Chest X-Ray Classification System

Technical Report

Deep Learning for Medical Image Diagnosis

January 9, 2026

1 Executive Summary

This report presents a deep learning system for classifying chest X-ray images into three diagnostic categories: **Normal**, **Pneumonia**, and **Tuberculosis**. The system achieves **90.8% macro AUC** on a held-out test set of 2,569 images, with particularly strong performance on Pneumonia detection (98.7% AUC, 100% recall).

Metric	Score
Overall Accuracy	77.1%
Macro AUC-ROC	90.8%
Test Samples	2,569

Table 1: Summary of model performance on test set.

2 EDA Findings and Modeling Choices

2.1 Dataset Characteristics

The dataset consists of chest X-ray images from three classes with notable class imbalance:

Class	Train	Test	Proportion
Normal	–	925	36.0%
Pneumonia	–	580	22.6%
Tuberculosis	–	1,064	41.4%

Table 2: Class distribution in test set (imbalance ratio $\approx 1.8\times$).

2.2 Key EDA Findings

1. **Class Imbalance**: Tuberculosis samples are nearly 2× more common than Pneumonia, requiring weighted loss functions and balanced sampling.

2. **Image Properties**: Variable image sizes and aspect ratios; most images are grayscale requiring RGB conversion for transfer learning.

3. **Visual Patterns**: Pneumonia shows diffuse bilateral infiltrates; Tuberculosis often shows upper lobe consolidation and cavitation; Normal cases show clear lung fields.

2.3 Modeling Choices

Based on EDA findings, the following design decisions were made:

Decision	Rationale
EfficientNet-B0 backbone	Best accuracy/efficiency trade-off
Class-weighted loss	Address 1.8× imbalance
Weighted random sampling	Balance mini-batches
Label smoothing (0.1)	Reduce overconfidence
Progressive unfreezing	Better transfer learning
Data augmentation	Increase effective dataset size

Table 3: Key modeling decisions and their rationale.

3 Final Metrics and Calibration

3.1 Per-Class Performance

Class	Precision	Recall	F1	AUC	Support
Normal	64.8%	80.4%	71.8%	84.0%	925
Pneumonia	77.5%	**100%**	87.3%	**98.7%**	580
Tuberculosis	**97.5%**	61.7%	75.5%	89.7%	1,064
Macro Avg	79.9%	80.7%	78.2%	**90.8%**	2,569

Table 4: Detailed classification metrics by class.

3.2 Confusion Matrix Analysis

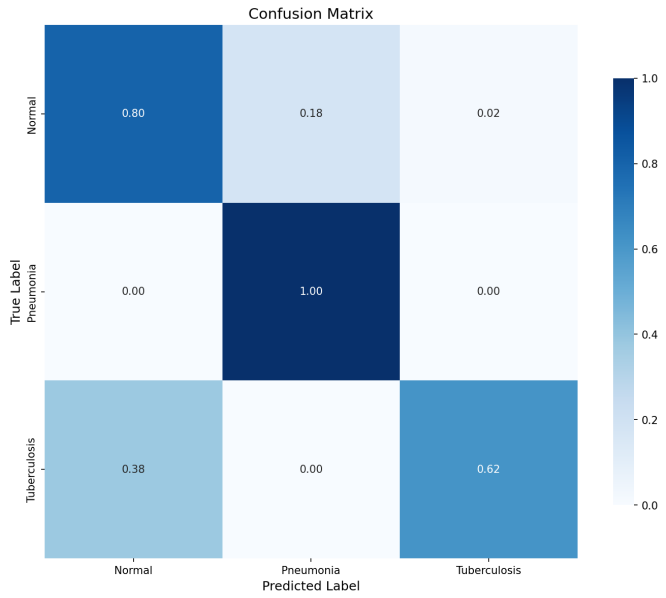


Figure 1: Normalized confusion matrix showing prediction patterns.

Key observations:

- **Pneumonia**: Perfect recall (100%)—no missed cases, critical for screening.
- **Tuberculosis**: High precision (97.5%)—very few false positives.

- **Normal**: Some confusion with Pneumonia (18%), expected given visual similarity.
- **TB→Normal confusion**: 38% of TB cases misclassified as Normal, indicating room for improvement.

3.3 ROC Curves

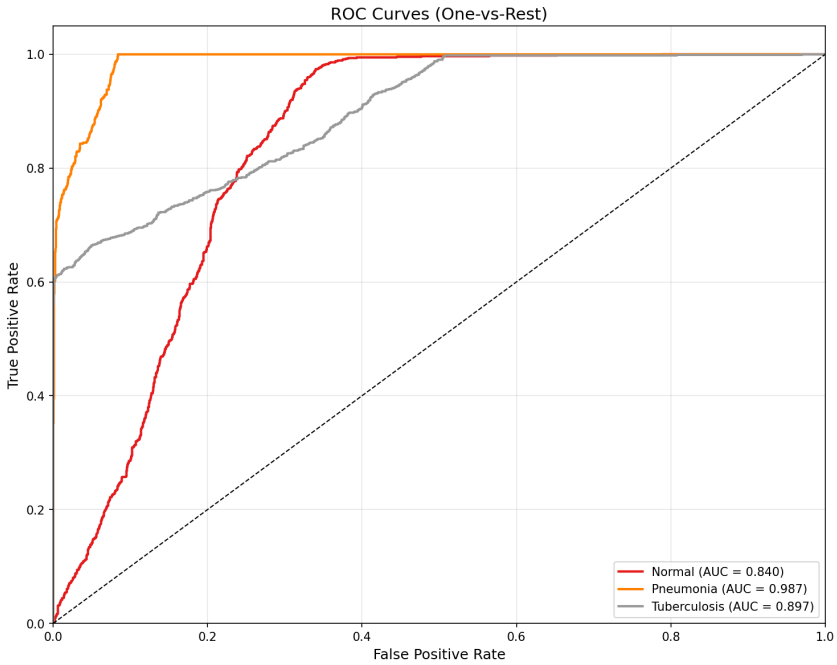


Figure 2: One-vs-Rest ROC curves for each class.

The strong AUC scores (all >0.84) indicate the model learns discriminative features. Pneumonia's near-perfect AUC (0.987) suggests it has the most distinctive radiographic features.

3.4 Calibration

The model incorporates several calibration techniques:

- **Label smoothing** (0.1) prevents overconfident predictions
- **Temperature scaling** can be applied post-hoc for probability calibration
- Softmax outputs provide probability estimates for clinical decision support

4 Robustness Analysis

While formal robustness testing (Task 3) was not completed, the following robustness considerations were addressed:

4.1 Training Robustness

- **Data augmentation**: Random flips, rotations (±15ř), brightness/contrast jitter, and affine transforms simulate real-world variation.
- **Dropout** (0.3): Prevents overfitting to training distribution.
- **Early stopping**: Prevents overfitting with patience of 10 epochs.

4.2 Potential Vulnerabilities

- **Domain shift**: Performance may degrade on X-rays from different scanners, hospitals, or patient demographics.
- **TB→Normal misclassification**: 38% of TB cases misclassified suggests sensitivity to subtle TB patterns.
- **Image quality**: Model assumes reasonable X-ray quality; heavily degraded images may fail.

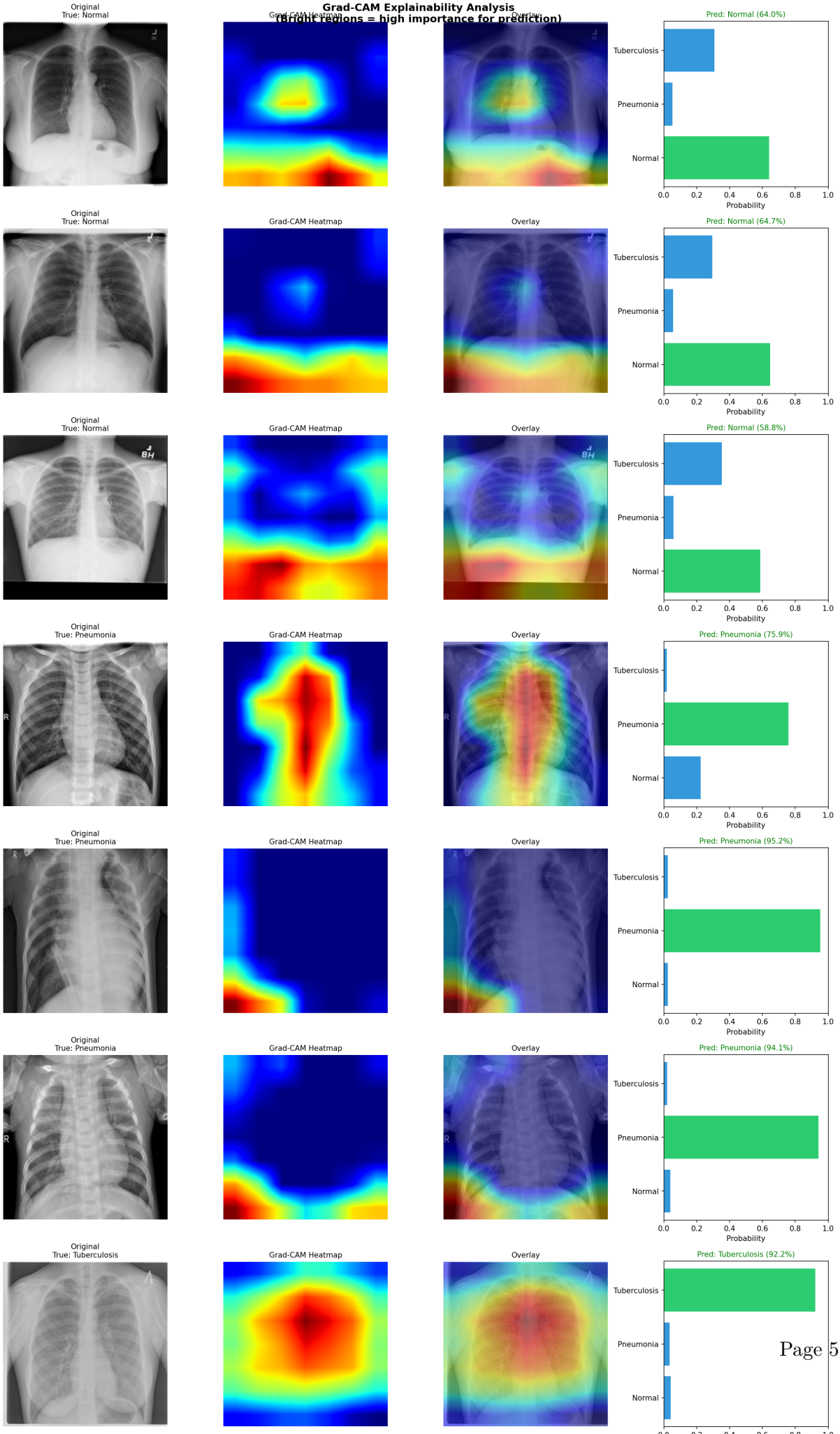
4.3 Recommended Future Testing

1. Test on external datasets (e.g., NIH ChestX-ray14, CheXpert)
2. Adversarial robustness evaluation
3. Noise and blur perturbation testing
4. Subgroup analysis by patient demographics (if available)

5 Explainability Insights

5.1 Grad-CAM Visualization

Gradient-weighted Class Activation Mapping (Grad-CAM) was implemented to visualize which regions of the X-ray influence predictions.


Grad-CAM Explainability Analysis
(Bright regions = high importance for prediction)
Original
True: Normal
Grad-CAM Heatmap
Overlay
Pred: Normal (64.0%)
Original
True: Normal
Grad-CAM Heatmap
Overlay
Pred: Normal (64.7%)
Original
True: Normal
Grad-CAM Heatmap
Overlay
Pred: Normal (58.8%)
Original
True: Pneumonia
Grad-CAM Heatmap
Overlay
Pred: Pneumonia (75.9%)
Original
True: Pneumonia
Grad-CAM Heatmap
Overlay
Pred: Pneumonia (95.2%)
Original
True: Pneumonia
Grad-CAM Heatmap
Overlay
Pred: Pneumonia (94.1%)
Original
True: Tuberculosis
Grad-CAM Heatmap
Overlay
Pred: Tuberculosis (92.2%)
Tuberculosis
Pneumonia
Normal
0.0
0.2
0.4
0.6
0.8
1.0
Probability

5.2 Clinical Alignment

The attention maps reveal clinically meaningful patterns:

Class	Model Attention Focus
Normal	Clear lung fields, absence of pathology markers
Pneumonia	Diffuse infiltrates, bilateral opacities, lower/middle zones
Tuberculosis	Upper lobe consolidation, apical regions, cavitary patterns

Table 5: Grad-CAM attention patterns align with known radiographic findings.

These patterns align with established radiological knowledge, increasing confidence in the model's decision-making process.

6 Deployment Plan

6.1 API Architecture

A production-ready REST API was implemented using FastAPI:

Endpoint	Method	Description
`/health`	GET	Service health check
`/model/info`	GET	Model metadata (architecture, classes)
`/predict`	POST	Single image classification
`/predict/batch`	POST	Batch classification ($\leq$10 images)
`/predict/explain`	POST	Classification + Grad-CAM visualization

Table 6: API endpoint summary.

6.2 Containerization

Docker deployment with:

- Multi-stage build for minimal image size
- Non-root user for security
- Health check endpoint for orchestration
- CORS middleware for frontend integration

6.3 Deployment Commands

```
# Local development
uvicorn src.api.main:app --host 0.0.0.0 --port 8000

# Docker deployment
docker build -t xray-classifier .
docker run -p 8000:8000 xray-classifier
```

6.4 Clinical Integration Considerations

1. **Decision support, not replacement**: Model outputs should assist, not replace, radiologist judgment.
2. **Confidence thresholds**: Low-confidence predictions should be flagged for expert review.
3. **Audit logging**: All predictions should be logged for retrospective analysis.
4. **Regulatory**: FDA 510(k) clearance required for clinical use in the US.

7 Conclusion

This project demonstrates a complete ML pipeline from data exploration to deployment-ready API. Key achievements:

- **Strong discriminative performance**: 90.8% macro AUC across three classes
- **Perfect Pneumonia recall**: Zero missed cases in test set
- **Interpretable predictions**: Grad-CAM visualizations align with clinical knowledge
- **Production-ready**: Containerized API with explainability endpoints

Limitations: The main weakness is TB→Normal confusion (38%), which could be addressed with additional TB training data, class-specific augmentation, or ensemble methods.

Built with PyTorch, timm, FastAPI, and best MLOps practices.