# Chest X-ray Classification: Model Development & Evaluation Report

## 1. Introduction

This project addresses the classification of chest X-ray images into five categories: Atelectasis, Effusion, Infiltration, No Finding, and Other Disease. The NIH Chest X-ray dataset was used as the primary data source. The goal is to build a robust deep learning model that not only performs well across multiple metrics but is also explainable through visualizations such as Grad-CAM.

## 2. Data Preparation

### 2.1 Image Preprocessing and Augmentation

* Resize to 224x224: Resizing all images to a fixed dimension allows the use of pre-trained models (like ResNet50) that expect specific input sizes.
* Augmentations Used:
  + Random Horizontal Flip
  + Random Rotation (10 degrees)
  + Color Jitter (brightness and contrast) These augmentations aim to simulate the variability of real-world scans, helping the model generalize better.

### 2.2 Stratified Sampling for Train/Test Split A stratified 80/20 train-test split was used to ensure proportional class representation across splits, preserving class distribution and mitigating data leakage.

## 3. Model Development

### 3.1 CNN Architecture

* We used ResNet50 with ImageNet pre-trained weights.
* The final fully connected layer was modified to output logits for 5 classes.

### 3.2 Transfer Learning and Fine-tuning

* The backbone was initialized with ImageNet weights to benefit from general feature extraction.
* The full model was fine-tuned, allowing updates to all layers to better capture domain-specific patterns in chest X-rays.

### 3.3 Weighted Cross Entropy Loss Due to class imbalance, class\_weight='balanced' from scikit-learn was used to compute weights inversely proportional to class frequencies. This adjustment:

* Penalized majority classes less
* Encouraged the model to focus on minority classes

### 3.4 Early Stopping Training monitored validation loss with a patience of 3 epochs. This helped avoid overfitting by halting training once performance plateaued.

## 4. Model Evaluation

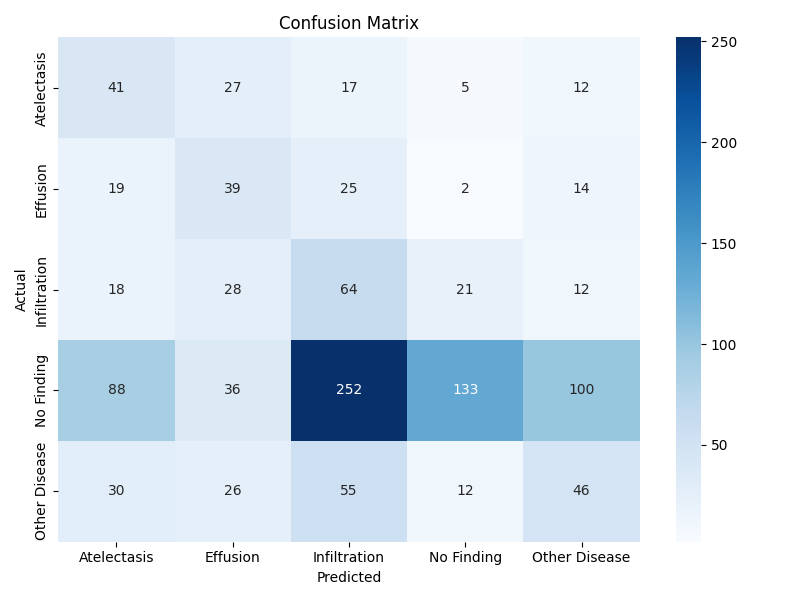
### 4.1 Metrics Reported

* Accuracy: 28.8% (macro avg)
* F1 Score (macro): 28.2%
* Precision (macro): 32.6%
* Recall (macro): 34.7%
* These numbers show modest performance. However, the model significantly outperformed random guessing for certain classes (Atelectasis, Effusion).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| Atelectasis | 0.2092 | 0.4020 | 0.2752 | 102 |
| Effusion | 0.2500 | 0.3939 | 0.3059 | 99 |
| Infiltration | 0.1550 | 0.4476 | 0.2302 | 143 |
| No Finding | 0.7688 | 0.2184 | 0.3402 | 609 |
| Other Disease | 0.2500 | 0.2722 | 0.2606 | 169 |
| **Accuracy** |  |  | **0.2879** | **1122** |
| **Macro Avg** | 0.3266 | 0.3468 | 0.2824 | 1122 |
| **Weighted Avg** | 0.5158 | 0.2879 | 0.3052 | 1122 |

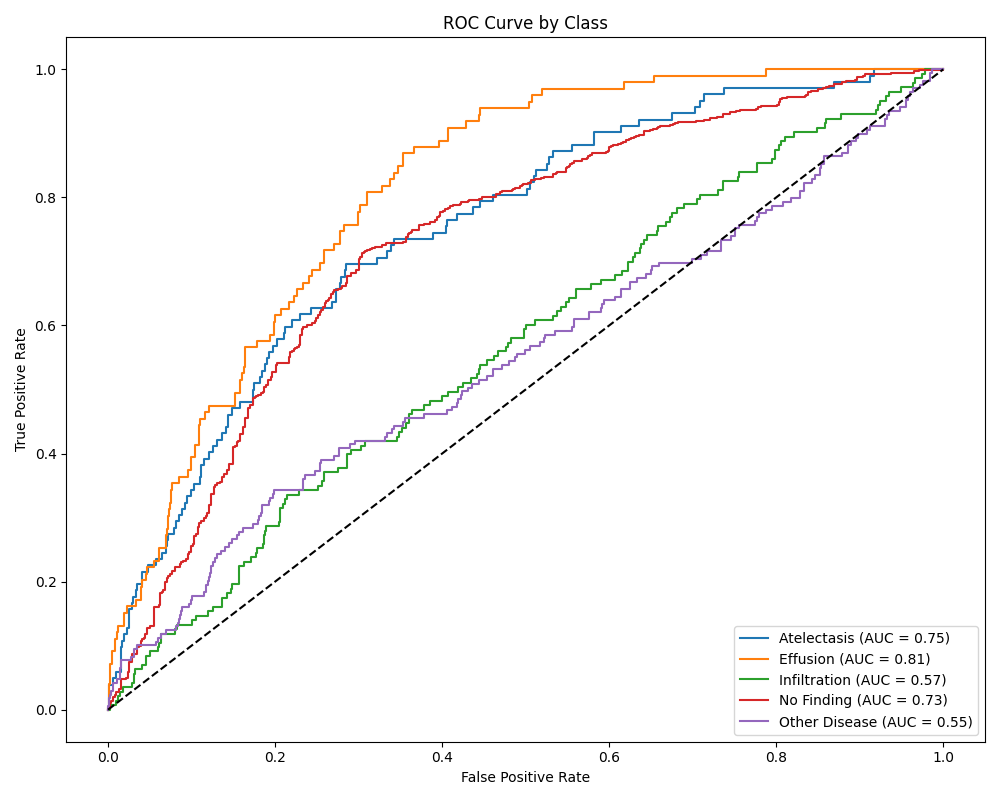
Detailed classification with class-wise precision, recall, and F1-Score Table

### 4.2 Confusion Matrix Analysis The confusion matrix revealed strong confusion among Infiltration and "No Finding" categories, suggesting overlapping radiological patterns or label noise.



### 4.3 ROC Curve Insights

* Best AUC was for Effusion (0.81) and Atelectasis (0.75)
* Weakest AUCs were Other Disease (0.55) and Infiltration (0.57)
* AUC > 0.5 suggests the model captures some signal beyond random classification



## 5. Grad-CAM Explainability

### 5.1 Single Sample Grad-CAM

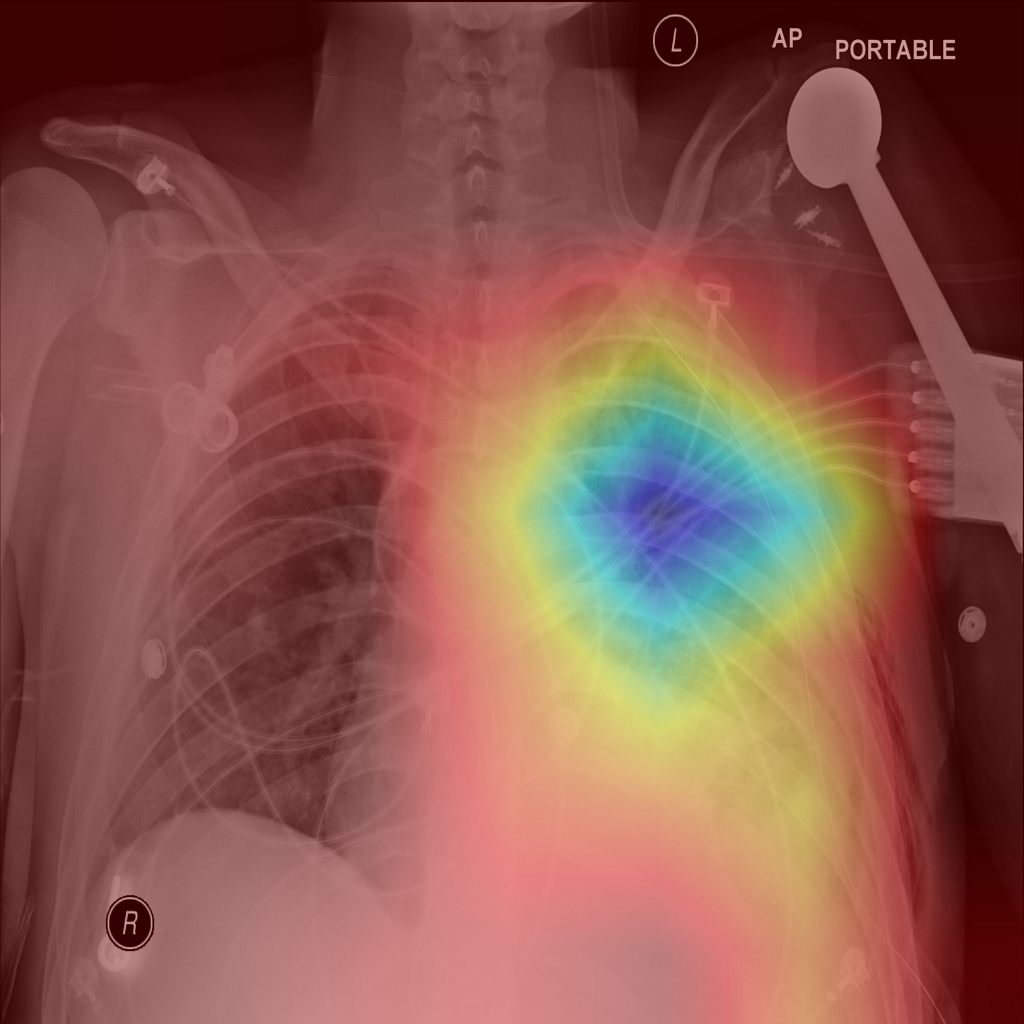
* Grad-CAM was generated on a manually chosen test image.
* The heatmap correctly focused on the lung areas.

A x-ray of a chest

AI-generated content may be incorrect.

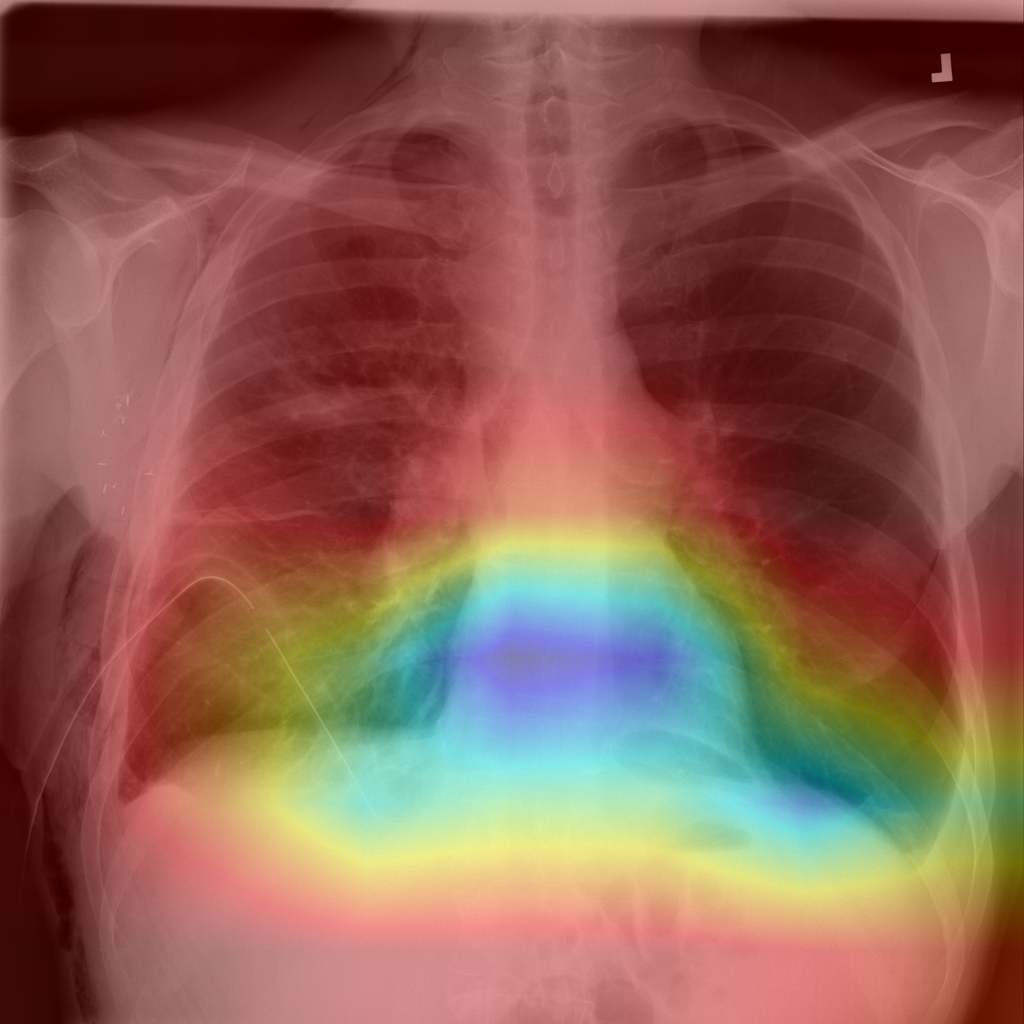
### 5.2 Batch Grad-CAM Evaluation

* 10 test images were used.
* Heatmaps were overlaid on input X-rays.
* We split outputs into two sets:
  + Correct Predictions: The model highlighted plausible pathology locations.



correct\_Atelectasis\_Pred

* + Incorrect Predictions: Grad-CAM often misfocused, suggesting model uncertainty or insufficient feature separability.



wrong\_Atelectasis\_Pred

These visualizations were crucial for identifying blind spots in the model's attention and serve as justification for future improvements.

## 6. Challenges and Lessons Learned

* Label Overlap and Ambiguity: The NIH dataset often contains multi-label cases reduced to a single class.
* Noisy Labels: "No Finding" can coexist with subtle pathologies.
* Class Imbalance: Despite rebalancing, minority class performance remained limited.

## 7. Beyond This Exercise: Potential Future Enhancements

While not part of the current task requirements, the following ideas may help enhance model performance and clinical applicability in future work:

* Multi-Label Classification: Align the model with the dataset’s nature by predicting multiple conditions per image instead of a single class.
* Ensemble Learning: Combine predictions from multiple architectures to reduce variance and improve robustness.
* Pathology-Specific Augmentations: Introduce augmentations tailored to simulate radiographic features (e.g., Gaussian blur to mimic haziness or opacities).

## 8. Deliverables

* Classification Report: Includes confusion matrix, ROC curves, and per-class metrics. Saved under the /outputs directory.
* Grad-CAM Visualizations: Both single-sample and batch overlays demonstrating model interpretability, located in /outputs/ and /outputs/gradcam\_samples/.
* Code Repository: Full codebase, model weights, and scripts are [publicly available on GitHub](github.com/H00zy/cms-provider-data-analysis.)
* Presentation Slides: PowerPoint deck summarizing methodology, results, and Grad-CAM examples for effective communication and review.