Sema-ChestX-Former: A Parameter-Efficient Hybrid Transformer-CNN for Robust Thoracic Disease Classification with XAI

Touhid Alam[1,4] 22-46330-1@student.aiub.edu
Sayedur Rahman Anik[1,4] 22-46418-1@student.aiub.edu
Nafiz Fahad[2,3] nafiz.fahad@student.mmu.edu.my
Md Tanzib Hosain[1,4] 20-42737-1@student.aiub.edu
Md Kishor Morol[3,4] kishormorol@ieee.org
Md. Jakir Hossen[2,4] jakir.hossen@mmu.edu.my
[1] *Department of Computer Science, American International University-Bangladesh, Dhaka, Bangladesh*
[2] *Faculty of Engineering and Technology, Multimedia University, Melaka, Malaysia*
[3] *College of Computing and Information Science, Cornell University, Ithaca, NY, USA*
[4] *ELITE Research Lab,17010 Cedarcroft Rd, Queens, NY, USA*

Editors: Under Review for MIDL 2026

Abstract

Developing reliable deep learning systems for thoracic disease diagnosis faces a critical trade-off. Convolutional Neural Networks (CNNs) excel at local feature extraction but often fail to model the global context necessary for complex diagnoses, while Vision Transformers (ViTs) capture long-range spatial dependencies but are notoriously parameter-heavy and computationally expensive. Addressing these challenges requires architectures that are robust to the unique complexities of medical imaging data—such as high class imbalance and visual similarity between pathologies—while remaining efficient enough for practical clinical deployment. In this work, we present **Sema-ChestX-Former**, a novel, parameter-efficient (∼1.84 million parameters) hybrid architecture. Our model synergistically integrates a Transformer-based backbone for semantic spatial feature extraction with specialized CNN-based attention blocks for fine-grained feature refinement. We conducted a comprehensive evaluation on three large-scale public datasets—Chest X-Ray (Pneumonia), COVID-19 Radiography, and NIH ChestX-ray14. Sema-ChestX-Former established comparative performance on the NIH dataset with a mean AUC of 0.846, while achieving 99.69% accuracy on the Pneumonia dataset and 98.34% on the COVID-19 dataset. Furthermore, we employed Explainable AI (XAI) using Gradient-weighted Class Activation Mapping (Grad-CAM) to ensure model transparency. Our findings demonstrate that a carefully designed, parameter-efficient hybrid architecture can outperform larger, more complex models (∼250M parameters), offering a promising and practical solution for automated Chest X-ray analysis.

Keywords: Hybrid Transformer-CNN, Parameter Efficiency, Explainable AI, Thoracic Disease, Medical Image Analysis.

1. Introduction

The development of deep learning models in healthcare settings has the potential to transform current medical practices in disease diagnosis, biomarker discovery, and personalized treatment. However, clinical deployment requires robust models—a standard that remains largely unmet due to the inherent complexities of medical imaging data. Thoracic diseases, encompassing a range of conditions from lower respiratory infections (LRIs) to lung cancer, represent a staggering global health and economic burden. The diagnostic complexity of conditions such as pneumonia and COVID-19 is particularly challenging. These diseases often caused by a diverse spectrum of pathogens, frequently produce overlapping radiological findings like ground-glass opacities, making definitive diagnosis from imaging alone notoriously difficult (Soltani et al., 2021).

In the clinical pathway, the chest X-ray (CXR) remains the most widely utilized frontline tool, prized for its speed and cost-effectiveness (Jones et al., 2021). However, the traditional diagnostic method—visual interpretation by a radiologist—is fraught with challenges. Retrospective studies have shown that clinically significant findings are missed in 20-30% of cases, often due to perceptual error or fatigue (Pesapane et al., 2024). Consequently, Artificial Intelligence (AI) has emerged as a crucial force in automated analysis.

Convolutional Neural Networks (CNNs) have historically been the backbone of this revolution, demonstrating remarkable capability in identifying fine-grained textures and local patterns (Wei et al., 2021; Sharma and Guleria, 2024). Models such as ResNet, DenseNet, and EfficientNet have achieved performance comparable to expert radiologists in specific tasks. Despite their success, CNNs possess an inherent limitation: their architecture, which excels at local feature extraction via small receptive fields, inherently struggles to model global, long-range spatial dependencies (Younesi et al., 2024). A CNN may identify a local infiltrate but fail to understand its contextual relationship to a distant finding in the opposite lung, a skill crucial for differential diagnosis in complex multi-label scenarios.

This gap has been addressed by the rise of Vision Transformers (ViTs) for medical imaging (Dosovitskiy, 2020). Adapted from natural language processing, ViTs treat an image as a sequence of patches and use a self-attention mechanism to weigh the importance of all patches relative to each other, allowing them to excel at capturing the global context of the entire image (Takahashi et al., 2024). However, this power comes at a significant cost. ViTs are often data-hungry and parameter-heavy. Recent hybrid architectures, such as ConvFormer and CaFormer, have attempted to bridge this gap, achieving state-of-the-art (SOTA) performance on benchmarks like the NIH ChestX-ray14 dataset (Yanar et al., 2025). Yet, these models often contain 28M to 60M parameters. More complex graph-based approaches, such as the MSASG model, reach over 250M parameters (Wang et al., 2025), making them computationally prohibitive for deployment in resource-constrained clinical environments or on edge devices like portable X-ray machines.

This presents a clear research gap: there is a need for a model that synergizes the local feature power of CNNs and the global relational modeling of Transformers but is explicitly designed for **parameter efficiency**. In this work, we propose the **Sema-ChestX-Former**, a novel hybrid architecture designed to overcome these limitations. Our model integrates a powerful Transformer-based backbone for semantic spatial feature extraction with a series of specialized CNN-based attention blocks for fine-grained feature refinement and localization. Crucially, we achieve SOTA competitive performance with only **1.84 million parameters**, a reduction of over 90% compared to comparable models. We validate our approach on three diverse datasets (Binary, Multi-class, and Multi-label) and address the ”black box” problem by integrating Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) to provide visual justification for the model's predictions.

2. Methodology

The Sema-ChestX-Former architecture is designed to hierarchically process CXR images, moving from coarse, global semantic understanding to fine-grained, localized feature refinement. The overall workflow of our proposed approach is illustrated in Figure 1. The model is composed of four distinct stages: (1) Semantic Spatial Backbone, (2) Attention CNN, (3) Squeeze-and-Excitation (SE) block, and (4) a final Classification Head.

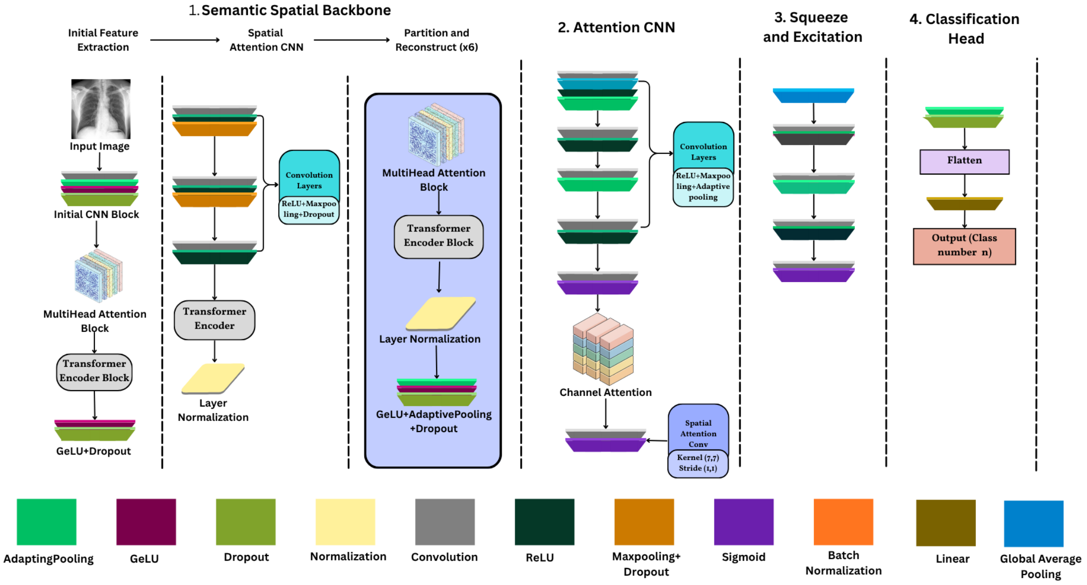


Figure 1: The end-to-end architecture of the proposed Sema-ChestX-Former. The model employs a Transformer backbone for global context and an Attention CNN for local refinement.

2.1. Stage 1: Semantic Spatial Backbone

The backbone is responsible for transforming the input image into a rich feature map that encodes both local textures and global spatial context. It utilizes a Partition and Reconstruct strategy using Transformer Encoder Blocks.

Image Patching and Embedding. An input image $X \in \mathbb{R}^{H \times W \times C}$ is first deconstructed into a sequence of N flattened 2D patches X_p. To retain crucial spatial information, which is otherwise lost in the patching process, a learnable positional embedding E_{pos} is added to the patch embeddings.

$$Z_0 = [x_p^1 E; x_p^2 E; \dots ; x_p^N E] + E_{pos} \tag{1}$$

The resulting sequence Z_0 serves as the input to the attention mechanism.

Multi-Head Attention (MHSA). The Multi-Head Attention block allows the model to weigh the importance of different regions of the lung relative to each other. For each attention head i, the input is linearly projected into Query (Q_i), Key (K_i), and Value (V_i) matrices:

$$Q_i = Z_0 W_i^Q, \quad K_i = Z_0 W_i^K, \quad V_i = Z_0 W_i^V \tag{2}$$

The core of the mechanism is the scaled dot-product attention function, which computes a weighted sum of the values:

$$\text{Attention}(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \tag{3}$$

The outputs from all h heads are concatenated and passed through a final linear projection to produce the block output:

$$\text{MHSA}(Z_0) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W^O \tag{4}$$

This mechanism enables the model to capture long-range dependencies, such as correlating an enlarged heart silhouette with pleural effusion in the lung bases.

2.2. Stage 2: Attention CNN

While the backbone captures global context, CNNs are superior for refining fine-grained local features. The globally-aware feature map is passed to the Attention CNN stage. This stage explicitly identifies ”what” (channel-wise) and ”where” (spatial-wise) is important using custom attention modules.

Channel Attention Module. This block recalibrates the channel-wise feature responses to selectively amplify informative features (e.g., texture patterns of pneumonia) and suppress less useful ones (e.g., background noise). It aggregates spatial information using both Average Pooling and Max Pooling, followed by a shared Multi-Layer Perceptron (MLP) and a sigmoid activation σ:

$$M_c(F) = \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \tag{5}$$

The input feature map F is then modulated via element-wise multiplication: $F' = M_c(F) \otimes F$.

Spatial Attention Module. Complementary to channel attention, the Spatial Attention Module focuses on ”where” the informative regions are located. It compresses the channel dimension by applying average-pooling and max-pooling along the channel axis. These descriptors are concatenated and convolved with a standard 7x7 convolution layer to generate a 2D spatial attention map M_s:

$$M_s(F) = \sigma(f^{7\times 7}([\text{AvgPool}(F); \text{MaxPool}(F)])) \tag{6}$$

The final refined feature map is obtained by: $F_{refined} = M_s(F) \otimes F$.

2.3. Stage 3 & 4: Squeeze-and-Excitation and Classification

After the Attention CNN has refined the features, they undergo a final channel-wise recalibration via a Squeeze-and-Excitation (SE) block. The final stage is the Classification Head. The refined 3D feature map is flattened into a 1D feature vector and passed through a fully connected (Linear) layer to produce the final output logits. The model is trained using a weighted Binary Cross-Entropy (BCE) loss to handle class imbalance effectively.

2.4. Explainable AI (Grad-CAM)

To provide visual justification for the predictions, we employ Gradient-weighted Class Activation Mapping (Grad-CAM). The importance weight α_k^c for a feature map A^k is calculated by global-average-pooling the gradients of the target class score y^c. The final heatmap is a weighted combination of these feature maps:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{7}$$

By overlaying these heatmaps onto the original radiographs, we can visually verify the model's reasoning.

3. Experiments and Results

3.1. Experimental Setup

To rigorously assess the performance of Sema-ChestX-Former, we utilized three large-scale datasets:

- **Chest X-Ray (Pneumonia) (Kermany, 2018):** A binary dataset containing 5,863 pediatric images.
- **COVID-19 Radiography (Rahman et al., 2021):** A multi-class dataset with 21,165 images across 4 classes.
- **NIH ChestX-ray14 (Wang et al., 2017):** A complex multi-label dataset with 112,120 images and 14 pathologies.

All images were resized to 224×224. We utilized the AdamW optimizer with a learning rate of 1×10^{-4} and a ReduceLROnPlateau scheduler. Training was conducted for 50 epochs on dual NVIDIA Tesla T4 GPUs.

3.2. Performance on Pneumonia Dataset

On the binary classification task using the Chest X-Ray (Pneumonia) dataset, Sema-ChestX-Former demonstrated exceptional discriminative capability. As detailed in Table 1, our model achieved a validation accuracy of 99.69% and a nearly perfect ROC AUC of 0.9992. This performance not only surpasses standard benchmarks like VGG16 (95.40%) but also exceeds computationally expensive ensemble methods, such as those proposed by Kundu et al. (2021) which reached 98.81%.

Table 1: Performance comparison on the binary Pneumonia dataset.

Model / Method	Accuracy	Precision	Recall	F1-Score	AUC
Ensemble (ResNet-50, Xception) (Ayan et al., 2022)	95.83%	95.73%	97.76%	96.70%	95.21%
EfficientNet-B2 (Jahan et al., 2021)	96.33%	98.00%	97.00%	97.00%	99.10%
Ensemble (GoogLeNet, ResNet) (Kundu et al., 2021)	98.81%	98.82%	98.80%	98.79%	98.35%
QCSA Network (Singh et al., 2023)	94.53%	93.56%	98.86%	96.14%	89.00%
VGG16 + NN (Sharma and Guleria, 2023)	95.40%	95.40%	95.40%	95.40%	98.80%
Sema-ChestX-Former (Ours)	**99.69%**	**99.63%**	**99.71%**	**99.67%**	**0.9992**

3.3. Performance on COVID-19 Radiography Dataset

For the more challenging multi-class task, Sema-ChestX-Former achieved an overall validation accuracy of 98.34% and a macro-averaged F1-Score of 98.37%. The comparative results in Table 2 show that our model outperforms specialized models like ResNet-34 (98.33%) and Xception (97.97%), indicating better overall class balance and separability.

Table 2: Performance comparison on the multi-class COVID-19 dataset.

Model / Method	Accuracy	Precision	Recall	F1-Score	AUC
Xception (Jain et al., 2021)	97.97%	99.00%	92.00%	95.00%	N/A
Modified MobileNetV2 (Akter et al., 2021)	98.00%	97.00%	98.00%	97.00%	N/A
Custom CNN (Bhandari et al., 2022)	94.31%	93.55%	94.20%	93.67%	N/A
ResNet-34 (Nayak et al., 2021)	98.33%	96.77%	100.00%	98.36%	0.9836
DCNN (Custom) (Hou and Gao, 2021)	96.03%	97.00%	95.00%	96.00%	N/A
Sema-ChestX-Former (Ours)	**98.34%**	**98.33%**	**98.41%**	**98.37%**	**0.9923**

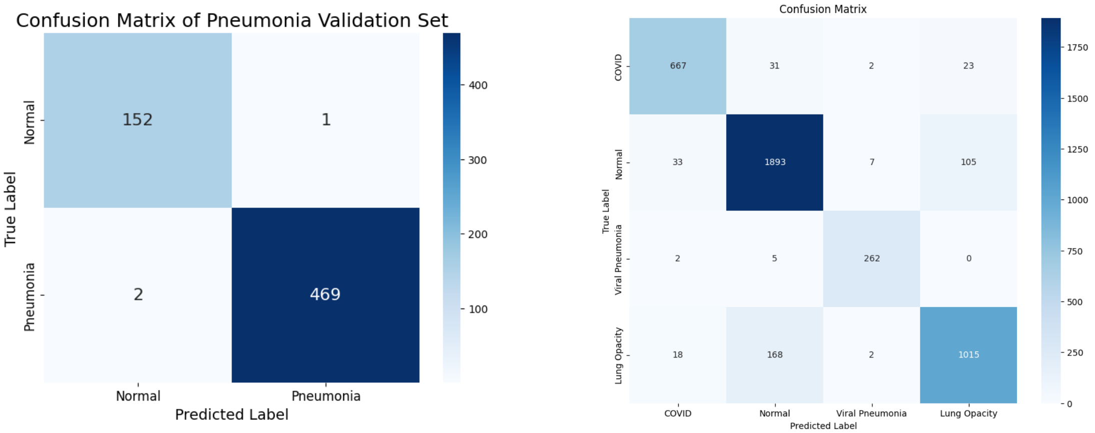
3.4. Combined Confusion Matrix Analysis

To illustrate the model's precision, we present the confusion matrices for the Pneumonia and COVID-19 datasets in Figure 2. For Pneumonia, the model made only 3 errors out of 624 samples. For COVID-19, confusion was minimal and primarily occurred between "Normal" and "Lung Opacity," which is clinically plausible given their visual similarity.

3.5. Performance on NIH ChestX-ray14 Dataset

The multi-label NIH dataset represents the most rigorous test. Table 3 compares Sema-ChestX-Former against leading Graph Convolutional Network (GCN) and Transformer-based models. Our model achieves a Mean AUC of 0.846.

This result is significant because it outperforms the massive MSASG model (Mean AUC 0.842), which utilizes over 251 million parameters and complex attribute-aware graphs (Wang et al., 2025). It also surpasses ConvFormer (Mean AUC 0.837), a leading Hybrid-ViT identified as SOTA by Yanar et al. (Yanar et al., 2025). Our model excels particularly in pathologies with distinct structural features, such as Cardiomegaly (0.913) and Effusion (0.904), validation of our hybrid approach where the Transformer captures the global organ



Pneumonia Dataset

COVID-19 Dataset

Figure 2: Confusion matrices for binary (Pneumonia) and multi-class (COVID-19) tasks.

shape and the CNN refines the boundaries. The lower performance on Infiltration (0.671) is consistent with the literature, as diffuse textures are notoriously difficult for all architectures.

Table 3: Comparison with SOTA models on the NIH ChestX-ray14 dataset (AUC scores). The best mean AUC is highlighted in bold.

Model	**Card**	**Emp**	**Effu**	**Her**	**Inf**	**Mass**	**Nod**	**Atel**	**P1**	**P2**	**PT**	**Edem**	**Fib**	**Cons**	**Mean**
MSASG (Wang et al., 2025)	0.936	0.915	0.881	**0.982**	**0.711**	0.800	0.707	0.786	0.828	**0.880**	**0.815**	**0.915**	**0.852**	0.815	0.842
ImageGCN (Mao et al., 2022)	0.890	0.915	0.874	0.943	0.702	0.843	0.768	0.802	0.715	0.883	0.791	0.900	0.825	0.796	0.832
GWSA-LCD (Xu et al., 2024)	0.877	0.924	0.827	0.921	0.701	0.822	0.790	0.770	0.732	0.870	0.782	0.847	0.839	0.746	0.818
CheXGAT (Lee et al., 2022)	0.879	**0.944**	0.837	0.931	0.699	0.839	**0.793**	0.786	0.741	0.879	0.794	0.851	0.842	0.754	0.826
ConvFormer (Yanar et al., 2025)	0.900	0.930	0.880	0.920	0.710	0.850	0.770	0.820	0.760	0.780	0.880	0.890	0.830	0.800	0.837
VMamba (Yanar et al., 2025)	0.860	0.790	0.840	0.880	0.680	0.770	0.630	0.770	0.700	0.720	0.810	0.860	0.750	0.770	0.774
Ours	**0.913**	0.935	**0.904**	0.864	0.671	**0.885**	0.789	**0.834**	**0.904**	0.815	0.808	0.914	0.786	**0.818**	**0.846**

3.6. Ablation Studies

To rigorously validate our architectural design choices, we conducted ablation studies on the NIH dataset. Table 4 summarizes the results. The most significant performance degradation occurred upon removing the Semantic Spatial Backbone (Transformer), with the mean AUC dropping to 0.806. This unequivocally confirms the critical importance of the Transformer-based core for capturing global, long-range spatial dependencies. Removing the Attention CNN led to a drop to 0.816, showing that the backbone alone is insufficient without fine-grained local refinement.

3.7. Computational Efficiency

A primary objective of this study was parameter efficiency. Table 5 compares our model with key SOTA architectures. Sema-ChestX-Former consists of only 1.84 million trainable parameters. This represents a reduction of over 90% compared to ConvFormer (∼28M) and over 99% compared to MSASG (∼251M). Training on the NIH dataset took only 13.5 hours on 2x Tesla T4 GPUs.

Table 4: Ablation study of architectural components on NIH ChestX-ray14.

Configuration	Card	Effu	Mass	Nod	Atel	Mean AUC
Without Squeeze-and-Excitation	0.900	0.890	0.860	0.770	0.810	0.825
Without Attention CNN	0.890	0.880	0.850	0.760	0.800	0.816
Without Transformer Backbone	0.880	0.870	0.840	0.750	0.790	0.806
Sema-ChestX-Former (Full)	**0.913**	**0.904**	**0.885**	**0.789**	**0.834**	**0.846**

Table 5: Computational Cost and Efficiency Comparison.

Model	Total Parameters	Training Time (NIH)
MSASG (Wang et al., 2025)	$\sim$ 251.0 M	High
ConvFormer (Yanar et al., 2025)	$\sim$ 28.0 M	Medium
EfficientNet-B4	$\sim$ 19.0 M	Medium
Sema-ChestX-Former	**1.84 M**	**13h 32m**

3.8. XAI Analysis for Model Interpretability

To address the "black box" nature of deep learning, we employed Grad-CAM. Figure 3 confirms that the model has learned to localize findings consistent with radiological practice. For Cardiomegaly, the activation correctly focuses on the cardiac silhouette. For Pleural Thickening, attention is directed to the lung periphery.

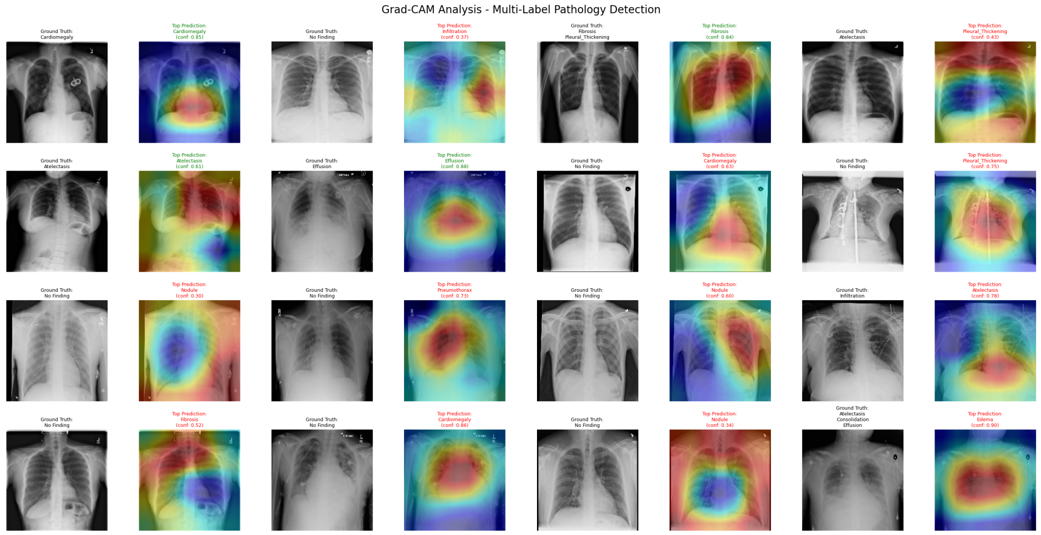

Grad-CAM Analysis - Multi-Label Pathology Detection
Ground Truth: Cardiomegaly
Top Prediction: Cardiomegaly (conf: 0.85)
Ground Truth: No Finding
Top Prediction: Infiltration (conf: 0.37)
Ground Truth: Fibrosis Pleural_Thickening
Top Prediction: Fibrosis (conf: 0.84)
Ground Truth: Atelectasis
Top Prediction: Pleural_Thickening (conf: 0.43)
Ground Truth: Atelectasis
Top Prediction: Atelectasis (conf: 0.61)
Ground Truth: Effusion
Top Prediction: Effusion (conf: 0.88)
Ground Truth: No Finding
Top Prediction: Cardiomegaly (conf: 0.63)
Ground Truth: No Finding
Top Prediction: Pleural_Thickening (conf: 0.75)
Ground Truth: No Finding
Top Prediction: Nodule (conf: 0.30)
Ground Truth: No Finding
Top Prediction: Pneumothorax (conf: 0.73)
Ground Truth: No Finding
Top Prediction: Nodule (conf: 0.60)
Ground Truth: Infiltration
Top Prediction: Atelectasis (conf: 0.78)
Ground Truth: No Finding
Top Prediction: Fibrosis (conf: 0.52)
Ground Truth: No Finding
Top Prediction: Cardiomegaly (conf: 0.86)
Ground Truth: No Finding
Top Prediction: Nodule (conf: 0.34)
Ground Truth: Atelectasis Consolidation Effusion
Top Prediction: Edema (conf: 0.90)

Figure 3: Grad-CAM analysis for multi-label pathology detection on the NIH dataset. The heatmaps confirm the model's ability to localize specific findings, such as focusing on the heart for Cardiomegaly.

4. Discussion and Conclusion

The experimental results demonstrate that Sema-ChestX-Former is a robust, efficient, and interpretable model. By achieving a state-of-the-art mean AUC of 0.846 on the NIH ChestX-ray14 dataset, along with exceptional accuracy on Pneumonia and COVID-19 datasets, we have shown that a carefully designed model with only 1.84 million parameters can outperform models that are orders of magnitude larger. The ablation studies confirmed the necessity of our hybrid design: the Transformer backbone provides the essential global context, while the Attention CNN refines the local features.

The development of such a lightweight model has significant practical implications. Its efficiency makes it a viable candidate for deployment in diverse clinical settings, including those with limited computational infrastructure or on edge devices like portable X-ray machines. Furthermore, the integration of extensive XAI analysis addresses the critical need for transparency in medical AI, fostering clinician trust.

Despite these strengths, the model's performance on diffuse pathologies like Infiltration (AUC 0.671) suggests room for improvement, likely requiring specialized attention mechanisms for undefined boundaries. Future work will focus on validating the model in prospective clinical trials and extending the architecture to other modalities.

Acknowledgments

The authors acknowledge ELITE Research Lab for providing the computational resources and support necessary for this research.

References

Shamima Akter, FM Javed Mehedi Shamrat, Sovon Chakraborty, Asif Karim, and Sami Azam. Covid-19 detection using deep learning algorithm on chest x-ray images. *Biology*, 10(11):1174, 2021.

Enes Ayan, Bergen Karabulut, and Halil Murat Ünver. Diagnosis of pediatric pneumonia with ensemble of deep convolutional neural networks in chest x-ray images. *Arabian Journal for Science and Engineering*, 47(2):2123–2139, 2022.

Mohan Bhandari, Tej Bahadur Shahi, Birat Siku, and Arjun Neupane. Explanatory classification of cxr images into covid-19, pneumonia and tuberculosis using deep learning and xai. *Computers in Biology and Medicine*, 150:106156, 2022.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jie Hou and Terry Gao. Explainable dcnn based chest x-ray image analysis and classification for covid-19 pneumonia detection. *Scientific Reports*, 11(1):16071, 2021.

Nusrat Jahan, Md Shamim Anower, and Rakibul Hassan. Automated diagnosis of pneumonia from classification of chest x-ray im ages using efficientnet. In *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, pages 235–239. IEEE, 2021.

Rachna Jain, Meenu Gupta, Soham Taneja, and D Jude Hemanth. Deep learning based detection and analysis of covid-19 on chest x-ray images. *Applied Intelligence*, 51(3): 1690–1700, 2021.

Catherine M Jones, Quinlan D Buchlak, Luke Oakden-Rayner, Michael Milne, Jarrel Seah, Nazanin Esmaili, and Ben Hachey. Chest radiographs and machine learning–past, present and future. *Journal of Medical Imaging and Radiation Oncology*, 65(5):538–544, 2021.

Daniel Kermany. Labeled optical coherence tomography (oct) and chest x-ray images for classification. *Mendeley data*, 2018.

Rohit Kundu, Ritacheta Das, Zong Woo Geem, Gi-Tae Han, and Ram Sarkar. Pneumonia detection in chest x-ray images using an ensemble of deep learning models. *PloS one*, 16 (9):e0256630, 2021.

Yan-Wei Lee, Sheng-Kai Huang, and Ruey-Feng Chang. Chexgat: A disease correlation-aware network for thorax disease diagnosis from chest x-ray images. *Artificial Intelligence in Medicine*, 132:102382, 2022.

Chengsheng Mao, Liang Yao, and Yuan Luo. Imagegcn: Multi-relational image graph convolutional networks for disease identification with chest x-rays. *IEEE transactions on medical imaging*, 41(8):1990–2003, 2022.

Soumya Ranjan Nayak, Deepak Ranjan Nayak, Utkarsh Sinha, Vaibhav Arora, and Ram Bilas Pachori. Application of deep learning techniques for detection of covid-19 cases using chest x-ray images: A comprehensive study. *Biomedical Signal Processing and Control*, 64:102365, 2021.

Filippo Pesapane, Giulia Gnocchi, Cettina Quarrella, Adriana Sorce, Luca Nicosia, Luciano Mariano, Anna Carla Bozzini, Irene Marinucci, Francesca Priolo, Francesca Abbate, et al. Errors in radiology: A standard review. *Journal of Clinical Medicine*, 13(15):4306, 2024.

Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier, Muhammad Salman Khan, et al. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in biology and medicine*, 132: 104319, 2021.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Shagun Sharma and Kalpna Guleria. A deep learning based model for the detection of pneumonia from chest x-ray images using vgg-16 and neural networks. *Procedia Computer Science*, 218:357–366, 2023.

Shagun Sharma and Kalpna Guleria. A systematic literature review on deep learning approaches for pneumonia detection using chest x-ray images. *Multimedia Tools and Applications*, 83(8):24101–24151, 2024.

Sukhendra Singh, Manoj Kumar, Abhay Kumar, Birendra Kumar Verma, and S Shitharth. Pneumonia detection with qcsa network on chest x-ray. *Scientific Reports*, 13(1):9025, 2023.

Saber Soltani, Armin Zakeri, Milad Zandi, Mina Mobini Kesheh, Alireza Tabibzadeh, Mahsa Dastranj, Samireh Faramarzi, Mojtaba Didehdar, Hossein Hafezi, Parastoo Hosseini, et al. The role of bacterial and fungal human respiratory microbiota in covid-19 patients. *BioMed research international*, 2021(1):6670798, 2021.

Satoshi Takahashi, Yusuke Sakaguchi, Nobuji Kouno, Ken Takasawa, Kenichi Ishizu, Yu Akagi, Rina Aoyama, Naoki Teraya, Amina Bolatkan, Norio Shinkai, et al. Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review. *Journal of Medical Systems*, 48(1):84, 2024.

Qian Wang, Zhijuan Wu, Jiyu Gao, Hongnian Yu, and Yongqiang Cheng. A multi-label chest x-ray image classification agorithm based on multi-scale and attribute-aware semantic graph. *Expert Systems with Applications*, page 129898, 2025.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8927–8948, 2021.

Yujia Xu, Hak-Keung Lam, Xinqi Bao, and Yuhan Wang. Learning group-wise spatial attention and label dependencies for multi-task thoracic disease classification. *Neurocomputing*, 573:127228, 2024.

Erdem Yanar, Furkan Kutan, Kubilay Ayturan, Uğurhan Kutbay, Oktay Algın, Fırat Hardalaç, and Ahmet Muhteşem Ağıldere. A comparative analysis of the mamba, transformer, and cnn architectures for multi-label chest x-ray anomaly detection in the nih chestx-ray14 dataset. *Diagnostics*, 15(17):2215, 2025.

Abolfazl Younesi, Mohsen Ansari, Mohammadamin Fazli, Alireza Ejlali, Muhammad Shafique, and Jörg Henkel. A comprehensive survey of convolutions in deep learning: Applications, challenges, and future trends. *IEEE Access*, 12:41180–41218, 2024.

Appendix A. Detailed Dataset Analysis

This section provides a granular breakdown of the datasets used in this study to highlight the challenges of class imbalance.

A.1. Distributions and Imbalance Ratios

The utilized datasets exhibit varying degrees of imbalance.

- **Pneumonia Dataset:** As shown in Figure 4, there is a significant imbalance between Pneumonia ($N = 4,273$) and Normal ($N = 1,583$) cases. This required the use of weighted loss functions to prevent the model from biasing towards the majority class.
- **COVID-19 Radiography:** This dataset (Figure 5) is dominated by Normal cases ($N = 10,192$) compared to Viral Pneumonia ($N = 1,345$).
- **NIH ChestX-ray14:** This dataset presents the most severe ”long-tail” distribution (Figure 6). Common conditions like Infiltration and Effusion have thousands of samples, while Hernia has fewer than 300. This extreme imbalance is the primary reason for the lower performance on rare classes across all SOTA models.

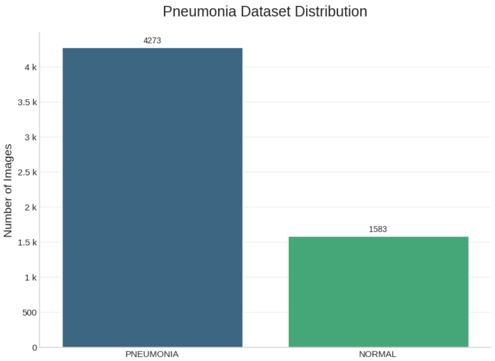


Figure 4: Distribution of Pneumonia Dataset.

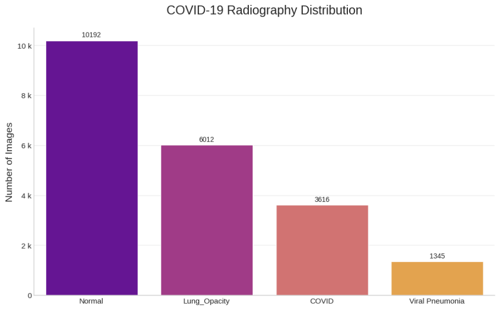


Figure 5: Distribution of COVID-19 Dataset.

Appendix B. Implementation Details and Algorithms

B.1. Training Algorithm

Algorithm 1 details the training loop employed for the Sema-ChestX-Former. We utilize Mixed Precision Training (AMP) to accelerate convergence and reduce memory usage, allowing for larger batch sizes on the Tesla T4 GPUs.

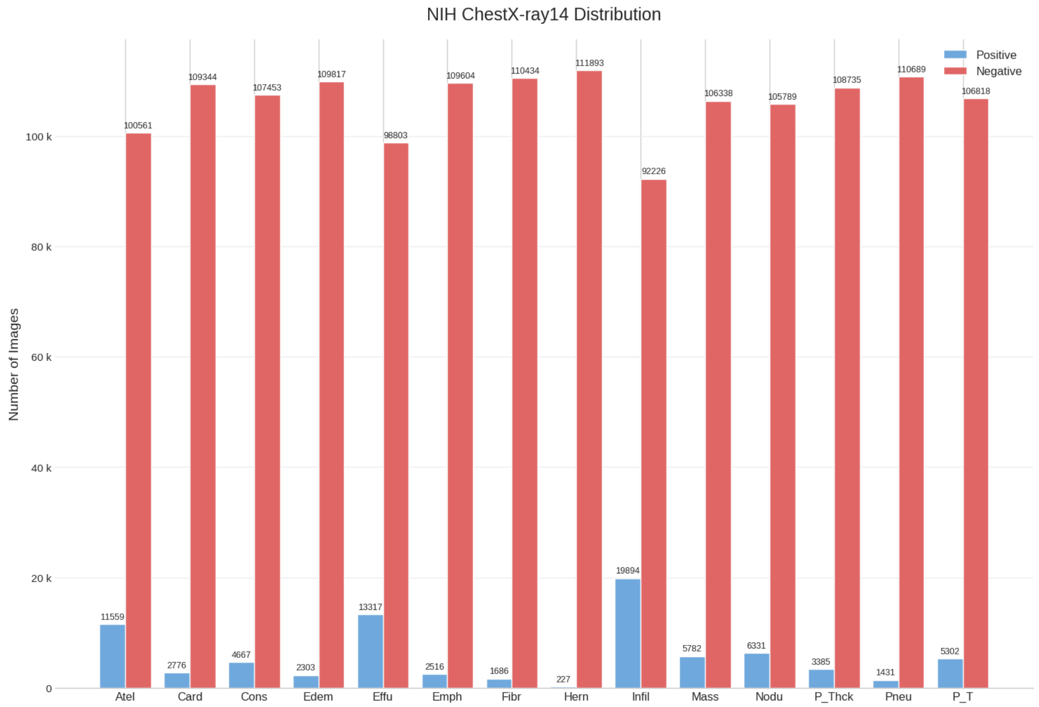


Figure 6: Detailed class distribution of the NIH ChestX-ray14 dataset, highlighting the severe class imbalance challenge.

```
Algorithm 1: Training Loop for Sema-ChestX-Former
Input: Dataset $D$, Model $M$, Optimizer $Opt$, Loss Function $L$, Scaler $S$
Output: Trained Model parameters $\theta$
Initialize $\theta$ randomly; Set learning rate $\eta = 1e - 4$; for epoch $e \leftarrow 1$ to 50 do
    for batch $(x, y)$ in $D_{train}$ do
        $x \leftarrow$ Augment($x$) Random Flip, Rotation Autocast $logits \leftarrow M(x)$;
        $loss \leftarrow L(logits, y)$; $S$.scale($loss$).backward(); $S$.step($Opt$); $S$.update();
        $Opt$.zero_grad();
    end
    Validate on $D_{val}$; if val_loss did not improve then
        $Scheduler$.step();
    end
end
```

B.2. Code Listings for Attention Modules

Code Listing 1 provides the PyTorch implementation of our custom Channel Attention module. This module uses a shared MLP to learn relationships between channels, amplifying informative features.

```
import torch
import torch.nn as nn
```

```

class ChannelAttention(nn.Module):
    def __init__(self, in_planes, ratio=16):
        super(ChannelAttention, self).__init__()
        self.avg_pool = nn.AdaptiveAvgPool2d(1)
        self.max_pool = nn.AdaptiveMaxPool2d(1)

        # Shared MLP
        self.fc1 = nn.Conv2d(in_planes, in_planes // ratio, 1, bias=False)
        self.relu1 = nn.ReLU()
        self.fc2 = nn.Conv2d(in_planes // ratio, in_planes, 1, bias=False)
        self.sigmoid = nn.Sigmoid()

    def forward(self, x):
        avg_out = self.fc2(self.relu1(self.fc1(self.avg_pool(x))))
        max_out = self.fc2(self.relu1(self.fc1(self.max_pool(x))))
        out = avg_out + max_out
        return self.sigmoid(out) * x
```

Listing 1: PyTorch Implementation of Channel Attention

Code Listing 2 shows the Spatial Attention module, which compresses channel information to highlight salient spatial regions.

```
class SpatialAttention(nn.Module):
    def __init__(self, kernel_size=7):
        super(SpatialAttention, self).__init__()
        assert kernel_size in (3, 7), ’kernel size must be 3 or 7’
        padding = 3 if kernel_size == 7 else 1

        # Large kernel convolution for spatial context
        self.conv1 = nn.Conv2d(2, 1, kernel_size, padding=padding, bias=
    False)
        self.sigmoid = nn.Sigmoid()

    def forward(self, x):
        avg_out = torch.mean(x, dim=1, keepdim=True)
        max_out, _ = torch.max(x, dim=1, keepdim=True)
        x = torch.cat([avg_out, max_out], dim=1)
        x = self.conv1(x)
        return self.sigmoid(x)
```

Listing 2: PyTorch Implementation of Spatial Attention

Appendix C. Detailed Architectural Specifications

This section provides the granular architectural details, layer configurations, and parameter counts for each block of Sema-ChestX-Former, as implemented in the final model.

C.1. Multi-Head Attention Process

Table 6 details the configuration of the attention mechanism used in the Semantic Spatial Backbone.

Table 6: Architectural Details of the Multi-Head Attention Process.

Layer / Operation	**Configuration / Parameters**
Input	Image ($X \in \mathbb{R}^{224\times224\times3}$)
1. Image Patching	Patch Size (P) = 16, Num Patches (N) = 196
2. Linear Projection	Projects patches to Model Dimension (D) = 192
3. Positional Embedding	Adds learnable spatial information
— **Multi-Head Attention** —	
4. Linear Projections	Generate Q, K, V for each of the h heads
5. Scaled Dot-Product	Parallel computation for each head
6. Concatenation	Concatenate outputs of all h heads
7. Final Linear Projection	Maps concatenated features back to dimension D
Hyperparameters	Model Dimension (D) = 192, Num Heads (h) = 6

C.2. Transformer Encoder Block

Table 7 outlines the layers within the Transformer Encoder, including the feed-forward network and normalization steps.

Table 7: Architectural Details of the Transformer Encoder Block.

Layer / Operation	**Configuration / Activation**
— **Attention Sub-layer** —	
1. Layer Normalization	Epsilon = 1e-6
2. Multi-Head Attention Block	As detailed in Table 6
3. Residual Connection	Element-wise Addition
— **Feed-Forward Sub-layer** —	
4. Layer Normalization	Epsilon = 1e-6
5. Linear Layer (Expansion)	Output Dim = 384, Activation = GELU
6. Dropout	Rate = 0.1
7. Linear Layer (Contraction)	Output Dim = 192
8. Dropout	Rate = 0.1
9. Residual Connection	Element-wise Addition

C.3. Channel Attention Block

Table 8 details the CNN-based Channel Attention block used for feature recalibration.

C.4. Attention CNN Stage

Table 9 breaks down the sequence of layers in the Attention CNN stage, which processes the output from the Transformer backbone.

Table 8: Architectural Details and Parameter Count of the Custom Channel Attention Block.

Layer / Operation	**Configuration / Activation**	**Parameters**
Input	Feature Map ($U \in \mathbb{R}^{H \times W \times C}$)	-
— **Convolutional Sequence** —		
1. Convolution Layer 1	Units = 16, Kernel = (5,5), Activation = ReLU	51,216
2. Convolution Layer 2	Units = 16, Kernel = (5,5), Followed by Pooling	6,416
3. Convolution Layer 3	Units = 32, Kernel = (3,3), Activation = ReLU	4,640
4. Convolution Layer 4	Units = 32, Kernel = (3,3), Activation = Sigmoid	9,248
Output	Rescaled Feature Map ($\tilde{X}$)	-
Total Trainable Parameters		**71,520**

Table 9: Detailed architecture and parameter count of the Attention CNN stage.

Layer / Block	**Configuration**	**Parameters**
Input Feature Map	From Backbone ([-1, 192, 14, 14])	-
1. Conv Layer	192 → 128 channels, (3,3) kernel, ReLU	221,312
2. Conv Layer	128 → 128 channels, (3,3) kernel, ReLU	147,584
3. Conv Layer	128 → 64 channels, (3,3) kernel, ReLU	73,792
4. Conv Layer	64 → 32 channels, (3,3) kernel, ReLU	18,464
5. Pooling	Adaptive/Max Pooling + Dropout	0
6. Channel Attention	Custom Block (Table 8)	71,520
7. Spatial Attention	(7,7) kernel convolution on pooled features	99
Total Parameters	**Attention CNN Block**	**532,771**

C.5. Full Model Summary

Table 10 provides a complete summary of the Sema-ChestX-Former architecture, showing the output shape at each stage.

Table 10: Detailed architectural summary and final parameter distribution of the Sema-ChestX-Former.

Stage	**Layer / Block**	**Output Shape**	**Parameters**
1. Semantic Spatial Backbone	Initial CNN Block	[-1, 192, H/4, W/4]	112,576
	Spatial Attention CNN	[-1, 192, H/8, W/8]	221,440
	Partition & Reconstruct	[-1, 192, H/8, W/8]	891,072
Subtotal for Stage 1			*1,225,088*
2. Attention CNN	Convolutional Layers	[-1, 32, 14, 14]	461,152
	Channel Attention	[-1, 32, 14, 14]	71,520
	Spatial Attention	[-1, 32, 14, 14]	99
Subtotal for Stage 2			*532,771*
3. Squeeze and Excitation	Channel Attention Block	[-1, 32, 14, 14]	71,520
Subtotal for Stage 3			*71,520*
4. Classification Head	Flatten	[-1, 6272]	0
	Final Classifier	[-1, n_classes]	6,273+
Subtotal for Stage 4			*6,273+*
Total	**Sema-ChestX-Former**		**∼1.84 Million**

Appendix D. Extended Quantitative Analysis

D.1. Loss Function Ablation

The choice of loss function is critical for imbalanced datasets. We compared Weighted Binary Cross-Entropy (BCE) against L1, L2, and Focal Loss. As shown in Table 11, BCE consistently yielded the highest AUC across datasets. Focal Loss, while designed for imbalance, sometimes led to unstable training dynamics in our hybrid architecture.

Table 11: Impact of Loss Functions on NIH Dataset Performance.

Loss Function	**Mean AUC**	**F1-Score**	**Recall**
Weighted BCE	**0.846**	**0.818**	**0.808**
Focal Loss	0.840	0.812	0.791
L2 Loss (MSE)	0.815	0.785	0.760

D.2. Threshold Optimization

For multi-label classification, the decision threshold τ must be tuned. We evaluated the Mean AUC on the NIH dataset for $\tau \in [0.2, 0.5]$. As illustrated in Figure 7, the optimal performance was found at $\tau = 0.4$. Lower thresholds increased recall but severely penalized precision, while higher thresholds missed subtle findings.

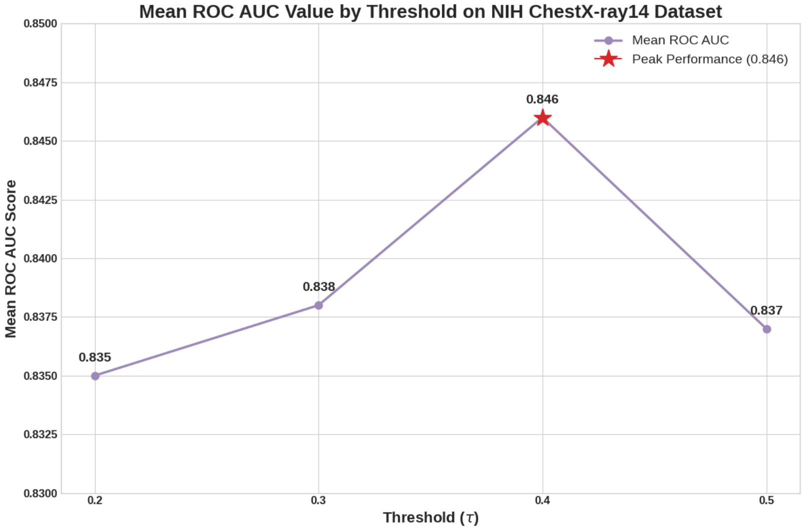


Figure 7: Mean ROC AUC on the NIH dataset as a function of the classification threshold (τ). The peak performance of 0.846 is achieved at $\tau = 0.4$.

Appendix E. Failure Analysis and Limitations

A critical aspect of robust AI development is understanding failure modes. While SemaChestX-Former performs exceptionally well on pathologies with clear boundaries (e.g., Cardiomegaly, Mass), it struggles with **Infiltration** (AUC 0.671).

Infiltration Analysis. Infiltration is characterized by diffuse, ill-defined opacities in the lung parenchyma. As shown in Figure 8, the Grad-CAM activation for a misclassified Infiltration case is scattered and weak. This suggests that the current attention mechanisms, while excellent for localized features, may struggle to aggregate the subtle, widespread textural changes associated with infiltration. Future work will investigate multi-scale attention pyramids to capture these diffuse patterns better.

Appendix F. Additional XAI Visualizations

We provide additional visualizations to demonstrate the model's reliability across different diseases. Figure 9 shows the model correctly identifying pneumonia consolidations. Figure 10 demonstrates the model's focus on peripheral opacities typical of COVID-19.

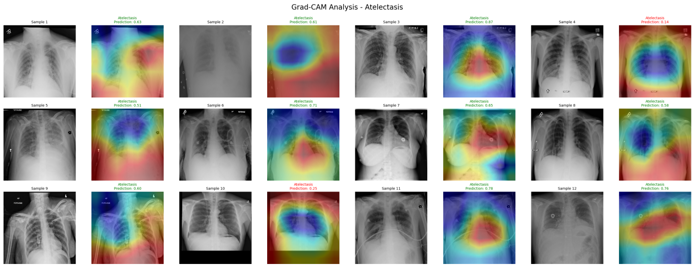


Figure 8: Analysis of a challenging case (Infiltration/Atelectasis). The heatmaps show diffuse activation, indicating model uncertainty compared to the sharp activations seen in Cardiomegaly.

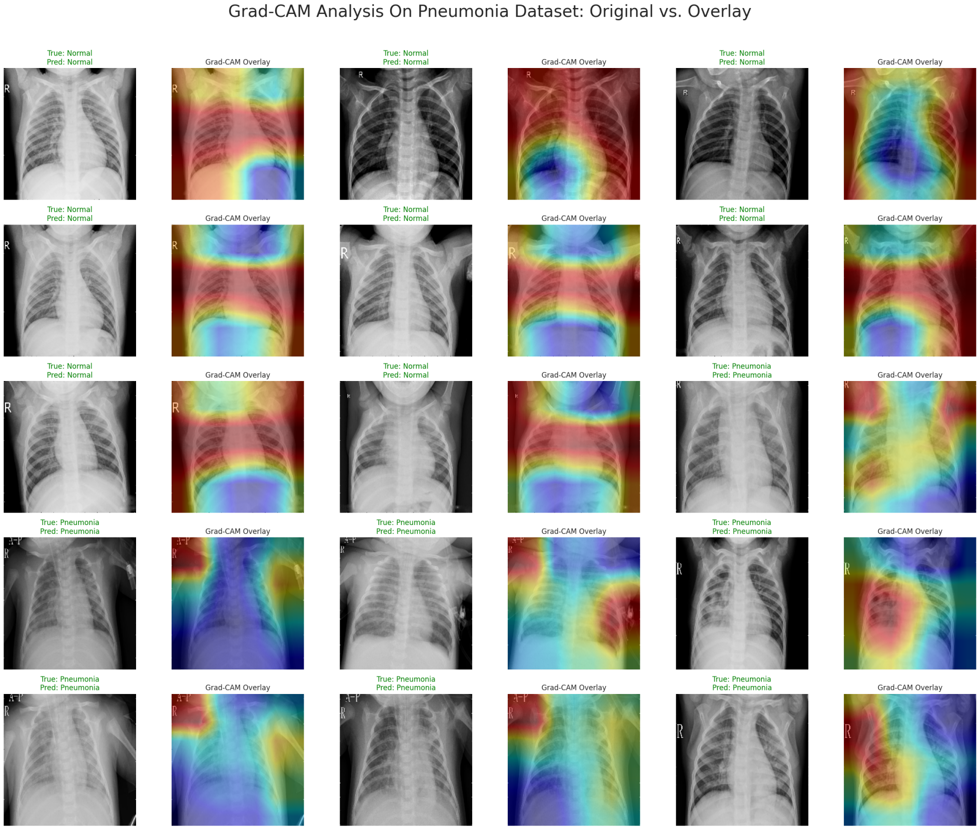


Figure 9: Grad-CAM analysis on the Pneumonia dataset. The model correctly focuses on areas of consolidation (bottom row) versus diffuse attention in normal cases (top row).

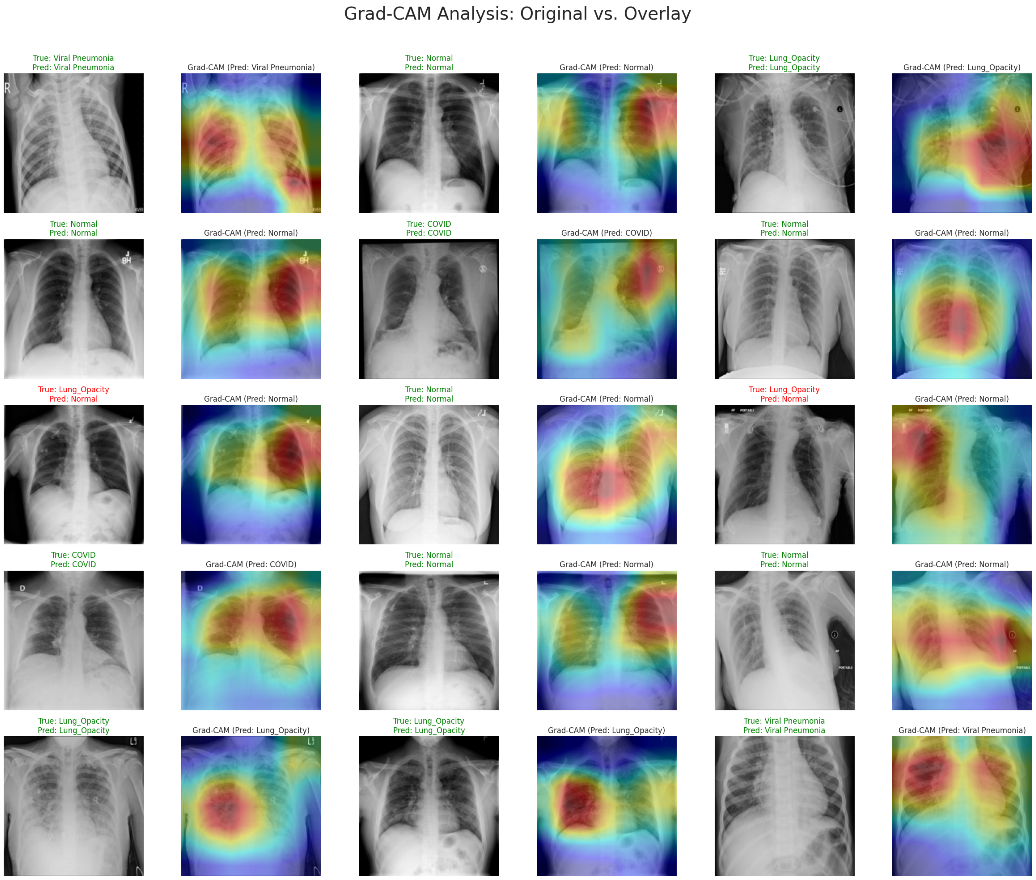


Figure 10: Grad-CAM analysis on the COVID-19 Radiography dataset. The heatmaps demonstrate that the model distinguishes between different pathologies by focusing on distinct patterns and locations.