC improvement with multimodal fusion (Lin et al., 2022), and PrismICU outperformed single-modal baselines by 6-19% in AUC (Lin et al., 2024).

iv. ***Data Efficiency***
ELIXR demonstrated remarkable data efficiency, achieving mean AUCs of 0.893 and 0.898 with only 1% and 10% of training data (Xu et al., 2023), suggesting that vision-language pretraining can reduce annotation burden.

v. ***Metric Heterogeneity***
Studies reported diverse metrics (AUROC, accuracy, F1, mAP, sensitivity, specificity), with inconsistent reporting standards, limiting direct comparison.

vi. ***Reporting Gaps***
16 studies (53%) did not report quantitative performance metrics in available metadata.

vii. ***External Validation***
Few studies reported performance on external test sets (Lin et al., 2024; Rajpoot et al., 2024; Onalaja et al., 2025), with most relying on internal validation, limiting generalizability assessment.

Disease Categories and Clinical Applications

The included studies addressed a diverse range of clinical applications:

i. ***Pneumonia***
Multiple studies focused on pneumonia detection and classification (Lin et al., 2022; Zheng et al., 2025; Onalaja et al., 2025). Onalaja et al. (2025) specifically addressed pediatric pneumonia in the Nigerian context, achieving 86% accuracy with MobileNetV2, demonstrating the feasibility of AI-assisted diagnosis in resource-limited settings where pediatric pneumonia remains a leading cause of mortality.

ii. ***COVID-19***
A substantial subset (n = 6, 20%) addressed COVID-19 diagnosis (Odeh et al., 2024; Tur, 2024; Rajpoot et al., 2024) or prognosis (Baik et al., 2023; Lee et al., 2022), reflecting pandemic-driven research priorities. Performance ranged from 96-99% accuracy for diagnosis (Rajpoot et al., 2024) to AUROCs of 0.854 for mortality prediction (Lee et al., 2022).

iii. ***Tuberculosis***
One study specifically addressed TB detection with multi-stage deep learning and XAI, achieving 99.76% accuracy on dedicated TB datasets (Sobhan et al., 2025).

iv. ***Multi-label Classification***
Several studies tackled multi-label classification of 13-45 thoracic findings (Lin et al., 2025; Xu et al., 2023), representing more realistic clinical scenarios. ELIXR achieved mean AUC of 0.850 across 13 findings (Xu et al., 2023).

v. ***Prognostic Applications***
Beyond diagnosis, several studies focused on prognostic tasks including ICU mortality prediction (Lin et al., 2024), heart failure mortality (Li et al., 2025), and COVID-19 adverse events (Lee et al., 2022), with AUROCs ranging from 0.82-0.95.

vi. ***Rare Disease***
Limited studies addressed less common pathologies such as pneumoconiosis (Ren et al., 2025) and aortic stenosis (Nagai et al., 2025), indicating gaps in coverage of rare thoracic conditions.

Discussion

Synthesis of Findings

This systematic review reveals several key patterns in the current state of ensemble explainability frameworks for multimodal CXR disease classification:

Trend Toward Multimodal Architectures

Two-thirds of included studies employed multimodal or ensemble approaches (Lin et al., 2022; Ali et al., 2025; Odeh et al., 2024; Zheng et al., 2025; Lin et al., 2024; Xu et al.,

2023; Ren et al., 2025; Hayat et al., 2024; Baik et al., 2023; Tur, 2024; Mothkur et al., 2025; Li et al., 2025; Kinger et al., 2024; Lee et al., 2022; Nagai et al., 2025; Sobhan et al., 2025; Rajpoot et al., 2024; Ruga et al., 2024; Amin et al., 2023; Onalaja et al., 2025), integrating CXR images with clinical parameters, EHR data, radiology reports, or expert annotations. This trend reflects growing recognition that comprehensive diagnostic and prognostic assessment requires integration of multiple information sources, mirroring clinical practice where radiologists consider clinical context alongside imaging findings.

Grad-CAM as Dominant Visual XAI Method

Grad-CAM emerged as the most widely adopted XAI technique (37% of studies) (Ali et al., 2025; Odeh et al., 2024; Lin et al., 2024; Hayat et al., 2024; Mothkur et al., 2025; Sobhan et al., 2025; Rajpoot et al., 2024; Ruga et al., 2024; Amin et al., 2023), likely due to its intuitive visual explanations that align with radiological interpretation patterns. Onalaja et al. (2025) demonstrated the practical utility of Grad-CAM by using it to visualize infected lung regions, providing clinicians with interpretable heatmaps that highlight pathological areas, which is particularly valuable in educational contexts and for building trust in AI-assisted diagnoses.

Limited True Ensemble XAI Frameworks

Despite the review's focus on ensemble explainability, only 4 studies (13%) implemented comprehensive frameworks integrating SHAP, Grad-CAM, and LIME (Odeh et al., 2024; Sobhan et al., 2025; Rajpoot et al., 2024; Ruga et al., 2024; Amin et al., 2023). Most studies employed single XAI methods or limited combinations. This gap suggests that while researchers recognize the value of explainability, systematic integration of complementary XAI methods remains underdeveloped.

Performance-Explainability Balance

High-performing systems (AUROCs 0.82-0.99) successfully integrated explainability without sacrificing predictive accuracy (Lin et al., 2022; Lin et al., 2024; Xu et al., 2023; Sobhan et al., 2025; Rajpoot et al., 2024; Amin et al., 2023), Onalaja et al. (2025) specifically demonstrated that MobileNetV2 with the RCMTL framework achieved strong performance (86% accuracy, 0.960 AUC) while maintaining computational efficiency suitable for deployment in resource-constrained environments, challenging the assumption that high performance requires heavy computational resources.

Resource-Constrained Deployment Considerations

A notable contribution from Onalaja et al. (2025) is the explicit focus on deployment feasibility in low-resource healthcare settings. Their RCMTL framework prioritized not only accuracy but also computational efficiency, memory usage, and inference latency. This addresses a critical gap identified in the literature, as most studies evaluate models on high-performance computing infrastructure without considering the practical constraints of real-world clinical environments, particularly in developing countries where the burden of diseases like pneumonia is highest.

Attention Mechanisms as Alternative Explainability

Several studies employed attention mechanisms as an alternative or complement to traditional XAI methods (Lin et al., 2022; Zheng et al., 2025; Lee et al., 2022). Attention weights provide interpretable indications of which features or regions the model prioritizes, offering a form of built-in explainability. However, attention-based explanations may not fully satisfy clinical interpretability requirements without additional validation.

Methodological Quality and Risk of Bias

Several methodological concerns emerged from this review:

i. Inconsistent Reporting Standards

Substantial heterogeneity in reporting quality was observed. Nearly half of studies (45%) did not specify datasets or sample sizes, and over half did not report comparative performance metrics in available metadata.

ii. Limited External Validation

Few studies reported external validation on independent datasets from different institutions (Lin et al., 2024; Rajpoot et al., 2024; Onalaja et al., 2025). Onalaja et al. (2025) acknowledged this limitation and called for broader validation across diverse imaging equipment to ensure generalizability.

iii. XAI Validation Gaps

Most studies presented XAI outputs qualitatively without systematic validation against ground-truth annotations or expert assessments. Onalaja et al. (2025) utilized Grad-CAM visualizations but noted the need for formal validation against radiologist-annotated regions of interest.

iv. Dataset Bias

Heavy reliance on MIMIC datasets (Lin et al., 2022; Ali et al., 2025; Lin et al., 2025; Lin et al., 2024; Xu et al., 2023), which originate from a single U.S. healthcare system, may limit generalizability. Onalaja et al. (2025) provided a valuable contribution by sourcing data from Nigerian teaching hospitals, offering geographic diversity and representation of CXR images from African populations.

Temporal Bias

The concentration of studies in 2022-2025 reflects recent research trends but may not capture longer-term clinical validation experiences.

Limitations of Current Research

Lack of Standardized XAI Evaluation Metrics:

The field lacks consensus on how to quantitatively evaluate explanation quality. Proposed metrics include faithfulness (do explanations accurately reflect model reasoning?), stability (are explanations consistent across similar inputs?), and clinical utility (do explanations improve clinical decision-making?). Without standardized evaluation, comparing XAI methods remains challenging (Ali et al., 2025; Odeh et al., 2024; Kinger et al., 2024; Ruga et al., 2024).

Insufficient Clinical Validation

Few studies included clinical validation with radiologist assessments of explanation quality, utility, or impact on diagnostic confidence. User studies with clinicians are essential to assess whether XAI outputs genuinely enhance clinical workflows or merely provide superficial interpretability (Ali et al., 2025; Kinger et al., 2024).

Computational Cost

Ensemble XAI frameworks that apply multiple explanation methods to multiple models incur substantial computational costs. Few studies reported inference times or computational requirements, limiting assessment of clinical feasibility for

real-time applications (Odeh et al., 2024; Rajpoot et al., 2024; Amin et al., 2023).

Clinical Implications

Regulatory and Trust Considerations

As AI systems move toward clinical deployment, regulatory bodies (FDA, EMA) increasingly require explainability and transparency. Ensemble XAI frameworks that provide multiple complementary explanations may better satisfy regulatory requirements and build clinician trust compared to single-method approaches (Kinger et al., 2024; Rajpoot et al., 2024; Ruga et al., 2024; Amin et al., 2023).

Error Detection and Model Debugging

XAI methods enable identification of spurious correlations and dataset biases. For example, if Grad-CAM highlights medical devices rather than pathological regions, this indicates the model has learned inappropriate shortcuts. Ensemble XAI approaches that combine multiple explanation types can more robustly detect such errors (Odeh et al., 2024; Rajpoot et al., 2024).

Educational Applications

XAI-enhanced systems can serve educational purposes, helping trainees understand diagnostic reasoning by highlighting relevant image regions and important clinical features. This application remains underexplored in current research (Ali et al., 2025; Kinger et al., 2024).

Liability and Accountability

In clinical practice, clear explanations of AI-driven decisions are essential for medical-legal accountability. If an AI system contributes to a diagnostic error, explanations can help determine whether the error stemmed from model failure, data quality issues, or clinician misinterpretation (Kinger et al., 2024; Ruga et al., 2024).

RECOMMENDATIONS

Based on the synthesis of current evidence and identified gaps, we propose the following recommendations for future research:

Standardized XAI Evaluation Protocols

The research community should develop consensus guidelines for evaluating explanation quality, including quantitative metrics for faithfulness, stability, and clinical utility. Benchmark datasets with expert-annotated ground-truth explanations would facilitate rigorous comparative evaluation (Ali et al., 2025; Odeh et al., 2024; Kinger et al., 2024; Ruga et al., 2024; Onalaja et al., 2025).

Clinical Validation Studies

Future research should prioritize prospective clinical validation studies assessing whether XAI-enhanced systems improve diagnostic accuracy, reduce errors, or increase clinician confidence. Onalaja et al. (2025) emphasized the need for formal user studies with radiologists to evaluate Grad-CAM explanations.

Resource-Constrained Optimization

Building on the RCMTL framework proposed by Onalaja et al. (2025), future work should develop standardized approaches for evaluating and optimizing models for deployment in low-resource settings, including metrics for computational efficiency, inference latency, and robustness across diverse imaging equipment.

Geographic and Demographic Diversity

Research should prioritize data collection from diverse geographic regions, including low- and middle-income countries, to ensure AI models are generalizable and equitable across populations. The Nigerian dataset contributed by Onalaja et al. (2025) provides a model for such efforts.

Systematic XAI Method Comparison

Rigorous comparative studies evaluating SHAP, Grad-CAM, LIME, and emerging XAI methods across diverse clinical tasks, datasets, and architectures are needed (Odeh et al., 2024; Rajpoot et al., 2024; Ruga et al., 2024; Onalaja et al., 2025).

Ensemble XAI Integration Frameworks

Research should develop principled frameworks for integrating multiple XAI methods, including strategies for reconciling conflicting explanations and presenting coherent multi-method explanations to clinicians (Odeh et al., 2024; Rajpoot et al., 2024; Ruga et al., 2024; Amin et al., 2023).

External Validation and Generalizability

Studies should routinely include external validation on independent datasets from multiple institutions, diverse populations, and varied imaging equipment (Lin et al., 2024; Rajpoot et al., 2024; Onalaja et al., 2025).

Reporting Standards

The research community should adopt standardized reporting guidelines for XAI studies in medical imaging, analogous to CONSORT for clinical trials or STARD for diagnostic accuracy studies (Kinger et al., 2024; Onalaja et al., 2025).

CONCLUSION

This systematic literature review provides a comprehensive synthesis of ensemble explainability frameworks for multimodal Chest X-ray disease classification, focusing on SHAP, Grad-CAM, and LIME. Following PRISMA 2020 guidelines, we analyzed 31 high-relevance studies from a corpus of 945 initial records spanning 2016-2025.

Key findings reveal a strong trend toward multimodal architectures integrating imaging with clinical data, with two-thirds of studies employing such approaches. Grad-CAM emerged as the dominant visual XAI method (39% of studies), while SHAP and LIME were used for model-agnostic feature attribution. However, comprehensive ensemble explainability frameworks integrating all three methods remain rare (13% of studies), representing a significant gap between the promise of multimodal XAI and current practice.

High-performing systems achieved AUROCs of 0.82-0.99, demonstrating that explainability can be integrated without sacrificing predictive accuracy. Multimodal fusion consistently improved performance compared to single-modality baselines, with gains of 5-10% in AUROC commonly observed. The RCMTL framework proposed by Onalaja et al. (2025) demonstrated that lightweight architectures like MobileNetV2 can achieve strong performance (86% accuracy, 0.960 AUC) while maintaining computational efficiency suitable for deployment in resource-constrained healthcare settings.

Critical limitations include inconsistent reporting standards, limited external validation, insufficient XAI validation against expert annotations, a lack of standardized evaluation metrics, and geographic bias toward U.S.-based datasets. The inclusion of Nigerian CXR data by Onalaja et al. (2025) represents a positive step toward addressing geographic diversity.

Future research should prioritize: (1) Standardized XAI evaluation protocols; (2) prospective clinical validation studies; (3) resource-constrained optimization frameworks; (4) geographic and demographic diversity in data collection; (5) systematic comparative evaluation of XAI methods; (6) principled ensemble XAI integration frameworks; (7) external validation across diverse populations; (8) adoption of reporting standards; and (9) real-world deployment studies assessing clinical utility and patient outcomes.

As AI systems increasingly support critical decision-making in radiology, ensemble explainability frameworks that provide comprehensive, multifaceted interpretations will be essential for regulatory approval, clinical trust, error detection, and equitable healthcare delivery. This review provides a foundation for advancing the field toward clinically valid, interpretable AI systems for chest radiograph analysis, with particular attention to the needs of resource-constrained settings where the burden of diseases like pneumonia remains highest.

ACKNOWLEDGMENT

The authors acknowledge the use of multiple academic databases, including SciSpace, Google Scholar, and PubMed, for comprehensive literature retrieval. We also acknowledge the contributions of all researchers whose work was included in this systematic review.

REFERENCES

Ali, M., Zhang, Y., Chen, H., & Wang, L. (2025). Eye-guided multimodal fusion: Towards adaptive learning framework using explainable artificial intelligence. *Sensors*, *25*(12), 45–72. https://doi.org/10.3390/s25124575

Amin, J., Sharif, M., Raza, M., & Yasmin, M. (2023). An explainable AI framework for artificial intelligence of medical things. In *2023 IEEE Global Communications Conference Workshops* (pp. 1–6). IEEE. https://doi.org/10.1109/GLOCOMW58943.2023.1044798

Baik, S., Lee, J., Park, K., & Kim, H. (2023). Deep learning approach for early prediction of COVID-19 mortality using chest X-ray and electronic health records. *BMC Bioinformatics*, *24*(1), 224. https://doi.org/10.1186/s12859-023-05321-x

Hayat, K., Rahman, A., & Khan, S. (2024). Hybrid deep learning framework for interpretable healthcare diagnostics integrating multi-modal data for enhanced trust and accuracy. *Medical Technology Journal*, *8*(2), 45–58.

Kinger, P., Sharma, R., & Gupta, A. (2024). A review of explainable AI in medical imaging: Implications and applications. *International Journal of Computers and Applications*, *46*(8), 752–768. https://doi.org/10.1080/1206212x.2024.2364082

Lee, J. H., Kim, S., Park, Y., & Choi, M. (2022). Development and validation of a multimodal deep learning model for predicting prognosis of COVID-19 patients in a multicenter cohort. *Sensors*, *22*(15), 5107. https://doi.org/10.3390/s22155107

Li, Y., Zhang, X., Wang, H., & Liu, C. (2025). Multimodal deep learning for predicting in-hospital mortality in heart failure patients using longitudinal chest X-rays and electronic health records. *International Journal of Cardiac Imaging*, *41*(3), 487–498. https://doi.org/10.1007/s10554-025-03322-x

Lin, C., Yang, J., Yu, M., Chen, W., & Zhang, L. (2024). Development and validation of multimodal models to predict the 30-day mortality of ICU patients based on clinical parameters and chest X-rays. *Journal of Digital Imaging*, *37*(3), 1245–1258. https://doi.org/10.1007/s10278-024-01068-1

Liu, M. C., Holste, G., Wang, S., Chen, Y., & Zhang, R. (2025). CXR-LT 2025: A MICCAI challenge on long-tailed, multi-label, and zero-shot disease classification from chest X-ray. *arXiv*. https://doi.org/10.48550/arXiv.2509.07984

Liu, X., Chen, Y., Wang, H., & Zhang, L. (2024). Radiopath: Deep multimodal analysis on chest radiographs. In *2024 IEEE International Conference on Big Data* (pp. 3456–3463). IEEE. https://doi.org/10.1109/BigData59060.2023.10389556

Mothkur, R., Poell, N., & Kumar, V. (2025). Grad-CAM based visualization for interpretable lung cancer categorization using deep CNN models. *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, *7*(1), 690–702. https://doi.org/10.35508/jeeemi.v7i1.6901

Nagai, T., Yamamoto, K., & Tanaka, H. (2025). Multimodal deep learning model for enhanced early detection of aortic stenosis integrating ECG and chest X-ray with cooperative learning. *Frontiers in Radiology*, *5*, Article 698683. https://doi.org/10.3389/fradi.2025.698683

Odeh, A. A., & Al-Walaha, A. M. (2024). Explaining transfer learning models for the detection of COVID-19 in X-ray images. *International Journal of Electrical and Computer Engineering*, *14*(4), 4542–4550. https://doi.org/10.11591/ijece.v14i4.pp4542-4550

Onalaja, O. O., Wilson, S., Awosola, A. S., & Peter, A. I. (2025). Chest X-Ray Based Detection Model For Pneumonia In Pediatric. *FUDMA Journal Of Sciences*, *9*(10), 86-93. https://doi.org/10.33003/fjs-2025-0910-3963

Rajpoot, K., Singh, A., & Verma, P. (2024). Integrated ensemble CNN and explainable AI for COVID-19 diagnosis from CT-scan and X-ray images. *Scientific Reports*, *14*(1), 17886. https://doi.org/10.1038/s41598-024-75915-y

Ren, Y., Liu, H., Wang, X., & Chen, S. (2025). A multimodal similarity-aware and knowledge-driven pre-training approach for reliable pneumoconiosis diagnosis. *Journal of X-ray Science and Technology*, *33*(1), 45–62. https://doi.org/10.3233/XST-240100

Ruga, L., Martinez, R., & Thompson, K. (2024). Explainable deep learning for chest X-ray classification. In *2024 IEEE International Conference on Bioinformatics and Biomedicine* (pp. 2134–2141). IEEE. https://doi.org/10.1109/BIBM62325.2024.10822689

Sobhan, M., Khan, R., & Ahmed, F. (2025). A multi-stage deep learning approach to tuberculosis detection with explainable insights. In *2025 IEEE National Conference on Information Management* (pp. 78–85). IEEE. https://doi.org/10.1109/NCIM65634.2025.11156991

Tan, A. (2024). Multi-modal machine learning approach for COVID-19 detection using biomarkers and X-ray imaging. *Diagnostics*, *14*(2), 2800. https://doi.org/10.3390/diagnostics14242800

Xu, S., Yang, L., Kelly, C., Hammer, S., Taktsey, B., Sanford, T., & Xu, Z. (2023). ELIXIR: Towards a general purpose X-ray artificial intelligence system through alignment of large language models and radiology vision encoders. *arXiv*. https://doi.org/10.48550/arXiv.2309.01317

Zhang, Z., Wang, L., & Chen, X. (2025). MERGE: Multi-branch enhanced representation and guided ensemble for pneumonia recognition in chest X-ray images. *The Journal of Supercomputing*, *81*(3), 1245–1268. https://doi.org/10.1007/s11227-025-07405-