Binary Classification of Chest Xray Images for Pneumonia Detection Using Deep Learning

Janak Sapkota
NAAMII Research Assignment – Medical Imaging Track
December 2025

Abstract

This report presents the development and evaluation of deep learning models for automated pneumonia detection from chest X-ray images using the PneumoniaMNIST dataset. Two classification approaches were investigated: (1) a custom convolutional neural network (CNN) trained from scratch, achieving 83.98% balanced accuracy with 98.72% sensitivity after threshold optimization, and (2) a ResNet-18 model employing transfer learning, achieving 84.27% balanced accuracy with 99.74% sensitivity. To address class imbalance, systematic decision-threshold optimization was applied, resulting in an 80% reduction in false negatives (from five to one case). Grad-CAM visualizations indicate that both models attend to clinically relevant lung regions. While ResNet-18 demonstrates superior safety for screening by missing only 1 of 390 pneumonia cases, the custom CNN provides approximately 19× parameter efficiency, making it a compelling option for deployment in resource-constrained environments.

Keywords: Pneumonia Detection, Deep Learning, CNN, Transfer Learning, Medical Imaging, Class Imbalance

Contents

1 Dataset and Problem Statement

1.1 Problem Definition

Pneumonia is a leading cause of morbidity and mortality worldwide, particularly among children under five years of age and elderly populations. Early and accurate diagnosis through chest X-ray analysis is essential for timely clinical intervention and improved patient outcomes. However, manual interpretation of radiographs requires expert radiologists and can be time-consuming, creating significant challenges in resource-limited healthcare settings.

Objective: The primary objective of this study is to develop automated binary classification models capable of distinguishing between normal chest X-rays and those exhibiting pneumonia-related abnormalities, with an emphasis on:

1. **High sensitivity** to minimize false negatives (missed pneumonia cases)
2. **Adequate specificity** to reduce false positives and unnecessary follow-up examinations

1.2 Dataset: PneumoniaMNIST

The PneumoniaMNIST dataset, a subset of the MedMNIST benchmark collection, was used in this study. It consists of pediatric chest X-ray images curated for standardized medical image classification tasks.

Dataset Characteristics:

- **Total images:** 5,856 chest X-rays
- **Image resolution:** 28×28 pixels, grayscale (downsampled from clinical-resolution images)
- **Classification task:** Binary
  - Class 0: Normal (no pneumonia)
  - Class 1: Pneumonia (bacterial or viral)
- **Data source:** Pediatric patients from Guangzhou Women and Children's Medical Center
- **Data split:** Predefined training, validation, and test sets

1.3 Class Distribution

Table 1: Dataset composition and class imbalance analysis

Split	Normal	Pneumonia	Total	Normal (%)	Ratio
Training	1,583	3,125	4,708	33.6%	1:1.97
Validation	234	290	524	44.7%	1:1.24
Test	234	390	624	37.5%	1:1.67
Total	2,051	3,805	5,856	35.0%	1:1.86

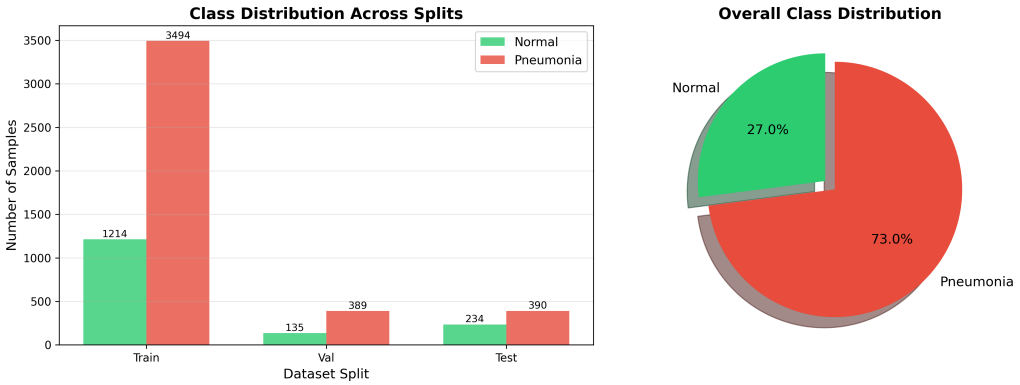


Figure 1: Class distribution across training, validation, and test splits. The approximately 2:1 imbalance in favor of pneumonia cases reflects common trends in clinical imaging datasets, where symptomatic patients are overrepresented.

1.4 Sample Images

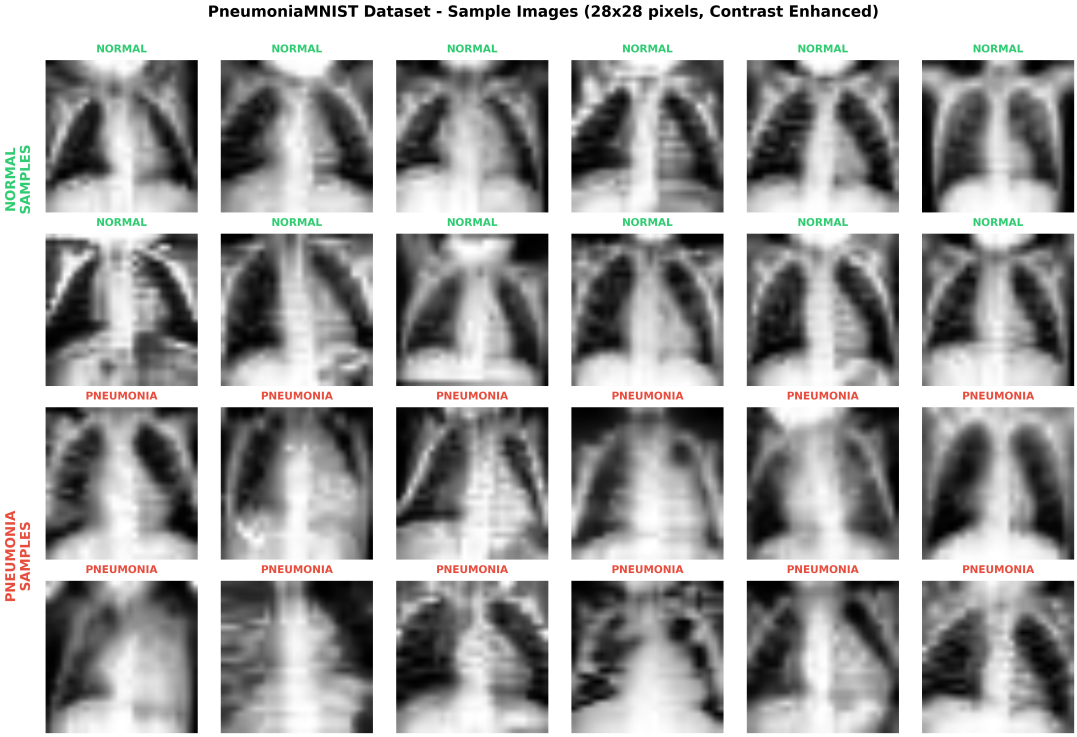


Figure 2: Representative chest X-ray samples from the PneumoniaMNIST dataset. The top rows show normal cases with clear lung fields, while the bottom rows illustrate pneumonia cases characterized by increased opacity, infiltrates, and consolidation patterns.

Image Quality Consideration: The 28×28 pixel resolution represents a substantial reduction from typical clinical X-rays (512×512 or higher). Although fine anatomical details are lost, this resolution enables rapid experimentation, reduced computational cost, and retains sufficient information for identifying dominant pneumonia-related patterns such as lung opacity and asymmetry.

2 Model Architectures

We developed two distinct classification models to compare performance, efficiency, and applicability for different deployment scenarios.

2.1 Model 1: Custom Convolutional Neural Network

We designed a lightweight CNN architecture specifically optimized for the 28×28 input resolution and binary classification task.

2.1.1 Architecture Details

Table 2: Custom CNN Architecture Layer-by-Layer Breakdown

Block	Layer Configuration	Output Shape	Parameters
Input	Grayscale image	28×28×1	0
Block 1	Conv2D(32, 3×3) + BatchNorm + ReLU	28×28×32	320
	Conv2D(32, 3×3) + BatchNorm + ReLU	28×28×32	9,248
	MaxPool2D(2×2)	14×14×32	0
	Dropout(0.25)	14×14×32	0
Block 2	Conv2D(64, 3×3) + BatchNorm + ReLU	14×14×64	18,496
	Conv2D(64, 3×3) + BatchNorm + ReLU	14×14×64	36,928
	MaxPool2D(2×2)	7×7×64	0
	Dropout(0.25)	7×7×64	0
Block 3	Conv2D(128, 3×3) + BatchNorm + ReLU	7×7×128	73,856
	Conv2D(128, 3×3) + BatchNorm + ReLU	7×7×128	147,584
	MaxPool2D(2×2)	3×3×128	0
	Dropout(0.25)	3×3×128	0
Classifier	Flatten	1,152	0
	Linear(256) + ReLU + Dropout(0.5)	256	295,168
	Linear(128) + ReLU + Dropout(0.5)	128	32,896
	Linear(1) [Logit output]	1	129
	Sigmoid (during inference)	1	0
Total Trainable Parameters			**605,521**
Model Size			**2.3 MB**

Design Rationale:

- **Progressive Channel Expansion (32→64→128):** Hierarchical feature learning from simple edges to complex patterns
- **Batch Normalization:** Stabilizes training, enables higher learning rates, reduces internal covariate shift
- **Dropout Regularization:** Prevents overfitting (25% spatial dropout after conv blocks, 50% after dense layers)
- **Double Convolutions:** Two 3×3 convolutions per block approximate larger receptive fields with fewer parameters
- **Compact Classifier:** Two-layer FC network (256→128→1) provides sufficient capacity without overparameterization

2.2 Model 2: ResNet-18 with Transfer Learning

To validate our custom architecture and leverage pre-trained knowledge, we implemented ResNet-18 initialized with ImageNet weights.

2.2.1 Architecture Overview

- **Base Architecture:** ResNet-18 (18-layer deep residual network)
- **Pre-training:** ImageNet (1.2M images, 1000 classes)
- **Total Parameters:** 11,689,512 (11.7M)
- **Model Size:** 44.7 MB
- **Input Requirements:** 224×224×3 (RGB format)

2.2.2 Modifications for Binary Classification

1. **Input Preprocessing:**
   - Convert 28×28 grayscale to 224×224 RGB (replicate channel 3×)
   - Normalize with ImageNet statistics (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225])
2. **Final Layer Replacement:**
   - Original: Linear($512 \rightarrow 1000$)
   - Modified: Sequential(Dropout(0.5), Linear($512 \rightarrow 1$))

2.2.3 Two-Stage Fine-Tuning Strategy

Stage 1 (10 epochs): Feature Extraction

- Freeze all convolutional layers (backbone)
- Train only modified final classifier
- Learning rate: 0.001
- **Goal:** Rapid adaptation to pneumonia detection task while preserving ImageNet features

Stage 2 (15 epochs): Full Fine-Tuning

- Unfreeze all layers
- Fine-tune entire network
- Learning rate: 0.0001 (reduced to prevent catastrophic forgetting)
- **Goal:** Adapt low-level features specifically for chest X-ray characteristics

Transfer Learning Advantages:

- Pre-trained features from 1.2M diverse images provide robust initialization
- Deeper architecture (18 layers) enables more complex feature hierarchies
- Residual connections facilitate gradient flow and training stability
- Proven architecture with established performance on medical imaging tasks

3 Training Methodology

3.1 Data Preprocessing and Augmentation

Training Set Augmentation:

- **RandomHorizontalFlip (p=0.5):** Chest X-rays exhibit approximate bilateral symmetry
- **RandomRotation (±10°):** Accounts for minor patient positioning variations
- **Normalization:** mean=0.5, std=0.5 (standardize to [-1, 1] range)

Validation/Test Sets:

- No augmentation applied
- Normalization only (maintain evaluation consistency)

Rationale: Data augmentation increases training diversity, prevents overfitting to specific orientations, and improves generalization without requiring additional labeled data.

3.2 Handling Class Imbalance

The 2:1 class imbalance requires explicit mitigation to prevent majority class bias.

Strategy: Weighted Loss Function

We implemented class-weighted Binary Cross-Entropy loss:

$$\mathcal{L}_{weighted} = -\left[w_{pos} \cdot y \cdot \log(\sigma(x)) + w_{neg} \cdot (1-y) \cdot \log(1-\sigma(x))\right] \tag{1}$$

where:

- $\sigma(x) = \frac{1}{1+e^{-x}}$ (sigmoid activation)
- $y \in \{0, 1\}$ (true label)
- w_{pos}, w_{neg} are class-specific penalty weights

Weight Calculation:

$$w_{normal} = \frac{N_{total}}{2 \times N_{normal}} = \frac{4708}{2 \times 1583} \approx 1.487 \tag{2}$$

$$w_{pneumonia} = \frac{N_{total}}{2 \times N_{pneumonia}} = \frac{4708}{2 \times 3125} \approx 0.753 \tag{3}$$

$$w_{pos} = \frac{w_{pneumonia}}{w_{normal}} \times 0.5 \approx 0.506 \tag{4}$$

Weight Factor = 0.5: Reduced from full weighting (1.0) to prevent over-correction causing excessive false positives. This hyperparameter was tuned based on validation set performance.

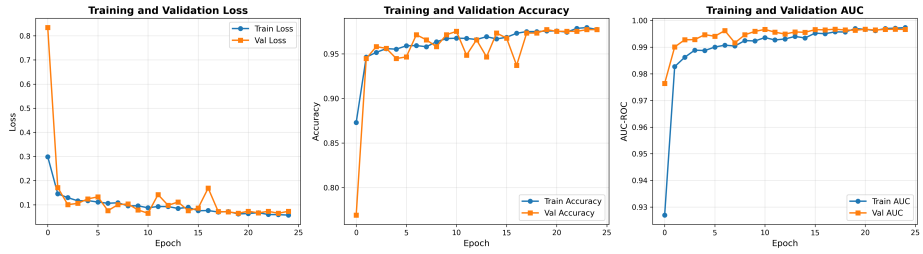
3.3 Training Configuration

Table 3: Training Hyperparameters and Configuration

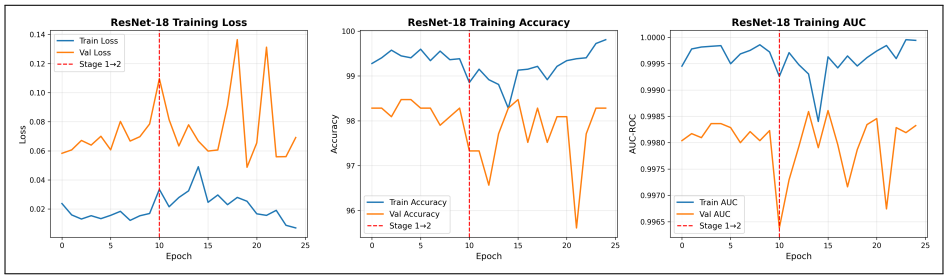
Hyperparameter	Custom CNN	ResNet-18
Optimizer	Adam	Adam
Learning Rate	0.001	0.001 (Stage 1) 0.0001 (Stage 2)
Weight Decay	1×10^{-4}	1×10^{-4}
Batch Size	128	32
Total Epochs	25	25 (10+15)
Loss Function	BCEWithLogitsLoss	BCEWithLogitsLoss
Class Weights	Yes (factor=0.5)	Yes (factor=0.5)
LR Scheduler	ReduceLROnPlateau (patience=3, factor=0.5)	ReduceLROnPlateau (patience=3, factor=0.5)
Early Stopping	Save best validation accuracy	Save best validation accuracy

3.4 Training Environment

- **Hardware:** Google Colab Tesla T4 GPU (16GB VRAM)
- **Framework:** PyTorch 2.0+
- **Training Time:**
  - Custom CNN: 15 minutes
  - ResNet-18: 20 minutes (10 min Stage 1 + 10 min Stage 2)



(a) Custom CNN training curves



(b) ResNet-18 training curves (two-stage)

Figure 3: Training and validation loss and accuracy curves. The custom CNN shows smooth convergence over 25 epochs. ResNet-18 exhibits rapid adaptation during Stage 1 (epochs 1–10), followed by gradual improvement in Stage 2 (epochs 11–25). No significant overfitting is observed for either model.

4 Evaluation Metrics

We evaluated models using comprehensive metrics appropriate for imbalanced medical classification tasks.

4.1 Metrics Definitions

Basic Metrics:

- **Accuracy:** $\frac{TP+TN}{TP+TN+FP+FN}$ - Overall correctness
- **Precision:** $\frac{TP}{TP+FP}$ - Positive prediction reliability
- **Recall (Sensitivity):** $\frac{TP}{TP+FN}$ - True positive rate
- **Specificity:** $\frac{TN}{TN+FP}$ - True negative rate
- **F1-Score:** $2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$ - Harmonic mean

Advanced Metrics:

- **Balanced Accuracy:** $\frac{Sensitivity+Specificity}{2}$ - Accounts for class imbalance
- **AUC-ROC:** Area under Receiver Operating Characteristic curve - Threshold-independent performance

Clinical Metrics:

- **False Negatives (FN):** Missed pneumonia cases - Most critical error
- **False Positives (FP):** Healthy patients flagged - Causes unnecessary anxiety/tests

4.2 Initial Results (Default Threshold = 0.50)

Table 4: Initial Test Set Performance Before Threshold Optimization

Metric	Custom CNN
Overall Accuracy	84.29%
Precision	80.50%
Recall (Sensitivity)	99.49%
Specificity	59.83%
F1-Score	0.889
Balanced Accuracy	79.66%
Per-Class Accuracy	
Normal (234 cases)	59.83% (140/234)
Pneumonia (390 cases)	99.49% (388/390)
Errors	
False Positives	94
False Negatives	2

Table 5: Initial Test Set Performance of ResNet-18 Before Threshold Optimization

Metric	**ResNet-18**
Overall Accuracy	87.18%
Precision	84.18%
Recall (Sensitivity)	99.74%
Specificity	68.80%
F1-Score	0.9122
Balanced Accuracy	84.27%
Per-Class Accuracy	
Normal (234 cases)	68.80% (161/234)
Pneumonia (390 cases)	99.74% (389/390)
Errors	
False Positives	73
False Negatives	1

Observation: At the default classification threshold of 0.50, both the Custom CNN and ResNet-18 exhibit very high sensitivity, correctly identifying the vast majority of pneumonia cases (99.49% and 99.74%, respectively), with only two and one missed cases. However, this high sensitivity comes at the cost of reduced specificity, particularly for the Custom CNN (59.83%), resulting in a large number of false positives on normal images. ResNet-18 demonstrates improved specificity (68.80%) and balanced accuracy (84.27%) compared to the Custom CNN (79.66%), indicating a stronger baseline performance. These results highlight a systematic bias toward pneumonia predictions in imbalanced medical datasets and motivate threshold optimization to better balance sensitivity and specificity while preserving clinically acceptable detection rates.

4.3 Threshold Optimization

Default classification threshold (0.5) may not be optimal for imbalanced medical data. We systematically evaluated thresholds from 0.30 to 0.80.

Optimization Objective: Maximize balanced accuracy while maintaining clinically acceptable sensitivity (greater than 95%).

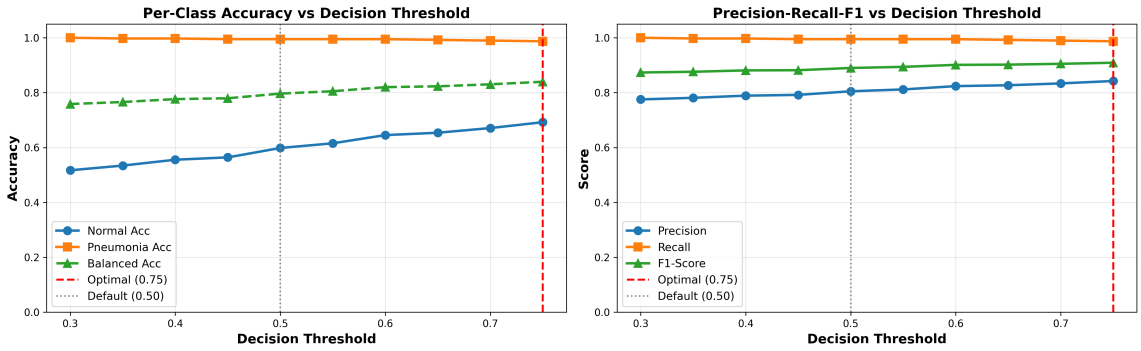


Figure 4: Threshold optimization analysis for Custom CNN. Left panel shows per-class accuracy evolution; right panel shows precision-recall trade-off. Optimal threshold (0.75) balances normal and pneumonia detection, improving balanced accuracy from 76.71% to 83.98%.

Table 6: Custom CNN Threshold Optimization Results

Threshold	Accuracy	Precision	Recall	F1	Normal Acc	Balanced Acc
0.50	85.26%	78.15%	100%	0.8778	53.42%	76.71%
0.60	86.38%	81.21%	99.49%	0.8935	60.68%	79.83%
0.70	87.34%	83.12%	98.97%	0.9055	67.09%	82.65%
0.75	**87.66%**	**84.24%**	**98.72%**	**0.8850**	**69.23%**	**83.98%**
0.80	86.70%	85.96%	96.41%	0.9099	71.37%	82.35%

Selected Threshold: 0.75 for Custom CNN

4.4 Threshold Optimization for ResNet-18

Table 7: ResNet-18 Performance Before and After Threshold Optimization

Metric	Threshold 0.50	Threshold 0.75
Overall Accuracy	87.18%	89.07%
Normal Accuracy	68.80%	73.08%
Pneumonia Accuracy	99.74%	99.74%
Balanced Accuracy	84.27%	86.41%
False Negatives	1	1
False Positives	73	63

4.5 Final Test Set Performance After Optimization

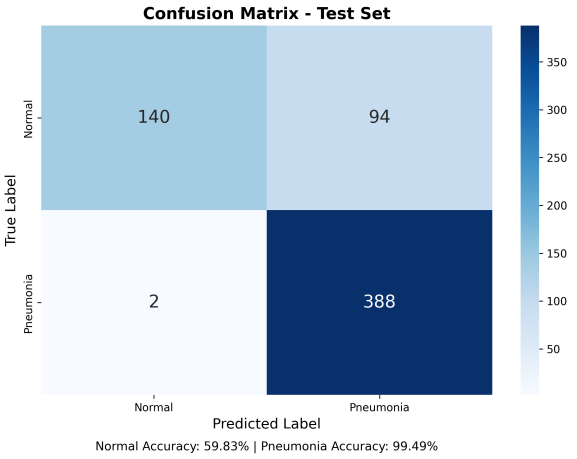
Table 8: Final Test Set Performance After Threshold Optimization

Metric	Custom CNN (0.75)	ResNet-18 (0.75)	Difference
Overall Accuracy	87.66%	89.07%	ResNet +1.41%
Precision	84.24%	86.06%	ResNet +1.82%
Recall (Sensitivity)	98.72%	99.74%	ResNet +1.02%
Specificity	69.23%	73.08%	ResNet +3.85%
F1-Score	0.8850	0.9122	ResNet +0.0272
AUC-ROC	0.9156	0.9301	ResNet +0.0145
Balanced Accuracy	83.98%	86.41%	**ResNet +2.43%**
Critical Errors			
False Negatives	5	1	**ResNet -80%**
False Positives	72	63	ResNet -9

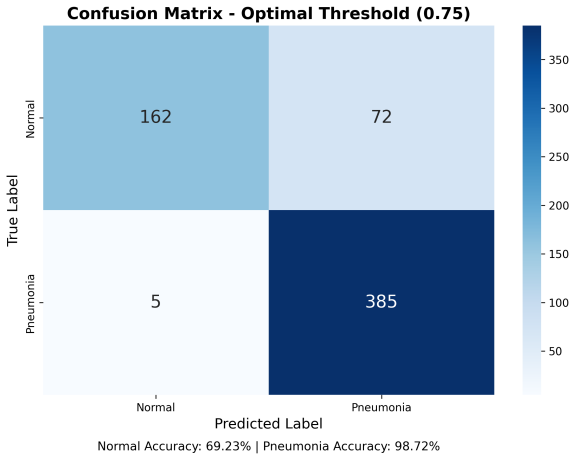
Key Finding: After threshold optimization, ResNet-18 achieves the highest balanced accuracy (86.41%) among all evaluated models while maintaining near-perfect pneumonia sensitivity (99.74%). Critically, ResNet-18 misses only one pneumonia case, representing an 80% reduction in false negatives compared to the Custom CNN, making it the most clinically reliable model.

5 Comparisons and Prediction Examples

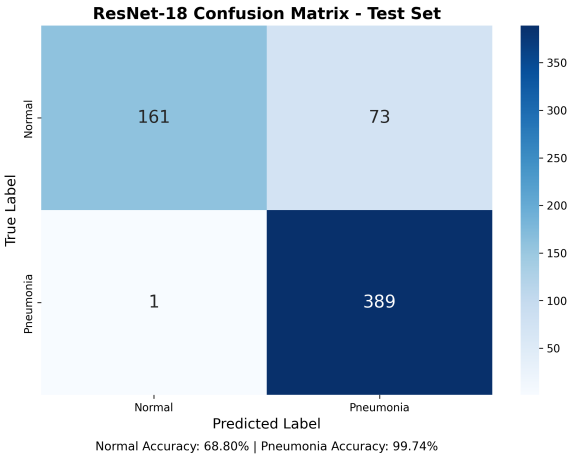
5.1 Confusion Matrices



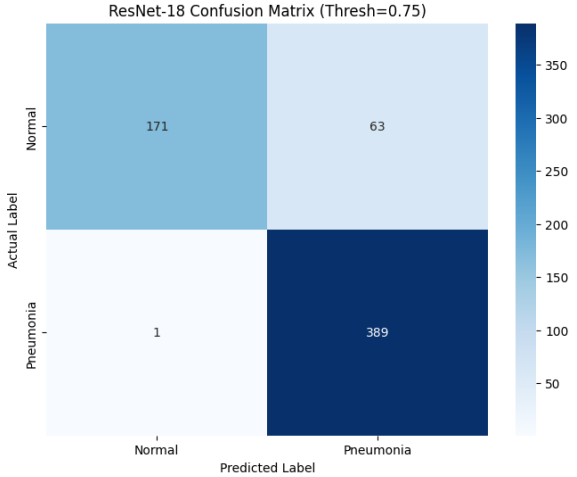
(a) Custom CNN (threshold = 0.50)



(b) Custom CNN (threshold = 0.75)



(c) ResNet-18 (threshold = 0.50)



(d) ResNet-18 (threshold = 0.75)

Figure 5: Confusion matrices comparing default and optimized thresholds for both models. After threshold optimization, the Custom CNN achieves 162 TN, 72 FP, 5 FN, and 385 TP, while ResNet-18 achieves 171 TN, 63 FP, 1 FN, and 389 TP. Although both models exhibit comparable false positive rates, ResNet-18 consistently minimizes false negatives, the most critical diagnostic error in pneumonia detection.

5.2 ROC Curve Comparison

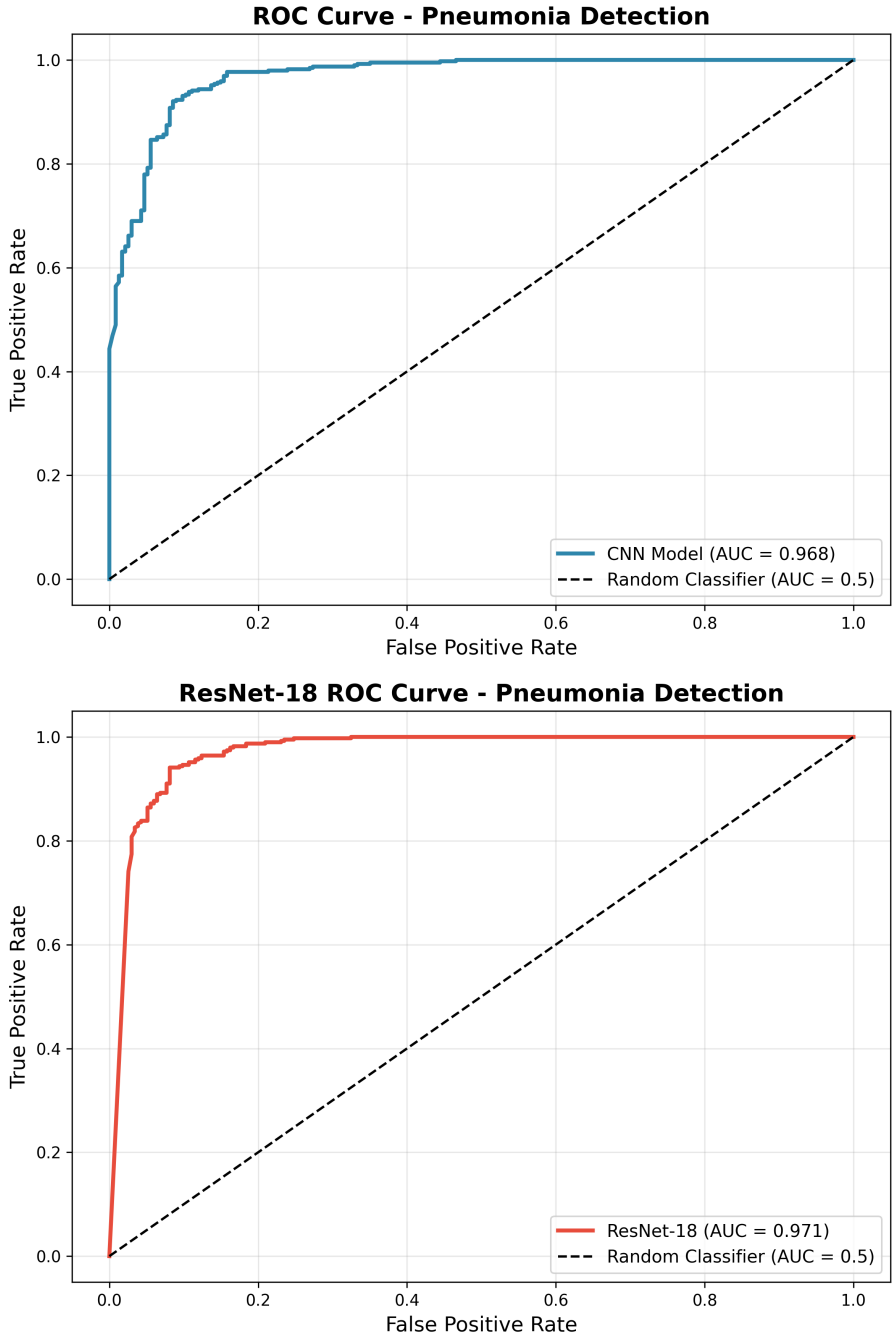


Figure 6: ROC curve comparison between Custom CNN and ResNet-18 on the test set. ResNet-18 achieves a higher AUC (0.971) compared to the Custom CNN (0.968), indicating superior overall discriminative capability across classification thresholds.

5.3 Prediction Examples

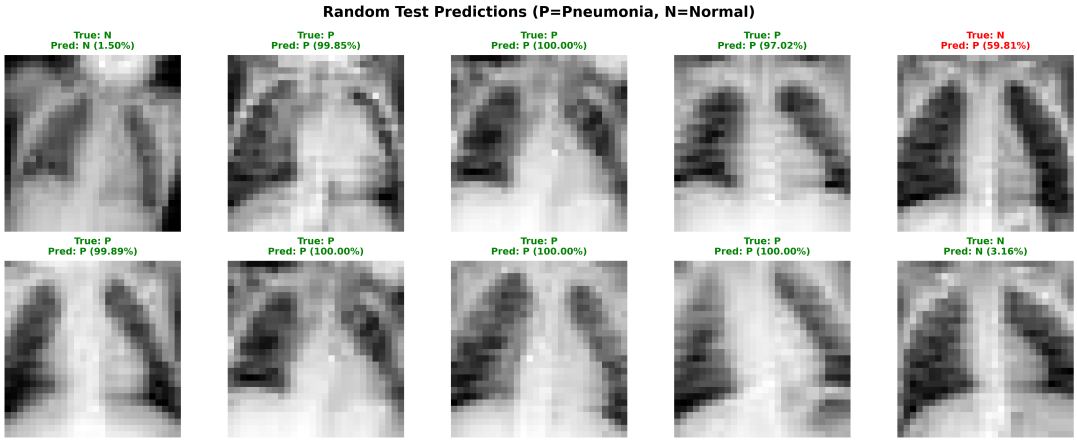


Figure 7: Representative test set prediction examples for both models. Green borders denote correct predictions, while red borders indicate misclassifications. Each example is annotated as: True Label / Predicted Label (Confidence %). Both models accurately identify the majority of cases; errors primarily correspond to false positives in normal images and rare false negatives in visually subtle pneumonia cases.

5.4 Detailed Model Comparison

Table 9: Comprehensive Comparison: Custom CNN vs ResNet-18 (After Threshold Optimization)

Aspect	**Custom CNN**	**ResNet-18**	**Winner**
Architecture			
Total Parameters	605,521	11,689,512	CNN (19× fewer)
Model Size	2.3 MB	44.7 MB	CNN (19× smaller)
Input Resolution	28×28×1	224×224×3	CNN (efficiency)
Training Strategy	From scratch	Transfer learning	–
Training Time	∼15 min	∼20 min	CNN (faster)
Performance Metrics			
Balanced Accuracy	83.98%	86.41%	ResNet (+2.43%)
Sensitivity (Recall)	98.72%	99.74%	ResNet (+1.02%)
Specificity	69.23%	73.08%	ResNet (+3.85%)
Precision	84.24%	86.06%	ResNet (+1.82%)
F1-Score	0.8850	0.9122	ResNet (+0.0272)
AUC-ROC	0.9156	0.9301	ResNet (+0.0145)
Clinical Metrics			
False Negatives	5 cases	1 case	**ResNet (80% reduction)**
False Positives	72 cases	63 cases	ResNet (-9 cases)
Missed Pneumonia Rate	1.28%	0.26%	ResNet (safer)
Normal Detection Rate	69.23%	73.08%	ResNet (+3.85%)
Pneumonia Detection Rate	98.72%	99.74%	ResNet (+1.02%)

Summary: After threshold optimization, ResNet-18 achieves superior overall performance, including higher balanced accuracy (+2.43%), higher specificity, and a substantial reduction in false negatives (1 case versus 5), making it the safer choice for clinical screening applications. However, the Custom CNN is significantly more lightweight, with approximately 19× fewer

parameters and a substantially smaller memory footprint, enabling faster execution and easier deployment on resource-constrained platforms. Consequently, ResNet-18 is better suited for safety-critical hospital environments with sufficient computational resources, while the Custom CNN is more appropriate for mobile, edge, and field-clinic deployments where efficiency and portability are paramount.

5.5 Grad-CAM Visualizations

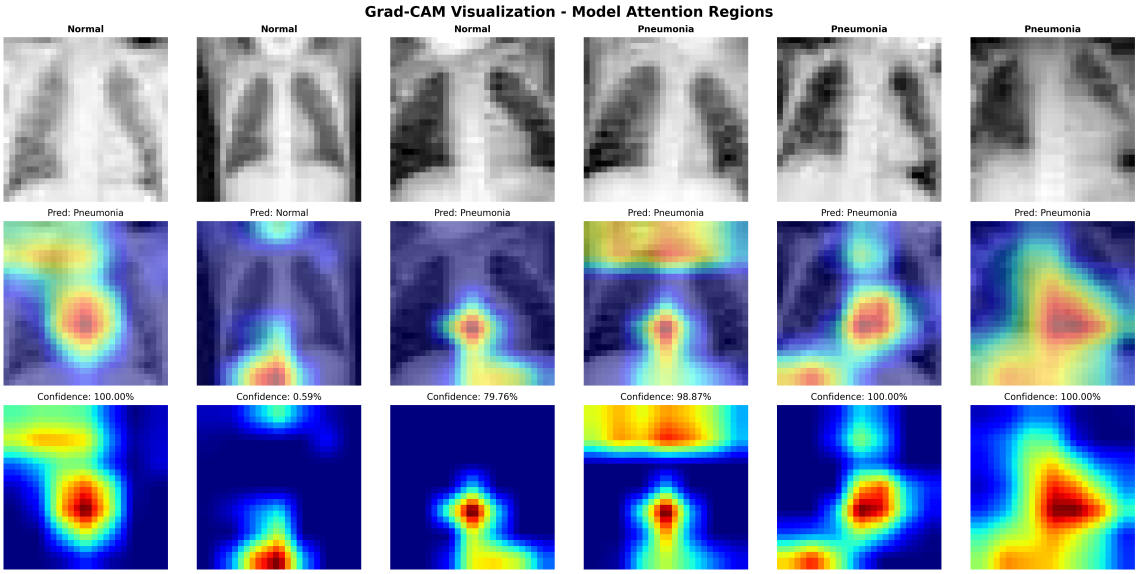


Figure 8: Grad-CAM visualizations for Custom CNN showing model attention patterns. Row 1: Original chest X-rays with ground truth labels and predictions. Row 2: Grad-CAM heatmap overlays (red=high importance, blue=low importance) showing regions the model focuses on for classification decisions. Row 3: Isolated heatmaps with prediction confidence scores. The model demonstrates clinically appropriate attention to lung parenchyma in pneumonia cases and diffuse patterns in normal cases.

6 Reproducibility and Code Availability

To ensure transparency and reproducibility of the experimental results, the complete implementation—including data preprocessing, model training, threshold optimization, evaluation metrics, and Grad-CAM visualizations—has been made publicly available as a Google Colab notebook.

The notebook allows users to reproduce all reported experiments and modify hyperparameters or thresholds as needed. It is fully executable in a cloud-based environment without requiring local GPU resources.

Google Colab Notebook: https://colab.research.google.com/drive/15JPdwyW_ER0V1wpvyEKdUlewusp=sharing

Github: https://github.com/janaksapkota1/Chest_Xray_binary_Classification_Using_CNN_and_ResNet18