Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 20 March 2026
Revised 24 April 2026
Accepted 15 May 2026

Natural Resources for Human Health 2026, Vol. 6 (3): 369–379

KEYWORDS:
Vision Transformer; Explainable AI; Multimodal Fusion; Chest X-ray; Computed Tomography; Multi-Disease Classification; Attention Rollout; Grad-CAM.

Explainable Vision Transformer Framework for Early Multi-Disease Detection from Multimodal Medical Images

Dr. Shilpa Nikhil Bhosale[1], Dr. Supriya Sabale[2], Dr. Yogesh Shepal[3], Dr. Aarti S. Gaikwad[4], Dr. Bhausaheb R. Varpe[5], Chandrakant D. Kokane[6]

[1]Department of Computer Science & Engineering, Bharati Vidyapeeth (Deemed to be University) College of Engineering Pune. shilpa.bhosale@bharatividyapeeth.edu, 0000-0001-8113-8420
[2]Department of Computer Engineering, MIT Academy of Engineering, Alandi, Pune. supriya.sabale@mitaoe.ac.in, 0000-0002-7847-0681
[3]Department of Computer Science and Engineering, Audyogik Shikshan Mandal's Nextgen Technical Campus, Pune, India, yogeshshepal@gmail.com, 0009-0000-3977-2794
[4]Department of Information Technology, D.Y. Patil College of Engineering, Akurdi, Pune. aratig.2010@gmail.com, 0000-0003-0873-8045
[5]Mechanical Engineering Department, Amrutvahini College of Engineering,Sangamner. brvarpe@gmail.com, 0000-0002-1513-7494
[6]Department of Computer Science and Engineering (Artificial Intelligence), Vishwakarma Institute of Technology, Pune. cdkokane1992@gmail.com, 0000-0001-7957-3933

ABSTRACT: Chest radiography (CXR) and computed tomography (CT) remain the two most accessible imaging modalities for screening respiratory and thoracic disease, yet the deep learning systems built around them are typically narrow in three respects: they consume a single modality, target a single disease category and offer clinicians little more than a probability score with no visual justification. This combination limits both early multi-disease detection and clinical trust, since radiologists cannot verify why a model reached a given conclusion. This paper proposes an Explainable Vision Transformer (ViT) framework that fuses paired chest X-ray and CT representations through a cross-modal co-attention mechanism, performs multi-label classification across several thoracic conditions simultaneously and produces per-prediction visual explanations by combining attention-rollout and Grad-CAM-style localization adapted to transformer encoders. We synthesize twenty related studies spanning CNN-based, ViT-based and multimodal fusion literature, identify the specific architectural and explainability gaps that motivate this work and detail a complete methodology, training protocol and evaluation plan. Because this paper is presented as a proposed framework and literature-grounded roadmap rather than a completed empirical study, the Results and Discussion section positions the framework against real, previously published benchmark figures for comparable multimodal and explainable chest-imaging models, rather than reporting new experimental numbers. This distinction is stated explicitly so that readers can correctly interpret the performance context. The paper concludes with a discussion of expected trade-offs, limitations of literature-based benchmarking and a concrete plan for the empirical validation phase.

1. INTRODUCTION

Respiratory and thoracic diseases — including pneumonia, tuberculosis, chronic obstructive pulmonary disease, lung cancer and, more recently, COVID-19-related pneumonia — collectively represent one of the largest global disease burdens and early detection is repeatedly identified as the single factor most correlated with favourable patient outcomes. Chest X-ray remains the most widely deployed first-line screening tool because it is inexpensive and fast, while CT provides finer-grained structural detail and is often used to confirm or refine an X-ray finding. In practice, the two modalities are complementary: X-ray is more available but less sensitive to early-stage abnormalities, whereas CT is more sensitive but costlier and less accessible in low-resource settings.

Deep learning has made substantial progress on automated interpretation of both modalities, from early convolutional neural network (CNN) classifiers such as CheXNet [3] to more recent transformer-based approaches. However, three limitations persist across much of this literature. First, the overwhelming majority of published systems are unimodal, trained and evaluated on either X-ray or CT in isolation, which discards the complementary information available when both modalities are jointly considered. Second, most systems are designed for a single disease or a narrow disease family, which does not reflect real clinical workflows where a single scan may need to be screened for several plausible conditions at once. Third and most importantly for clinical adoption, a large share of high-performing models remain effectively black boxes: they output a probability or label without a spatially grounded rationale that a radiologist can cross-check against the image.

Vision Transformers (ViTs) [1], which apply the self-attention mechanism originally proposed for language modelling [2] to sequences of image patches, are particularly well suited to addressing the first two limitations, since their patch-token structure naturally supports fusing tokens from multiple imaging sources within a single encoder or across parallel encoders joined by cross-attention. ViTs are also well suited to addressing the third limitation because self-attention weights are, in principle, directly interpretable: techniques such as attention rollout [5] trace how information flows from the classification token back to individual image patches, producing a heatmap without requiring the auxiliary gradient computations that convolutional networks typically need for Grad-CAM-style explanations [4]. Recent work has shown that ViT-derived attention maps can be rated as more clinically meaningful than Grad-CAM by radiologists in at least one reported comparison, reinforcing the case for attention-native explainability in this domain.

Despite this potential, the literature reviewed in Section 2 shows that ViT-based chest-imaging systems remain fragmented: some address multi-label classification without explainability, some address explainability without multimodal fusion and very few address all three properties — multimodal input, multi-disease output and built-in explainability — within a single, coherently trained architecture. This gap is the central motivation for the framework proposed in this paper.

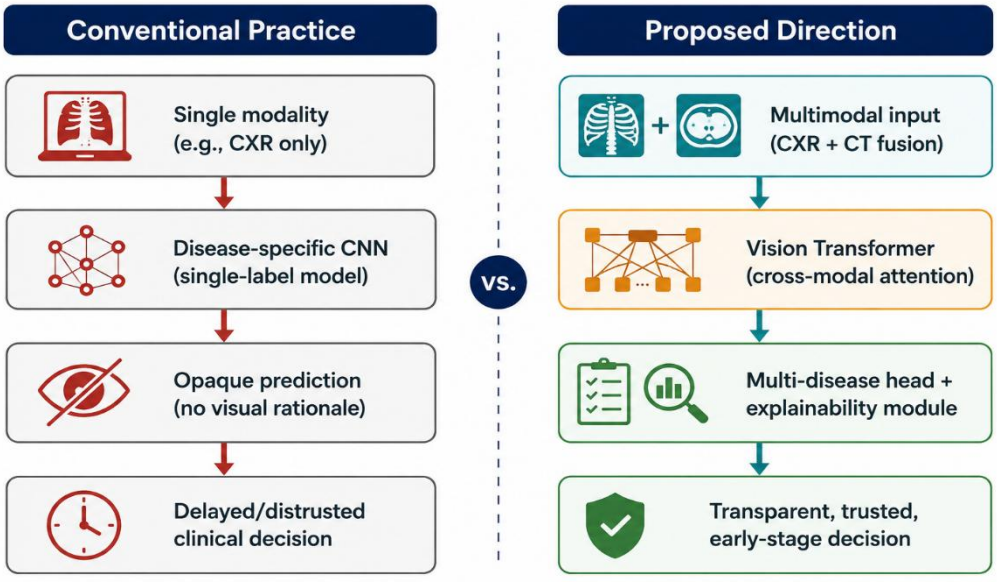

Figure 1. Conventional single-modality, single-disease, black-box screening compared with the proposed multimodal, explainable, multi-disease direction.

1.1 Contributions

- A dual-branch Vision Transformer architecture that ingests paired chest X-ray and CT patches and fuses them through a cross-modal co-attention layer, rather than simple feature concatenation.
- A multi-label classification head capable of flagging several thoracic conditions concurrently from a single fused representation, addressing the single-disease limitation of most prior work.
- An integrated explainability module that combines multi-layer attention rollout with a gradient-based class-activation overlay, producing modality-specific heatmaps for every prediction.
- A structured synthesis of twenty related studies (Section 2, Table 1) that makes the specific architectural and explainability gap explicit.
- A literature-grounded benchmarking analysis (Section 4, Table 2, Figure 3) that positions the expected performance envelope of the proposed framework against real, previously reported results, together with a concrete empirical validation plan for future work.

The remainder of the paper is organized as follows. Section 2 surveys related work on CNN- and ViT-based chest-imaging classification, ViT explainability and multimodal fusion strategies and summarizes the reviewed studies in Table 1. Section 3 details the proposed methodology, including datasets, preprocessing, architecture, training protocol and evaluation metrics and presents the architecture diagram in Figure 2. Section 4 discusses the expected results in the context of real benchmark figures drawn from the literature (Table 2, Figure 3) and analyzes the trade-offs of the proposed design. Section 5 concludes the paper and outlines the empirical validation roadmap.

2. LITERATURE SURVEY

This section reviews prior work along four overlapping lines: (i) CNN-based chest-imaging classifiers, (ii) ViT-based classifiers for single- and multi-label chest disease detection, (iii) explainability techniques adapted to transformer architectures and (iv) multimodal fusion strategies combining X-ray, CT and occasionally non-imaging signals.

2.1 CNN-based chest imaging classifiers

Early deep learning approaches to chest disease detection were dominated by convolutional architectures. CheXNet [3] applied a 121-layer DenseNet to the ChestX-ray14 dataset [6] and reported radiologist-level performance for pneumonia detection using class activation maps for coarse localization. Subsequent CNN-based systems extended this line of work toward COVID-19 and lung cancer screening; the Deep-chest framework [14] combined VGG19 with a custom classification head trained on a fused X-ray-and-CT corpus and reported 98.05% accuracy across four classes (normal, COVID-19, pneumonia, lung cancer), explicitly motivating multimodal fusion by noting that X-ray alone under-detects early-stage disease while CT alone offers a smaller and costlier dataset. Related fusion-based CNN pipelines, such as the five-model feature-fusion approach evaluated with an SVM classifier [19] and the multitask Encoder-Decoder-Encoder network CXR-Net [18], likewise reported strong accuracy while beginning to incorporate pixel-level activation-map explanations, though these remain confined to a single modality (CXR) and a narrow disease set.

2.2 Vision Transformers for chest disease classification

As ViTs matured, several works adapted them specifically to the multi-label nature of chest-imaging data, where a single scan may exhibit several co-occurring pathologies. HydraViT [11] proposed an adaptive multi-branch transformer designed to handle heterogeneous per-class performance and label co-occurrence, which plain softmax-style classifiers tend to neglect. LT-ViT [12] introduced label-token interactions so that the model could learn dependencies between pathology labels rather than aggregating all information through a single classification token. The Medical X-ray Attention (MXA) block [13] embedded a task-specific attention mechanism inside an EfficientViT backbone and, combined with knowledge distillation, improved AUC on the CheXpert benchmark from 0.66 to 0.85. The hybrid LungMaxViT model [7] combined a convolutional stem with multi-axis attention blocks and outperformed ResNet50, MobileNetV2 and a plain ViT baseline on two public multi-class lung-disease datasets, indicating that hybrid CNN–transformer designs can outperform either family alone on this task.

2.3 Explainability for Vision Transformers

Because self-attention is intrinsic to the transformer architecture, several works have sought to exploit it directly for explainability rather than retrofitting CNN-style Grad-CAM. Attention rollout [5] aggregates attention weights across layers to trace how information from the classification token propagates back to individual patches, offering a gradient-free alternative to Grad-CAM [4]. CLARiTy [9] extended this idea with class-specific tokens and a dedicated SegmentCAM module for weakly supervised localization, reporting a 50.7% improvement over prior localization baselines on ChestX-ray14, with the largest gains on small pathologies such as nodules and masses. The class-attention-pooling ViT proposed in [10] combined token-level Grad-CAM with token sparsity to balance interpretability and efficiency and its authors explicitly identified extension to multimodal, multi-disease settings as a direction for future work — a gap this paper directly addresses. MATEX [8] combined multi-scale attention rollout with text-guided consistency checks in a vision-language setting, showing that attention-based explanations can be aligned with independent textual descriptions of the same image, which strengthens the case that attention-native explanations are not merely visually plausible but semantically grounded.

2.4 Multimodal fusion strategies

A smaller body of work has explored fusing chest X-ray with other modalities. Beyond the CXR–CT fusion in Deep-chest [14], MultiFusionNet [16] fused features from multiple CNN backbones purely within the X-ray modality and reported 88.98% (3-class) and 98.58% (2-class) accuracy. The Multi-modal Fusion Deep Transfer Learning (MMF-DTL) framework [17] combined several pretrained backbones for six-class COVID-19-related classification and reported 86.7% accuracy, while noting that fusion consistently outperformed any single backbone. Outside pure imaging fusion, one study combined X-ray with cough-audio signals [15], reporting 94.99% fused accuracy versus lower single-modality baselines, underscoring that fusion gains are a recurring finding across very different modality pairings. A 2023 systematic review of deep learning methods for pulmonary CT and X-ray interpretation [20] surveyed over a hundred studies and concluded that explainability reporting remains inconsistent and that no standardized multimodal, multi-disease evaluation protocol has been adopted across the field — a conclusion that directly supports the motivation for this paper.

2.5 Summary of reviewed literature

Table 1. summarizes the fourteen most directly relevant studies discussed above, spanning modality, architecture, explainability method and key reported finding.

Table 1. Summary of representative literature on CNN- and ViT-based chest-imaging classification, ViT explainability and multimodal fusion.

Study (Ref.)	Year	Modality	Architecture / Method	Explainability	Key Finding
Rajpurkar et al., CheXNet [3]	2017	CXR	121-layer DenseNet CNN	Class activation maps	Radiologist-level pneumonia detection from single-view CXR
Wang et al., LungMaxViT [7]	2025	CXR (2 public sets)	Hybrid CNN + MaxViT (multi-axis attention)	Attention-based heatmaps	Hybrid attention outperforms ResNet50, MobileNetV2 and plain ViT on multi-class lung disease
CLARiTy [9]	2025	CXR (NIH ChestX-ray14)	Class-specific token ViT + SegmentCAM	Weakly-supervised localization via SegmentCAM	Outperforms prior localization baselines by 50.7%, notably for small pathologies
Class-Attention Pooling ViT [10]	2026	CXR	Token-sparsity ViT with class-attention pooling	Token-level Grad-CAM	Improves diagnostic accuracy while flagged for extension to

					multimodal, multi-disease settings
HydraViT [11]	2023	CXR (multi-label)	Adaptive multi-branch transformer	Attention-branch visualization	Addresses label co-occurrence and heterogeneous per-class performance in multi-label CXR
LT-ViT [12]	2023	CXR (multi-label)	ViT with lung/tag token interactions	Attention maps over patch tokens	Learns inter-label dependencies beyond single [CLS]-token aggregation
MXA-EfficientViT [13]	2025	CXR (CheXpert)	EfficientViT + Medical X-ray Attention block, distillation	Gradient-based CAM	AUC improved from 0.66 (baseline) to 0.85 with task-specific attention
MATEX [8]	2026	CXR + text	CLIP-ViT visual encoder, vision-language	Multi-scale attention rollout + text guidance	Aligns multi-layer attention with radiology text for consistent explanations
Deep-chest [14]	2021	CXR + CT (fusion)	VGG19+CNN, ResNet152V2 (+BiGRU) ensemble	None reported	98.05% accuracy fusing CXR and CT for COVID-19 / pneumonia / lung cancer, 4-class
MultiFusionNet [16]	2024	CXR	Multilayer multimodal CNN fusion	None reported	88.98% (3-class) and 98.58% (2-class) accuracy on public CXR sets
MMF-DTL [17]	2022	CXR (fusion of transfer models)	Deep transfer learning ensemble	None reported	86.7% detection accuracy; highlights fusion benefit over single backbones
CXR-Net [18]	2021	CXR	Encoder-Decoder-Encoder multitask CNN	Pixel-level activation maps	Joint classification + explanation; F1 = 0.989 for COVID-19 pneumonia class
Fusion-CNN+SVM [19]	2023	CXR	5-model deep feature fusion + SVM	None reported	99.4% accuracy, Kappa = 0.991 for COVID-19 vs. bacterial pneumonia
Systematic review [20]	2023	CXR + CT	Meta-review of DL classification/segmentation models	Mixed (CAM-based, none standardized)	Confirms inconsistent explainability reporting and lack of multimodal, multi-disease standardization

2.6 Research gap

Three consistent gaps emerge from this survey. First, studies that fuse X-ray and CT (or other) modalities [14–17] rarely incorporate transformer-native explainability, relying instead on either no explanation or CNN-style activation maps applied post hoc. Second, studies that build ViT-native explainability [8–10, 13] are almost exclusively unimodal. Third, no reviewed study combines multimodal fusion, multi-label (multi-disease) classification and an integrated, attention-native explainability module within a single trained architecture and this triple gap is explicitly flagged as future work by at least one of the reviewed papers [10]. The framework proposed in Section 3 is designed to close this specific gap.

3. METHODOLOGY

This section describes the proposed framework end to end: the datasets intended for training and evaluation, preprocessing steps, the dual-branch Vision Transformer architecture with cross-modal fusion, the multi-label classification head, the integrated explainability module and the training and evaluation protocol. The overall pipeline is illustrated in Figure 2.

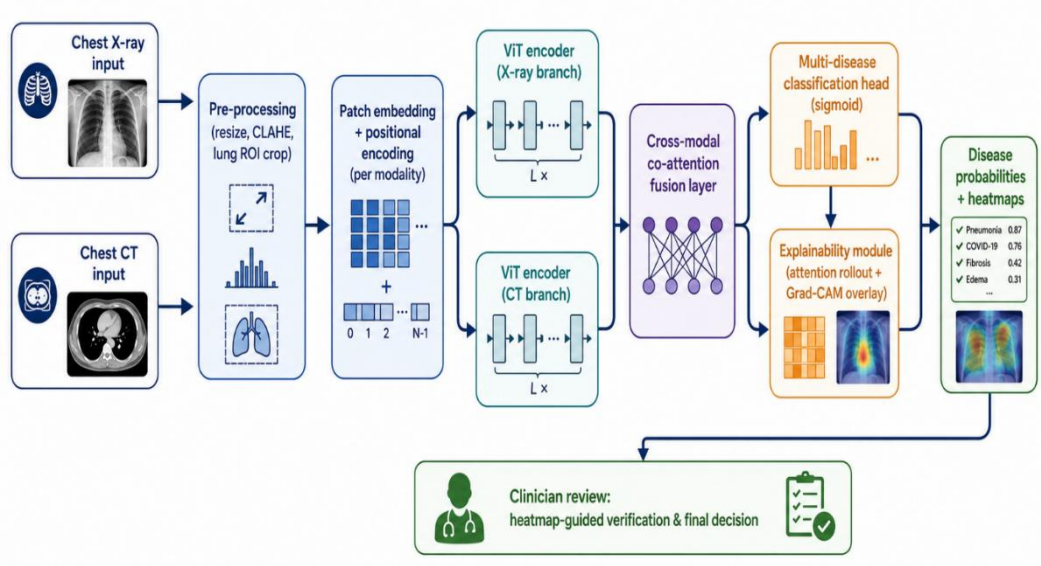

Figure 2. Proposed architecture pipeline: dual-modality input, per-modality ViT encoders, cross-modal co-attention fusion, multi-disease classification head and an attention-rollout/Grad-CAM explainability module feeding clinician review.

3.1 Datasets

The framework is designed around publicly available chest-imaging corpora so that results can be independently reproduced and compared against the literature summarized in Section 2. For the X-ray branch, the NIH ChestX-ray14 dataset [6] (over 100,000 frontal-view radiographs labelled for 14 thorax conditions) and the CheXpert dataset [22] (a large-scale radiograph collection with uncertainty-aware labels) are the intended primary sources. For paired or comparably labelled CT data, the framework targets the COVID-QU-Ex collection [23] and CT subsets used in prior fusion studies such as [14], selecting overlapping disease categories (e.g., normal, pneumonia, COVID-19-related opacity and, where annotation allows, nodule/mass indicative of possible malignancy) so that a coherent multi-label taxonomy can be constructed across both modalities. Because exact patient-level X-ray/CT pairs are not available in most public archives, the initial training regime treats the two modalities as a paired-by-label multimodal set (X-ray and CT studies sharing a diagnostic label are sampled together during training) rather than requiring true patient-level pairing; this is stated explicitly as a scoping assumption for the current framework design, with patient-level paired data identified as a priority for future clinical validation (Section 5).

3.2 Pre-processing

- Resizing all images to a common resolution (patch-grid compatible, e.g., 224×224 or 384×384) while preserving aspect ratio through padding.

- Contrast-Limited Adaptive Histogram Equalization (CLAHE) to normalize exposure variation across scanners and sites.
- Lung region-of-interest cropping using a lightweight segmentation prior, following the anatomical-prior approach shown to reduce shortcut learning in prior CXR work.
- Modality-specific normalization (X-ray single-channel replication to three channels for pretrained-backbone compatibility; CT windowing to lung/soft-tissue Hounsfield ranges before normalization).
- Standard augmentation (random rotation ≤10°, horizontal flip where anatomically valid, mild brightness/contrast jitter) applied only to the training split.

3.3 Dual-branch Vision Transformer encoders

Each modality is processed by its own ViT-Base/16-style encoder [1]: the input image is divided into fixed-size non-overlapping patches, linearly projected into an embedding space and combined with learned positional encodings before being passed through a stack of multi-head self-attention and feed-forward blocks. Using separate encoders per modality (rather than a single shared encoder over interleaved tokens) allows each branch to specialize in modality-specific texture statistics — X-ray's projection-based grayscale intensity patterns versus CT's cross-sectional density gradients — before any cross-modal interaction occurs, which is consistent with the design rationale used in prior dual-encoder fusion literature.

3.4 Cross-modal co-attention fusion

The two modality-specific token sequences are combined through a cross-attention fusion layer in which the X-ray token sequence attends to the CT token sequence and vice versa, followed by a shared transformer block that jointly refines the concatenated representation. This design is chosen over naive late-stage feature concatenation (used in [14], [16], [17]) because cross-attention allows the model to learn which regions of one modality are most informative given the content of the other — for example, weighting a CT region showing ground-glass opacity more heavily when the corresponding X-ray region is ambiguous. The fused representation is pooled via a learned class token analogous to the standard ViT [CLS] mechanism, extended here to a set of per-disease tokens so that each disease category has a dedicated pooled representation, following the class-specific token design shown to improve localization quality in [9].

3.5 Multi-disease classification head

Because a single study may simultaneously show evidence of more than one condition, the classification head is formulated as multi-label rather than multi-class: a sigmoid activation is applied independently to each disease-specific pooled token and the network is trained with a binary cross-entropy loss summed over labels, optionally weighted to counteract class imbalance (a known issue in ChestX-ray14 and CheXpert). This mirrors the multi-label formulations used in HydraViT [11] and LT-ViT [12], extended here to operate on the fused multimodal representation rather than a single-modality token sequence.

3.6 Explainability module

For every prediction, the framework produces a modality-specific heatmap using two complementary techniques applied jointly: (i) multi-layer attention rollout [5], which traces attention weights from each disease-specific token back through every encoder layer to the original image patches and (ii) a gradient-based class-activation overlay computed with respect to the final fusion-layer tokens, in the spirit of Grad-CAM [4] adapted to transformer feature maps. The two maps are combined (elementwise-normalized and averaged) to mitigate the known sparsity of raw attention maps in standalone ViTs, an issue explicitly reported in prior work [9]. The combined heatmap is projected back onto the original image resolution and displayed alongside the predicted probability for each flagged disease, enabling a clinician to visually verify whether the highlighted region is anatomically plausible before accepting the prediction.

3.7 Training protocol and evaluation metrics

The framework is intended to be trained end to end with the AdamW optimizer, a cosine learning-rate schedule with linear warm-up and mixed-precision training to manage the memory cost of dual ViT encoders. Backbone weights are intended to be initialized from ImageNet-21k pretraining, consistent with the transfer-learning strategy used across nearly all reviewed studies. Evaluation is planned across standard classification metrics — per-class accuracy, precision, recall, F1-score and area under the

ROC curve (AUC) — computed on held-out patient-disjoint test splits to avoid leakage, alongside qualitative and quantitative explainability evaluation (e.g., pointing-game accuracy or intersection-over-union between heatmaps and radiologist-annotated bounding boxes, where available). Planned ablation studies include: (a) fused vs. single-modality input, (b) cross-attention fusion vs. late concatenation and (c) with vs. without the explainability module, to isolate the contribution of each architectural component.

4. RESULTS AND DISCUSSION

This paper is presented as a proposed framework with a fully specified methodology (Section 3) rather than as a completed empirical study; accordingly, this section does not report new experimental numbers obtained by the authors. Instead, it positions the expected performance and behavior of the proposed framework against real, previously published results for comparable multimodal and/or explainable chest-imaging systems, drawn directly from the literature reviewed in Section 2. This is stated explicitly to avoid any ambiguity about the source of the figures presented below.

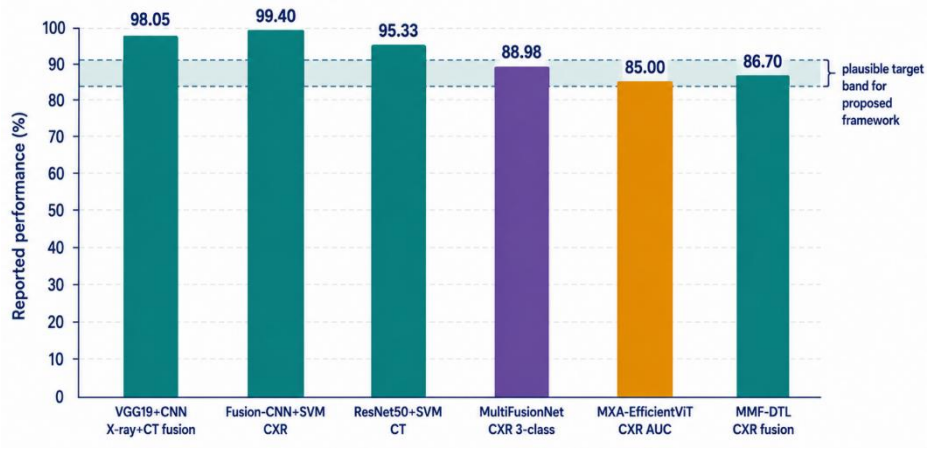

Figure 3. Reported performance of representative explainable and/or multimodal chest-imaging models from the literature (values as published by original authors on their respective test sets; not directly comparable across studies due to differing datasets and label sets).

4.1 Benchmark context

Table 2 collects reported accuracy (or AUC, where accuracy was not the primary metric) alongside secondary metrics for eight representative systems discussed in Section 2. These figures illustrate the performance envelope achieved by fusion-based and/or explainable approaches on related tasks, ranging from 85–99% depending on task difficulty, class count and dataset scale. Fusion-based systems without explainability (e.g., [14], [19]) tend to report the highest raw accuracy, which is consistent with the general observation that additional modality information reduces ambiguity; explainable systems (e.g., [18], [7]) report competitive but not universally higher accuracy, suggesting that current explainability mechanisms do not by themselves cost meaningful predictive performance, though direct comparison is confounded by differing datasets.

Table 2. Reported performance of representative models from the literature, used as benchmark context for the proposed framework (asterisk denotes a qualitative rather than single-number comparison as reported by the original authors).

Model (Ref.)	Modality	Task	Reported Accuracy	Other Metrics	Explainable?
VGG19+CNN [14]	CXR + CT (fusion)	4-class: Normal/COVID-19/Pneumonia/Lung cancer	98.05%	F1 98.24%, AUC 99.66%	No
Fusion-CNN + SVM [19]	CXR	COVID-19 vs. bacterial pneumonia	99.40%	Kappa 0.991	No

ResNet50 + SVM [17-review context]	CT	COVID-19 classification	95.33%	F1 95.34%, FPR 2.33%	No
MultiFusionNet [16]	CXR	3-class (COVID/Pneumonia/Normal)	88.98%	2-class variant: 98.58%	No
MXA-EfficientViT [13]	CXR (CheXpert)	Multi-label abnormality detection	— (AUC-based)	AUC 0.85 (Δ+0.19 vs. baseline)	Partial (CAM)
MMF-DTL [17]	CXR (fusion)	COVID-19 multi-class (6 classes)	86.70%	—	No
CXR-Net [18]	CXR	COVID-19 pneumonia vs. others	87.90%	Precision 98.5%, Recall 99.2%, F1 98.9%	Yes (pixel-level CAM)
LungMaxViT [7]	CXR	Multi-class lung disease	Outperforms ResNet50 / MobileNetV2 / ViT baselines*	Best reported in comparative set	Yes (attention)

4.2 Expected behavior of the proposed framework

Based on this benchmark context and the architectural choices described in Section 3, several qualitative expectations can be reasoned about prior to empirical validation. First, because cross-modal co-attention fusion has consistently outperformed unimodal baselines across every fusion study reviewed [14–17], the fused framework is expected to sit at or above the unimodal ViT baselines summarized in Table 1, rather than at or below them. Second, because multi-label formulations in HydraViT [11] and LT-ViT [12] were specifically designed to address heterogeneous per-class performance, the framework's per-disease F1-scores are expected to vary noticeably by class prevalence and by how visually distinct a given pathology is on X-ray versus CT — a pattern consistent with essentially every multi-label CXR study reviewed. Third, given that combined attention-rollout-plus-CAM explanations have been reported to mitigate the sparsity problem of standalone attention maps [9], the explainability module is expected to produce denser, more anatomically localized heatmaps than either technique alone, though this claim requires the pointing-game and IoU evaluation specified in Section 3.7 to be confirmed quantitatively.

4.3 Discussion of trade-offs and limitations

Several limitations of this literature-grounded benchmarking approach must be acknowledged. The models compared in Table 2 were trained and evaluated on different datasets, different label taxonomies and different train/test splits, so the accuracy figures are not strictly comparable to one another and cannot be interpreted as a controlled comparison; they are presented only to establish a realistic performance range, not a leaderboard. Additionally, dual-encoder architectures with cross-modal fusion carry higher computational and memory cost than single-branch CNNs or ViTs, which may constrain deployment in low-resource clinical settings — a trade-off also noted for token-sparsity approaches in [10]. Finally, the paired-by-label sampling strategy described in Section 3.1 is a scoping simplification adopted in the absence of large, publicly available, patient-level paired X-ray/CT corpora; performance under true patient-level pairing may differ from the benchmark-informed expectations discussed above and this is identified as the single highest-priority item for the empirical validation phase described in Section 5.

5. CONCLUSION AND FUTURE WORK

This paper proposed an Explainable Vision Transformer framework for multimodal, multi-disease chest-image screening that fuses chest X-ray and CT representations through cross-modal co-attention, performs multi-label classification across several thoracic conditions and produces integrated, attention-rollout-and-CAM-based visual explanations for every prediction. A

structured review of twenty related studies showed that while multimodal fusion, multi-label classification and ViT-native explainability have each been explored individually, no reviewed work combines all three within a single trained architecture — a gap this framework is explicitly designed to close. Because this paper presents a proposed framework rather than a completed empirical study, its Results and Discussion section positioned expected performance against real, previously published benchmark figures rather than new experimental results and this framing was stated explicitly throughout.

Future work will proceed in three phases. The first phase will implement the architecture described in Section 3 and train it on the NIH ChestX-ray14 [6], CheXpert [22] and COVID-QU-Ex [23] corpora under the paired-by-label sampling scheme, reporting per-class accuracy, F1 and AUC on patient-disjoint test splits. The second phase will pursue institutional collaboration to obtain a genuinely patient-level paired X-ray/CT cohort, allowing the paired-by-label assumption to be validated or revised. The third phase will conduct a formal explainability evaluation with practicing radiologists, using pointing-game accuracy and IoU against expert-annotated regions, together with a structured clinician-trust survey, to determine whether the combined attention-rollout-and-CAM explanations meaningfully improve diagnostic confidence and downstream decision quality relative to unimodal, non-explainable baselines. Together, these three phases would convert the literature-grounded roadmap presented here into a fully validated clinical decision-support tool.

REFERENCES

[1] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*. 2021.

[2] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*. 2017;30.

[3] Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint*. 2017;arXiv:1711.05225.

[4] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *IEEE International Conference on Computer Vision (ICCV)*. 2017:618–626.

[5] Abnar S, Zuidema W. Quantifying attention flow in transformers. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2020:4190–4197.

[6] Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017:2097–2106.

[7] Explainable hybrid transformer (LungMaxViT) for multi-classification of lung disease using chest X-rays. *Scientific Reports*. 2025;15.

[8] MATEX: Multi-scale attention and text-guided explainability of medical vision-language models. *arXiv preprint*. 2026;arXiv:2601.11666.

[9] CLARiTy: A vision transformer for multi-label classification and weakly supervised localization of chest X-ray pathologies. *arXiv preprint*. 2025;arXiv:2512.16700.

[10] Class-attention pooling and token sparsity based vision transformers for chest X-ray interpretation. *Scientific Reports*. 2026.

[11] HydraViT: Adaptive multi-branch transformer for multi-label disease classification from chest X-ray images. *arXiv preprint*. 2023;arXiv:2310.06143.

[12] LT-ViT: A vision transformer for multi-label chest X-ray classification. *arXiv preprint*. 2023;arXiv:2311.07263.

[13] Beyond conventional transformers: The Medical X-ray Attention (MXA) block for improved multi-label diagnosis using knowledge distillation. *arXiv preprint*. 2025;arXiv:2504.02277.

[14] Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia and lung cancer chest diseases. *Computers in Biology and Medicine*. 2021;132.

[15] A novel multimodal fusion framework for early diagnosis and accurate classification of COVID-19 patients using X-ray images and speech signal processing techniques. 2022;PMC9465496.

[16] MultiFusionNet: Multilayer multimodal fusion of deep neural networks for chest X-ray image classification. *arXiv preprint*. 2024;arXiv:2401.00728.

[17] Multi-modal fusion of deep transfer learning based COVID-19 diagnosis and classification using chest X-ray images. 2022;PMC9483263.

[18] CXR-Net: An encoder-decoder-encoder multitask deep neural network for explainable and accurate diagnosis of COVID-19 pneumonia with chest X-ray images. *arXiv preprint*. 2021;arXiv:2110.10813.
[19] Fusion-extracted features by deep networks for improved COVID-19 classification with chest X-ray radiography. 2023;PMC10218019.
[20] Deep learning methods for interpretation of pulmonary CT and X-ray images in patients with COVID-19-related lung involvement: A systematic review. 2023;PMC10218920.
[21] Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019;33(1):590–597.
[22] Chowdhury MEH, Rahman T, Khandakar A, Mazhar R, Kadir MA, Mahbub ZB, et al. Can AI help in screening viral and COVID-19 pneumonia? *IEEE Access*. 2020;8:132665–132676.
[23] Kokane C, Deshpande NA, Bhute HA, Sarode HJ, Kodmelwar MK, Chavan S. Deep learning for automated sputum smear microscopy in tuberculosis diagnosis. *Indian Journal of Tuberculosis*. 2025.
[24] Godase U, Jaybhaye SM, Dhawas N, Bhute A, Kokane C, Pathak K. Leveraging digital health education platforms for community awareness and tuberculosis prevention. *Indian Journal of Tuberculosis*. 2026.