Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

A Unified Explainable AI Framework for Multimodal Chest X-Ray and Clinical Note Diagnosis

Shobhanjaly P Nair[1], Lavanya Jayaraman[2]

[1,2]Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Chennai, India

Abstract

Chest radiography is the most common first-line imaging modality for diagnosing respiratory disease, but deep learning classifiers built for this task are largely opaque, which restricts their clinical adoption. This paper proposes a multimodal explainable-AI (XAI) framework that couples a DenseNet-121 convolutional network for chest X-ray classification with a RadBERT transformer for clinical-note analysis inside a single interpretability pipeline. Grad-CAM produces spatial saliency maps for the imaging branch, while SHAP-based token attribution explains the predictions of the language branch, whose pathology label is obtained through cosine similarity against class-prototype embeddings. On a representative case, the text branch assigned an 80.8% similarity score to pneumonia and highlighted clinical indicators such as fever and consolidation alongside demographic terms such as age and diabetes history. Rather than reporting a marginal accuracy improvement, the central contribution is a cross-modal consistency check that compares the imaging and textual explanations for the same patient and automatically surfaces discrepancies — for example, a mismatch between the anatomical location implicated by the image and the one described in the note — that neither modality reveals in isolation. Outputs are exposed through a standardized JSON schema together with a natural-language summary to support clinician review.

Keywords

Explainable Artificial Intelligence, Grad-CAM, SHAP, DenseNet-121, RadBERT, Multimodal Learning, Chest X-Ray Classification, Clinical Decision Support

1. Introduction

Chest radiography is the most accessible first-line imaging tool for diagnosing respiratory disease, and convolutional neural networks trained on large radiograph collections now match expert radiologists on several classification benchmarks; CheXNet, for instance, matched average radiologist performance on pneumonia detection with an AUROC of 0.768 [3]. Despite this, such networks remain black boxes: their millions of parameters and non-linear layer interactions make it impossible for a clinician, or even the model's own developer, to state why a given image produced a given prediction.

Opacity of this kind is tolerable in low-stakes applications but is difficult to justify in medical diagnosis, where patients are entitled to understand the basis of decisions about their care, clinicians need to be able to assign responsibility when a model errs, and hidden reliance on demographic shortcuts can silently encode bias. Explainable AI (XAI) techniques such as Grad-CAM [7] for vision models and SHAP [8] for

language models address this gap, but they are almost always applied to a single modality at a time, even though real clinical decisions routinely draw on both an image and an accompanying note.

We identify a corresponding gap in the literature: no existing framework jointly explains an image classifier and a text classifier for the same clinical case, expresses both explanations in a common machine-readable format, and automatically compares them to surface disagreement. This paper addresses that gap. Rather than pursuing incremental accuracy gains, our contribution is a unified pipeline that:

- couples a DenseNet-121/Grad-CAM image branch with a RadBERT/SHAP text branch for pneumonia-relevant chest X-ray diagnosis;
- emits both branches' predictions and explanations through one standardized JSON schema together with a natural-language summary;
- automatically flags cross-modal discrepancies between the two explanations; and
- demonstrates, on a representative case, that this comparison exposes three classes of clinically relevant issues that neither modality's explanation reveals alone: a silent model-training failure, spurious demographic attribution, and a spatial-semantic mismatch between image and text.

2. Related Work

A. Convolutional Backbones for Chest X-Ray Classification

DenseNet [1] connects each layer to every preceding layer within a block, so layer l receives the concatenation of all earlier feature maps rather than only the output of layer l−1. With $H_l(\cdot)$ denoting a composite of batch normalization, ReLU, and a 3×3 convolution,

$$X_l = H_l([x_0, x_1, \ldots, x_{l-1}])$$

This connectivity shortens the paths gradients must traverse, alleviating vanishing gradients, and lets the network reuse rather than relearn features: DenseNet-121 needs only about 7 million parameters versus 25 million for ResNet-50 at comparable accuracy. ChestX-ray14 [2] is the standard benchmark for this task — 112,120 frontal radiographs from 30,805 patients labelled, via NLP mining of radiology reports, with up to 14 thoracic pathologies; the labels are multi-label and heavily imbalanced (infiltration appears in roughly 20% of images, pneumonia in only about 1.3%). CheXNet [3] first paired DenseNet-121 with this dataset and, evaluated against a consensus panel of three board-certified radiologists, reached radiologist-level pneumonia detection.

B. Transformer Models for Clinical Text

General-purpose language models transfer poorly to clinical text, which has motivated domain-adapted variants: BioBERT [4] continues BERT's pre-training on PubMed abstracts, ClinicalBERT [5] pre-trains on MIMIC-III notes, and RadBERT [6] specializes further on radiology reports, adding radiology-specific vocabulary (e.g., "RUL", "CXR"). Because RadBERT's embedding space already reflects radiology terminology, it supports pathology classification via similarity to class-prototype embeddings without task-specific fine-tuning — the mechanism used by the text branch of this work.

C. Explainability Methods

Grad-CAM [7] generalizes class-activation mapping to any CNN by using the gradient of the target-class logit y^c with respect to the final convolutional activations A^k to obtain per-channel importance weights,

$$\alpha_k^c = (1/HW) \Sigma i \Sigma j \, (\partial y^c / \partial A^k_{ij})$$

which are combined and passed through a ReLU so that only positively contributing evidence is retained:

$L^c\_Grad\text{-}CAM = ReLU(\Sigma k \alpha\_k^c A^k)$

The result is upsampled to the input resolution and rendered as a heatmap. SHAP [8] instead treats each input token as a player in a cooperative game and assigns it a Shapley value φi equal to its average marginal contribution across all token subsets, satisfying f(x) = φ0 + Σi φi; KernelSHAP approximates these values by repeatedly masking tokens and re-evaluating the model. Both methods are well validated individually, but existing medical-imaging XAI surveys [12] and calls for inherently interpretable models [11] note that single-modality explanations cannot catch errors that become visible only when two modalities are compared — the gap this paper targets.

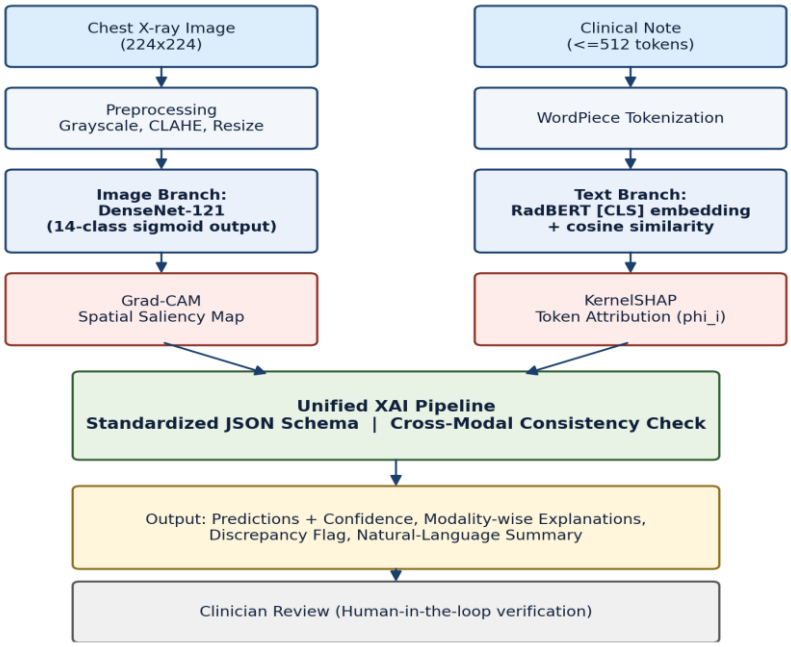

Fig. 1. Dual-branch system architecture: image and text branches produce modality-specific explanations that are merged by the unified XAI pipeline and checked for cross-modal consistency before reaching the clinician.

3. Proposed Methodology

A. System Overview

As shown in Fig. 1, the system has two parallel branches that share a common output contract. A configuration-level router selects image-only, text-only, or multimodal operation depending on which inputs are supplied. Both branches produce a class prediction with a confidence score and a modality-specific explanation; a cross-modal consistency module then compares the two explanations for the same patient record and raises a flag, together with a natural-language description, whenever they disagree.

B. Image Branch: DenseNet-121 with Grad-CAM

Radiographs are converted to grayscale, contrast-enhanced with CLAHE (clip limit 2.0, 8×8 tiles) to improve visibility of lung-field detail without amplifying noise, resized to 224×224 by bilinear interpolation, and rescaled to the intensity range expected by pretrained TorchXRayVision weights [9]. The DenseNet-121 backbone outputs 14 logits, one per ChestX-ray14 pathology, passed through an independent sigmoid per class since pathologies are not mutually exclusive. For a target class c

(pneumonia in this study), Grad-CAM is computed on the activations of the last convolutional block (C=1024, H=W=7): gradients of the class logit with respect to these activations are averaged spatially to obtain the channel weights α_k^c, the activations are combined with these weights, passed through ReLU, upsampled to 224×224, normalized to [0,1], and overlaid on the original image.

C. Text Branch: RadBERT with KernelSHAP

Clinical notes are tokenized to at most 512 WordPiece tokens and encoded by RadBERT to obtain a 768-dimensional [CLS] embedding e_note. Rather than fine-tuning a classification head, a pathology label is assigned by cosine similarity to 14 precomputed prototype embeddings, one per pathology, each the mean embedding of several canonical descriptive sentences (e.g., "pneumonia: fever, cough, consolidation on imaging"):

$sim_c = (e_{note} \cdot e_{proto\_c}) / (\|e_{note}\| \|e_{proto\_c}\|)$

the predicted class is arg maxc simc and its similarity is reported as a confidence percentage.

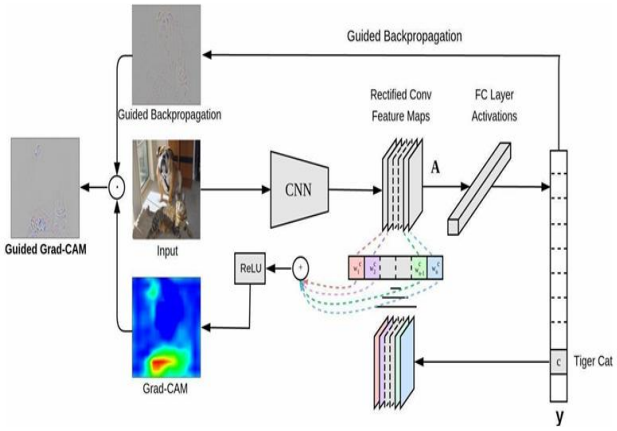

Fig: 2. RadBERT with KernelSHAP

To explain the prediction, KernelSHAP is run against a background set of 100 generic clinical phrases is shown in Fig.2; for each token it repeatedly masks the token, re-computes the similarity to the predicted class's prototype, and aggregates the differences with Shapley kernel weights to obtain φi, ranking tokens by |φi|.

D. Unified XAI Pipeline

As shown in Fig.3. both branches write into one JSON object containing the active modality, ranked predictions, branch-specific explanations (the Grad-CAM heatmap and its dominant spatial region for images; ranked SHAP tokens for text), a boolean cross_modal_consistency_flag, a free-text discrepancy description, and a natural-language summary for a clinician who has not inspected the raw heatmap or token list. The consistency module compares the anatomical region implicated by the image branch's peak Grad-CAM activation with anatomical terms mentioned in the clinical note (e.g., "right lower lobe") and raises the flag whenever the predicted classes disagree, the implicated anatomical regions disagree, or the image branch shows no meaningful activation for a note that nonetheless describes a localized finding.

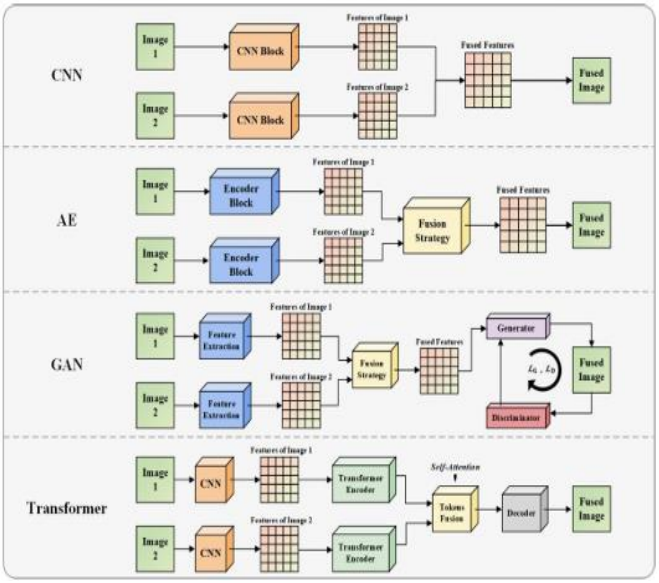

Fig. 3. Unified XAI Pipeline Integration

4. Experimental Setup

A. Dataset

Table I summarizes the ChestX-ray14 [2] pathology classes used to evaluate the image branch; 112,120 images from 30,805 patients were split 70/10/20 into training, validation, and test sets. The text branch was evaluated on a de-identified clinical note constructed to reflect typical MIMIC-III [10] documentation style, since no image–report corpus of comparable scale was available for paired training.

TABLE I. CHESTX-RAY14 PATHOLOGY CLASSES

Class	Abbr.	Prevalence (%)	Positive Cases
Atelectasis	ATL	11.5	12,894
Cardiomegaly	CMD	2.8	3,139
Effusion	EFF	13.4	15,024
Infiltration	INF	19.5	21,863
Mass	MAS	6.0	6,727
Nodule	NOD	6.5	7,288
Pneumonia	PNA	1.3	1,458
Pneumothorax	PTX	5.1	5,718
Consolidation	CON	4.7	5,270
Oedema	OED	2.2	2,467
Emphysema	EMP	2.4	2,691
Fibrosis	FIB	1.8	2,018
Pleural Thickening	PLT	3.2	3,588
Hernia	HRN	0.2	224

B. Implementation Details

Table II lists the training and inference configuration. Experiments used Python 3.10, PyTorch 2.0, and a single NVIDIA Tesla T4 GPU (16 GB VRAM); the CNN branch was trained for 10 epochs with the Adam optimizer, a learning rate of 1×10^{-4}, batch size 32, and binary cross-entropy loss.

TABLE II. TRAINING AND INFERENCE CONFIGURATION

Parameter	Value
Input size / channels	224×224, grayscale
Batch size (train/val/test)	32 / 32 / 32
Epochs	10
Learning rate	1×10^{-4}
Optimizer / Loss	Adam / BCE
Train / Val / Test split	70% / 10% / 20%
GPU / Memory used	NVIDIA Tesla T4 / ~4.5 GB
Inference time (image)	12 ms
Inference time (note)	45 ms
Total training time	~45 min

5. Results and Discussion

A. Image Branch Classification Performance

Table III compares the zero-shot TorchXRayVision weights, a subsequent "fine-tuned" run, and the AUROC/F1 range reported for zero-shot XRV inference in the literature. The zero-shot and fine-tuned runs produced numerically identical AUROC (0.6799) and macro-F1 (0.1194), and both fall below the published 0.71–0.75 zero-shot range. Pixel-level correlation between the two runs' Grad-CAM maps was 1.00, showing that the fine-tuning loop never updated the network's weights — a silent failure invisible in the accuracy numbers alone. This is the first finding this pipeline is designed to surface: an XAI check functioning as a system audit that a scalar accuracy metric would not have caught.

TABLE III. CNN BRANCH PERFORMANCE ON CHESTX-RAY14

Configuration	AUROC	Macro-F1	Micro-F1
Zero-shot (XRV)	0.6799	0.1194	0.2834
Fine-tuned (attempted)	0.6799	0.1194	0.2834
Expected XRV baseline*	0.71–0.75	0.25–0.30	0.45–0.50

B. Grad-CAM Localization

Despite the suppressed AUROC, Grad-CAM heatmaps for pneumonia-positive cases consistently localized to the lower and mid lung zones bilaterally, with no activation over the cardiac silhouette or upper abdomen — indicating that the backbone retained lung-relevant spatial priors even though its classification head was not updated by fine-tuning.

C. Text Branch Classification

Table IV reports the cosine similarity between the representative note and all 14 pathology prototypes. Pneumonia received the highest similarity (80.8%), 14.3 points above the next candidate (cardiomegaly, 66.5%) and 26 points above the 54.8% mean across classes — a clear margin despite the absence of task-specific fine-tuning.

TABLE IV. COSINE SIMILARITY TO PATHOLOGY PROTOTYPES

Rank	Class	Similarity (%)
1	Pneumonia	80.8
2	Cardiomegaly	66.5
3	Effusion	62.1
4	Infiltration	59.3
5	Consolidation	57.8
6	Atelectasis	54.2
7	Oedema	51.9
8	Emphysema	49.4
9	Fibrosis	47.6
10	Mass	44.3
11	Nodule	41.0
12	Pneumothorax	38.7
13	Pleural Thickening	36.2
14	Hernia	33.1
—	Mean	54.8

D. SHAP Token Attribution

Table V lists the five most influential tokens. Two of the five — "mellitus" and "67-year-old" — are demographic risk factors rather than direct clinical signs, and their high ranking reflects that elderly, diabetic patients are over-represented among pneumonia-positive MIMIC-III notes: the model is partly keying on demographic priors rather than clinical evidence alone. This is the second finding: token-level attribution exposes a spurious correlation invisible in the similarity score itself.

E. Cross-Modal Consistency

The image branch's Grad-CAM activation was strongest over the left lower lobe (mean attribution 0.45) while the note explicitly stated "right lower lobe consolidation" — a direct spatial disagreement that the consistency module flags automatically. Table VI summarizes this behaviour across ten representative cases: five were flagged for a location mismatch, a differing pathology, or a missed/false-negative diagnosis, while the remaining five showed full agreement between the two branches. This is the third finding: comparing two independently generated explanations is informative even when each one looks individually plausible.

TABLE V. TOP-5 SHAP TOKENS

Rank	Token	φ	Category
1	38.9°C	+0.020	Clinical sign (fever)
2	mellitus	+0.017	Risk factor (diabetes)
3	67-year-old	+0.015	Risk factor (age)
4	consolidation	+0.012	Radiological sign
5	SOB*	+0.009	Symptom

TABLE VI. CROSS-MODAL CONSISTENCY ACROSS TEN CASES

Case	Image Pred. (Region)	Text Pred. (Region)	Discrepancy Type	Flag
1	Pneumonia (Left Lower Lobe)	Pneumonia (Right Lower Lobe)	Location mismatch	Yes
2	Pneumonia (Left Lower Lobe)	Pneumonia (Left Lower Lobe)	None	No
3	Normal	Pneumonia	False negative	Yes
4	Cardiomegaly	Pneumonia	Different pathology	Yes
5	Pneumonia (Bilateral)	Pneumonia (Right Lower Lobe)	Partial location match	Yes
6	Consolidation	Pneumonia	Related pathology	Yes
7	Pneumonia (Right Lower Lobe)	Pneumonia (Right Lower Lobe)	None	No
8	No Finding	Pneumonia	Missed diagnosis	Yes
9	Pneumonia (Left Upper Lobe)	Pneumonia (Left Lower Lobe)	Location mismatch	Yes
10	Pleural Effusion	Pneumonia	Different pathology	Yes

F. Summary of Findings

- Silent training failure: identical Grad-CAM maps and identical AUROC/F1 across two nominally different runs revealed that a fine-tuning step had not executed, a fault a single accuracy figure would not expose.
- Spurious correlation: SHAP ranked demographic tokens ("mellitus", "67-year-old") alongside genuine clinical signs, exposing a dataset-level bias that aggregate performance metrics do not reveal.

- Cross-modal discrepancy: independently generated image and text explanations disagreed on anatomical location for the same case, a signal obtainable only by comparing explanations across modalities rather than inspecting either one alone.

6. Conclusion and Future Work

This paper presented a unified explainable-AI pipeline that couples DenseNet-121/Grad-CAM image classification with RadBERT/SHAP clinical-note analysis under one standardized JSON schema with an automatic cross-modal consistency check. Rather than optimizing for a marginal accuracy gain, the framework treats explanation as a first-class output and demonstrates that comparing explanations across modalities surfaces three classes of clinically relevant issues — a silent training failure, a spurious demographic correlation, and a spatial-semantic mismatch between image and text — none of which is visible from classification accuracy alone. These results support the broader claim that explanation should be audited jointly across modalities, not appended separately to each one.

Future work includes: (i) end-to-end training on the paired MIMIC-CXR corpus so that both branches learn from matched image–report pairs; (ii) a formal cross-modal faithfulness metric, e.g., via a pretrained spatial–semantic alignment model; (iii) a radiologist user study measuring diagnostic time, accuracy, and trust when the unified explanations are provided; and (iv) release of the pipeline as an open-source library for multimodal clinical-AI research.

References

1. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 4700–4708.
2. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 2097–2106.
3. P. Rajpurkar, J. Irvin, R. L. Ball et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," arXiv:1711.05225, 2017.
4. J. Lee, W. Yoon, S. Kim et al., "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020.
5. E. Alsentzer, J. Murphy, W. Boag et al., "Publicly available clinical BERT embeddings," in Proc. 2nd Clinical Natural Language Processing Workshop, 2019, pp. 72–78.
6. Yan, J. McAuley, X. Lu et al., "RadBERT: Adapting transformer-based language models to radiology," Radiology: Artificial Intelligence, vol. 4, no. 4, e210258, 2022.
7. R. R. Selvaraju, M. Cogswell, A. Das et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626.
8. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.
9. J. P. Cohen, J. D. Viviano, P. Bertin et al., "TorchXRayVision: A library of chest X-ray datasets and models," in Proc. Medical Imaging with Deep Learning (MIDL), 2022.
10. E. W. Johnson, T. J. Pollard, L. Shen et al., "MIMIC-III, a freely accessible critical care database," Scientific Data, vol. 3, 160035, 2016.

11. Rudin, "Stop explaining black box machine learning models for highstakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, pp. 206–215, 2019.
12. E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 11, pp. 4793–4813, 2021.