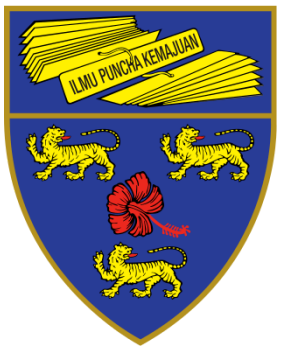


UNIVERSITY OF MALAYA

Faculty of Computer Science & Information Technology

WQF7009 Explainable Artificial Intelligence

Semester 1, 2025/2026

Report for Assignment 3 - XAI on Pneumonia Detection

Lecturer

Prof. Dr. Loo Chu Kiong

Prepared By

Hong Jia Herng - U2005313

Table of Contents

Task 1: Data Preparation & Pre-processing

Data Preparation

The dataset used is the Chest X-Ray Images (Pneumonia) dataset from Kaggle, consisting of 5,863 X-ray images (JPEG) categorized into two classes: PNEUMONIA and NORMAL. The data is pre-organized into three standard splits: train, val, test.

I downloaded the dataset and copied it into the project directory as shown below:

```
wqf7009-a3/
└── data/
    └── chest_xray
        ├── train
        │   ├── NORMAL
        │   └── PNEUMONIA
        ├── val
        │   ├── NORMAL
        │   └── PNEUMONIA
        └── test
            ├── NORMAL
            └── PNEUMONIA
```

Data Pre-processing

Preprocessing Pipeline

To ensure compatibility with pre-trained models (ResNet152/VGG16), we implemented a transformation pipeline using torchvision.transforms.

1. **Resize & Crop:** Images are resized to `256x256` and then center-cropped to `224x224`. This preserves the aspect ratio of the central lung region while strictly adhering to the input dimension requirements of standard CNN architectures.
2. **Normalization:** We normalized pixel values using the ImageNet standards (`mean=[0.485, 0.456, 0.406]`, `std=[0.229, 0.224, 0.225]`). This is critical for Transfer Learning, as it aligns the input distribution with what the pre-trained weights expect, leading to faster convergence.
3. **Tensor Conversion:** Images are converted to PyTorch tensors for GPU acceleration.

Code implementation:

```
# Transform pipeline
transform = transforms.Compose([
    transforms.Resize(256),
    transforms.CenterCrop(224),
    transforms.ToTensor(),
    transforms.Normalize(mean=[0.485, 0.456, 0.406], std=[0.229,
0.224, 0.225])
])
```

Class Imbalance Issue

As illustrated in **Figure 1**, the training dataset exhibits a significant class imbalance, containing **1,341 Normal** cases versus **3,875 Pneumonia** cases.

In medical image analysis, synthetic oversampling techniques are often avoided to prevent the introduction of artifacts that could be misinterpreted as pathology. Therefore, we adopted a **Weighted Binary Cross Entropy Loss** function during training, assigning a higher penalty to misclassifications of the minority class (Normal). This ensures the model learns to discriminate between both classes equally effectively, rather than biasing its predictions toward the majority class.

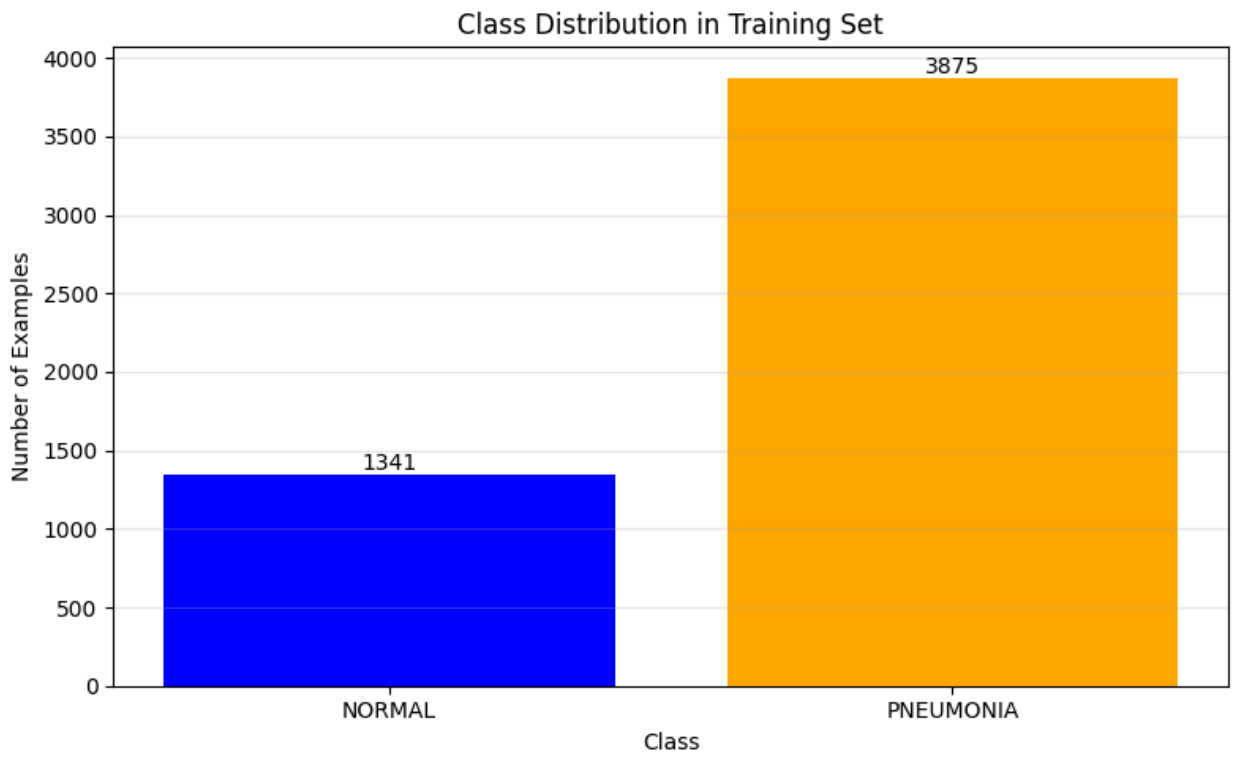


Figure 1: Class distribution in training set showing data imbalance issue.

Task 2: Model Training, Comparison & Evaluation

Model Training & Evaluation

To ensure a consistent benchmark, all five model configurations were trained using the following hyperparameters:

- **Training duration:** All models were trained for a fixed **10 epochs on Google Colab's A100 GPU**. The model checkpoint with the **highest Validation F1-Score** was selected for final testing to ensure the best performing weights were used.
- **Loss Function:** A **Weighted Binary Cross Entropy Loss** was used to penalize misclassifications of the minority class (Normal) more heavily.
- **Optimizer:** Adam (`lr=1e-4`).
- **Batch size:** 128

Disclaimer: *I didn't fully train every model until convergence due to time and computational constraints. The current configuration is a good default to show the idea works.*

SimpleCNN Baseline

A custom lightweight network consisting of 3 convolutional blocks (32, 64, 128 filters) followed by a classifier head. It is designed to test if a simple model can learn X-ray features from scratch without pre-training.

- **Total params:** 1,699,393
- **Trainable params:** 1,699,393
- **Non-trainable params:** 0
- **Total Training Time:** 0:04:03

Metric	Train	Validation	Test (Best Model)
Loss	0.0815	0.5103	0.4575
Accuracy	0.9406	0.7500	0.8093
F1 Score	0.9593	0.8000	0.8627

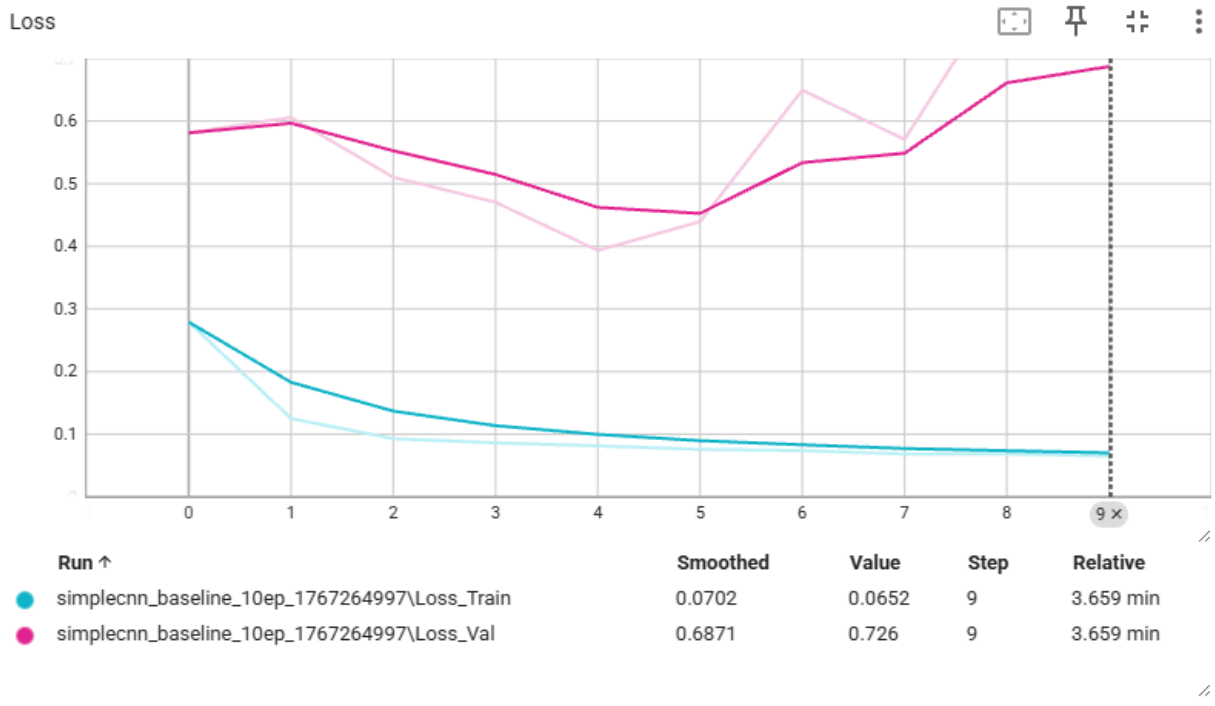


Figure 2: Loss curve for SimpleCNN (10 epochs), showing early convergence.

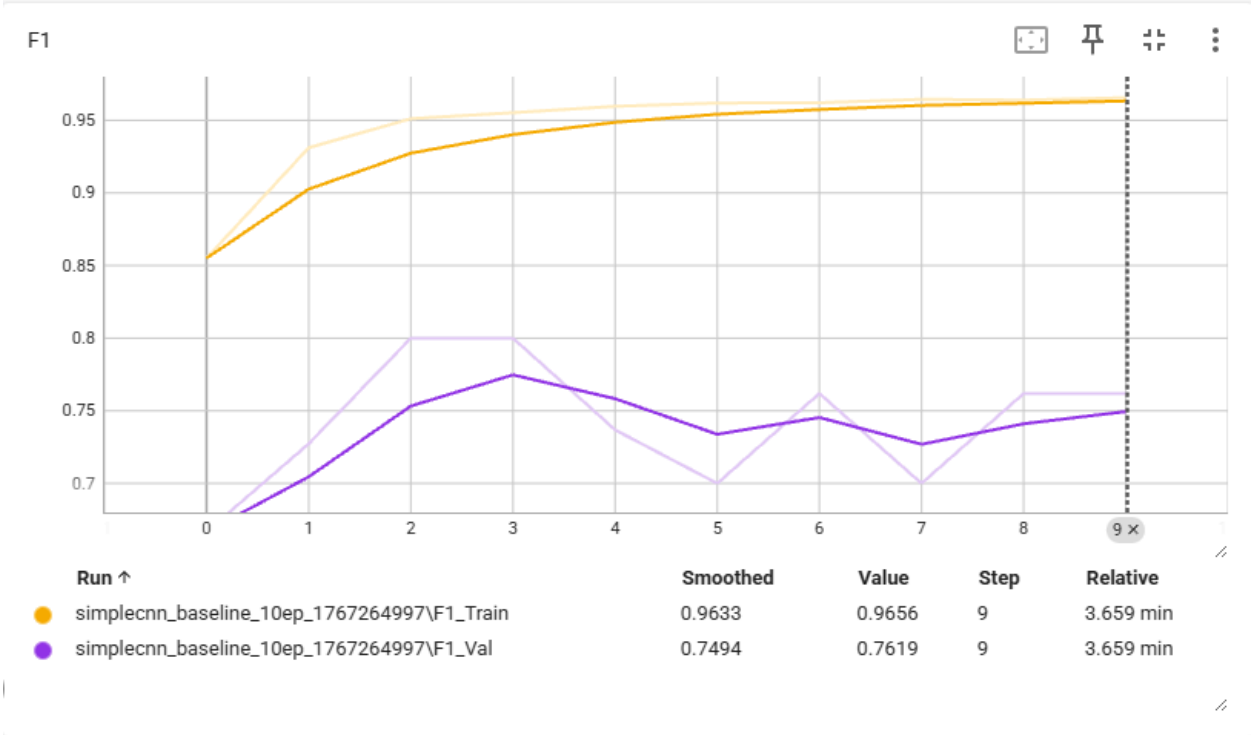


Figure 3: F1-score curve for SimpleCNN (10 epochs).

ResNet152 (Frozen Backbone)

A deep Residual Network where the feature extraction backbone (pre-trained on ImageNet) is frozen, and only the final classification layer is trained. This configuration tests the transferability of generic image features to X-rays with minimal computational cost.

- **Total params:** 58,145,857
- **Trainable params:** 2,049
- **Non-trainable params:** 58,143,808
- **Total Training Time:** 0:04:09

Metric	Train	Validation	Test (Best Model)
Loss	0.2131	0.4051	0.3216
Accuracy	0.8951	0.7500	0.8141
F1 Score	0.9260	0.8000	0.8599

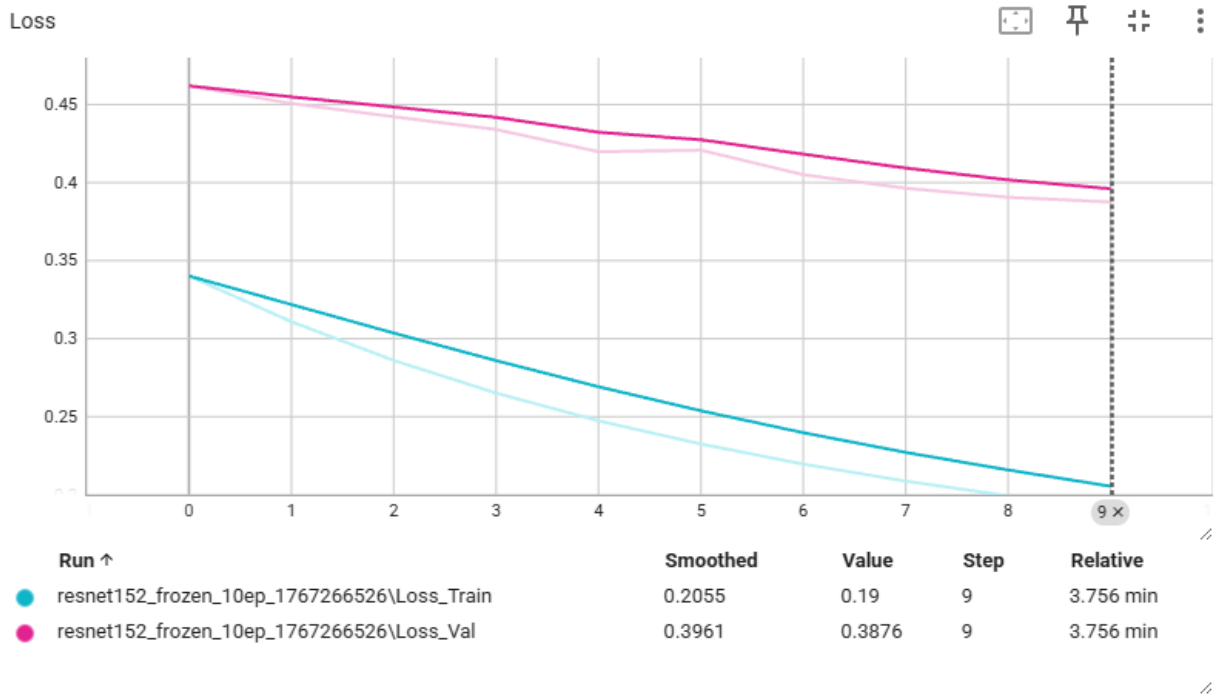


Figure 4: Loss curve for ResNet152 Frozen (10 epochs)

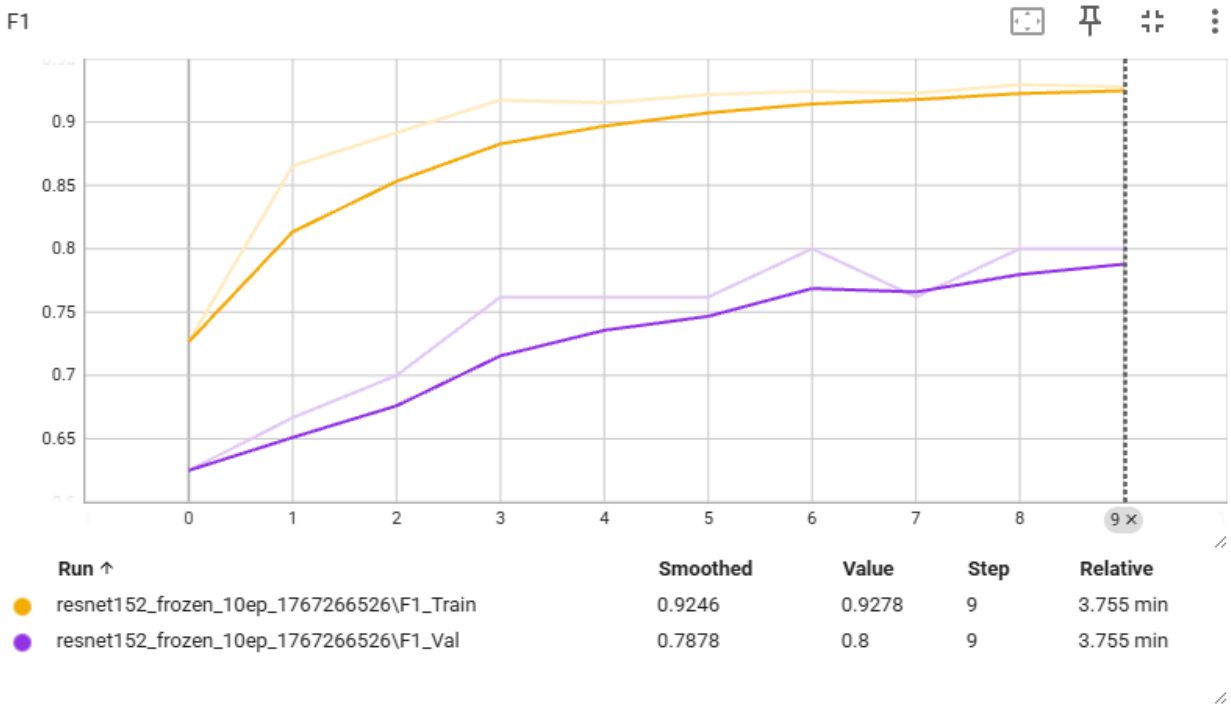


Figure 5: F1-score curve for ResNet152 Frozen (10 epochs).

ResNet152 (Unfrozen)

The same ResNet152 architecture but with all layers unfrozen, allowing the deep feature extractors to be fine-tuned specifically for X-ray pathology. This setup captures more subtle patterns but carries a higher risk of overfitting.

- **Total params:** 58,145,857
- **Trainable params:** 58,145,857
- **Non-trainable params:** 0
- **Total Training Time:** 0:04:15

Metric	Train	Validation	Test (Best Model)
Loss	0.0006	0.0010	0.5863
Accuracy	1.0000	1.0000	0.8622
F1 Score	1.0000	1.0000	0.8998

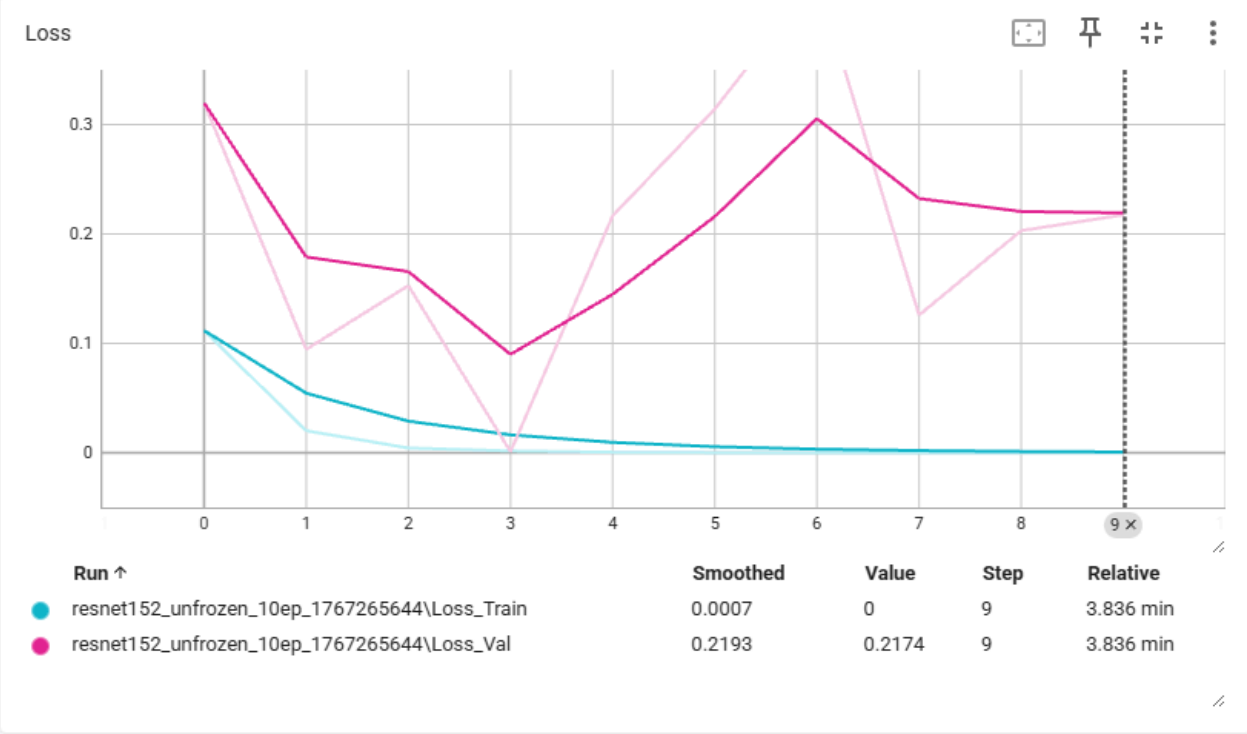


Figure 6: Loss curve for ResNet152 Unfrozen (10 epochs).

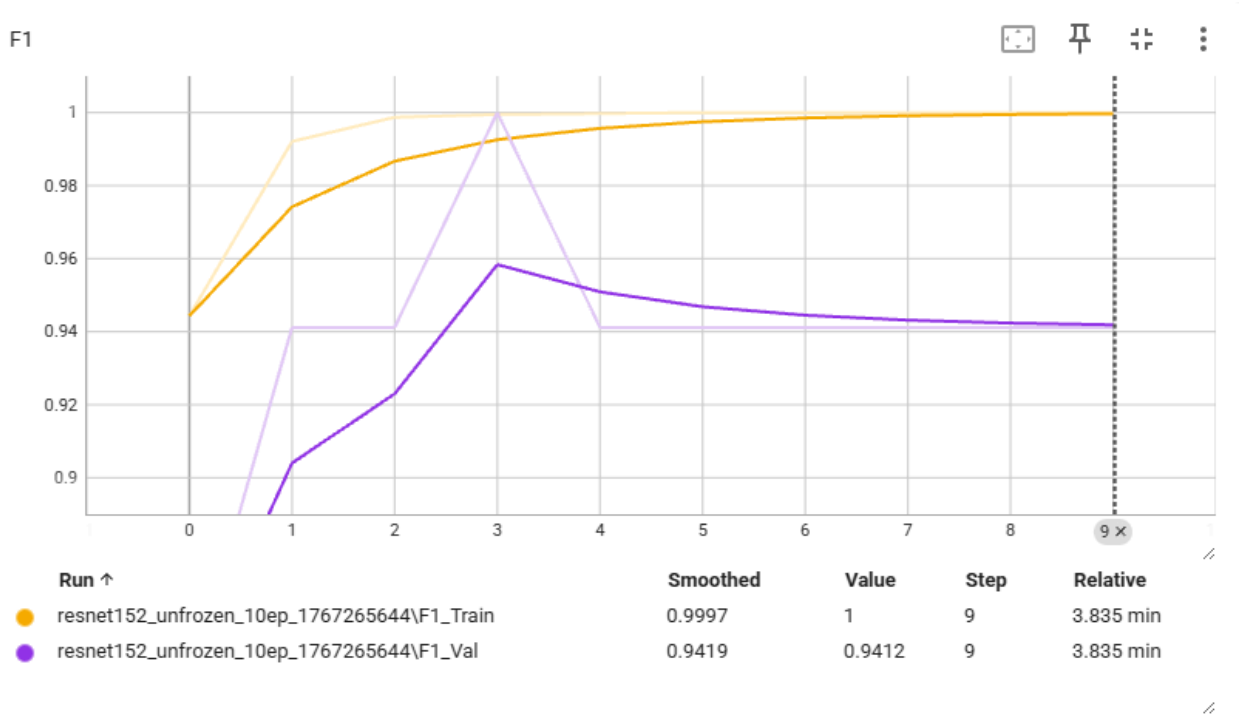


Figure 7: F1-score curve for ResNet152 Unfrozen (10 epochs).

VGG16 (Fronzen Backbone)

A classic deep CNN architecture with the convolutional base frozen. Due to VGG16's architecture involving large dense layers in the classifier head, the number of trainable parameters remains very high even with a frozen backbone.

- **Total params:** 134,273,089
- **Trainable params:** 119,549,953
- **Non-trainable params:** 14,723,136
- **Total Training Time:** 0:04:08

Metric	Train	Validation	Test (Best Model)
Loss	0.0255	0.1211	0.5849
Accuracy	0.9860	1.0000	0.8221
F1 Score	0.9906	1.0000	0.8749

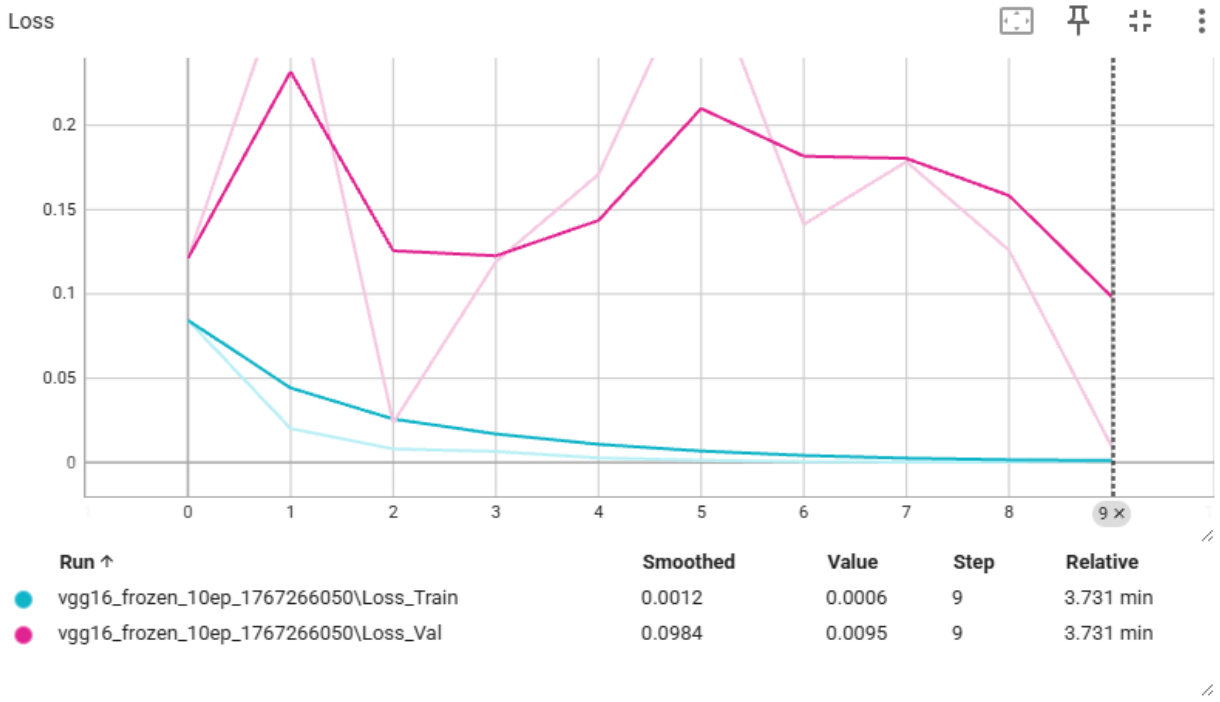


Figure 8: Loss curve for VGG16 Frozen (10 epochs).

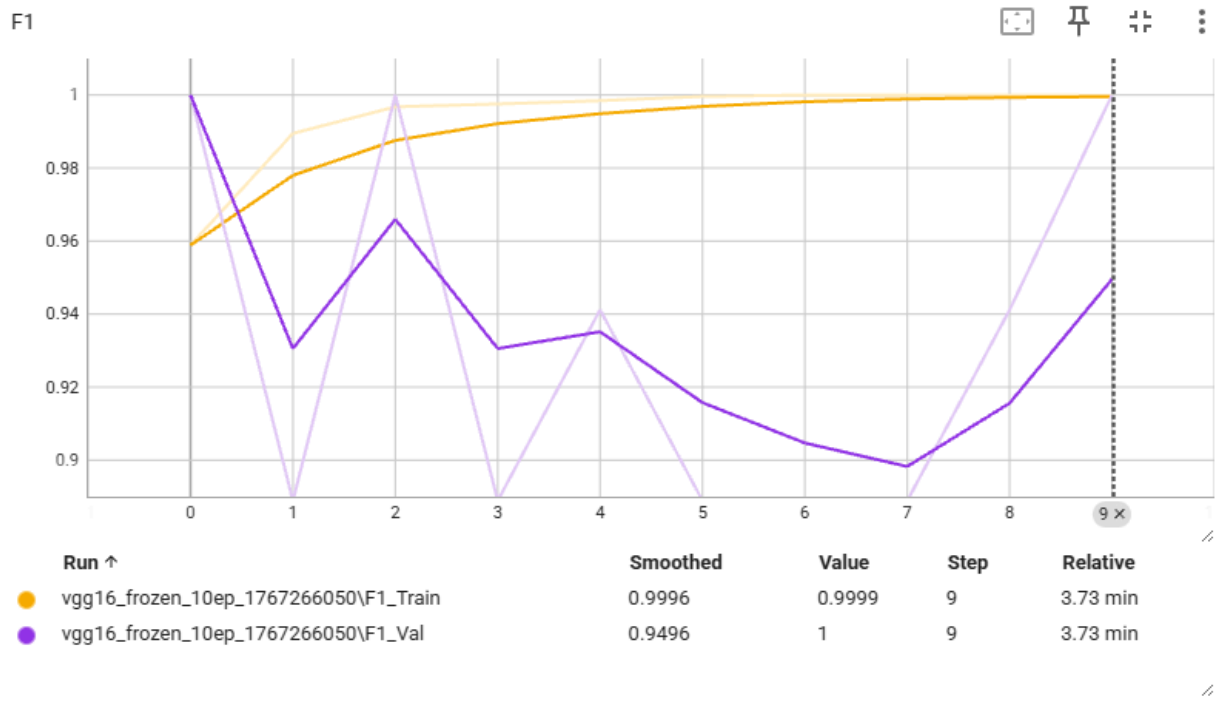


Figure 9: F1-score curve for VGG16 Frozen (10 epochs).

VGG16 (Unfrozen)

The VGG16 architecture with all layers fine-tuned.

- **Total params:** 134,273,089
- **Trainable params:** 134,273,089
- **Non-trainable params:** 0
- **Total Training Time:** 0:04:14

Metric	Train	Validation	Test (Best Model)
Loss	0.0003	0.0027	2.0636
Accuracy	0.9996	1.0000	0.8317
F1 Score	0.9997	1.0000	0.8811

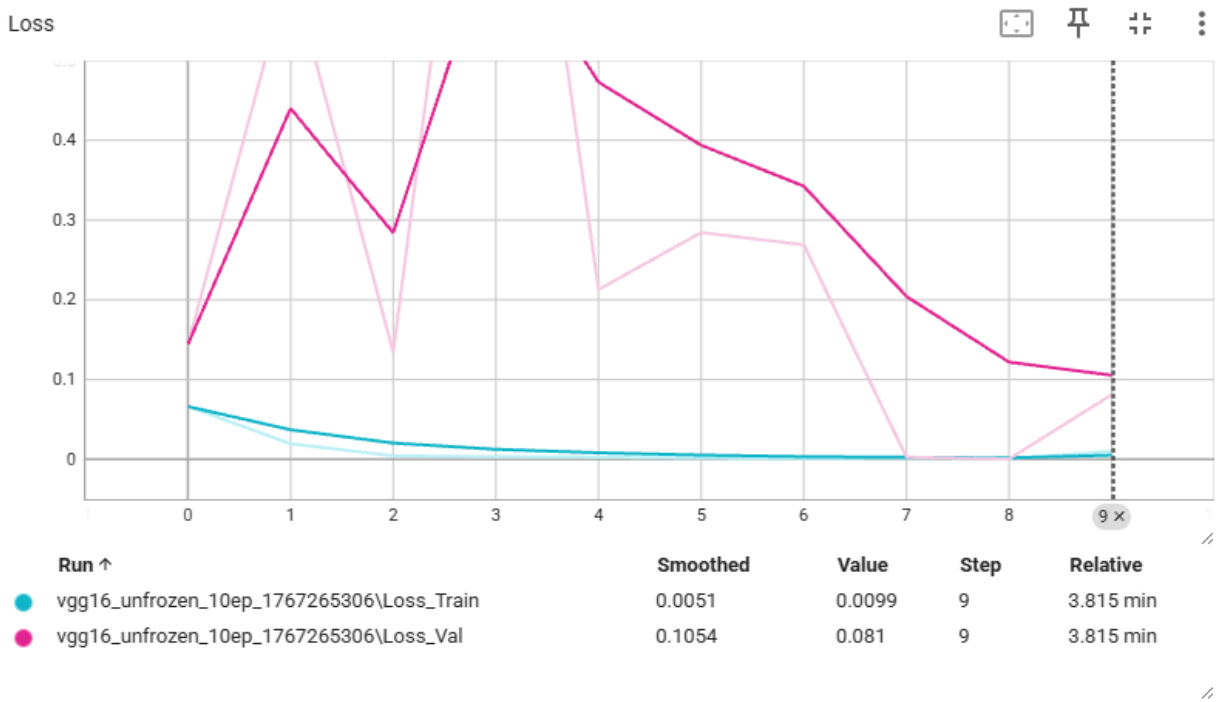


Figure 10: Loss curve for VGG16 Unfrozen (10 epochs).

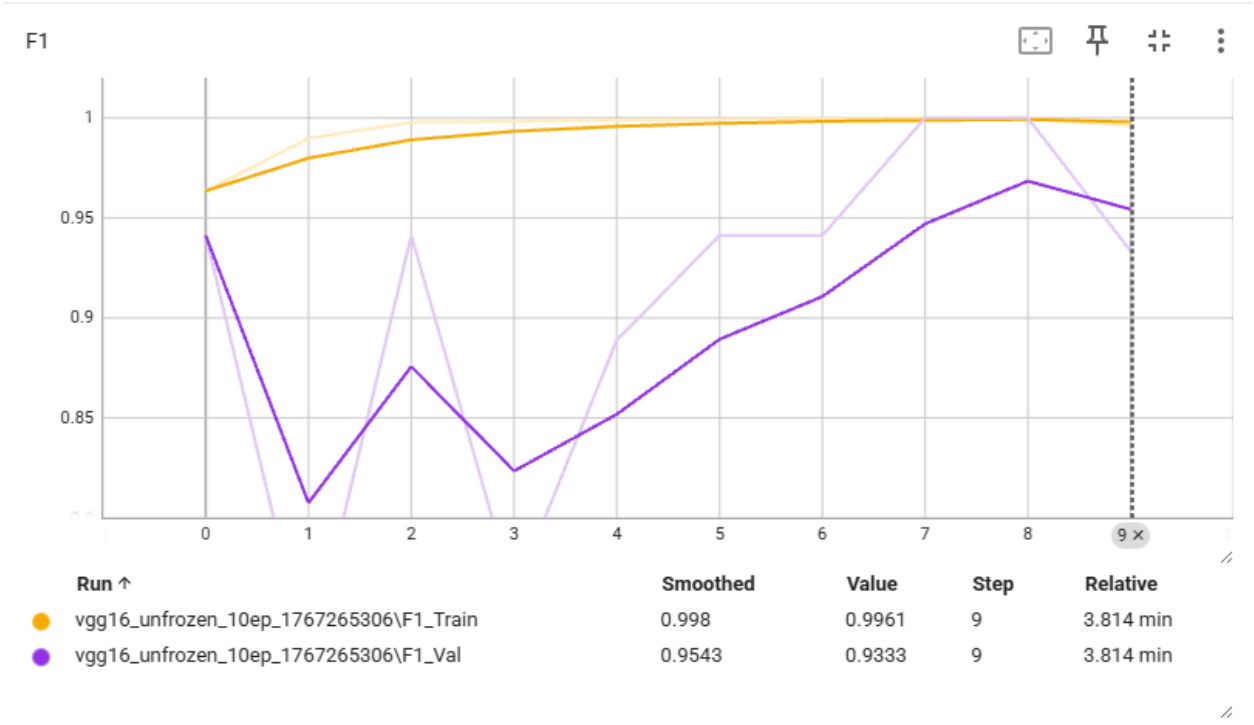


Figure 11: F1-score curve for VGG16 Unfrozen (10 epochs).

Model Comparison

The following table summarizes the performance of the best checkpoint (selected via Val F1) on the hold-out Test set. We've also plotted out the results in Figure 12 and Figure 13.

Model	Trainable Parameters	Test Loss	Test Accuracy	Test F1-Score
SimpleCNN	1.7M	0.4575	0.8093	0.8627
ResNet152 (Frozen)	2K	0.3216	0.8125	0.8585
ResNet152 (Unfrozen)	**58M**	**0.5864**	**0.8622**	**0.8998**
VGG16 (Frozen)	119M	0.5849	0.8221	0.8749

VGG16 (Unfrozen)	134M	2.0633	0.8317	0.8811

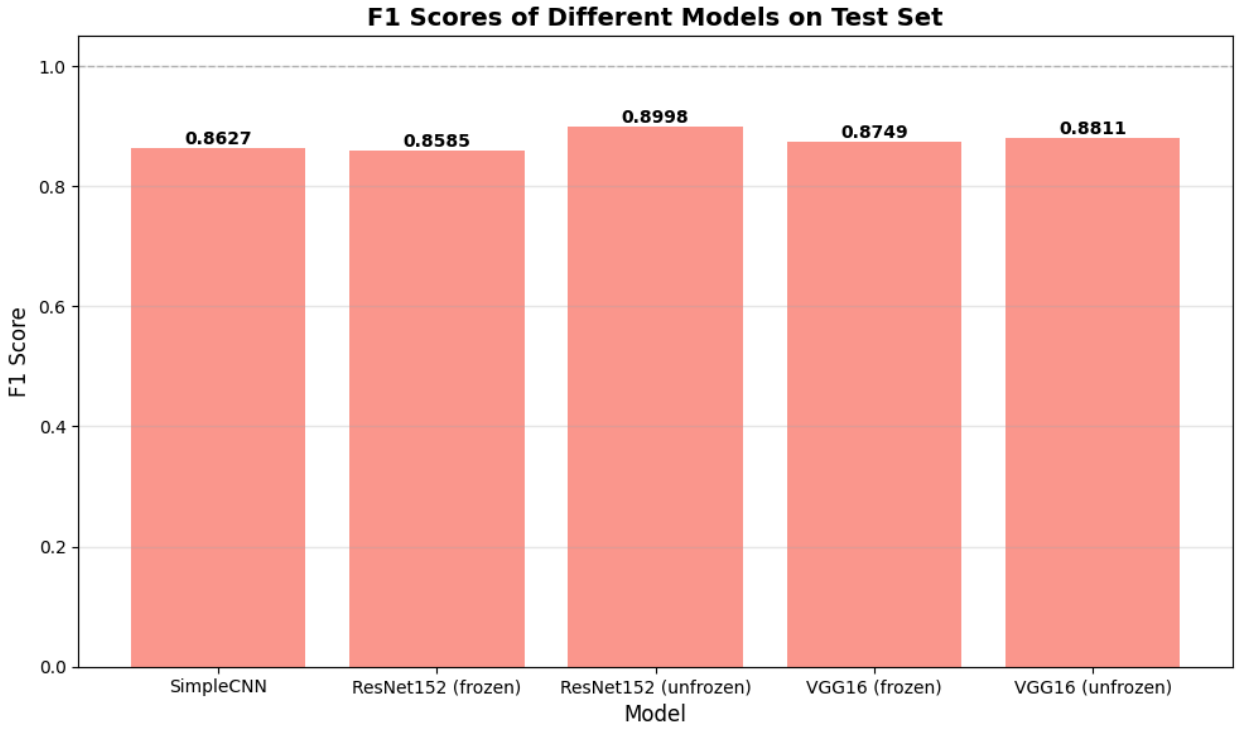


Figure 12: F1 Scores of Different Models on Test Set

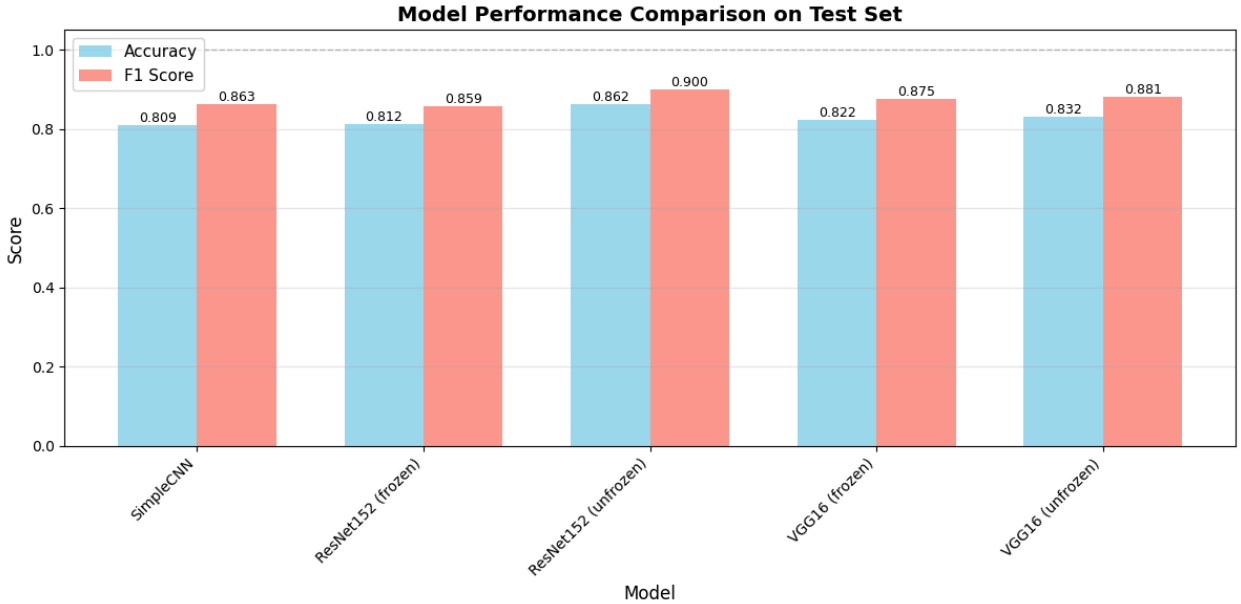


Figure 13: Model Performance Comparison on Test Set, accuracy and f1-score.

Conclusion: ResNet152 (Unfrozen) was selected as the best model. It achieved the highest F1-Score (0.8998) and Accuracy (86.22%), demonstrating that fine-tuning a deep residual network provides the best balance of feature extraction capability and generalization. While the Frozen ResNet152 was the most stable (lowest loss), it lacked the flexibility to fully capture the nuances of pneumonia pathology compared to the unfrozen version.

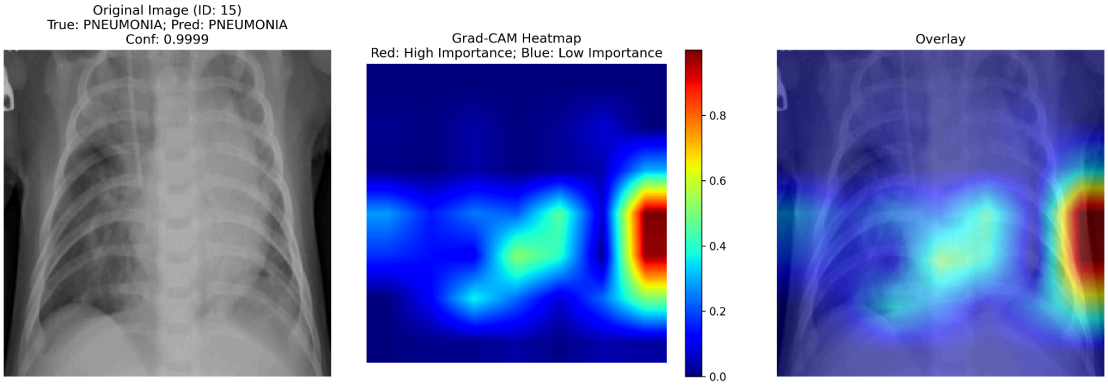
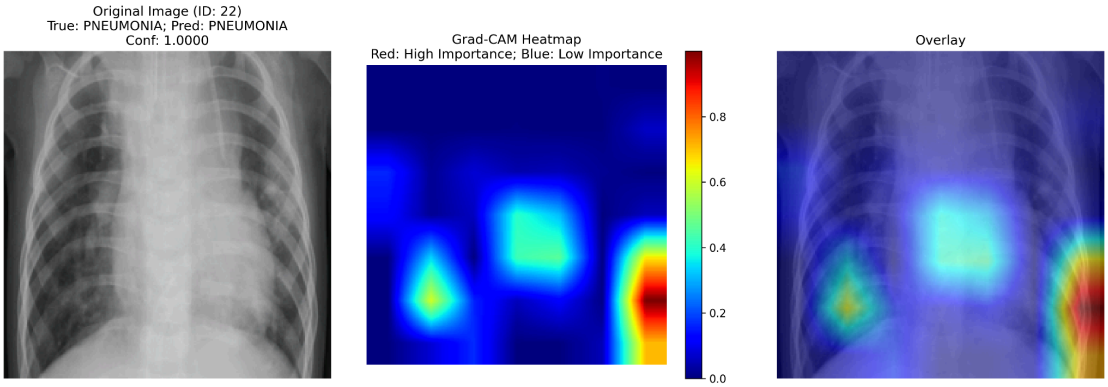
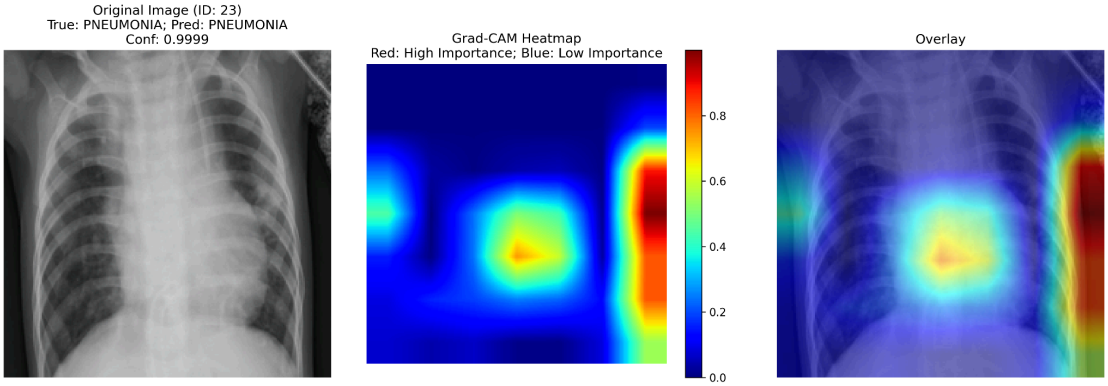
Task 3: XAI Method Implementation (GradCAM)

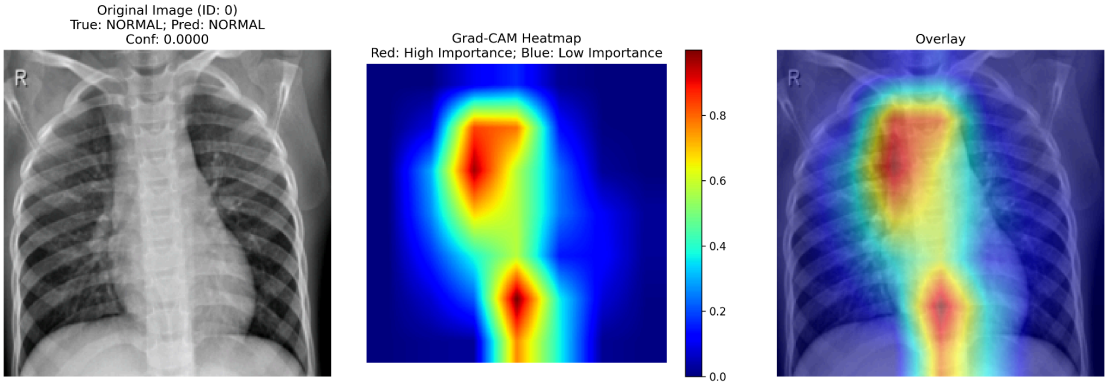
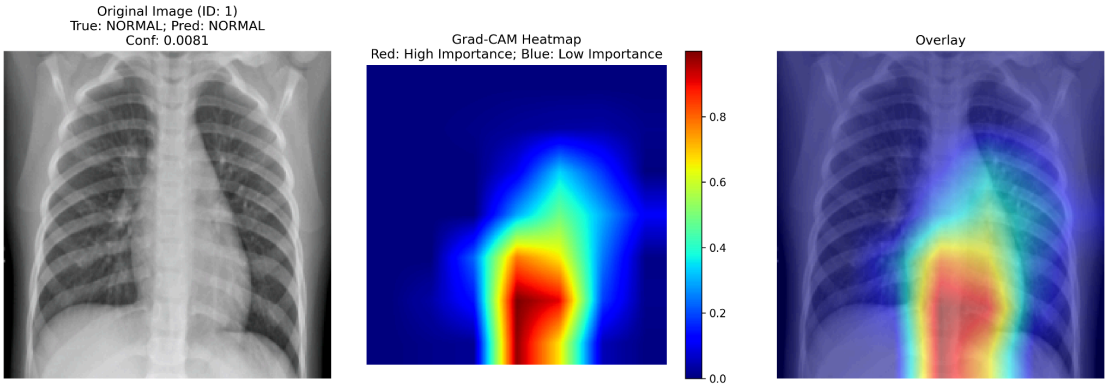
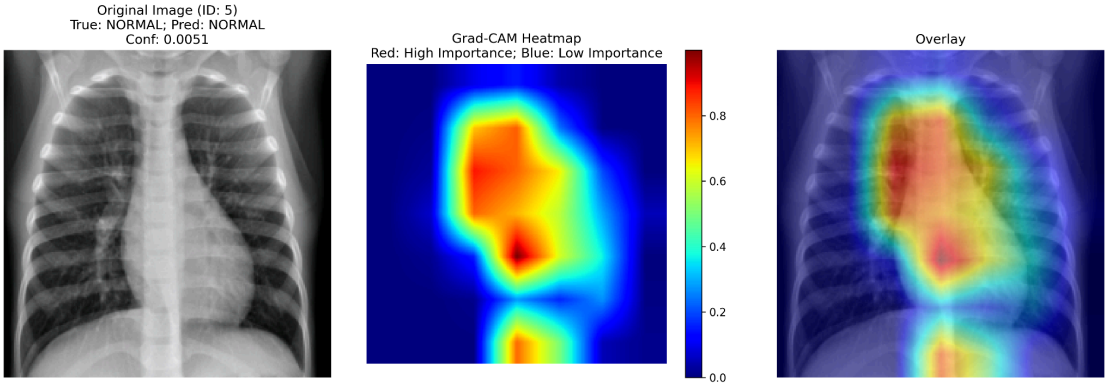
Method Selection For the primary Explainable AI (XAI) analysis, we selected **Grad-CAM (Gradient-weighted Class Activation Mapping)**. Grad-CAM is a widely adopted technique for Convolutional Neural Networks (CNNs) that visualizes the regions of an input image that are important for predictions. Unlike Class Activation Mapping (CAM), Grad-CAM does not require architectural changes (like removing fully connected layers) and works on any differentiable CNN.

Implementation Details

Instead of a custom implementation (since this is already assessed in assignment 2), we utilized the industry-standard `pytorch-grad-cam` library. This ensures a robust, verified implementation of the algorithm that handles gradient computation and hook registration efficiently.

- **Target Layer:** `layer4` (the final convolutional block of ResNet152).
- **Mechanism:** Gradients are pooled to obtain neuron importance weights, which are then combined with the feature maps to generate the heatmap.
- **Visualization:** The heatmap is upsampled and superimposed on the original X-ray image to visualize the specific anatomical regions driving the decision.

Pneumonia	 Original Image (ID: 15) True: PNEUMONIA; Pred: PNEUMONIA Conf: 0.9999 Grad-CAM Heatmap Red: High Importance; Blue: Low Importance Overlay 0.8 0.6 0.4 0.2 0.0
	 Original Image (ID: 22) True: PNEUMONIA; Pred: PNEUMONIA Conf: 1.0000 Grad-CAM Heatmap Red: High Importance; Blue: Low Importance Overlay 0.8 0.6 0.4 0.2 0.0
	 Original Image (ID: 23) True: PNEUMONIA; Pred: PNEUMONIA Conf: 0.9999 Grad-CAM Heatmap Red: High Importance; Blue: Low Importance Overlay 0.8 0.6 0.4 0.2 0.0

Normal	 Original Image (ID: 0) True: NORMAL; Pred: NORMAL Conf: 0.0000 Grad-CAM Heatmap Red: High Importance; Blue: Low Importance Overlay 0.8 0.6 0.4 0.2 0.0  Original Image (ID: 1) True: NORMAL; Pred: NORMAL Conf: 0.0081 Grad-CAM Heatmap Red: High Importance; Blue: Low Importance Overlay 0.8 0.6 0.4 0.2 0.0  Original Image (ID: 5) True: NORMAL; Pred: NORMAL Conf: 0.0051 Grad-CAM Heatmap Red: High Importance; Blue: Low Importance Overlay 0.8 0.6 0.4 0.2 0.0

Task 4: Written XAI Explanation for Task 3

1. Analysis of Normal Cases (IDs 0, 1, 5) In the correctly classified Normal cases, the Grad-CAM heatmaps predominantly focus on the **mediastinum** (the central region containing the heart and spine) and the clear lung fields.

- **Medical Interpretation:** This suggests the model is employing negative reasoning. By focusing on the sharp edges of the **cardiac silhouette** (the heart border), the model is confirming the **absence of the "Silhouette Sign."** In pneumonia patients, the heart border is often obscured by fluid; by verifying the heart border is crisp and defined, the model correctly deduces the lung is normal.
- **Conclusion:** The model correctly identifies "Normal" by recognizing the structural integrity of the central chest anatomy.

2. Analysis of Pneumonia Cases (IDs 15, 22, 23) For the Pneumonia cases, the model behavior reveals both correct pathological detection and reliability issues.

- **Correct Detection:** In several instances, the heatmap correctly highlights areas of **consolidation**, the "fluffy," radiopaque (white) patches where fluid has accumulated. This indicates the model has learned the primary visual feature of bacterial pneumonia.
- **Shortcut Learning:** However, a concerning pattern is observed where high-confidence regions appear **outside the thoracic cage** (e.g., background artifacts). This is a classic example of **"Shortcut Learning,"** where the model exploits non-medical correlations (like scanner tags or cables) common in the Pneumonia dataset. Despite using a Weighted Loss to fix class imbalance, the model still latched onto these background artifacts.

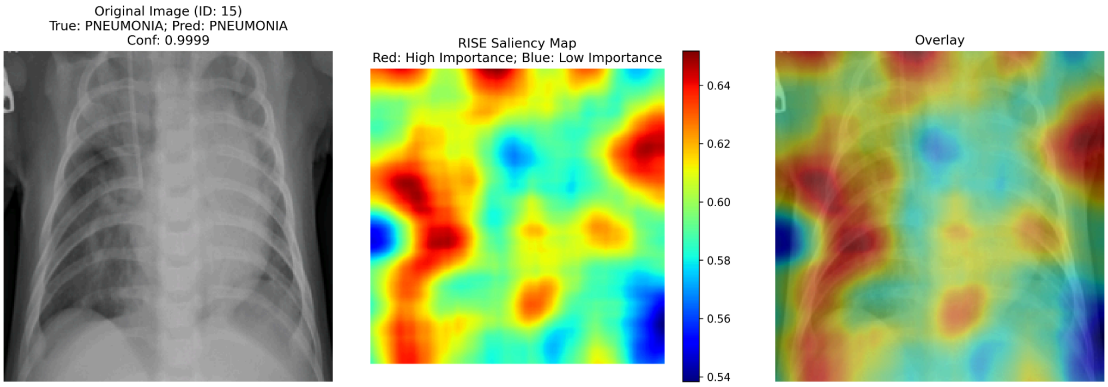
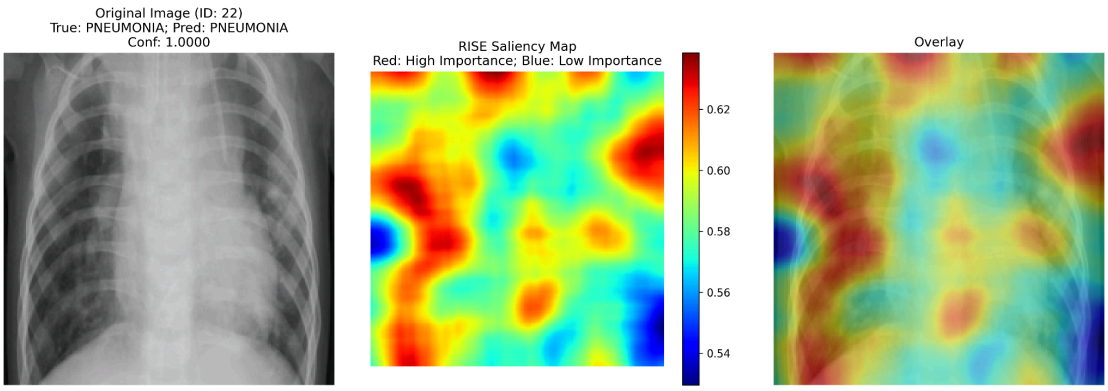
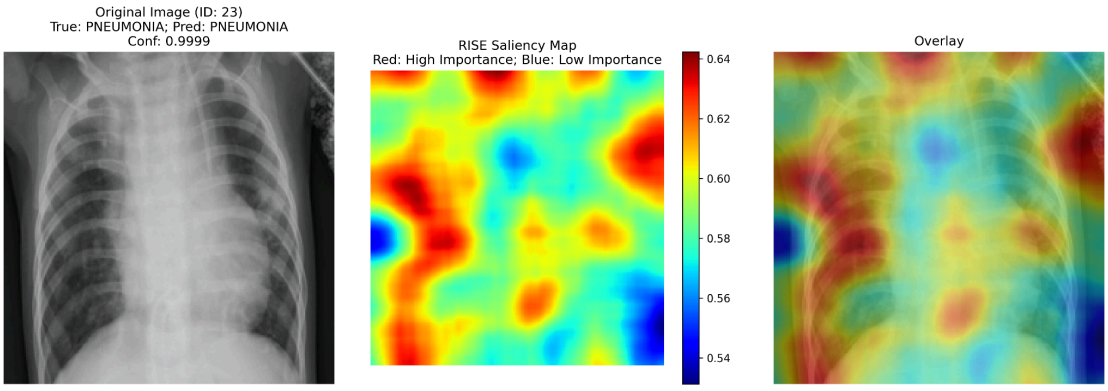
3. Conclusion on Reliability While the model achieves high F1 scores, the XAI analysis reveals it is **not fully reliable** for clinical deployment due to its reliance on background artifacts.

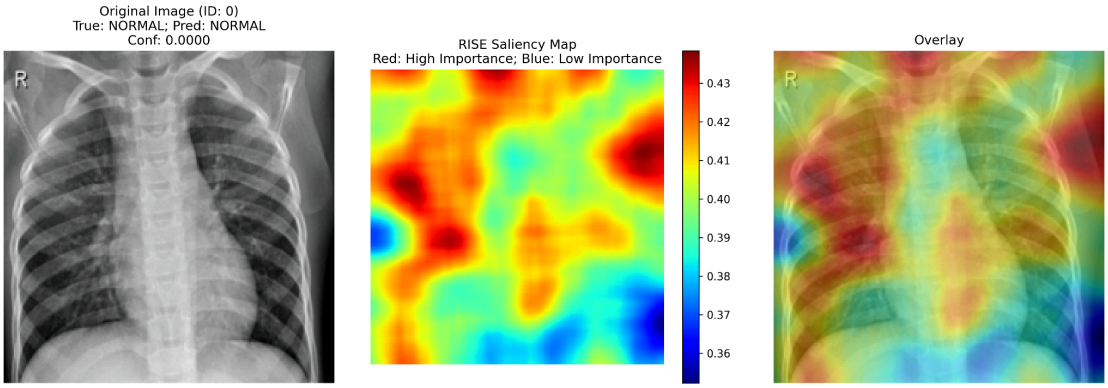
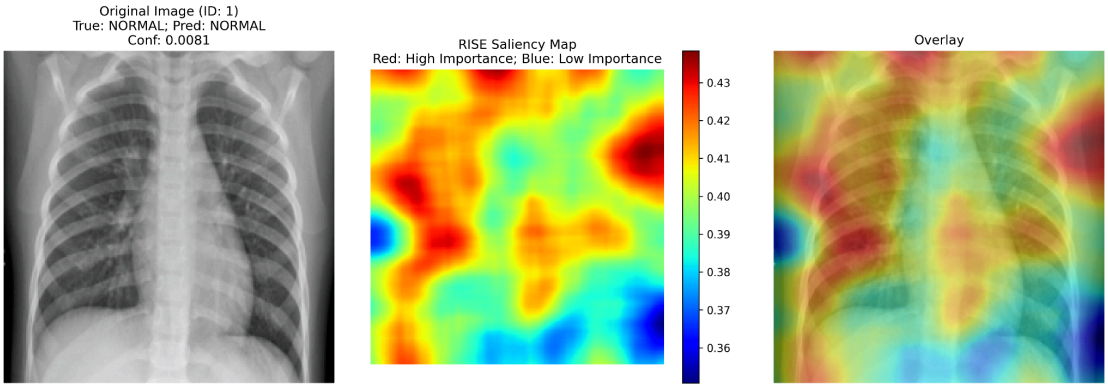
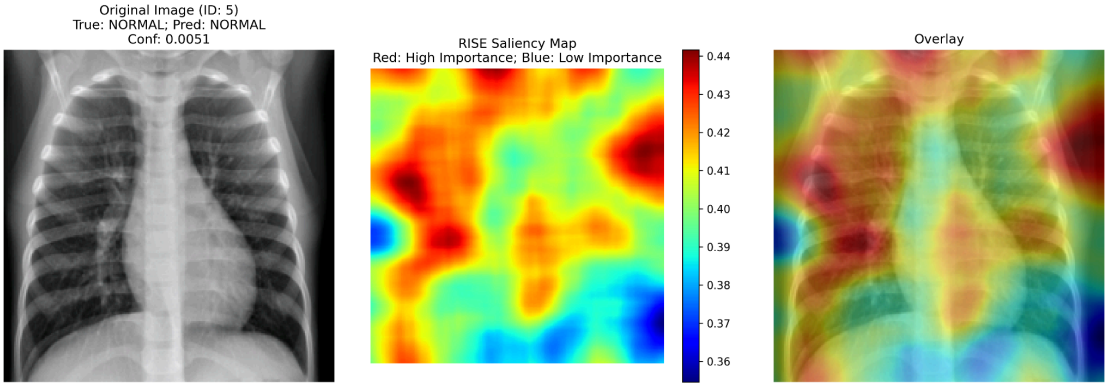
Task 5: Second XAI Method Implementation (RISE)

For the second method, we implemented **RISE (Randomized Input Sampling for Explanation)**. Unlike Grad-CAM, which relies on gradients, RISE is a black-box perturbation method.

Implementation Details:

- **Mechanism:** We generated **2,000 random binary masks** (upsampled from an 8 x 8 grid) for every image.
- **Scoring:** These masks were applied to the input image, and the occluded images were passed through the model. The final saliency map was computed as a weighted sum of the masks, where the weight is the model's confidence score for the target class.
- **Logic:** If masking a specific region causes the confidence to drop significantly, that region is considered highly important.

Pneumonia	  

Normal	 Original Image (ID: 0) True: NORMAL; Pred: NORMAL Conf: 0.0000 RISE Saliency Map Red: High Importance; Blue: Low Importance 0.43 0.42 0.41 0.40 0.39 0.38 0.37 0.36 Overlay  Original Image (ID: 1) True: NORMAL; Pred: NORMAL Conf: 0.0081 RISE Saliency Map Red: High Importance; Blue: Low Importance 0.43 0.42 0.41 0.40 0.39 0.38 0.37 0.36 Overlay  Original Image (ID: 5) True: NORMAL; Pred: NORMAL Conf: 0.0051 RISE Saliency Map Red: High Importance; Blue: Low Importance 0.44 0.43 0.42 0.41 0.40 0.39 0.38 0.37 0.36 Overlay

Task 6: Written Explanation for Second XAI Method

1. Interpretation of RISE Heatmaps

Unlike Grad-CAM, which relies on the gradients of the target class to identify discriminative regions, **Randomized Input Sampling for Explanation (RISE)** utilizes randomized masking (perturbation) to empirically test the model's sensitivity to every region of the input image.

- **Normal Cases (IDs 0, 1, 5):** The RISE heatmaps for normal cases exhibit broad, continuous areas of high importance (red regions) that cover the clear lung fields, particularly the right lung, and the mediastinum (heart and spine region).
    - **Medical Insight:** This pattern suggests the model employs a strategy of detecting textural uniformity. The continuous positive signal indicates the model relies on the presence of "radiolucency" (the characteristic blackness of air-filled lungs) across the entire lung field to predict a "Normal" status, rather than identifying a single focal feature.
- **Pneumonia Cases (IDs 15, 22, 23):** In contrast, the heatmaps for pneumonia cases are significantly more fragmented. They display scattered "blobs" of high importance that correspond to areas of consolidation (the "fluffy" white patches indicative of fluid or infection).
    - **Medical Insight:** This fragmentation implies the model is reacting to specific, localized irregularities in lung texture, confirming it detects pathology based on focal disruptions rather than global shape.

2. Comparison with Grad-CAM

A comparative analysis of RISE and Grad-CAM reveals fundamental differences in how the model's decision-making process is visualized:

- **Localization vs. Granularity:** Grad-CAM produces smooth, cohesive heatmaps because it operates on low-resolution feature maps (e.g., 7×7) which are then upsampled. Conversely, RISE generates fine-grained, pixel-level sensitivity maps. While Grad-CAM effectively highlights where the model is looking in a general sense, RISE reveals exactly which pixels cause the confidence score to drop when occluded.

- **Sensitivity to Noise & Artifacts:** The RISE heatmaps contain significantly more "artifacts," such as scattered red spots in the background corners (top-left and top-right). This visualization exposes the model's fragility rather than the method's flaw. Because RISE tests random occlusions, it proves that the model is chemically sensitive to background noise; altering pixels in the corners does materially affect the probability score, a behavior that Grad-CAM's smoothing effect largely obscured.

3. Critical Reflection on Model Trustworthiness

The RISE analysis reinforces and amplifies the "Shortcut Learning" hypothesis initially observed in the Grad-CAM analysis.

- **Trustworthiness Issues:** The fact that RISE highlights scattered pixels outside the thoracic cage confirms that the model is over-sensitive to background artifacts. While Grad-CAM showed the model attended to the background, RISE provides empirical evidence that these background pixels actively drive the classification decision.
- **Method Evaluation & Clinical Utility:** While Grad-CAM appears "cleaner" and allows for rapid assessment by a clinician, RISE provides a more honest, albeit noisier, representation of the model's pixel-level reliance. However, for this specific model, Grad-CAM is the superior visualization for clinical interpretability. The extreme granularity of RISE makes it difficult to distinguish between true pathological detection and the model's reliance on random noise, potentially confusing a diagnosis rather than aiding it.

References

- Chest X-Ray Images (Pneumonia): https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia/data
- RISE Implementation: https://github.com/eclique/RISE/tree/master
- GradCAM: https://jacobgil.github.io/pytorch-gradcam-book/introduction.html

Appendix

Full documentation and source code are available in the GitHub repository.

- **Source Code:** https://github.com/hongjiaherng/wqf7009-xai-pneumonia/tree/main/src/wqf7009_a3
- **Interactive Notebook:** solution.ipynb
- **Static Exports:** PDF and HTML (HTML recommended for better readability, download required)
- **Training logs:** https://github.com/hongjiaherng/wqf7009-xai-pneumonia/tree/main/runs