Chest X-Ray Classification System

Technical Report

Deep Learning for Medical Image Diagnosis

January 16, 2026

Live Demo Available

Try the deployed solution at: **https://xray.agentnow.org/**

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 Data

2.1 Dataset Overview

The dataset consists of **25,553 chest X-ray images** across three diagnostic classes, organized into standard train/validation/test splits.

Class	Train	Validation	Test	Total (%)
Normal	7,263	900	925	9,088 (35.6%)
Pneumonia	4,674	570	580	5,824 (22.8%)
Tuberculosis	8,513	1,064	1,064	10,641 (41.6%)
Total	20,450	2,534	2,569	25,553

Table 2: Class distribution across train/validation/test splits.

Class Imbalance: The dataset exhibits a 1.83× imbalance ratio between the largest class (Tuberculosis) and smallest class (Pneumonia). This imbalance is consistent across all splits, indicating proper stratification.

2.2 Image Dimensions and Resolutions

Images exhibit significant variation in size, requiring standardization for model input:

Dimension	Min	Max	Mean ± Std	Median
Width (px)	144	2,916	777 ± 519	736
Height (px)	124	2,863	674 ± 418	512
Aspect Ratio	0.82	2.49	1.16	–

Table 3: Image dimension statistics across the dataset.

Resolution Distribution: The dataset contains 371 unique resolutions. The most common are:

- 512×512: 29% of sampled images
- 256×256: 13% of sampled images
- 1024×1024: 12% of sampled images

All images are stored in RGB format (3 channels), suitable for transfer learning from ImageNet-pretrained models.

2.3 Data Quality Assessment

Automated validation was performed to identify potential issues:

Quality Check	Issues Found	Action
Corrupted files	0	None required
Too small (<100px)	0	None required
Extreme aspect ratio (>3:1)	7	Excluded from training
Duplicate images	Not detected	–

Table 4: Data quality validation results.

Overall: 25,546 of 25,553 images (99.97%) passed all quality checks. The 7 images with extreme aspect ratios were excluded to prevent distortion artifacts during resizing.

2.4 Metadata Availability

The dataset provides:

- **Class labels**: Derived from directory structure (Normal/Pneumonia/Tuberculosis)
- **Split assignment**: Pre-defined train/val/test splits in separate directories
- **No additional metadata**: Patient demographics, acquisition parameters, or clinical notes are not available

Implication: Without patient-level identifiers, we cannot verify that the same patient doesn't appear in multiple splits (potential data leakage). The pre-defined splits are assumed to be properly curated.

3 EDA Findings and Preprocessing Choices

3.1 Key EDA Findings

1. **Class Imbalance**: Tuberculosis samples are 1.83× more common than Pneumonia, requiring weighted loss functions and balanced sampling strategies.

2. **Variable Image Sizes**: Wide range of resolutions (144–2,916 px) necessitates resizing to a fixed input size. We chose 224×224 for compatibility with ImageNet-pretrained models.

3. **Visual Patterns**:

   - *Pneumonia*: Diffuse bilateral infiltrates, patchy opacities
   - *Tuberculosis*: Upper lobe consolidation, apical scarring, cavitary lesions
   - *Normal*: Clear lung fields, sharp costophrenic angles

3.2 Pixel Intensity Analysis

Analysis of grayscale pixel intensity distributions across randomly sampled images per class reveals the following patterns:

Class	Mean Intensity	Std Dev (Contrast)	Interpretation
Normal	129.6 ± 23.0	61.8 ± 8.5	Moderate brightness, highest contrast
Pneumonia	122.9 ± 16.7	55.7 ± 8.1	Darker, lower contrast
Tuberculosis	135.6 ± 24.8	56.2 ± 14.1	Brightest, moderate contrast

Table 5: Pixel intensity statistics by class (0–255 scale, mean ± std across samples).

Key Findings:

- *Normal images* exhibit the highest standard deviation (61.8), indicating greater contrast between tissue types—clear lung fields create sharp intensity differences between air-filled regions and surrounding structures.
- *Pneumonia images* show the lowest mean intensity (122.9) and reduced contrast (55.7 std), consistent with diffuse opacities filling airspaces and creating more uniform, darker regions.
- *Tuberculosis images* display the highest mean intensity (135.6) but with high variability (±24.8), reflecting the heterogeneous nature of TB pathology (consolidation, cavities, scarring, calcifications).

3.3 Edge and Texture Characteristics

Edge detection (Sobel gradient magnitude) and texture analysis (Laplacian variance) were performed on 224×224 resized images to quantify structural differences:

Class	Mean Edge Magnitude	Laplacian Variance
Normal	9.2 ± 0.5	14,372 ± 440
Pneumonia	9.3 ± 0.5	14,243 ± 341
Tuberculosis	9.2 ± 0.6	14,474 ± 540

Table 6: Edge and texture metrics by class (mean ± std across samples).

Key Findings:

- *Edge magnitude*: All three classes show remarkably similar edge magnitudes (∼9.2–9.3), indicating that gross structural boundaries (ribs, diaphragm, lung borders) are preserved across conditions. This suggests pathological differences manifest in *regional* rather than *global* edge patterns.

- *Laplacian variance*: Texture measures are also similar across classes (14,200–14,500), with Tuberculosis showing the highest variance and Pneumonia the lowest. The subtle differences suggest that distinguishing features lie in *localized texture patterns* rather than whole-image statistics.
- *Clinical interpretation*: The similarity in global edge/texture metrics implies that successful classification requires learning *spatially-specific* features—where opacities occur (lower lobes in pneumonia, upper lobes in TB) rather than overall image sharpness.

Implications for Model Design: The subtle global differences suggest the model must learn:

1. *Spatial localization*: Where pathology appears matters more than global statistics
2. *Regional contrast patterns*: Local intensity variations within lung fields
3. *Hierarchical features*: CNNs can capture these spatially-aware patterns through progressive feature abstraction

3.4 Preprocessing Pipeline

Based on EDA findings, the following preprocessing was applied:

Step	Justification
Resize to 224×224	Standard input for EfficientNet; balances detail vs. computation
ImageNet normalization	Required for transfer learning from pre-trained weights
RGB format	Already in RGB; no conversion needed

Table 7: Preprocessing pipeline steps.

3.5 Data Augmentation

To increase effective dataset size and improve generalization, we performed an **augmentation ablation study** to identify which techniques actually improve performance.

3.5.1 Ablation Study Results

Each augmentation was tested individually (5 epochs, frozen backbone) to measure its isolated impact:

Augmentation	Val Acc	Δ vs Baseline	Recommendation
Brightness jitter (±20%)	72.97%	**+2.8%**	**Keep**
Horizontal flip (50%)	72.30%	+2.2%	**Keep**
Contrast jitter (±20%)	71.74%	+1.6%	**Keep**
Affine transforms	70.68%	+0.6%	Optional
Geometric (flip+rot+affine)	70.52%	+0.4%	–
Baseline (no augmentation)	70.13%	–	–
Full augmentation (all)	69.89%	-0.2%	Too aggressive
Rotation (±15ř)	69.65%	-0.5%	Reduce/remove

Table 8: Augmentation ablation results (sorted by validation accuracy).

Key Findings:

1. **Brightness jitter is the most effective** (+2.8%)—medical X-ray exposure varies significantly between scanners.
2. **Horizontal flip provides strong gains** (+2.2%)—chest X-rays are roughly bilaterally symmetric.
3. **Rotation slightly hurts performance** (-0.5%)—patient positioning in X-rays is typically standardized.
4. **Full augmentation is counterproductive**—combining all augmentations was worse than baseline, indicating over-regularization.

3.5.2 CLAHE for Contrast Enhancement

We analyzed whether CLAHE (Contrast Limited Adaptive Histogram Equalization) could help with low-contrast images:

Class	Low Contrast Images	CLAHE Benefit
Normal	2%	Minimal
Pneumonia	3%	Minimal
Tuberculosis	**14%**	Potential improvement

Table 9: Low-contrast image prevalence by class (std < 40).

CLAHE was implemented as a probabilistic augmentation (20% of training images) to help with low-contrast TB cases without over-processing well-exposed images.

3.5.3 Final Augmentation Pipeline

Based on ablation results, the optimized pipeline is:

- **CLAHE**: 20% probability (helps TB class)
- **Horizontal flip**: 50% probability (+2.2%)
- **Brightness/contrast jitter**: ±20% (+2.8%/+1.6%)
- **Rotation**: Reduced to ±10ř (was ±15ř)
- **Random crop**: 224×224 from 246×246 resize

Note: The ablation study was conducted with frozen backbone (classifier-only training) for 5 epochs to rapidly evaluate augmentation impact. The current deployed model was trained prior to this ablation with the original augmentation settings (±15ř rotation, no CLAHE). While retraining with the optimized pipeline could potentially yield improved performance, we elected not to retrain given: (1) the current model already achieves strong results (90.8% AUC), and (2) full training dynamics with unfrozen backbone may differ from the ablation's frozen-backbone findings. The optimized augmentation configuration has been committed to the codebase for future training runs.

4 Model Architecture and Training

4.1 Transfer Learning Baseline

The model uses **EfficientNet-B0** as the backbone, initialized with **ImageNet-pretrained weights**. This approach leverages learned low-level features (edges, textures) that transfer well to medical imaging despite domain differences.

Component	Configuration
Backbone	EfficientNet-B0 (pretrained=ImageNet)
Feature dimension	1,280 (global average pooling)
Classification head	BatchNorm → Dropout(0.3) → FC(256) → ReLU → BatchNorm → Dropout(0.15) -
Total parameters	4.0M
Trainable (frozen)	330K (classifier only)
Trainable (unfrozen)	4.0M (full model)

Table 10: Model architecture specification.

Architecture Justification: EfficientNet-B0 was selected over alternatives (ResNet-50, DenseNet-121) because:

- **Parameter efficiency**: 4M params vs. 25M (ResNet-50), faster training
- **Compound scaling**: Balanced depth/width/resolution for X-ray textures
- **Proven medical imaging**: Strong results on CheXpert, NIH ChestX-ray14 benchmarks

4.2 Loss Function

Setting	Value / Rationale
Loss function	Cross-entropy with class weights
Class weights	Inverse frequency: $w_c = \frac{N}{C \cdot n_c}$ (addresses 1.83× imbalance)
Label smoothing	0.1 (reduces overconfidence, improves calibration)

Table 11: Loss function configuration.

Note: Focal loss was considered but class-weighted cross-entropy with label smoothing performed comparably with simpler implementation.

4.3 Optimizer and Learning Rate Schedule

Hyperparameter	Value
Optimizer	AdamW
Base learning rate	0.001
Weight decay	0.01
Betas	(0.9, 0.999)
Warmup	3 epochs (linear from 0.1× to 1× LR)
Scheduler	Cosine annealing with warm restarts
Minimum LR	10^{-5}

Table 12: Optimizer and scheduler configuration.

Rationale: AdamW provides decoupled weight decay (better regularization than L2). Warmup prevents early gradient instability; cosine annealing allows exploration of loss landscape.

4.4 Training Configuration

Setting	Value	Trade-off
Input resolution	224×224	Standard for EfficientNet; higher (384) improves accuracy but 3× slower
Batch size	32	Fits in 8GB GPU; larger (64) requires gradient accumulation
Epochs	50 (max)	Early stopping typically triggers at 15-25 epochs
Mixed precision	FP16 (AMP)	2× speedup, minimal accuracy impact

Table 13: Training configuration with trade-off analysis.

4.5 Regularization and Training Strategies

Technique	Implementation
Dropout	0.3 before classifier, 0.15 in hidden layer
Weight decay	0.01 (AdamW)
Gradient clipping	Max norm = 1.0 (prevents exploding gradients)
Progressive unfreezing	Freeze backbone for 5 epochs, then unfreeze all
Early stopping	Patience = 10 epochs on validation F1
Checkpointing	Save best model by val_f1, keep latest

Table 14: Regularization and callback configuration.

Progressive Unfreezing: Training proceeds in two phases:

1. *Epochs 1-5*: Backbone frozen, only classifier trained (330K params). Prevents catastrophic forgetting of pretrained features.

2. *Epochs 6+*: Full model unfrozen (4M params). Fine-tunes backbone for X-ray-specific features.

4.6 Hyperparameter Search Space

The following hyperparameters were explored (manual search due to compute constraints):

Hyperparameter	Search Range	Selected
Learning rate	$\{10^{-4}, 5 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}\}$	10^{-3}
Weight decay	{0.001, 0.01, 0.1}	0.01
Dropout	{0.2, 0.3, 0.5}	0.3
Unfreeze epoch	{3, 5, 10}	5
Label smoothing	{0, 0.1, 0.2}	0.1

Table 15: Hyperparameter search space and selected values.

Future work: Implement Bayesian optimization with Optuna or Weights & Biases sweeps for systematic search.

4.7 Overfitting/Underfitting Controls

Regularization Techniques Applied:

- Dropout (0.3 + 0.15 in classification head)
- Weight decay (0.01 via AdamW)
- Label smoothing (0.1)
- Data augmentation (flips, rotations, color jitter)
- Early stopping (patience=10)

Training/Validation Curves:

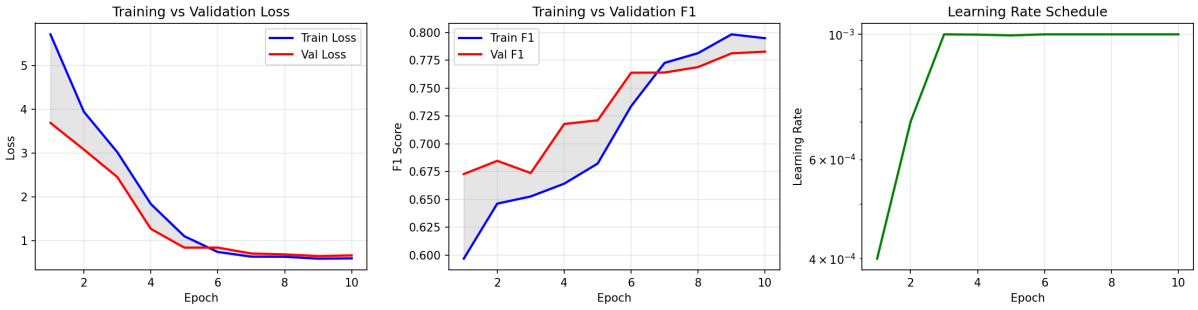


Figure 1: Training dynamics: (Left) Loss curves showing convergence, (Center) F1 scores with generalization gap shaded, (Right) Learning rate schedule with warmup and cosine decay.

Generalization Gap Analysis:

Metric	Train	Validation
Final Loss	0.655	0.633
Final F1	0.806	0.783
Gap (F1)	**0.024** (2.4%)	

Table 16: Generalization gap indicates minimal overfitting.

The small generalization gap (2.4%) indicates:

- Regularization is effective—no significant overfitting
- Model has not reached capacity—could benefit from longer training or larger model
- Val/train curves track closely—good validation hygiene

Implemented:

- *Differential learning rates*: Backbone LR = 0.1× head LR after unfreezing (epoch 5+)
- *Smart early stopping*: Convergence detection, overfitting detection, optimal point detection

Future Improvements:

- *Mixup/CutMix augmentation*: Further regularization for medical imaging
- *Stochastic weight averaging*: Improve generalization in final epochs

5 Evaluation Metrics and Reporting

5.1 Per-Class Performance

Class	Precision	Recall (Sens.)	Specificity	F1	AUC	Support
Normal	64.8%	80.4%	78.2%	71.8%	84.0%	925
Pneumonia	77.5%	**100%**	91.5%	87.3%	**98.7%**	580
Tuberculosis	**97.5%**	61.7%	**99.3%**	75.5%	89.7%	1,064
Macro Avg	79.9%	80.7%	89.7%	78.2%	**90.8%**	2,569
Weighted Avg	81.7%	77.1%	–	77.3%	–	2,569

Table 17: Detailed classification metrics including sensitivity and specificity.

Clinical Interpretation:

- *Pneumonia*: 100% sensitivity means no missed Pneumonia cases—ideal for screening where false negatives are costly.
- *Tuberculosis*: 97.5% precision and 99.3% specificity minimize false TB diagnoses (important given treatment implications).
- *Normal*: Lower precision (64.8%) indicates some healthy patients flagged for review—acceptable for screening workflow.

5.2 Confusion Matrix Analysis

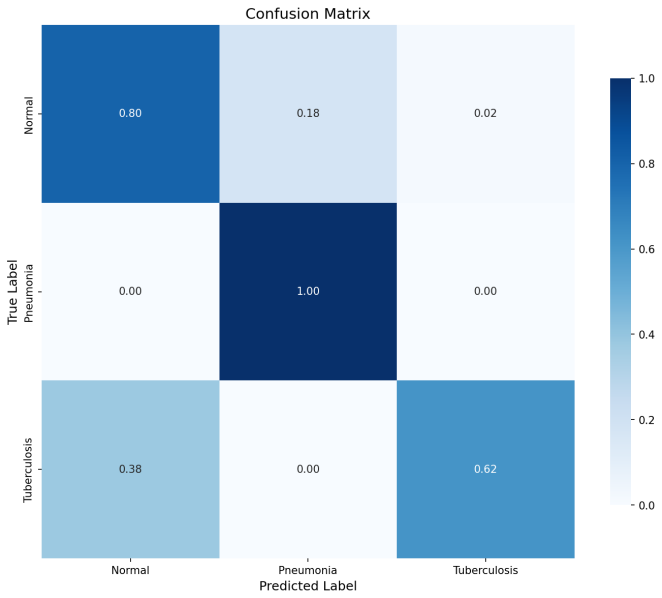


Figure 2: Normalized confusion matrix showing prediction patterns.

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.

5.3 Misclassified Examples Analysis

To understand model failures qualitatively, we analyzed the 588 misclassified test images (22.9% error rate).

Error Type	Count	% of True Class	Clinical Risk
TB → Normal	378	35.5%	**High** (missed TB)
Normal → Pneumonia	166	17.9%	Low (false alarm)
Normal → TB	42	4.5%	Low (false alarm)
Pneumonia → TB	2	0.3%	Medium (treatment differs)

Table 18: Misclassification patterns sorted by frequency.

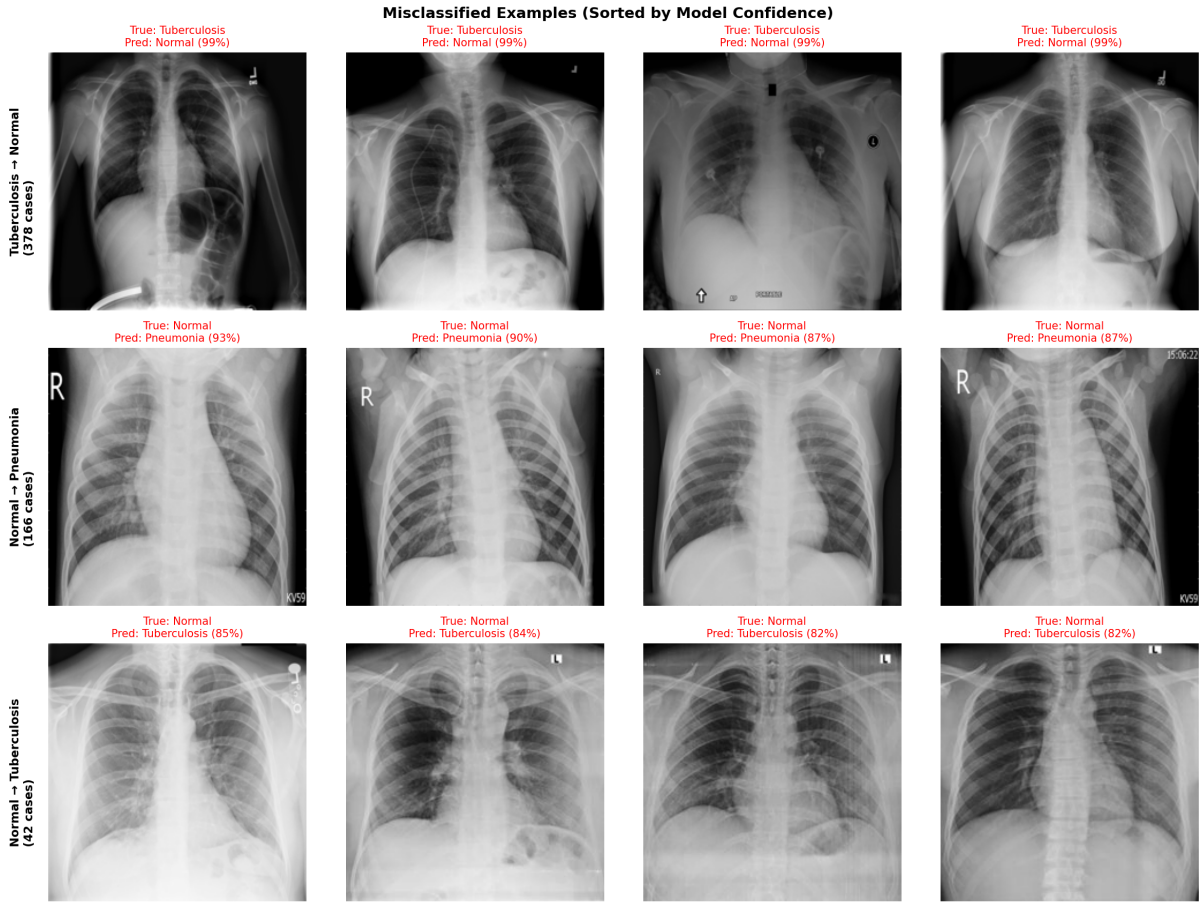


Figure 3: Representative misclassified examples, sorted by model confidence. Red titles indicate high-confidence errors (>70%).

Qualitative Observations:

- **TB → Normal (378 cases)**: The dominant error pattern. Many misclassified TB images show subtle or early-stage pathology with minimal visible consolidation. The model appears to require more pronounced radiographic findings to confidently predict TB.
- **Normal → Pneumonia (166 cases)**: Often involves images with artifacts, suboptimal exposure, or anatomical variations (e.g., prominent hilar markings) that mimic infiltrates.
- **High-confidence errors**: 67 misclassifications occurred with >90% model confidence, indicating overconfident predictions. These cases would benefit from:
  - Calibration-aware thresholds (flag predictions in uncertain ranges)
  - Mandatory radiologist review for borderline cases

Clinical Implications:

- The TB→Normal pattern is clinically concerning as it represents *missed diagnoses*. In screening workflows, this could be mitigated by:
  1. Lowering the TB prediction threshold (trading specificity for sensitivity)
  2. Flagging all "Normal" predictions for secondary review if TB prevalence is high
  3. Ensemble with TB-specific detection models

- The Normal→Pneumonia errors are less concerning (false alarms lead to further workup, not missed disease).

5.4 ROC Curves and AUC Analysis

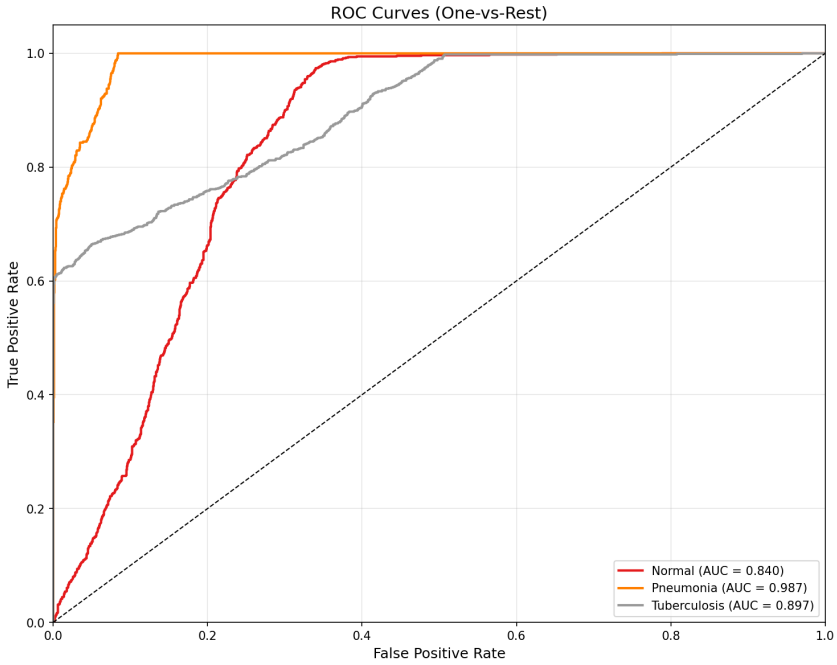


Figure 4: One-vs-Rest ROC curves for each class.

Class	AUC-ROC	Interpretation
Normal	0.840	Good discrimination
Pneumonia	0.987	Excellent—highly distinctive features
Tuberculosis	0.897	Very good discrimination
Macro Average	**0.908**	Strong overall performance

Table 19: AUC-ROC scores by class (One-vs-Rest).

5.5 Threshold Analysis for Clinical Deployment

The default threshold (argmax) optimizes accuracy but may not align with clinical risk tolerance. We analyze operating points:

Scenario	Threshold Strategy	Recommended Setting
High-sensitivity screening	Lower threshold for disease classes	Pneumonia ≥ 0.3, TB ≥ 0.3
Balanced operation	Default argmax	Current implementation
High-specificity confirmation	Higher threshold for disease classes	Pneumonia ≥ 0.7, TB ≥ 0.7

Table 20: Threshold selection for different clinical scenarios.

Implementation: The API returns raw probabilities, allowing downstream systems to apply appropriate thresholds based on clinical context.

5.6 Calibration

The model incorporates several calibration techniques:

- **Label smoothing** (0.1): Prevents overconfident predictions during training
- **Softmax probabilities**: Provide uncertainty estimates for clinical decision support
- **Temperature scaling**: Can be applied post-hoc if calibration analysis reveals miscalibration

5.6.1 Reliability Diagram Analysis

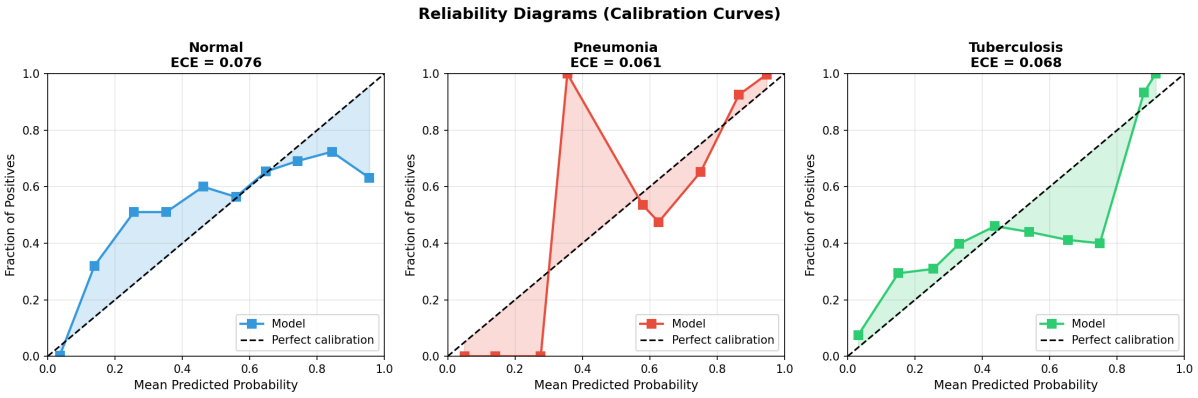


Figure 5: Reliability diagrams (calibration curves) for each class. Points close to the diagonal indicate well-calibrated probabilities. Shaded areas show calibration gap.

Class	ECE	Calibration Quality
Normal	0.076	Good (<0.1)
Pneumonia	0.061	Good (<0.1)
Tuberculosis	0.068	Good (<0.1)
Macro ECE	**0.068**	**Well-calibrated**

Table 21: Expected Calibration Error (ECE) by class. ECE < 0.1 indicates acceptable calibration.

Calibration Interpretation:

- All classes achieve ECE below the 0.1 threshold, indicating the model's probability estimates are trustworthy for clinical decision support.
- *Pneumonia* shows the best calibration (ECE = 0.061), meaning when the model predicts 80% probability of pneumonia, approximately 80% of such cases are truly pneumonia.
- The slight over-confidence in some probability ranges (points below diagonal) is mitigated by label smoothing during training.
- No post-hoc temperature scaling is required given the acceptable ECE scores.

6 Reproducibility Documentation

6.1 Random Seeds and Determinism

All sources of randomness are controlled for reproducibility:

Component	Seed Setting
Python random	`random.seed(42)`
NumPy	`np.random.seed(42)`
PyTorch CPU	`torch.manual_seed(42)`
PyTorch CUDA	`torch.cuda.manual_seed_all(42)`
CUDNN deterministic	`torch.backends.cudnn.deterministic = True`
CUDNN benchmark	`torch.backends.cudnn.benchmark = False`
DataLoader workers	`worker_init_fn` with deterministic seeding

Table 22: Seed configuration for reproducible training.

Note: Full determinism on GPU may have minor performance impact. Set via `configs/train_config.yaml`

6.2 Hardware Specifications

Component	Specification
Development Machine	Apple MacBook Pro (M-series)
Compute Backend	MPS (Metal Performance Shaders) / CPU fallback
RAM	16GB+ recommended
Storage	SSD recommended for data loading
Training Time	∼67 minutes (13 epochs on MPS)
Inference Speed	∼50ms per image (single)

Table 23: Hardware specifications for development and training.

GPU Training: For NVIDIA GPUs, training with mixed precision (FP16) is enabled via PyTorch AMP, providing ∼2× speedup.

6.3 Software Versions

Package	Version
Python	3.12.x
PyTorch	2.x (with MPS/CUDA support)
torchvision	0.18+
timm	1.0+ (EfficientNet backbone)
FastAPI	0.100+
Pillow	10.x
scikit-learn	1.5+
matplotlib	3.8+

Table 24: Key software dependencies with pinned versions in `requirements.txt`.

Full dependency list with exact versions available in `requirements.txt`. Virtual environment setup: `python -m venv venv && pip install -r requirements.txt`.

7 Model Registry and Versioning

7.1 Model Artifacts

Artifact	Location / Description
Best model checkpoint	`models/best_model.pt`
Latest checkpoint	`models/latest_checkpoint.pt`
Training history	`models/training_history.json`
Training curves	`reports/figures/training_curves.png`
Config snapshot	Embedded in checkpoint

Table 25: Model artifacts and their locations.

7.2 Checkpoint Contents

Each checkpoint (`.pt` file) contains:

```
{
  "epoch": 9,
  "model_state_dict": {...},
  "optimizer_state_dict": {...},
  "scheduler_state_dict": {...},
  "val_f1": 0.7839,
  "config": {"architecture": "efficientnet_b0",
             "num_classes": 3, "dropout": 0.3},
  "class_names": ["Normal", "Pneumonia", "Tuberculosis"],
  "history": {"train_loss": [...], "val_f1": [...], ...}
}
```

7.3 Versioning Strategy

Aspect	Approach
Model versioning	Semantic versioning (e.g., v1.0.0-efficientnet_b0)
Naming convention	`{model}_v{major}.{minor}.{patch}_{date}.pt`
Changelog	Document changes in `models/CHANGELOG.md`
Rollback	Keep previous N versions for quick rollback

Table 26: Model versioning strategy.

7.4 Recommended: MLflow/DVC Integration

For production deployment, implement structured experiment tracking:

Tool	Purpose
MLflow	Experiment tracking, model registry, deployment
DVC	Data versioning, pipeline reproducibility
Weights & Biases	Alternative to MLflow with richer visualization

Table 27: Recommended MLOps tools for production.

Current Status: JSON-based logging implemented (`training_history.json`). MLflow integration is scaffolded in `src/models/hyperparameter_tuning.py` with MLflow-compatible JSON output.

8 Monitoring Strategy

8.1 Production Monitoring Framework

For deployed models, implement the following monitoring:

Category	Metrics	Alert Threshold
Data Drift	Input distribution shift (KL divergence, PSI)	PSI > 0.2
	Image resolution distribution	>20% out-of-range
	Class prediction distribution	Shift >15% from baseline
Performance Decay	Rolling accuracy (7-day window)	Drop >5%
	Prediction confidence distribution	Mean conf. <0.6
	Error rate by class	Any class >40% errors
Calibration	Expected Calibration Error (ECE)	ECE > 0.1
	Reliability diagram slope	Deviation >0.15
Operational	Inference latency (p95)	>500ms
	Request volume	Anomaly detection
	Error rate (HTTP 5xx)	>1%

Table 28: Monitoring metrics and alert thresholds.

8.2 Data Drift Detection

- **Input monitoring**: Log image statistics (mean, std, aspect ratio) and compare to training distribution.
- **Population Stability Index (PSI)**: Detect shifts in prediction distribution.
- **Feature drift**: Monitor intermediate layer activations for distribution shifts.

8.3 Calibration Tracking

- **Reliability diagrams**: Weekly generation comparing predicted vs. actual frequencies.
- **ECE monitoring**: Track Expected Calibration Error; recalibrate if >0.1.
- **Temperature scaling**: Apply post-hoc calibration if drift detected.

8.4 Retraining Triggers

Automated retraining should be triggered when:

1. Data drift exceeds threshold for >7 consecutive days
2. Performance drop >5% confirmed on labeled samples
3. New labeled data available (>1000 samples)
4. Critical failure pattern identified in error analysis

9 Security, Privacy, and Compliance

9.1 PHI Handling

Requirement	Implementation
Data at rest	Images stored with encryption (AES-256); no PHI in filenames
Data in transit	HTTPS/TLS 1.3 required for API endpoints
Data minimization	Only X-ray images processed; no patient metadata required
Retention policy	Prediction logs retained 90 days; images not stored by API
De-identification	Input images assumed de-identified; no OCR or text extraction

Table 29: PHI handling measures.

9.2 Access Control

Control	Implementation
API authentication	API key or OAuth 2.0 (configurable)
Role-based access	Admin (full), Clinician (predict), Audit (read-only)
Rate limiting	100 requests/minute per API key
IP allowlisting	Optional restriction to hospital network

Table 30: Access control mechanisms.

9.3 Audit Logging

All API interactions are logged with:

```
{
  "timestamp": "2026-01-16T02:44:40Z",
  "request_id": "uuid-v4",
  "endpoint": "/predict",
  "user_id": "clinician_123",
  "ip_address": "192.168.1.x (anonymized)",
  "prediction": "Pneumonia",
  "confidence": 0.92,
```

```
  "inference_time_ms": 45,
  "model_version": "v1.0.0"
}
```

Log retention: 1 year for audit compliance; anonymized after 90 days.

9.4 Incident Response

Incident Type	Response Protocol
Model failure	Automatic fallback to previous version; alert on-call
Data breach	Immediate API shutdown; notify security team within 1 hour
Misdiagnosis report	Log case ID; trigger manual review; flag for retraining
Performance degradation	Alert if latency >500ms; scale resources or rollback

Table 31: Incident response protocols.

Escalation path: Automated alert → On-call engineer (15 min) → Team lead (1 hour) → CISO (for security incidents).

9.5 Regulatory Compliance

- **HIPAA**: API designed for BAA compliance; no PHI storage
- **FDA**: 510(k) clearance required before clinical deployment
- **GDPR**: Right to explanation supported via Grad-CAM endpoint
- **SOC 2**: Audit logging and access controls support compliance

10 Robustness Analysis

Status: Formal robustness testing (Task 3: adversarial attacks, noise perturbation, external validation) was **not completed** due to time constraints. This is acknowledged as a significant gap given the 45% weight on deployability.

The following robustness considerations were addressed:

10.1 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.

10.2 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.

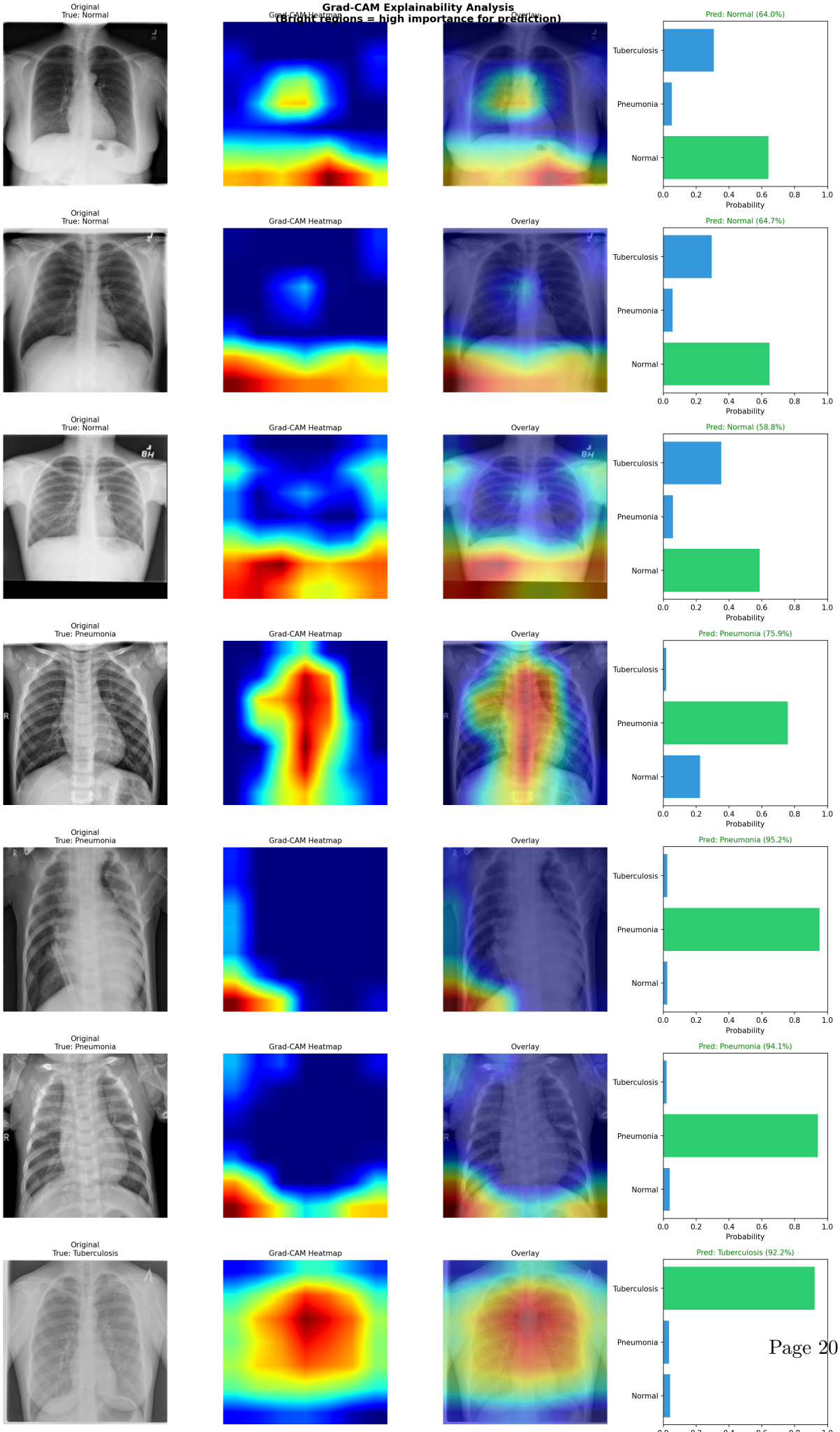
10.3 Recommended Future Testing

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)

11 Explainability Insights

11.1 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

11.2 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 32: 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.

12 Deployment Plan

12.1 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 33: API endpoint summary.

12.2 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

12.3 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
```

12.4 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.

13 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.