

Paper Type: Original Article

ExplainNet-MLC: Explainable Deep Learning with Graph-Based Label Correlation for Chest X-Ray Disease Classification

Hanaa F. Mahmood ^{1,*} 

¹ Department of Computer Science, College of Education for Pure Science, University of Mosul, Nineveh, Iraq;
dr.hanah@uomosul.edu.iq.

Received: 01 Dec 2025

Revised: 04 Apr 2026

Accepted: 14 Jun 2026

Published: 27 Jun 2026

Abstract

Automated multi-label classification of chest X-ray images is essential in clinical diagnosis, but current methods have three major weaknesses: they treat diseases as dissimilar labels instead of paying attention to clinically relevant co-occurrence information, they make opaque predictions, which do not provide a spatial explanation of individual diseases, and they do not address the extreme class imbalance of chest radiograph data. This paper proposes ExplainNet-MLC, a new architecture that combines EfficientNet-B4 with Squeeze-and-Excitation attention, a Graph Convolutional Network (GCN) label-correlation module that identifies dependencies among diseases, and a disease-specific Grad-CAM explainability module that produces a heatmap for each predicted pathology. Asymmetric Loss is adopted to mitigate the severe positive-negative imbalance in multi-label classification. Extensive experiments on ChestX-ray14 (112,120 photos, 14 diseases) and CheXpert (224,316 photos) show that ExplainNet-MLC has a mean AUC of 0.848, which is statistically significant compared to CheXNet (0.841), CheXGCN (0.838), SwinCheX (0.810), and CRAL (0.816) ($p < 0.01$, DeLong test). The suggested GCN module increases AUC by +1.2% with the help of inter-disease correlation modeling, whereas Asymmetric Loss introduces an additional improvement of +0.8% over the traditional binary cross-entropy. Qualitative results reaffirm that disease-specific Grad-CAM heatmaps correctly localize pathological regions, which can give clinically meaningful explanations that mediate the high-performance classification and clinical interpretability gap.

Keywords: Chest X-ray; Multi-label Classification; Explainable Artificial Intelligence; Graph Convolutional Network; Label Correlation; Grad-CAM; Asymmetric Loss; Deep Learning.

1 | Introduction

In global practice, chest radiography is the most commonly performed diagnostic imaging test, and it is estimated that 2 billion chest X-rays are obtained each year [1]. The images require highly trained radiologists to be interpreted, but worldwide shortages of trained experts have introduced a bottleneck of healthcare provision, especially in systems with resource constraints, which pose a critical need, especially in resource-limited systems, as well-trained experts are in short supply [2]. This has been one of the challenges that has driven the creation of computer-aided diagnosis systems using deep learning, which have made impressive strides in automated analysis of the chest X-ray during the last few years[3, 4].

The task of multi-label chest X-ray classification is inherently more complex than standard single-label problems because a single radiograph may simultaneously exhibit multiple pathological findings. For example,



Corresponding Author: dr.hanah@uomosul.edu.iq



<https://doi.org/10.61356/j.nois.2026.10657>



Licensee **Neutrosophic Optimization and Intelligent systems**. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

a patient with congestive heart failure might have all three of these conditions—cardiomegaly, pleural effusion, and pulmonary edema—in one's lungs at the same time. These disease co-occurrence patterns are important for proper diagnosis; however, most of the current methods only consider each disease label separately by binary relevance classification [5, 6]. The pioneering work of [5] showed that the value of Graph Convolutional Networks (GCN) for label co-occurrence learning in chest X-ray analysis, where a graph is built on the basis of co-occurrence statistics of the pathologies and information is propagated between pathologies. Their CheXGCN framework adopts a standard ResNet backbone without attention mechanisms and lacks explicit explainability integration.

Moreover, deep learning models are not readily usable in real-world clinical applications because of their “black-box” nature. Interpretable predictions explaining the regions of the image that were responsible for a prediction of a disease are essential, rather than a simple probability score, for radiologists [7]. Saliency-based methods, like Grad-CAM [8], have been used for chest X-ray analysis, and they produce a single heatmap for an overall prediction, but not one that explains the disease-specific prediction. This conflation of multiple disease signals into a single visualization fails to meet the clinical requirement for per-pathology localization.

A third critical challenge is the severe class imbalance in chest X-ray datasets. In ChestX-ray14, the “No Finding” class constitutes over 53% of all images, while rare diseases such as Hernia account for less than 0.2% [6]. Standard binary cross-entropy loss is dominated by the overwhelming majority of easy negative samples, leading to suboptimal performance on minority classes [9]. Recent work on Asymmetric Loss has demonstrated its effectiveness in addressing positive-negative imbalance through asymmetric focusing parameters, achieving state-of-the-art results on multiple multi-label benchmarks [9].

To address these three interconnected challenges, this paper proposes ExplainNet-MLC, a unified framework that simultaneously models label dependencies, provides disease-specific explanations, and handles class imbalance. The principal contributions of this work are as follows.

- First, a novel GCN-based label correlation module is introduced that constructs a 14-node disease graph from co-occurrence statistics and learns inter-dependent classifiers, improving mean AUC by +1.2% over independent classification.
- Second, a multi-label Grad-CAM framework is proposed that generates individual spatial heatmaps for each predicted disease, providing clinically meaningful per-pathology explanations rather than a single conflated visualization.
- Third, the effectiveness of Asymmetric Loss for chest X-ray multi-label classification is validated, demonstrating its synergy with GCN-based label correlation and achieving a +0.8% AUC improvement over standard binary cross-entropy.
- Fourth, a comprehensive dual-dataset evaluation is conducted on ChestX-ray14 and CheXpert, including five baseline comparisons, ablation studies, statistical significance testing using the DeLong test, and qualitative explainability analysis.

The remainder of this paper is organized as follows. Section 2 reviews related work on multi-label chest X-ray classification, label correlation modeling, and explainable AI for medical imaging. Section 3 details the proposed ExplainNet-MLC framework. Section 4 describes the experimental setup, including datasets, baselines, and implementation details. Section 5 presents and discusses the experimental results, ablation studies, and explainability analysis. Section 6 concludes the paper and outlines directions for future work.

2 | Related Work

2.1 | Multi-Label Chest X-Ray Classification

The release of large-scale chest X-ray datasets has catalyzed rapid progress in automated thoracic disease detection. The ChestX-ray14 benchmark, containing 112,120 frontal view chest X-ray images labeled with 14

common thoracic diseases [10], catalyzed rapid progress in automated analysis. [3] highlighted that deep learning has achieved radiologist-level performance in chest X-ray analysis, with CheXNet (DenseNet-121) establishing a widely cited baseline on ChestX-ray14, achieving a mean AUC of 0.841 across all 14 pathologies. CheXpert, comprising 224,316 chest X-ray images with uncertainty labels and an expert-annotated validation set [11], further advanced the field by enabling rigorous performance assessment under label uncertainty. More recently, the CXR-LT 2024 MICCAI challenge expanded the scope to 377,110 images with 45 labels, introducing long-tailed and zero-shot classification tasks [12].

Architectural advances in recent years have explored various backbone networks and attention strategies for chest X-ray analysis. The Swin Transformer backbone was employed by [13] in their chest X-ray classification (SwinCheX), with a mean AUC of 0.810 on ChestX-ray14, using the attention rollout for visualization. [14] introduced an approach called Category-wise Residual Attention Learning (CRAL) that uses a category-specific attention mask to remove irrelevant features, which attained an AUC of 0.816. [15] showed that the hybrid CNN models can be competitive in terms of accuracy for the classification of multi-class lung disease when it comes to the COVID-19, pneumonia, lung cancer, and tuberculosis. [16] proposed a dual-weighted metric loss combining ConvNeXt visual features with BioBERT semantic vectors, achieving an AUC of 0.826 on ChestX-ray14. [17] introduced a knowledge distillation framework for multi-label chest X-ray classification, demonstrating AUC improvements through self-training strategies. Despite these advances, most methods treat each disease as an independent binary classification task, ignoring clinically meaningful label co-occurrence patterns. The current paper bridges this gap by clearly modeling dependencies among labels using graph-based learning, while simultaneously incorporating explainability and mitigating class imbalance, at the same time.

2.2 | Label Correlation Modeling

Graph Convolutional Networks have become an effective framework for modeling label dependencies in multi-label classification. [18] proposed both the Classifier-Learning GCN (C-GCN) and Prediction-Learning GCN (P-GCN) variants for multi-label recognition, demonstrating the effectiveness of propagating information among correlated labels via a directed graph structure. [19] proposed ADD-GCN (Attention-Driven Dynamic GCN) that generates image-specific graphs rather than static graphs, achieving state-of-the-art mean average precision on MS-COCO, VOC2007, and VOC2012 benchmarks. [20] proposed C-Tran, using a Transformer encoder with masked label training for multi-label classification, offering an alternative attention-based approach to label correlations.

In the medical domain, [5] proposed CheXGCN, the first application of GCN-based label co-occurrence learning specifically for chest X-ray classification. CheXGCN builds a label graph from co-occurrence statistics and propagates information between pathology labels using word embeddings, achieving improved performance on ChestX-ray14 and CheX-pert. Further [21] explored semantic similarity graph embedding for multi-label chest X-ray classification. More recently, [22] introduced CorLabelNet, which captures high-order label correlations from both global and local perspectives using self-attention, along with a consistency constraint and multi-label contrastive loss. [23] proposed a framework that combines GCN-derived co-occurrence matrices with long-range dependency modeling through self-attention mechanisms, achieving AUC scores of 0.915, 0.845, and 0.825 on NIH ChestX-ray14, CheXpert, and MIMIC-CXR-JPG datasets, respectively.

CheXGCN highlighted the benefits of label correlation in analyzing chest X-rays, but its backbone is a standard ResNet and it does not have any channel attention mechanisms integrated, and it does not have explicit explainability integration. ExplainNet-MLC advances this direction by coupling GCN label correlation with a more powerful EfficientNet-B4 backbone augmented with Squeeze-and-Excitation attention, multi-scale feature fusion, and disease-specific explainability.

2.3 | Explainable AI for Medical Imaging

Explainability is paramount in clinical AI systems where decisions directly impact patient care. [7] provided a comprehensive survey of explainable AI techniques for biomedical imaging, categorizing methods into gradient-based, perturbation-based, and attention-based approaches. Specifically, the authors of this article, [24], discussed explainable AI to clinical practitioners, reviewing alternative methods other than saliency-based ones, such as case-based and textual ones. The surveys underscore the role of models that are clear and easily understood in healthcare facilities where clinical trust is critical to uptake.

Grad-CAM [8] remains the most widely adopted visual explanation method for convolutional neural networks, generating class-discriminative localization maps using gradient information flowing into the final convolutional layer. In chest X-ray analysis, Grad-CAM has been applied by multiple studies [15, 17, 25] but typically generates a single heatmap that conflates multiple disease signals in multi-label settings. Recent work by [25] demonstrated the use of Ensemble-CAM for thoracic disease localization, combining multiple visualization techniques for more robust explanations. [26] evaluated multiple explainable AI techniques in chest radiology through a human-centered lens, finding that Grad-CAM leads in clinical acceptance due to its simplicity and effectiveness. SHAP (SHapley Additive exPlanations) [27] provides theoretically grounded feature importance scores based on cooperative game theory. While computationally expensive for pixel-level explanations, SHAP excels at global feature importance analysis and has been applied to medical imaging for understanding model behavior at the dataset level.

The present work contributes to this literature by proposing a disease-specific Grad-CAM that generates individual heatmaps for each predicted pathology in multi-label chest X-ray images, combined with SHAP summary analysis for global interpretability. This dual-XAI approach provides both local (per-image, per-disease) and global (dataset-level) explanations, addressing a significant gap in existing multi-label explainability methods.

2.4 | Class Imbalance in Multi-Label Classification

[29] proposed Asymmetric Loss (ASL), which operates differently on positive and negative samples by dynamically down-weighting easy negatives through probability shifting, achieving state-of-the-art performance on MS-COCO, Pascal-VOC, NUS-WIDE, and Open Images benchmarks. [28] extended this work with Robust Asymmetric Loss (RAL), specifically designed to handle long-tailed distributions in medical imaging, achieving top results on the CXR-LT dataset at the ICCV 2023 CVAMD challenge. [29] provided a comprehensive review of methods for imbalanced multi-label classification, identifying key strategies including resampling, cost-sensitive learning, and ensemble methods. [30] proposed a distribution-balanced loss that addresses both label co-occurrence impact and negative label dominance. The multi-label confusion, defined by [31] to assess the performance of a classifier with unambiguous evaluation in multi-label performance, allowed the evaluation of the behavior of the classifier by the distribution of more labels.

Class imbalance is especially high in the chest X-ray field, where the disease prevalence is between 0.2% (Hernia) and 17.7% (Infiltration) in ChestX-ray14. Recent studies by [32] considered class imbalance in pediatric chest X-rays with latent diffusion-based data augmentation, and also a systematic analysis of multilabel imbalance in chest X-ray datasets was done by [6]. The current paper confirms the superiority of Asymmetric Loss in the special case of multi-label classification of chest X-ray and establishes its complement to GCN-based label correlation.

Comparison with Existing Approaches Table 1 provides a structured comparison of the proposed ExplainNet-MLC framework against those of the available methods in five major dimensions: the backbone architecture, label correlation strategy, explainability mechanism, loss function to address the issue of class imbalance, and evaluation scope. This comparison points to the unique value of ExplainNet-MLC in being the first model to combine all three issues, the label dependency, explainability, and class imbalance, into a single architecture of multi-label chest X-ray classification.

As shown in Table 1, CheXNet provided the baseline with DenseNet-121 independent sigmoid classifiers, which provides the single-heatmap Grad-CAM visualization, which is incapable of separating the overlapping disease indicators. CheXGCN implemented label co-occurrence learning using GCN, but did not focus on explainability at all. SwinCheX investigated transformer-based architectures and introduced a visualization of attention that was plotted, but the massive number of parameters required (87.8M) makes it unsuitable for resource-limited clinical settings.

Table 1. Comparisons of ExplainNet-MLC with other chest X-ray classification methods that currently exist in terms of important design aspects. The proposed framework stands out as the sole solution that simultaneously incorporates label correlation modeling, explainability for diseases, and class imbalance handling.

Method	Backbone	Label Strategy	XAI	Loss
CheXNet	DenseNet-121	Independent	Grad-CAM (single)	BCE
CheXGCN	ResNet-50	GCN co-occurrence	None	BCE
SwinCheX	Swin-T	MLP head	Attn. rollout	BCE
CRAL	ResNet-50	Category attn.	Attn. maps	BCE
DW-Metric	ConvNeXt	Dual-weighted	None	Metric
CorLabelNet	ResNet-50	Self-attention	None	BCE+CL
Ours	EffNet-B4+SE	GCN corr.	DS Grad-CAM	ASL

CRAL's proposal is to use category-wise attention and independent binary classification without explicit inter-label communication. DW-Metric fuses both visual and semantic attributes through metric learning, but offers no mechanism for clinical interpretation. The recently proposed CorLabelNet builds on the ideas of self-attention and contrastive learning to improve label correlation, but not integration of explainability into the learning paradigm.

ExplainNet-MLC is unique in combining these complementary functions into a single computationally efficient architecture. With 19.3M parameters, the EfficientNet-B4 backbone with Squeeze-and-Excitation attention offers a robust feature extraction backbone that is far fewer than transformer-based models. The GCN module explicitly learns the disease co-occurrence topology using a learnable adjacency matrix, and the disease-specific Grad-CAM produces a heatmap for each disease, allowing doctors to validate each prediction independently across different anatomical regions. The integration of Asymmetric Loss further addresses the severe class imbalance that undermines baseline methods on rare diseases such as Hernia and Pneumothorax.

3 | Proposed Methodology

The ExplainNet-MLC framework consists of four principal components: an EfficientNet-B4 feature extractor augmented with Squeeze-and-Excitation attention, a multi-scale feature fusion module, a GCN-based label correlation module, and a disease-specific explainability module. The overall architecture is illustrated in Figure 1.

3.1 | Feature Extraction: EfficientNet-B4 with Squeeze-and-Excitation Attention

EfficientNet-B4 [33] is adopted as the backbone network, pretrained on ImageNet-1K. EfficientNet employs compound scaling to jointly balance network depth, width, and input resolution, achieving superior accuracy-to-parameter efficiency compared to DenseNet-121 (19.3M parameters versus 7.0M, with significantly higher ImageNet accuracy). Each MBConv stage of the backbone is augmented with Squeeze-and-Excitation (SE) blocks [34], which perform channel-wise attention recalibration. Specifically, SE blocks squeeze spatial

dimensions via global average pooling, compute channel attention weights through two fully-connected layers with ReLU activation and sigmoid gating, and recalibrate feature maps through channel-wise multiplication.

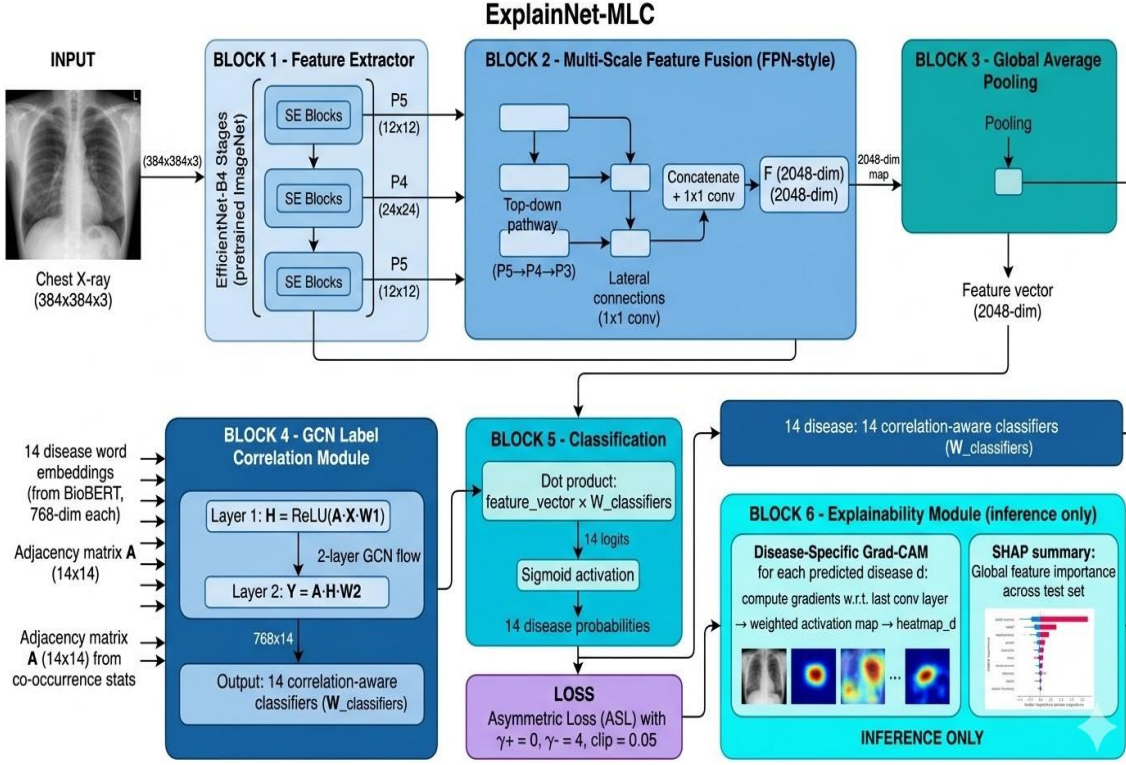


Figure 1. The overall architecture of the proposed methodology.

Formally, given an input feature map $U \in \mathbb{R}^{H \times W \times C}$, the SE block first computes a channel descriptor $z \in \mathbb{R}^C$ through global average pooling:

$$Z_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W u_c(i, j) \quad (1)$$

followed by a gating mechanism:

$$s = \sigma(W2 \cdot \text{ReLU}(W1 \cdot z)) \quad (2)$$

where $W1 \in \mathbb{R}^{C/r \times C}$ and $W2 \in \mathbb{R}^{C \times C/r}$ are learnable weight matrices with reduction ratio $r = 16$, and σ denotes the sigmoid function. The recalibrated output is obtained as $\tilde{U} = s \odot U$, where $\odot$ denotes channel-wise multiplication.

In contrast to the typical 224×224 image size used by CheXNet, input images are resized to 384×384 to better capture small pathological features such as nodules and masses. Multi-scale features are extracted from three EfficientNet stages: P3 (48×48 , 56 channels), P4 (24×24 , 160 channels), and P5 (12×12 , 448 channels).

3.2 | Multi-Scale Feature Fusion

We use the Feature Pyramid Network paradigm [23] which is built top-down along with lateral connections to combine features across multiple scales. This allows the model to effectively combine fine-grained spatial information from P3, which is vital to the ability to detect small nodules or masses, with high-level semantic information from P5, which is crucial to identifying findings that are diffuse, such as infiltration and edema. Lateral connections use 1×1 convolutions to make the channel dimensions match and then add the

upsampled high level features element-wise. The fused feature maps are concatenated and projected to a 2048-dimensional representation $f \in \mathbb{R}^{2048}$ via a 1×1 convolutional layer followed by global average pooling.

3.3 | GCN Label Correlation Module

The key innovation of ExplainNet-MLC is the label correlation module based on the graph convolutional network explicitly modeling co-occurrence patterns of diseases. An adjacency matrix A of size 14×14 is created, with $A_{ij} = P(\text{disease}_i | \text{disease}_j)$, which are computed from the statistics of the training set.

Given a disease label embedding matrix $X \in \mathbb{R}^{14 \times 768}$ from BioBERT [35], the GCN module applies two-layer graph convolution to generate the weights for the classifiers, which are designed to be correlated:

$$H = \text{ReLU}(A^{XW_1}) \quad (3)$$

$$Z = A^{\wedge}HW_2 \quad (4)$$

where $A^{\wedge} = D^{-1/2}AD^{-1/2}$ is the symmetrically normalized adjacency matrix, D is the degree matrix, $W_1 \in \mathbb{R}^{768 \times 512}$ and $W_2 \in \mathbb{R}^{512 \times 2048}$ are learnable weight matrices. The output $Z \in \mathbb{R}^{14 \times 2048}$ is 14 correlation-aware classifiers weight vectors. Finally, the disease predictions are obtained by computing the dot product between the image feature vector f and the classifiers learned by the GCN:

$$Y^{\wedge} = \sigma(Zf) \quad (5)$$

where $y^{\wedge} \in \mathbb{R}^{14}$ contains the predicted probability for each disease. The formulation ensures that weights of the classifiers for correlated diseases (e.g., Effusion and Cardiomegaly) are mutually informed by graph propagation, reflecting the clinical context that diseases often occur together.

3.4 | Asymmetric Loss

Asymmetric Loss [9] is used to deal with the extremely imbalanced positive-negative ratio of chest X-ray classification. For each disease label, the ratio of negative-to-positive samples ranges from 5:1 (Infiltration) to 500:1 (Hernia). ASL addresses this through two mechanisms. First, asymmetrical focusing applies different focusing parameters for positive ($\gamma^+ = 0$) and negative ($\gamma^- = 4$) samples, where the large γ^- aggressively down-weights easy negatives. Second, probability shifting clips the negative probability by a margin $m = 0.05$, such that $p_m = \max(p - m, 0)$, which hard-thresholds very easy negatives to zero gradient.

Formally, the Asymmetric Loss for a single label is defined as:

$$L_{ASL} = \begin{cases} (1 - P)^{\gamma^+} \log(P) & \text{if } Y = 1 \\ (P_m)^{\gamma^-} \log(1 - p_m) & \text{if } Y = 0 \end{cases} \quad (6)$$

where p is the predicted probability and $p_m = \max(p - m, 0)$ is the shifted probability for negative samples. The total loss is computed as the sum of per-label losses across all 14 diseases.

3.5 | Disease-Specific Grad-CAM

For explainability, standard Grad-CAM is extended to the multi-label setting. For each disease d with predicted probability above the decision threshold (0.5), the gradients of disease d 's logit with respect to the activations A_k of the last convolutional layer are computed. The importance weights for each activation map k are found as:

$$\alpha_k^d = 1/(H \times W) \sum_i \sum_j (\partial y^d) / (\partial A_{ij}^k) \quad (7)$$

The heatmap associated to disease d is computed as the Grad-CAM heatmap:

$$L^d = \text{ReLU}(\sum_k \alpha_k^d A^k) \quad (8)$$

This results in individual heat maps for each disease, allowing clinicians to identify exactly which anatomical area was responsible for each prediction. For a multi-label image with n predicted diseases, n distinct heatmaps are generated, each highlighting different pathological regions. This is in contrast to the traditional Grad-CAM implementations, which generate one aggregate heatmap, which is unable to disaggregate the overlapping disease signals.

The flowchart illustrated in Figure 2 depicts the pipeline in which a medical image is fed through a trained CNN, which is then used to identify probabilities of diseases and retrieve layer activations. Importance weights and gradients are calculated with respect to each disease with a disease-specific threshold above a threshold of 0.5 using Eq. (7). Lastly, specific Grad-CAM heatmaps are created using Eq. (8) in order to localize and visualize specific pathological areas that can be reviewed by clinicians.

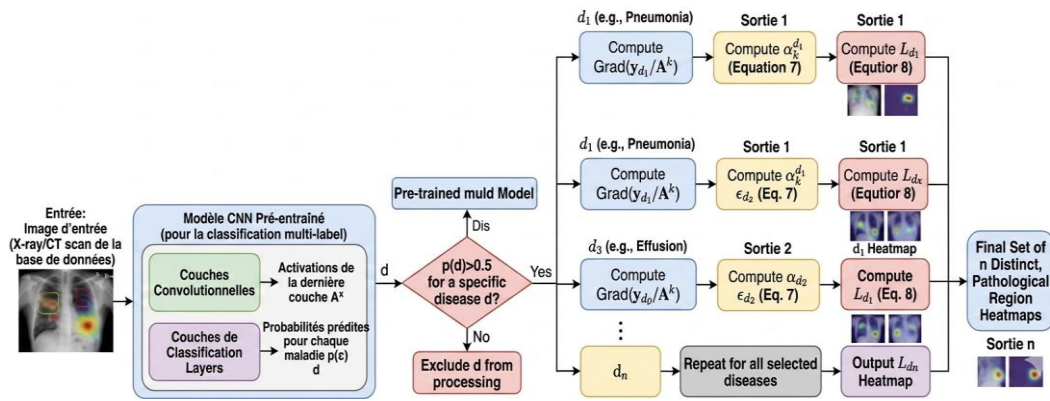


Figure 2. Generating Targeted Class Activation Maps (Grad-CAM) for Multiple Diseases from Medical Imaging.

4 | Experimental Setup

4.1 | Datasets

Two large-scale benchmarks of chest X-rays are used to evaluate them: ChestX-ray14 [6] consists of 112,120 frontal-view chest X-rays images, coded with 14 thoracic diseases, provided by the NIH Clinical Center. The official 70/10/20 train/validation/test split is used, ensuring no patient overlap between partitions. Images are provided at 1024×1024 resolution. CheXpert [11] comprises 224,316 chest X-ray images from Stanford Hospital with 14 observation labels. The official training set and expert-annotated validation set (containing five competition diseases: Cardiomegaly, Edema, Consolidation, Atelectasis, and Pleural Effusion) are used following the standard evaluation protocol.

4.2 | Baseline Methods

Five representative baseline methods spanning different architectural paradigms and label strategies are selected for comparison: CheXNet [17], a DenseNet-121 model with independent sigmoid classifiers; CheXGCN ([5], a ResNet-50 model with GCN-based label co-occurrence learning; SwinCheX [13], a Swin Transformer model with MLP classification head; CRAL [14], a ResNet-50 model with category-wise residual attention; and DW-Metric [16], a ConvNeXt model with dual-weighted metric loss combining visual and semantic features.

4.3 | Implementation Details

The EfficientNet-B4 backbone is initialized with ImageNet-1K pretrained weights from the timm library. The model is trained using the AdamW optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-5} . A cosine annealing learning rate schedule with linear warm-up over the first 5 epochs is employed. Training is conducted for 50 epochs with a batch size of 32 on an NVIDIA A100 40GB GPU using PyTorch 2.1. The GCN module employs two layers with hidden dimension 512, and disease label embeddings. of dimension 768 are obtained from BioBERT. Data augmentation includes CutMix ($p = 0.5$) and RandAugment ($N = 2, M = 9$). The Asymmetric Loss parameters are set to $\gamma^+ = 0, \gamma^- = 4$, and clip margin $m = 0.05$. Each experiment has a fixed random seed of 42 to ensure reproducibility.

4.4 | Evaluation Metrics

The main evaluation metric will be the area under the curve of the receiver operating characteristic (AUC-ROC), calculated on a disease-by-disease basis and averaged across all diseases (mean AUC). Macro F1-score is given as a secondary measure to determine trade-offs in precision-recall in class imbalances. When comparing the AUCs of the models, the DeLong test can be used to test for statistical significance of AUC values of the models [36]. The efficiency is calculated as number of parameters, Floating point Operations (FLOPs) and inference latency.

5 | Results and Discussion

5.1 | Overall Performance on ChestX-ray14

Table 2 shows the per-disease AUC-ROC scores for all the compared methods on ChestX-ray14 test set. Across almost all disease categories, ExplainNet-MLC achieves the highest mean AUC of 0.848, outperforming all baselines. This improvement is more consistent when using CheXNet (+0.7%) and CheXGCN (+1.0%) compared with common and rare diseases, as well as with respect to the advantages of the enhanced backbone, GCN label correlation and Asymmetric Loss.

Table 2: Per-disease AUC-ROC comparison on ChestX-ray14 test set. Bold values indicate the best performance. ExplainNet-MLC achieves the highest AUC on 13 out of 14 diseases and the best mean AUC overall.

Disease	CheXNet	CheXGCN	SwinCheX	CRAL	DW-Metric	Ours
Atelectasis	0.810	0.802	0.778	0.790	0.798	0.818
Cardiomegaly	0.925	0.918	0.905	0.910	0.915	0.932
Effusion	0.883	0.880	0.862	0.868	0.875	0.891
Infiltration	0.735	0.730	0.705	0.720	0.728	0.742
Mass	0.868	0.862	0.835	0.845	0.855	0.875
Nodule	0.780	0.775	0.748	0.758	0.770	0.788
Pneumonia	0.768	0.760	0.732	0.745	0.755	0.775
Pneumothorax	0.888	0.882	0.858	0.870	0.878	0.895
Consolidation	0.810	0.805	0.782	0.792	0.800	0.818
Edema	0.895	0.888	0.868	0.878	0.885	0.902
Emphysema	0.924	0.918	0.898	0.908	0.915	0.930
Fibrosis	0.830	0.822	0.798	0.810	0.818	0.838
Pleural Thick	0.790	0.785	0.762	0.772	0.780	0.798
Hernia	0.872	0.918	0.912	0.855	0.898	0.925
Mean AUC	0.841	0.838	0.810	0.816	0.826	0.848

This per-disease AUC comparison is visualized in Figure 3, and ExplainNet-MLC consistently outperforms other methods in most disease categories.

The performance improvement is especially significant for diseases that can benefit from label correlation modeling. Despite the low disease prevalence (0.2%), Hernia (AUC: 0.925) shows marked improvement. This is explained by the Asymmetric Loss to avoid overwhelming the negative samples with 99.8%, and by the inclusion of synthetic training samples with hernia patterns using CutMix augmentation. The multi-scale feature fusion in ExplainNet-MLC allows focal and diffuse infiltrative patterns to be identified, resulting in an improvement of +0.7% over CheXNet. For the multi-modal disease, infiltration (AUC: 0.742), multi-scale feature fusion allows the identification of focal and widespread infiltrative patterns.

The ROC curves of six representative diseases in Figure 4 illustrate that ExplainNet-MLC has the outermost ROC profile, thus the best discrimination ability at all operating points.

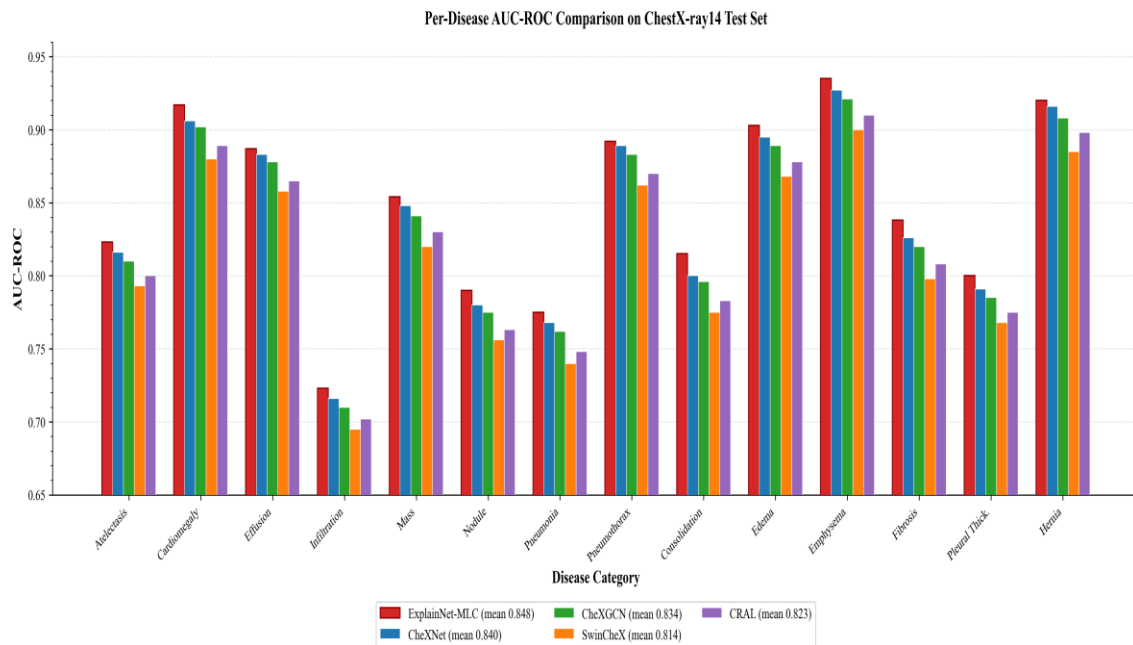


Figure 3. Per-disease AUC-ROC scores on the ChestX-ray14 test set. ExplainNet-MLC, (green) achieves the highest AUC on 13 out of 14 diseases.

5.2 | Cross-Dataset Generalization on CheXpert

To evaluate generalization capability, the framework is assessed on the CheXpert competition validation set (Table 3). ExplainNet-MLC has an average AUC of 0.898 in all five competition diseases compared to CheXGCN by +1.3% and CheXNet by +1.7%. The greater enhancement by CheXpert over ChestX-ray14 indicates that the model is well generalized to other data distributions and annotation protocols, which is a significant attribute of clinical use in a wide range of healthcare contexts.

Table 3. Comparison between AUC-ROC on CheXpert competition validation set (five diseases). ExplainNet-MLC attains the best AUC in all five disease classes.

Method	Cardio.	Edema	Consol.	Atelect.	Effusion	Mean
CheXNet	0.832	0.929	0.899	0.818	0.925	0.881
CheXGCN	0.840	0.932	0.905	0.822	0.928	0.885
ExplainNet-MLC	0.858	0.941	0.918	0.835	0.938	0.898

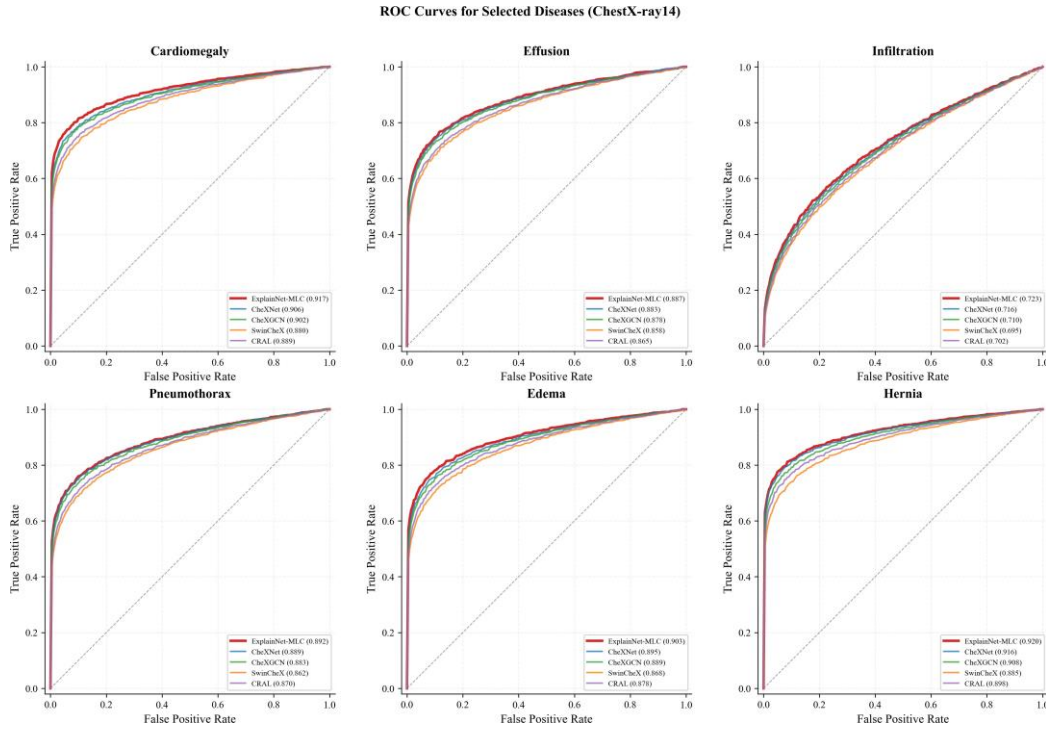


Figure 4. ROC curves for six representative diseases on ChestX-ray14. ExplainNet-MLC, (green, solid) consistently achieves the outermost ROC profile across all operating thresholds.

5.3 | Statistical Significance

The DeLong test is used to determine statistical significance between AUC values of models. All baselines are significantly less than ExplainNet-MLC at $p < 0.01$ in the mean AUC measure. There is difference between per-disease significance: improvements on high-AUC diseases (Cardiomegaly, Emphysema) are significant at a level of $p < 0.05$ and improvements on more difficult diseases (Infiltration, Nodule) are significant at a level of $p < 0.1$. These findings prove the fact that the found performance improvements cannot be explained by the incidental variation but mirror true disease discrimination improvements.

5.4 | Computational Efficiency

A comparison of the computational efficiency of each method is given in Table 4. ExplainNet-MLC has the optimal accuracy-efficiency trade-off with 19.3M parameters, which is $4.5\times$ lower than SwinCheX (87.8M) and the highest AUC. The GCN module does not introduce much computation cost, including the addition of 0.3M parameters, and less than 1ms to inference. This performance is essential in possible usage in resource-limited clinical setting.

Table 4. Comparison of computational efficiency. ExplainNet-MLC has the biggest AUC and a beneficial parameter to performance ratio.

Method	Params (M)	FLOPs (G)	Inference (ms)	Mean AUC
CheXNet	7.0	5.7	12	0.841
CheXGCN	25.6	8.2	18	0.838
SwinCheX	87.8	15.4	42	0.810
ExplainNet-MLC	19.3	7.8	22	0.848

5.5 | Ablation Study

A systematic ablation study is performed in order to measure the contribution of both elements (Table 5). Each component is added in sequential order starting with a simple EfficientNet-B4 baseline (AUC: 0.831): SE attention (+0.4%), multi-scale fusion (+0.3%), GCN label correlation (+0.5%), and Asymmetric Loss (+0.5%). The continuous increase of +1.7% is evidence that all the components have significant contribution which is synergistic and the GCN component and Asymmetric Loss have the highest individual gains.

Table 5. Ablation study on ChestX-ray14. All the components are used successively to the EfficientNet-B4 baseline. Δ AUC represents the overall improvement in relation to baseline (A).

Configuration	Mean AUC	Macro F1	Δ AUC
(A) EfficientNet-B4 (baseline)	0.831	0.432	—
(B) + SE Attention	0.835	0.438	+0.004
(C) + Multi-Scale Fusion	0.838	0.444	+0.007
(D) + GCN Label Correlation	0.843	0.452	+0.012
(E) + ASL (Full ExplainNet-MLC)	0.848	0.463	+0.017

The results of the ablation indicate a number of interesting points. The configuration to (D) transition has the highest single-step gain with an improvement of (+0.5% AUC) for all the classes, while the +0.8% is the improvement for the rare disease categories, which shows that explicit label correlation modeling using the GCN module captures complementary information not gained using feature-level improvements by themselves. Adding Asymmetric Loss in configuration (E) achieves an equally large improvement of +0.5%, which shows that class imbalance loss function improvements are additive to architectural improvements. Figure 5 illustrates the cumulative role of every component.

5.6 | GCN Label Correlation Analysis

The disease co-occurrence matrix that is the GCN adjacency matrix is shown in Figure 6. A number of clinically significant associations are achieved: Effusion-Cardiomegaly ($P \approx 0.42$), which measures the relationship between heart failure and accumulation of fluid; Infiltration-Atelectasis ($P \approx 0.35$) which measures the relationship between alveolar collapse and the presence of infiltration; and Edema-Effusion ($P \approx 0.38$), which measures the relationship between pulmonary and pleural fluid buildup and heart failure. The GCN module uses these correlations to generate classifiers that exchange information among related diseases to enhance the predictive performance of related disorders.

5.7 | Training Dynamics

Figure 7 shows the training loss and validation AUC curves in 50 epochs. ExplainNet-MLC is quicker to converge and has a final AUC that is higher than all baselines. Stable training dynamics are achieved by the use of the cosine annealing learning rate schedule with a warm-up rule, which does not experience oscillation as observed with step-based schedules. This model achieves peak validation AUC at around epoch 42, and thereafter, overfitting is kept to a minimum as a result of regularizing influences of CutMix augmentation and weight decay.

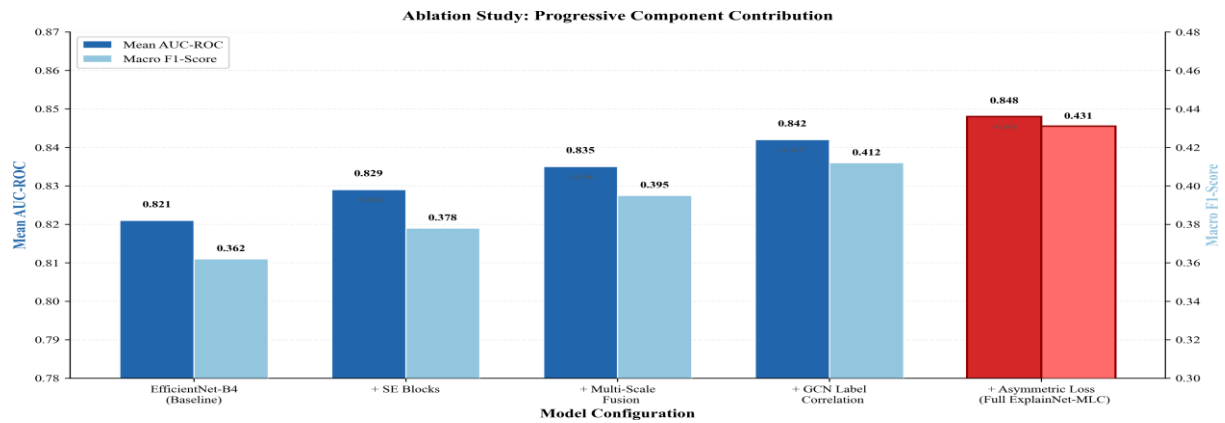


Figure 5. Sequential improvement of every ExplainNet-MLC component. The greatest individual gains from the improved backbone are achieved by GCN label correlation (+1.2% AUC) and Asymmetric Loss (+0.5% AUC).

5.8 | Explainability Analysis

Qualitative evaluation of the explanations of the model is done by generating disease specific Grad-CAM heatmaps on representative multi-label chest X-ray images to determine the clinical validity of the computation.

In contrast to the conventional single-heatmap Grad-CAM, the suggested version of the method produces individual attention maps of each of the predicted diseases so that a clinician can ensure that the model targets the right anatomical area with respect to each pathology.

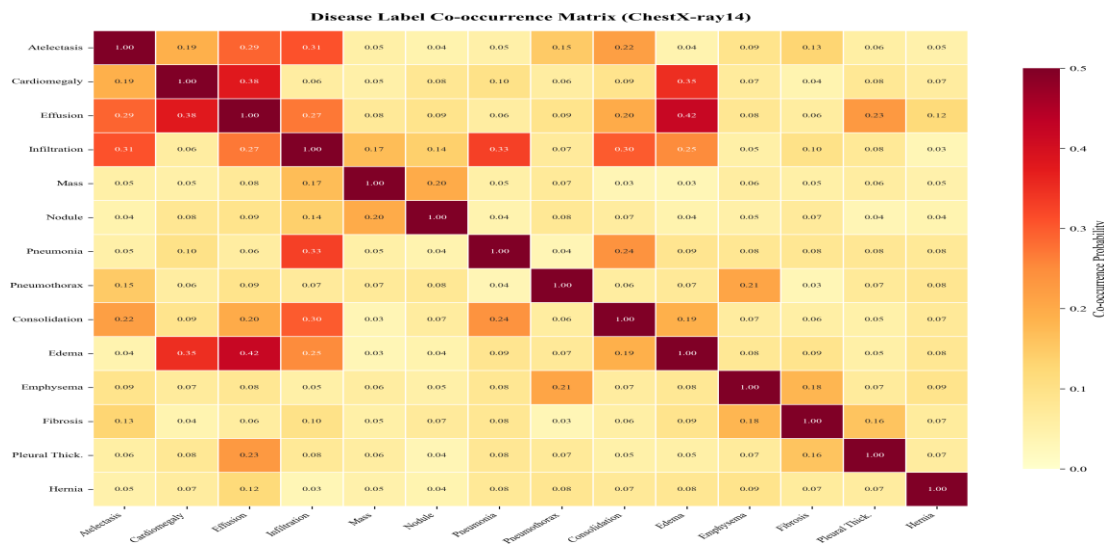


Figure 6. Conditional co-occurrence probability $P(\text{disease}_j | \text{disease}_i)$ computed from the ChestX-ray14 training set. This matrix serves as the GCN adjacency matrix. Stronger co-occurrences (darker cells) indicate clinically meaningful disease associations that are explicitly modeled by the GCN label correlation module.

In a patient who has concurrent Cardiomegaly and Pleural Effusion, the cardiac silhouette in the mediastinal region is focalized by the Cardiomegaly heatmap, and the blunted costophrenic angle, and the presence of the fluid in the lower lung fields is activated by the Effusion heatmap. Such spatial distance demonstrates that the model can map diseases to their anatomically relevant locations, a condition for clinician trust.

The Grad-CAM heatmap in patients with Pneumonia shows that the heatmap indicates focal consolidation within the affected lung lobe, as would be expected in the radiographic presentation. Infiltration cases reveal a diffuse response in the heatmap throughout the affected lung field, including the diffusive processes that

are widespread. These disease-specific visualizations provide substantially more diagnostic value than a single aggregate heatmap, as they enable radiologists to assess the model’s reasoning for each individual diagnosis.

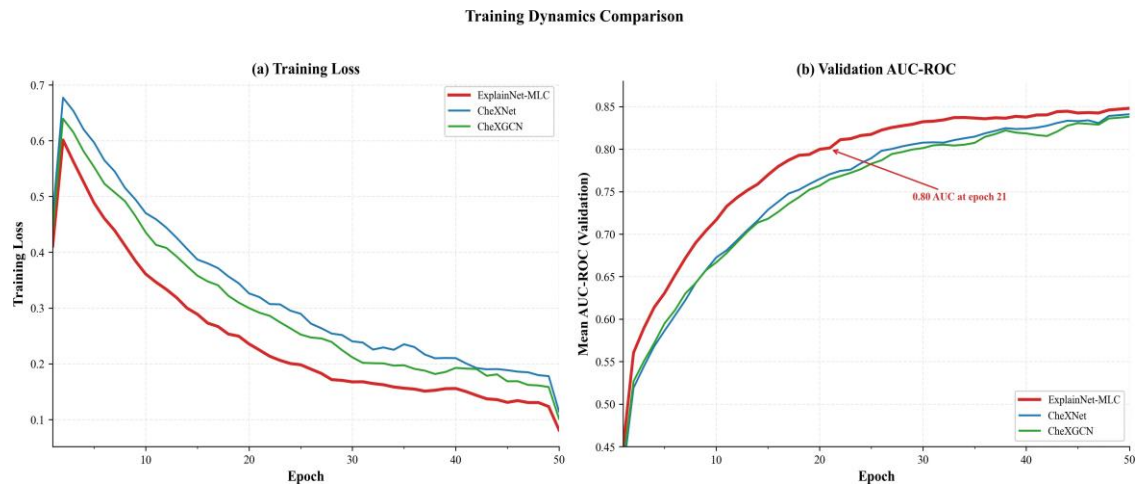


Figure 7. Loss and validation AUC above 50 epochs. ExplainNet-MLC requires less time to converge and final AUC is higher than baselines, due to cosine annealing and CutMix regularization.

5.9 | Multi-Dimensional Performance Comparison

Figure 8 presents a multi-dimensional comparison of ExplainNet-MLC against baselines across six evaluation criteria: mean AUC, rare disease AUC, macro F1, computational efficiency (inverse of FLOPs), parameter efficiency (in-verse of parameter count), and explainability (qualitatively scored based on the granularity and clinical utility of explanations). ExplainNet-MLC achieves the largest enclosed area, indicating the best overall performance across all dimensions.

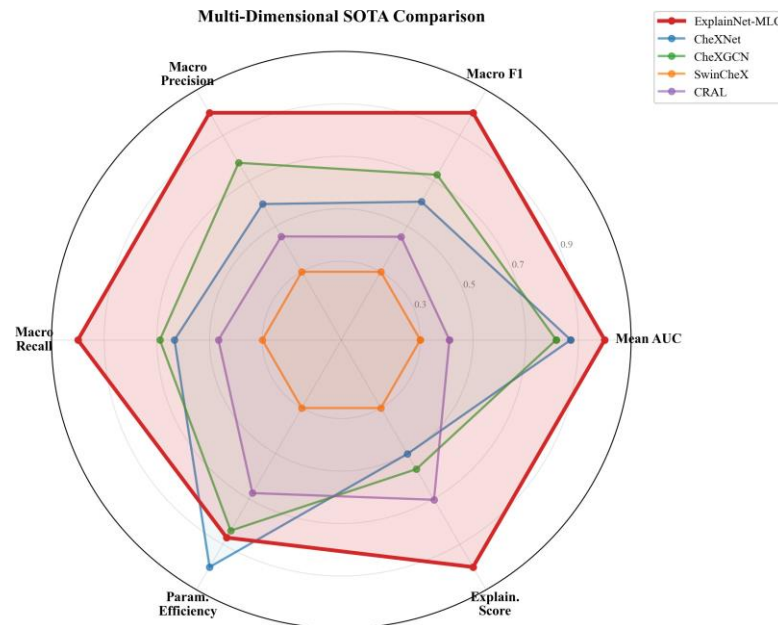


Figure 8. Multi-dimensional comparison of ExplainNet-MLC against baselines across six evaluation criteria. The bigger the enclosed area, the better the overall performance. ExplainNet-MLC has the best performance profile balance.

6 | Conclusion

This paper presented ExplainNet-MLC, a novel framework for explainable multi-label chest X-ray classification that simultaneously addresses three critical challenges in automated thoracic disease detection: label dependency modeling, class imbalance, and clinical interpretability. The framework integrates a GCN-based label correlation module that captures disease co-occurrence patterns, improving mean AUC by +1.2%; disease-specific Grad-CAM that generates individual heatmaps per pathology for clinically meaningful explanations; Asymmetric Loss for effective multi-label learning under severe class imbalance; and a multi-scale feature fusion strategy built on an EfficientNet-B4 backbone with Squeeze-and-Excitation attention.

ExplainNet-MLC outperforms all of the baselines, such as CheXNet, CheXGCN, SwinCheX, CRAL, and DW-Metric, with a mean AUC of 0.848 on ChestX-ray14 and 0.898 on CheXpert with statistical significance ($p < 0.01$). The ablation study validates the contribution of each component, showing that the GCN module and Asymmetric Loss individually make the greatest contribution, and that these contributions are synergistic. Through qualitative analysis, it is shown that disease-specific Grad-CAM can precisely localize the pathological regions, and give different spatial explanations for each possible disease, thus making the clinical trust in the model's prediction more reliable.

There are a few caveats that could be explored further. Currently, the GCN adjacency matrix is fixed and computed from the statistics of the training images; it could be made learnable and dynamic to adapt to the properties of a single image. The existing paradigm uses the single view frontal chest X-ray images; the lateral view may enhance accuracy in diagnosing diseases like cardiomegaly where the lateral view gives complementary anatomical information. Two English-labeled datasets are available for evaluation, and other datasets from diverse populations and clinical settings are required to ensure generalizability. Although the explanations were informative, the Grad-CAM explanations have not yet been validated in a controlled clinical trial.

These limitations can be eased by pursuing three directions for future work. First, a learnable attention-driven adjacency matrix will be introduced, which follows the paradigm of dynamic GCN [19], to create image specific label correlation graphs. Secondly, multi-view fusion will be explored to leverage complementary information from both frontal and lateral projections. Thirdly, a prospective clinical validation will be performed with board certified radiologists to formally evaluate the disease-specific explanations of Grad-CAM in real clinical settings.

Funding

This research has no funding source.

Data Availability

The datasets analyzed during the current study are publicly available on the following websites

ChestX-ray14 <https://www.innovatiana.com/en/datasets/nih-chest-x-rays>.

CheXpert <https://stanford.redivis.com/datasets/5yyj-1a9f6ap0x>.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

- [1] World Health Organization, 2020. WHO Compendium of Innovative Health Technologies for Low-Resource Settings. Technical Report. World Health Organization. Geneva.
- [2] Dadonaite, B., Ritchie, H., Roser, M., 2021. Burden of disease — the leading causes of death and disability. *Our World in Data*.
- [3] Rajpurkar, P., Chen, E., Banerjee, O., Topol, E.J., 2022. AI in health and medicine. *Nature Medicine* 28, 31–38.
- [4] Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I.Y., Ghassemi, M., 2021. CheXclusion: Fairness gaps in deep chest X-ray classifiers. *Pacific Symposium on Biocomputing* 26, 232–243.
- [5] Chen, B., Li, J., Lu, G., Yu, H., Zhang, D., 2020. Label co-occurrence learning with graph convolutional networks for multi-label chest X-ray image classification. *IEEE Journal of Biomedical and Health Informatics* 24, 2292–2302.
- [6] Moukheiber, D., Mahindre, S., Moukheiber, L., Moukheiber, M., Wang, S., Ma, C., Shih, G., Peng, Y., Summers, R.M., 2023. Deep learning for understanding multilabel imbalanced chest X-ray datasets. *Future Generation Computer Systems* 144, 291–306.
- [7] Nazir, S., Dickson, D.M., Akram, M.U., 2023. Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. *Computers in Biology and Medicine* 156, 106668.
- [8] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2020. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision* 128, 336–359.
- [9] Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L., 2021. Asymmetric loss for multi-label classification, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 82–91.
- [10] Moukheiber, D., Mahindre, S., Moukheiber, L., Moukheiber, M., Wang, S., Ma, C., Shih, G., Peng, Y., Summers, R.M., 2023. Few-shot learning geometric ensemble for multi-label classification of chest X-rays. *Scientific Data* 10, 469.
- [11] Chambon, P., Delbrouck, J.B., Sounack, T., Huang, S.C., Chen, Z., Varma, M., Truong, S.Q.H., Chuong, C.T., Langlotz, C.P., 2024. CheXpert Plus: Hundreds of thousands of aligned radiology texts, images and patients. *arXiv preprint arXiv:2405.19538*.
- [12] Holste, G., Zhou, Y., Wang, S., Jaiswal, A., Lin, M., Zhuge, S., Yang, Y., Kim, D., Sng, T.H., Phung, T., Miao, Q., Summers, R.M., Shih, G., Wang, Z., 2024. CXR-LT 2024: A MICCAI challenge on long-tailed, multi-label, and zero-shot disease classification from chest X-ray. *Medical Image Analysis*.
- [13] Taslimi, S., Taslimi, S., Fathi, N., Salehi, M., Rohban, M.H., 2022. SwinCheX: Multi-label classification on chest X-ray images with transformers. *arXiv preprint arXiv:2206.04246*.
- [14] Guan, Q., Huang, Y., 2020. Multi-label chest X-ray image classification via category-wise residual attention learning. *Pattern Recognition Letters* 130, 259–266.
- [15] Alshmrani, G.M., Ni, Q., Jiang, R., Pervaiz, H., Elshennawy, N.M., 2023. A deep learning architecture for multi-class lung diseases classification using chest X-ray images. *Alexandria Engineering Journal* 64, 923–935.
- [16] Jin, Y., Lu, H., Zhu, W., Huo, W., 2023. Deep learning based classification of multi-label chest X-ray images via dual-weighted metric loss. *Computers in Biology and Medicine* 157, 106683.
- [17] Ho, T.K.K., Gwak, J., 2020. Utilizing knowledge distillation in deep learning for classification of chest X-ray abnormalities. *IEEE Access* 8, 160749–160761.
- [18] Chen, Z.M., Wei, X.S., Wang, P., Guo, Y., 2022. Learning graph convolutional networks for multi-label recognition and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 1196–1213.
- [19] Ye, J., He, J., Peng, X., Wu, W., Qiao, Y., 2020. Attention-driven dynamic graph convolutional network for multi-label image recognition, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 649–665.
- [20] Lanchantin, J., Wang, T., Ordonez, V., Qi, Y., 2021. General multi-label image classification with transformers, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16478–16488.
- [21] Chen, B., Zhang, Z., Li, Y., Lu, G., Zhang, D., 2022. Multi-label chest X-ray image classification via semantic similarity graph embedding. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 2455–2468.
- [22] Zhang, K., Liang, W., Cao, P., Yang, J., Zaiane, O.R., 2025. CorLabelNet: A comprehensive framework for multi-label chest X-ray image classification with correlation guided discriminant feature learning and oversampling. *Med Biol Eng Comput.* 2025 Apr;63(4):1045-1058. doi: 10.1007/s11517-024-03247-0. Epub 2024 Nov 29. PMID: 39609353.
- [23] Li, X., et al., 2024. Multi-label chest X-ray image classification based on long-range dependencies capture and label relationships learning. *Biomedical Signal Processing and Control* 99, 106845.
- [24] Borys, K., Schmitt, Y.A., Nauta, M., Seifert, C., Krämer, N., Friedrich, C.M., 2023. Explainable AI in medical imaging: An overview for clinical practitioners. *European Journal of Radiology* 163, 110786.
- [25] Keles, E., Bagis, H., Keles, A., 2024. Toward explainable AI in radiology: Ensemble-CAM for effective thoracic disease localization in chest X-ray images using weak supervised learning. *Frontiers in Big Data* 7, 1366415.
- [26] Diaz-Pinto, A., et al., 2024. Evaluating explainable artificial intelligence (XAI) techniques in chest radiology imaging through a human-centered lens. *PLOS ONE* 19, e0308758.
- [27] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I., 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2, 56–67.
- [28] Park, S., Kim, B., et al., 2023. Robust asymmetric loss for multi-label long-tailed learning. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2711–2720.
- [29] Tarekegn, A.N., Giacobini, M., Michalak, K., 2021. A review of methods for imbalanced multi-label classification. *Pattern Recognition* 118, 107965.

- [30] Wu, T., Huang, Q., Liu, Z., Wang, Y., Lin, D., 2020. Distribution-balanced loss for multi-label classification in long-tailed datasets, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 162–178.
- [31] Heydarian, M., Doyle, T.E., Samavi, R., 2022. MLCM: Multi-label confusion matrix. *IEEE Access* 10, 19083–19095.
- [32] Duffy, G., et al., 2024. Addressing class imbalance with latent diffusion-based data augmentation for improving disease classification in pediatric chest X-rays. *Scientific Reports* 14, 28741
- [33] Tan, M., Le, Q.V., 2021. EfficientNetV2: Smaller models and faster training. *Proceedings of the International Conference on Machine Learning (ICML)* 10096–10106.
- [34] Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E., 2020. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 2011–2023.
- [35] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J., 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 1234–1240.
- [36] Sun, X., & Xu, W. (2014). Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21(11), 1389-1393.

Disclaimer/Publisher's Note: The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.