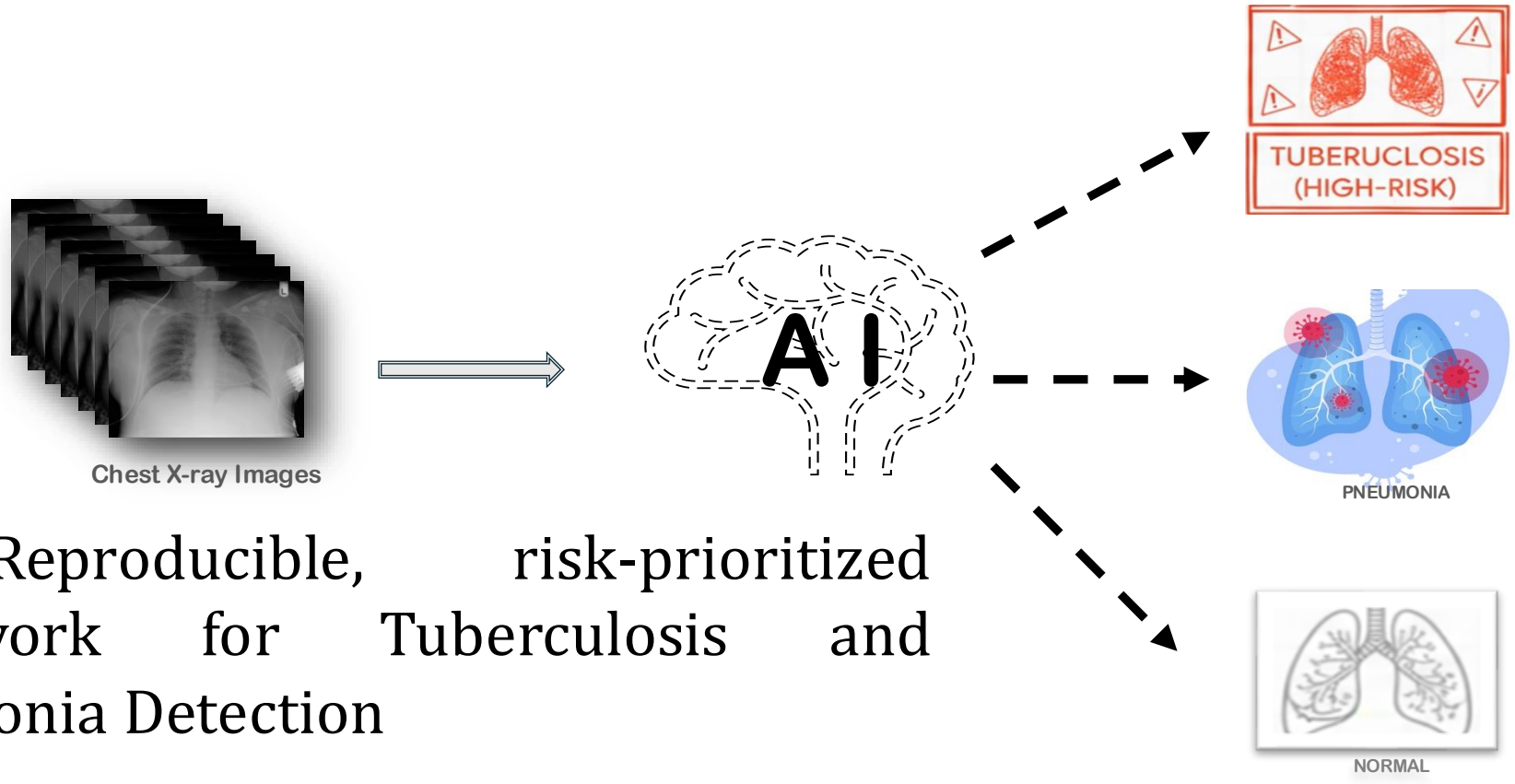
The Clinical Challenge

Chest X-ray is a frontline screening tool for Tuberculosis and Pneumonia detection

High-volume screening creates radiologist workload pressure

Build a risk-aware, explainable, and deployable AI screening system

AI-Augmented Chest X-Ray Screening

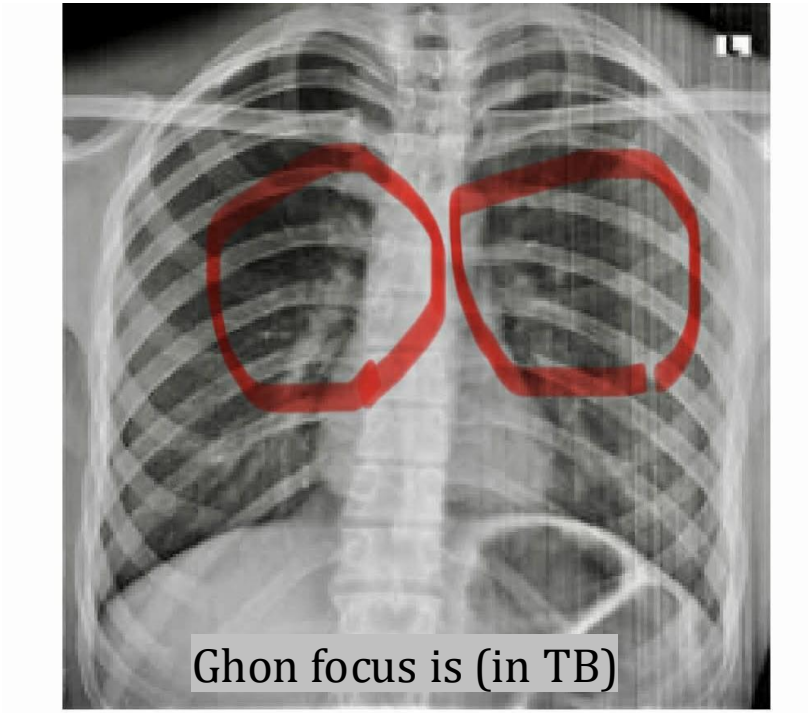


A Reproducible, risk-prioritized framework for Tuberculosis and Pneumonia Detection

Kaggle Dataset: *25,553* Images | **Stack**: *Pytorch*, *MLFlow*, *Docker*, *Fast API*

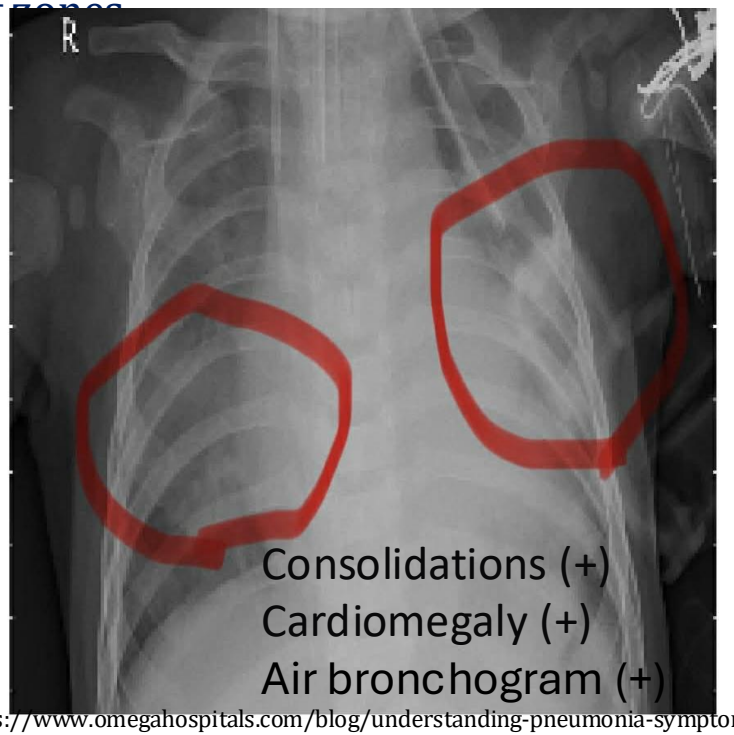
Tuberculosis and Pneumonia

- TB often more spatially distributed and can present with very subtle cues in its early stages and shows broad, bilateral lung activation (appearing on both sides) rather than being confined to a single lobe.
- Frequently involves the upper lung regions.



https://www.advocatehealth.com/health-services/lung-respiratory-care/tuberculosis-tb

- Pneumonia is typically characterized by localized infections that fill the air sacs with fluid or pus.
- Radiographic evidence of pneumonia is frequently concentrated in the mid-to-lower lung zones.



https://www.omegahospitals.com/blog/understanding-pneumonia-symptoms-causes-and-effective-treatments-for-better-health/

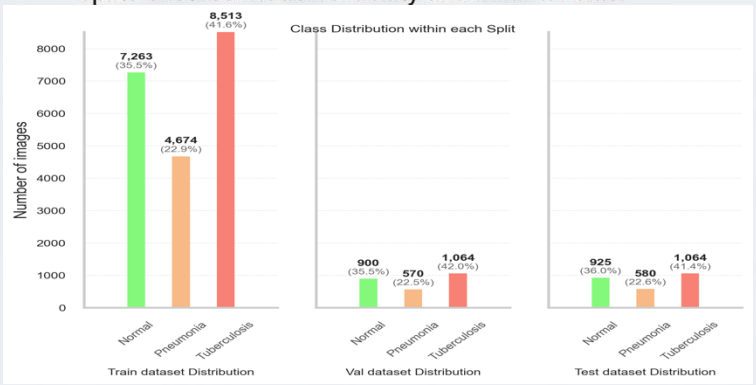
FROM RAW DATA TO DIAGNOSTIC FEATURES

Optimized for Grayscale Texture

Images were resized to 224x224 and enhanced with CLAHE to preserve subtle lung parenchymal patterns.

25,553 Validated Images

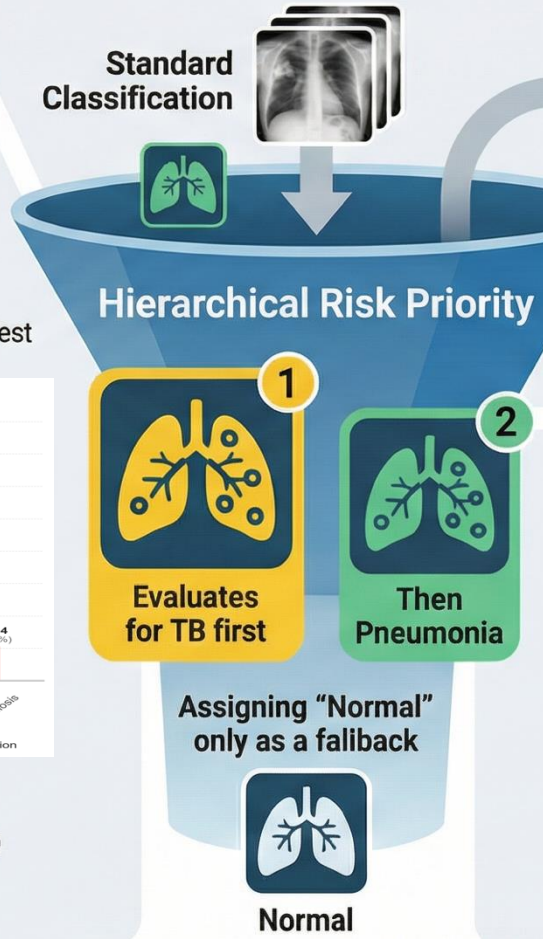
A consistent distribution across Train, Validation, and Test splits ensures model reliability and minimal bias.



Resource-Efficient Architecture

Custom 3-block CNNs and Linear Probing were used to maintain performance on consumer grade hardware.

CLINICAL IMPACT: THE "POLICY-AWARE" SOLUTION



Missed TB cases reduced by 73%

TB Sensitivity Boosted from 63% to 90%

Policy-oware thresholding reduced missed Tuberculosis cases by approximately 73% compared to standard classification.

Explainable & Deployment Ready

Grad-CAM Visualizations

Transparent, Localized Diagnostic Evidence

Integrates Grad-CAM visualizations and FastAPI to provide clinicians with transparent, localized diagnostic evidence.

** Generated with AI*

Engineering for Clinical Safety

OBJECTIVE

Robust chest X-ray classifiers under realistic compute constraints

CONTRIBUTION

Moved beyond standard accuracy maximation to implement a Risk-prioritized Decision Policy that trade specificity for safety

DEPLOYMENT

Fully containerized, reproducible MLOps pipeline (MLFlow -> Model Bundle -> Docker)

Key Results

~ 73%

Reduction in Tuberculosis false negatives

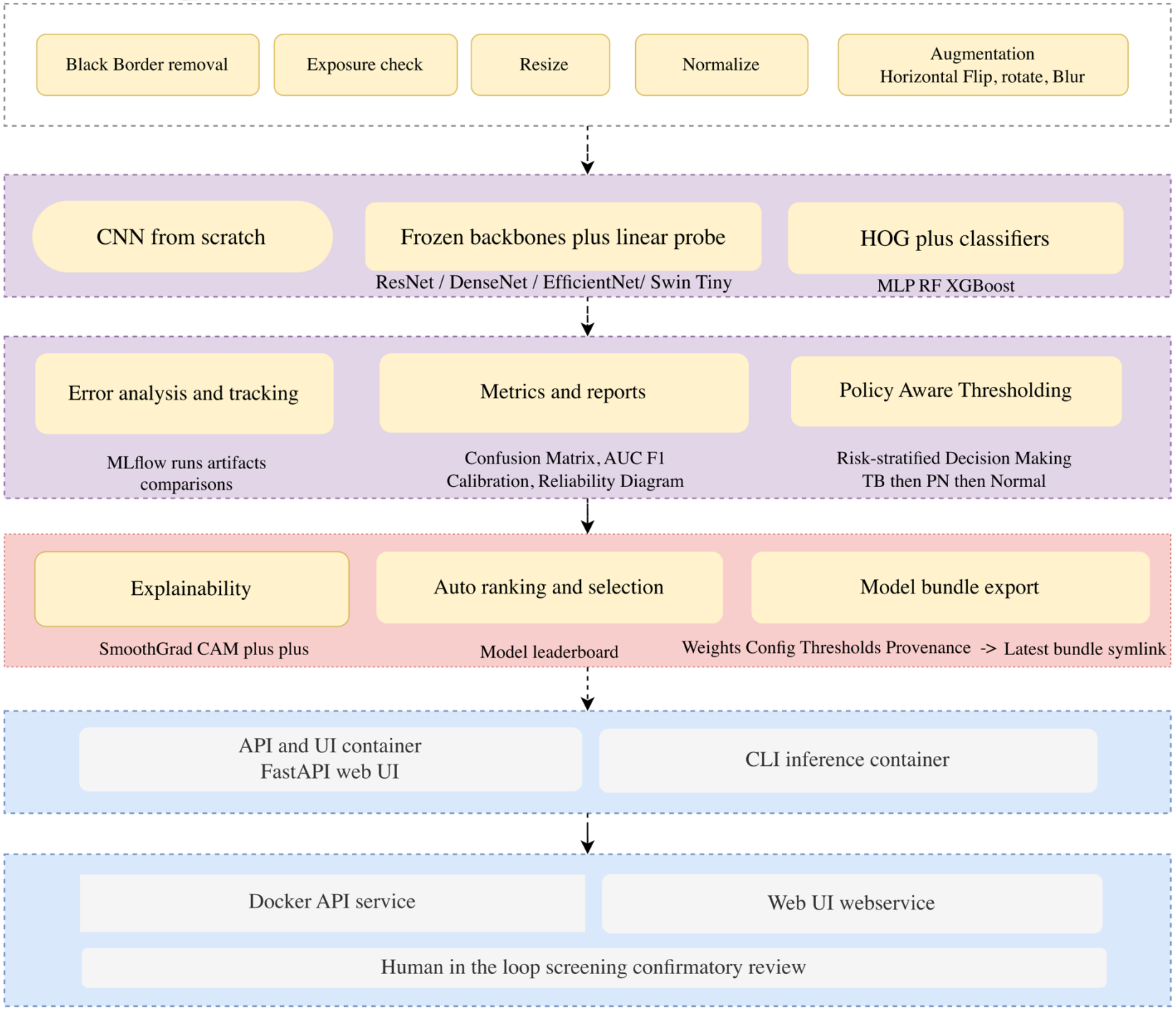
>90%

TB Sensitivity
(Up from 63.4%)

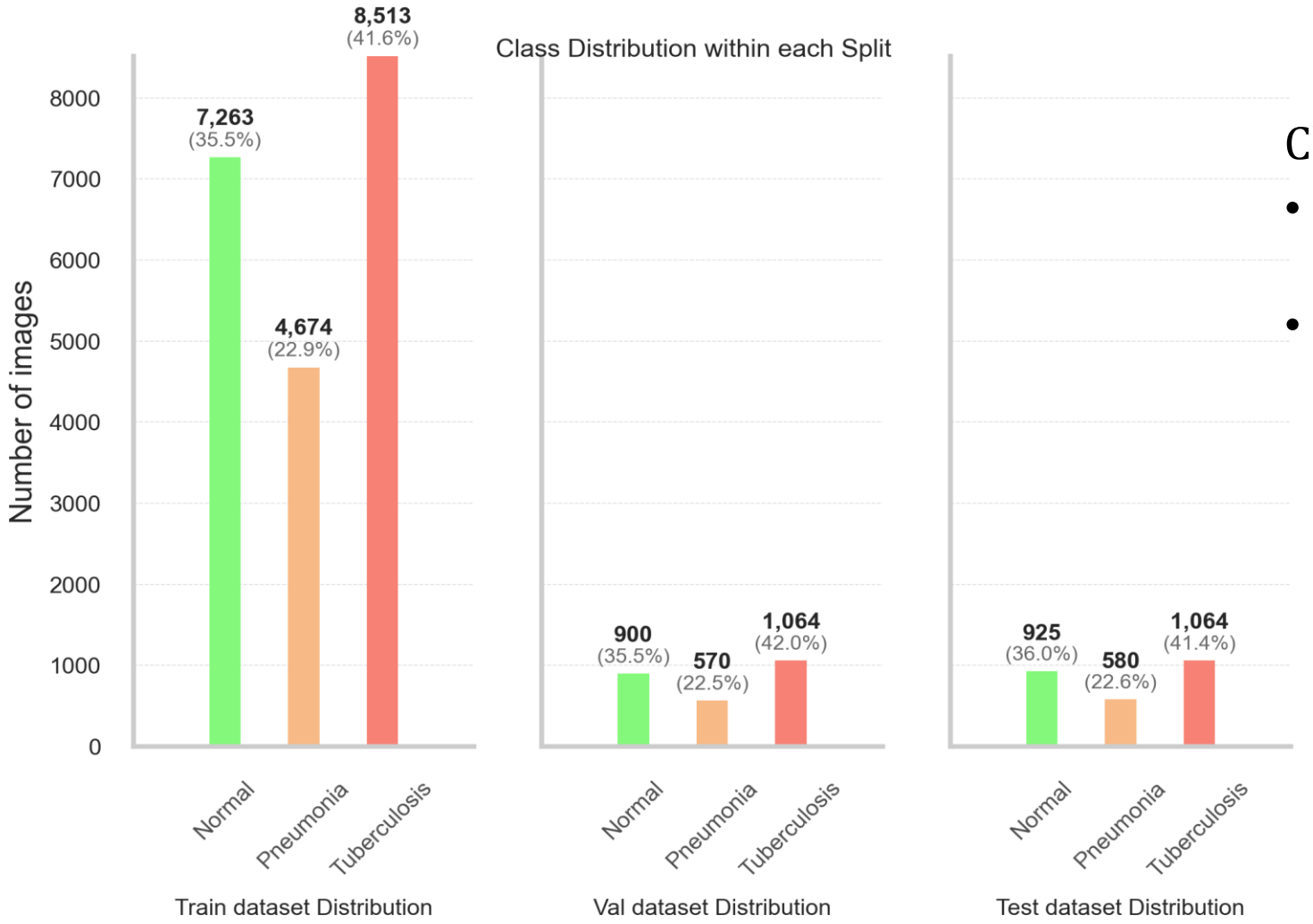
>85%

Maintained Pneumonia Sensitivity

Pipeline


Preprocessing
Black Border removal
Exposure check
Resize
Normalize
Augmentation
Horizontal Flip, rotate, Blur
Model Training
CNN from scratch
Frozen backbones plus linear probe
ResNet / DenseNet / EfficientNet/ Swin Tiny
HOG plus classifiers
MLP RF XGBoost
Model Evaluation
Error analysis and tracking
MLflow runs artifacts comparisons
Metrics and reports
Confusion Matrix, AUC F1 Calibration, Reliability Diagram
Policy Aware Thresholding
Risk-stratified Decision Making TB then PN then Normal
PostProcessing
Explainability
SmoothGrad CAM plus plus
Auto ranking and selection
Model leaderboard
Model bundle export
Weights Config Thresholds Provenance -> Latest bundle symlink
Dockerized System
API and UI container FastAPI web UI
CLI inference container
Deployment
Docker API service
Web UI webservice
Human in the loop screening confirmatory review

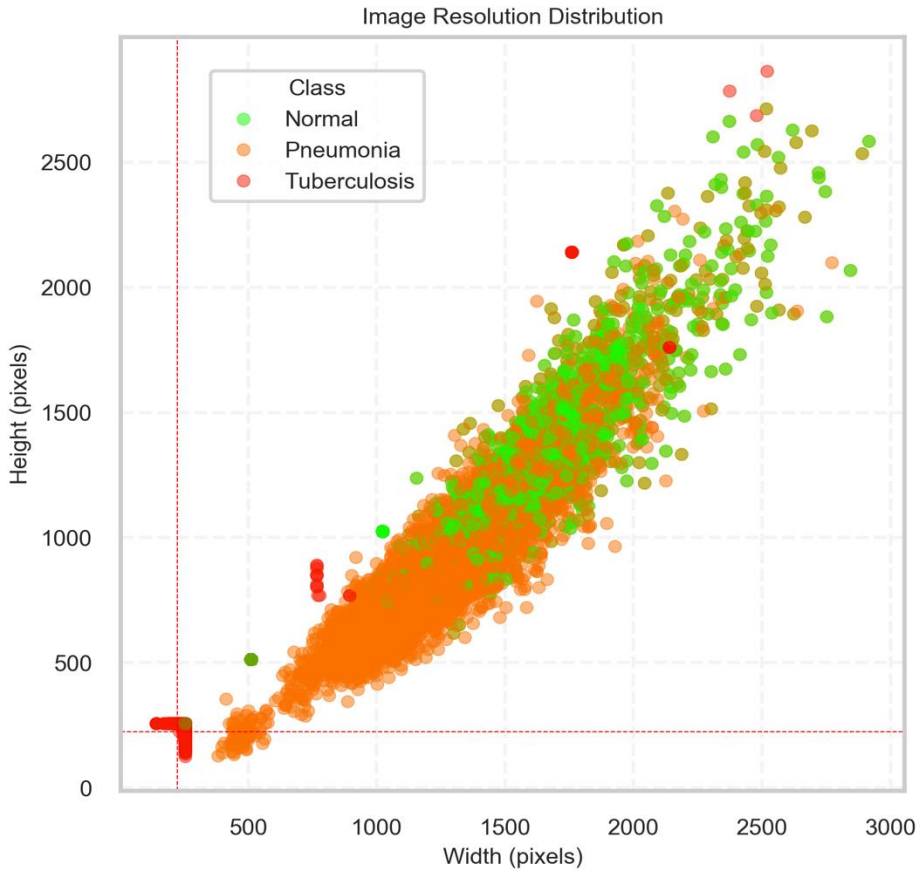
EDA: Class Distribution



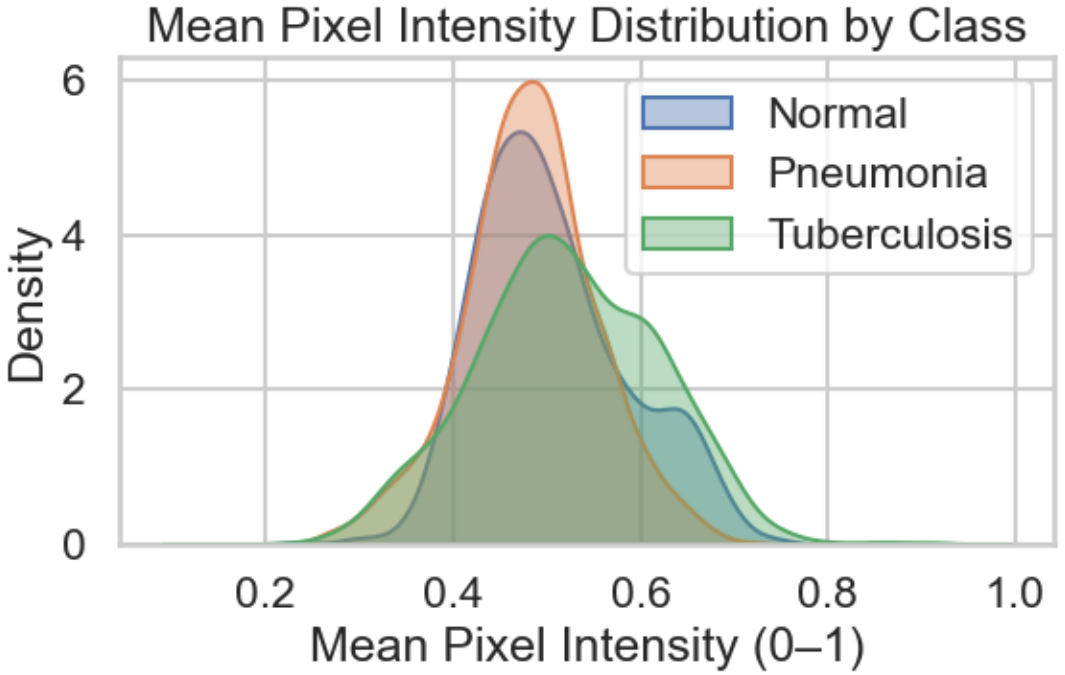
Class imbalance

- Pneumonia is the most underrepresented class
- Class proportions stable across splits

EDA: Resolution & Photometric Analysis



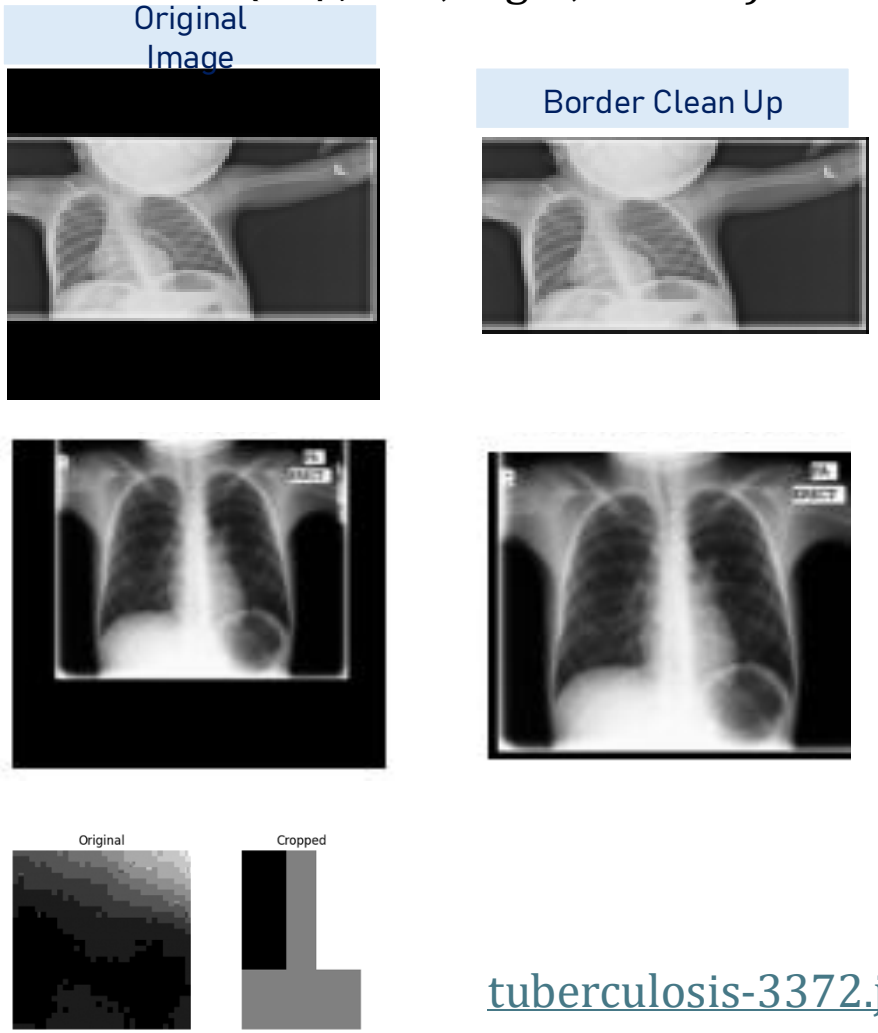
Low-resolution images concentrated in TB (~19%)



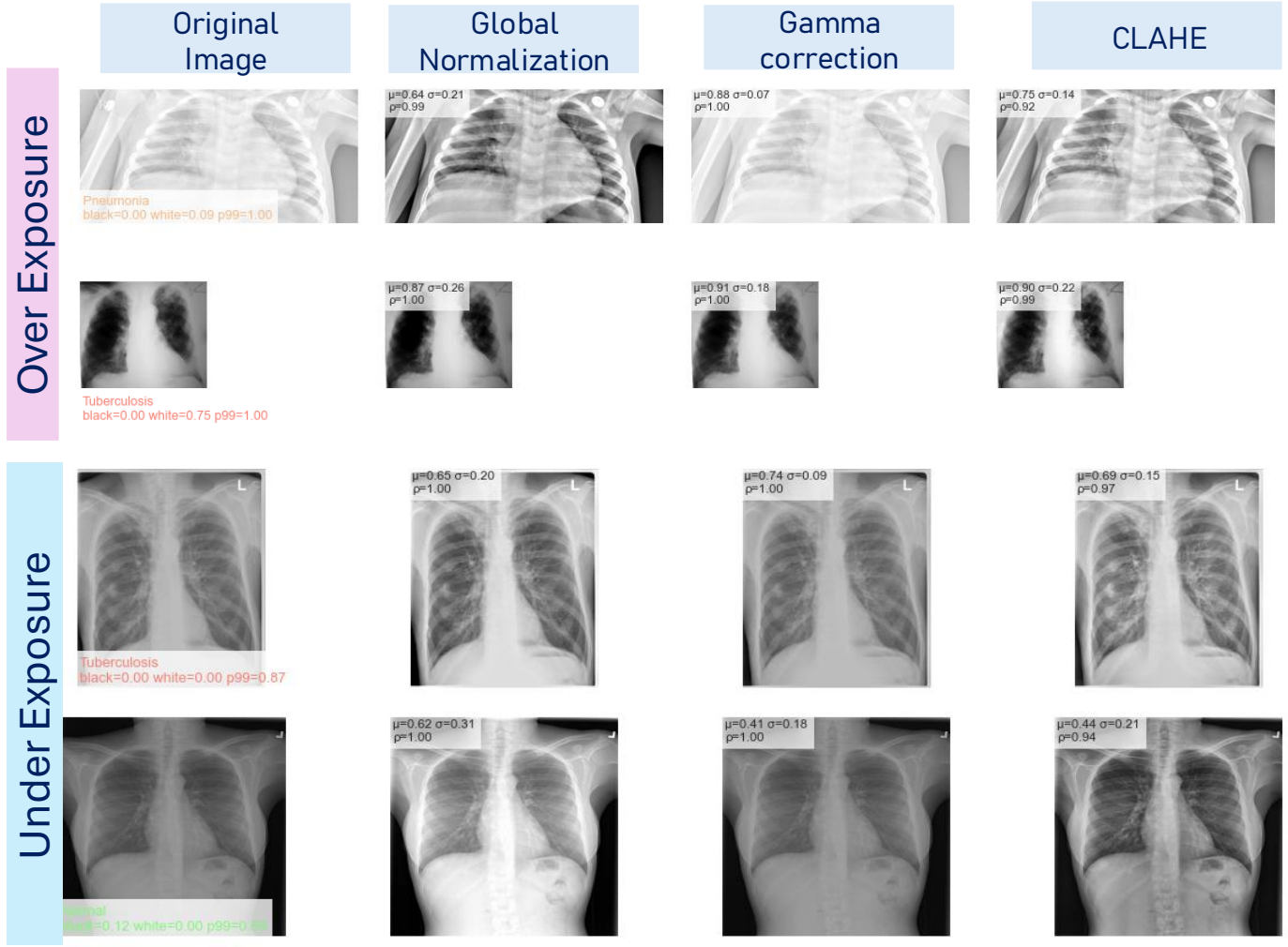
- TB images show higher exposure variability
- ~2% statistically under/over-exposed
  - Under-exposed (474 images)
  - Over-exposed (248 images)

EDA: Contrast Enhancement & Black Borders

Edge region intensity ratio
(Top, Left, Right, Bottom)

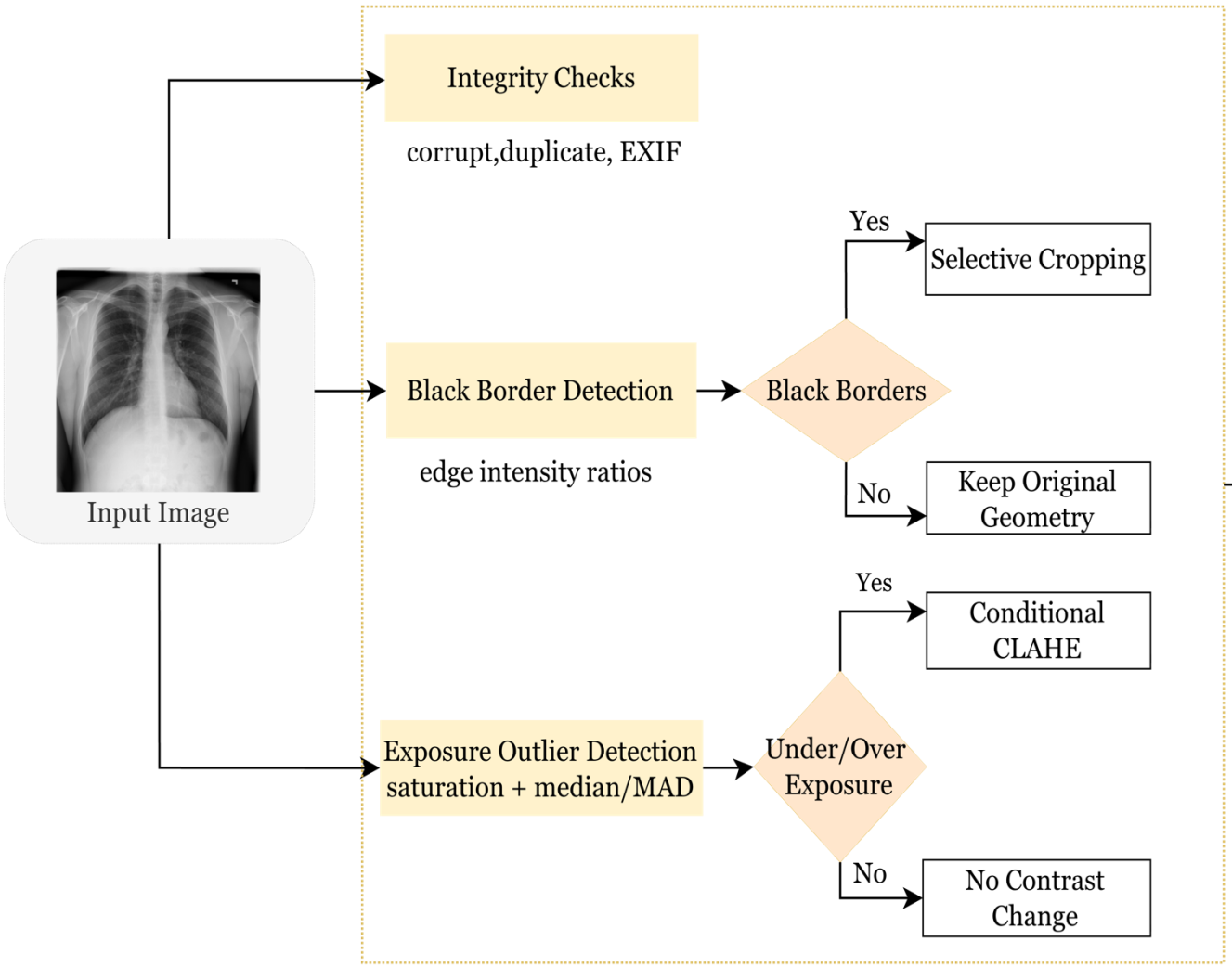


tuberculosis-3372.jpg



Contrast-Limited Adaptive Histogram Equalization (CLAHE) exhibits stronger enhancement without sever exposure degradation

Data Strategy: Preserving Anatomical Signal



Clinically plausible

- Horizontal flip
- Small rotations (≤10°)
- Minor translations
- Gaussian noise

Augmentations

Model Architecture

Model Architecture : Three paradigms

Linear Probing

Frozen backbones:

- ResNet-50
- DenseNet-121 *
- EfficientNet-B0
- Swin-Tiny

Result: Good baselines, but they struggled to separate subtle TB from Normal

Convolutional Neural Network

(Winner)

Architecture : 3 Block Sequential
(Conv → BN → ReLU → Pool)

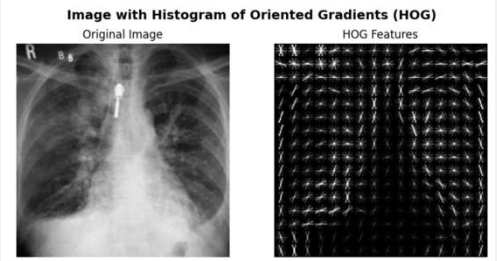
Rationale: Trained from Scratch.

Support: Grad-CAM++ Analysis

CNN Ablation Study : Progressive Build

1. Original Dataset
2. + Horizontal Flip
3. ++ conditional CLAHE *

Classical Features



Histogram of Oriented Gradients (HOG)

- MLP *
- Random Forest
- XGBoost

Result: Failed to capture spatial nuance.

Hardware Constraint: Mac MPS Backend (Batch Size <= 8)

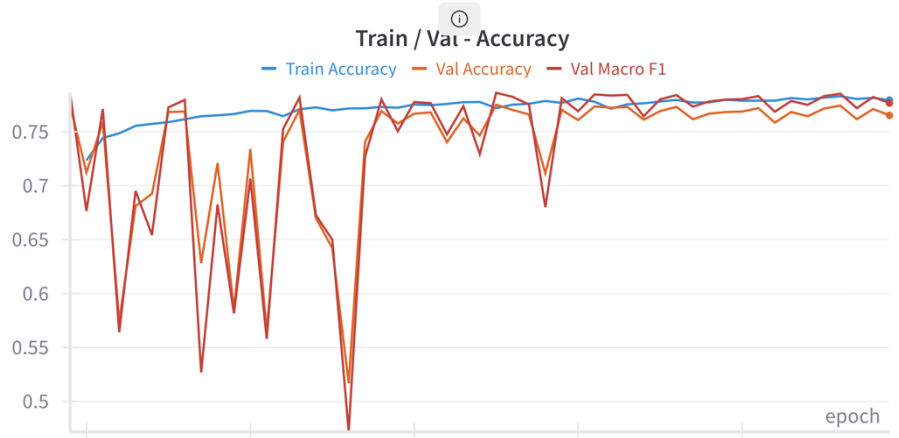
Training Setup

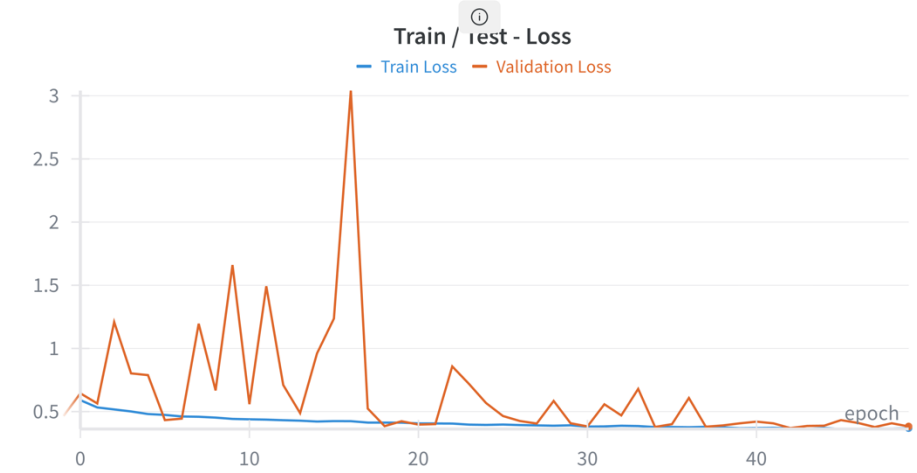
Configuration

Framework: PyTorch

1. Optimizer: AdamW
2. Weighted cross-entropy loss
3. Batch Size : 4-8
4. Cosine learning rate schedule
5. Gradient clipping (L2 norm = 1.0)
6. Early stopping : Patient of 10 on validation loss
7. Fixed random seed for reproducibility

CNN



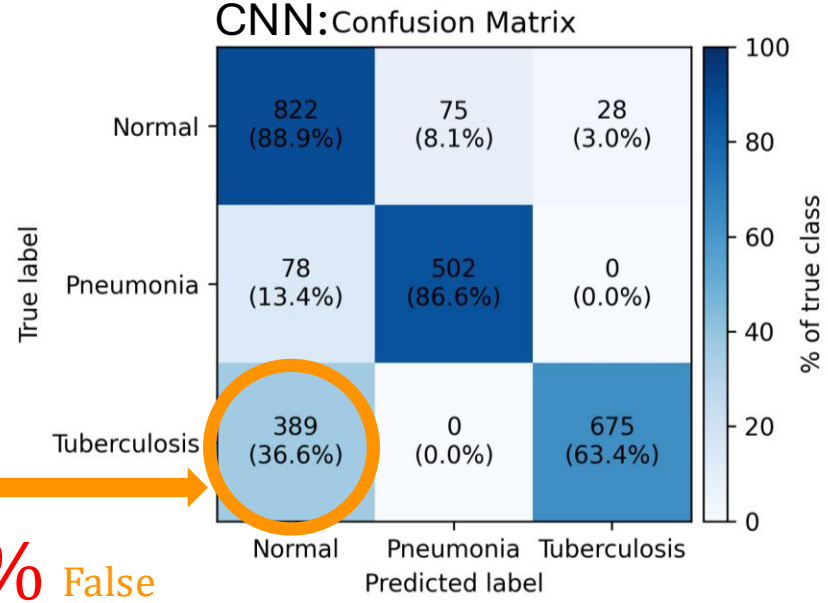


Model Comparison

Run Name	Cr	Da	Du	Source	Models	best_val_macro_f1
cnn-ft-30ep-clahe	·			train.py	best_model_macro_f1 +1	0.699889600276947
efficientnetb0-lp-20	·			train.py	best_model_macro_f1 +8	0.7688824534416199
resnet50-lp-20	·			train.py	best_model_macro_f1	0.7757332921028137
cnn-ft-30ep-clahe	·			train.py	best_model_macro_f1 +8	0.7790143489837646
cnn-ft50ep-flip-clahe	·			train.py	best_model_macro_f1 +7	0.783299446105957
densenet121-lp-20	·			train.py	best_model_macro_f1 +4	0.7839264869689941
swin-tiny-lp-20	·			train.py	best_model_macro_f1 +3	0.7845169901847839
cnn-ft50ep-clahe	·			train.py	best_model_macro... +11	0.7861547470092773

The ArgMax: High Accuracy, Critical Failures

The Issue: Argmax favors precision over sensitivity



36.6% False Negatives

78% Accuracy

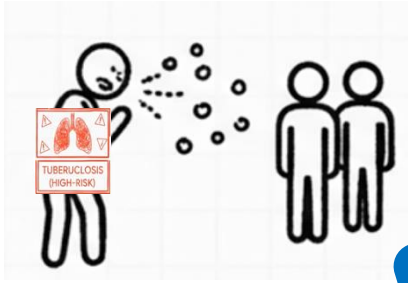
Classification Report

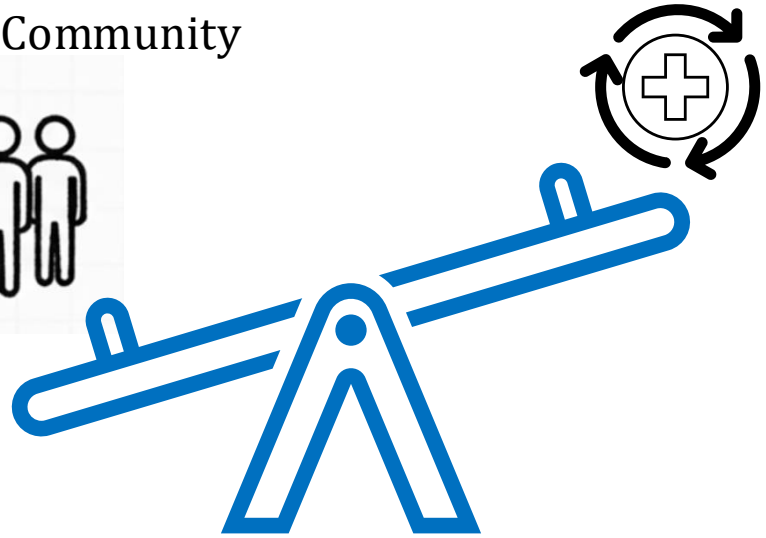
	precision	recall	f1-score	specificity	support
Normal	0.64	0.89	0.74	0.72	925
Pneumonia	0.87	0.87	0.87	0.96	580
Tuberculosis	0.96	0.63	0.76	0.98	1064
accuracy			0.78		
macro avg	0.82	0.8	0.79		2569
weighted avg	0.82	0.78	0.78		2569

The Consequence:

Added Cost to workflow

Spread disease to Community





Accepting more false positives is far safer than high false-negative rates for TB ? But by how much ?

Risk-Prioritized Policy

Policy-Aware Decision Framework

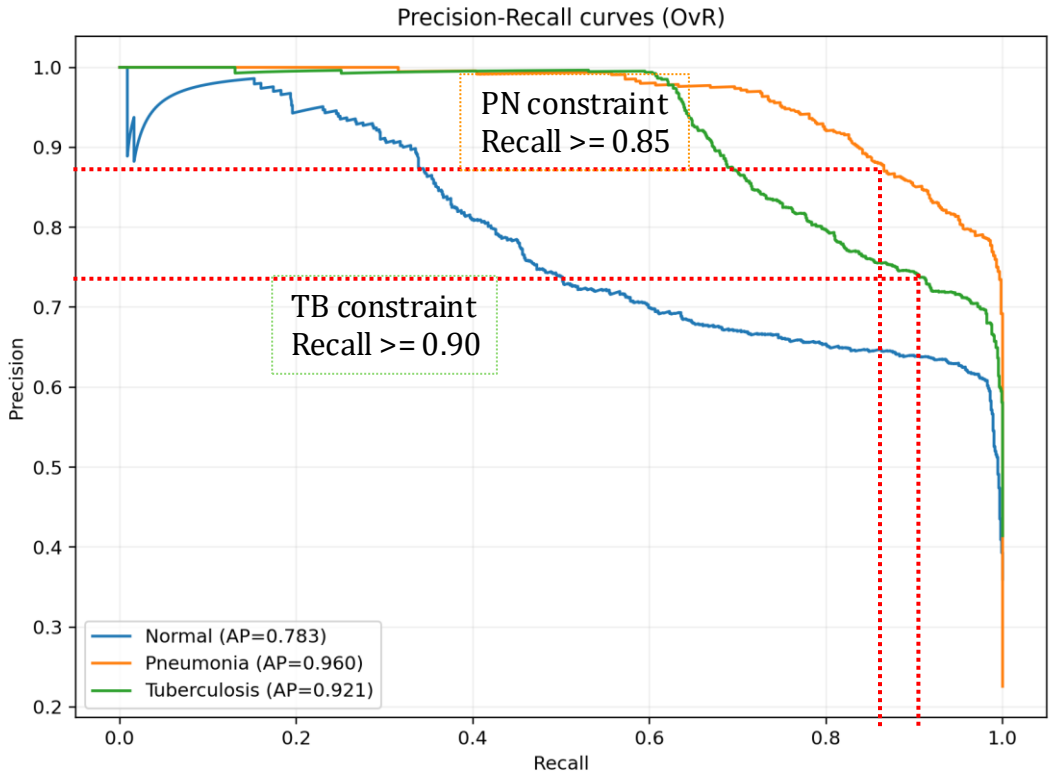
$$f(x) = [p_{TB}(x), p_{PN}(x), p_N(x)]$$

Class-specific operating thresholds τ_c, $c \in \{TB, PN\}$,
one-vs-rest (OvR) threshold sweeps

$$Sensitivityc(\tau c) \geq \alpha c, \alpha TB = 0.90, \alpha PN = 0.85$$

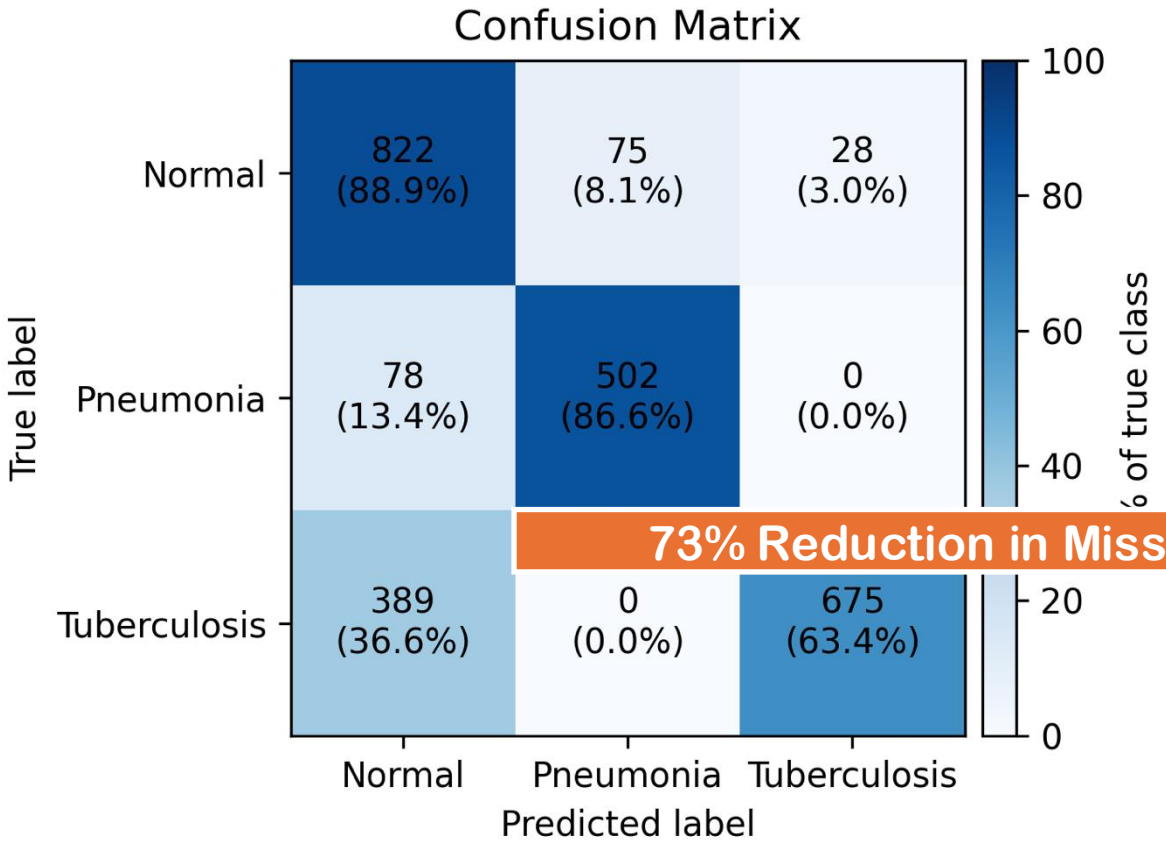
Hierarchical order

- $$\hat{y}(x) = \begin{cases} \text{TB}, & p_{\text{TB}}(x) \geq \tau\text{TB}, \\ \text{PN}, & p_{\text{PN}}(x) \geq \tau\text{PN} \ \wedge \ p_{\text{TB}}(x) < \tau\text{TB}, \\ \text{N}, & \text{otherwise} \end{cases}$$

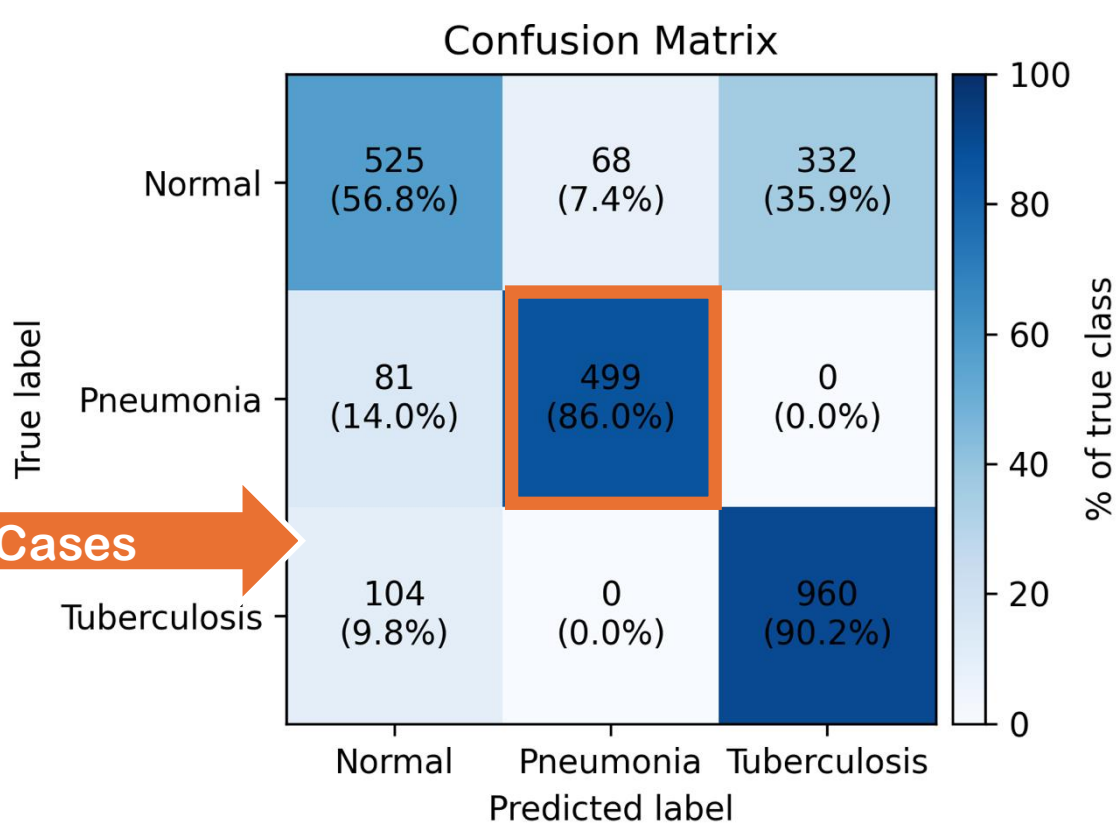


Policy Impact: Trading Precision for Safety

Standard Argmax



Risk-Prioritized Policy



73% Reduction in Missed Cases

TB Sensitivity: 63.4% ➔ **90.2%**

Normal Recall: 88.9% ➔ 56.8% (Intentional Over-Triage)

Run time @ Batch processing

```
cnn_model = load_run_model_pytorch(cnn_run_id)
stats = benchmark_inference(cnn_model, cnn_test_loader, device)
print(stats)
```

✓ 15.7s

```
Downloading artifacts:   0% 0/1 [00:00<?, ?it/s]
Downloading artifacts: 100% 6/6 [00:00<00:00, 354.63it/s]
{'num_images': 2569, 'total_time_sec': 15.140673166999477, 'latency_ms_per_image': 5.8936057481508275, 'throughput_img_per_sec': 169.67541480251865}
```

num_images = 2569
total_time_sec = 15.14 s
latency_ms_per_image = 5.89 ms
throughput_img_per_sec = 169.7 img/s

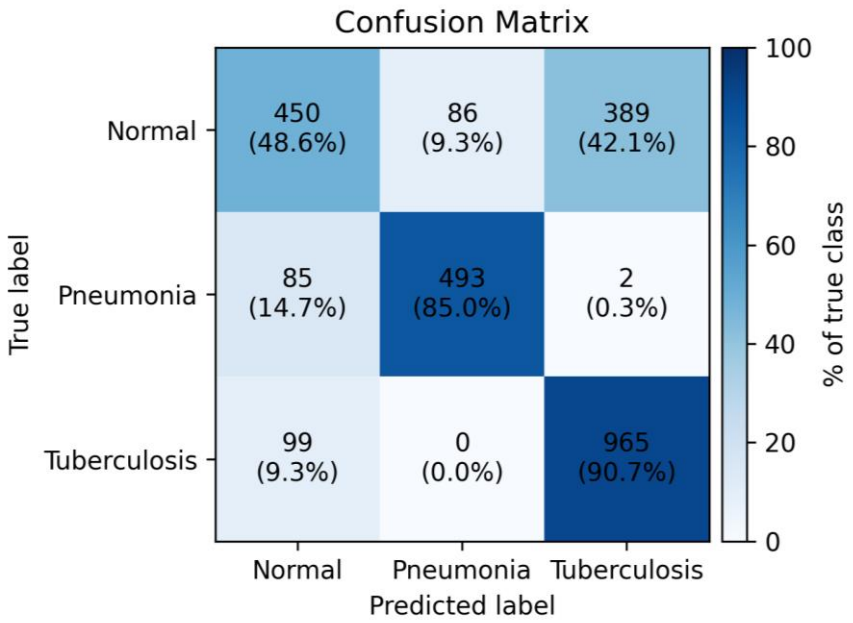
Benchmarked inference with proper warm-up and device synchronization.

The CNN runs at ~6 milliseconds per image, with a throughput of ~170 images per second on the test set.
This comfortably supports near–real-time screening workflows without becoming a bottleneck in clinical pipelines.

Comparative Analysis: Why CNN Prevailed

DenseNet-121

Linear Probe Classification Report (Policy)



Classification Report

	precision	recall	f1-score	specificity	support
Normal	0.71	0.49	0.58	0.89	925
Pneumonia	0.85	0.85	0.85	0.96	580
Tuberculosis	0.71	0.91	0.8	0.74	1064
accuracy			0.74		
macro avg	0.76	0.75	0.74		2569
weighted avg	0.74	0.74	0.73		2569

DenseNet-121 (Linear Probe)

- Even with policy adjustments, the frozen backbone produced more False Positives among Normal cases.

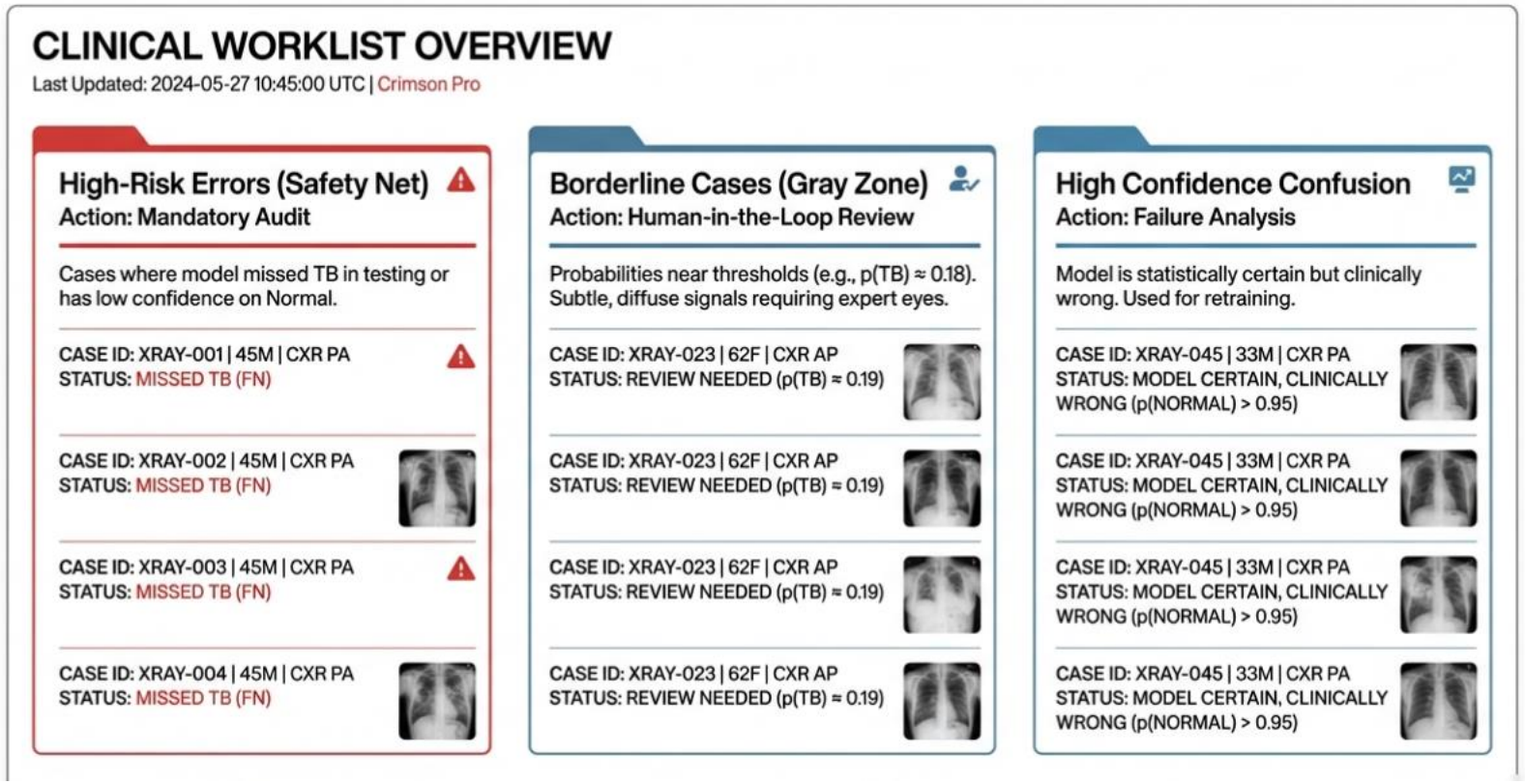
The Limitation

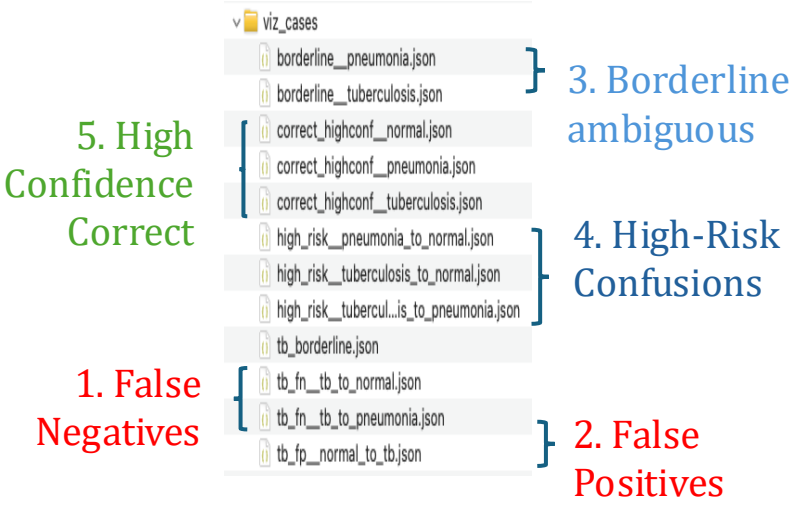
Frozen backbones encode generic texture features Without fine-tuning (due to compute constraints), they couldn't separate “atypical Normal” from “subtle TB” as effectively as the trained-from-scratch CNN.

HOG Features

- Proved insufficient for safety-critical tasks due to poor calibration and lack of spatial awareness.

Beyond Labels: Curated Clinical Worklists





We will use SmoothGrad-CAM++ to audit the model

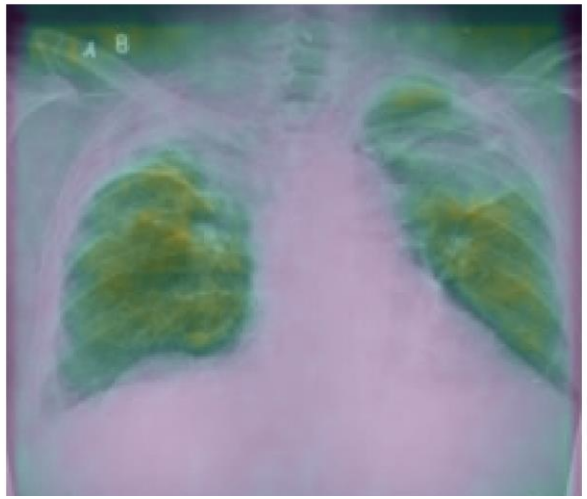
Verification: Validation Feature Attention

Correct cases

True: Tuberculosis | Pred: Tuberculosis | Confidence: 0.983 | Explaining: Tuberculosis (0.983)

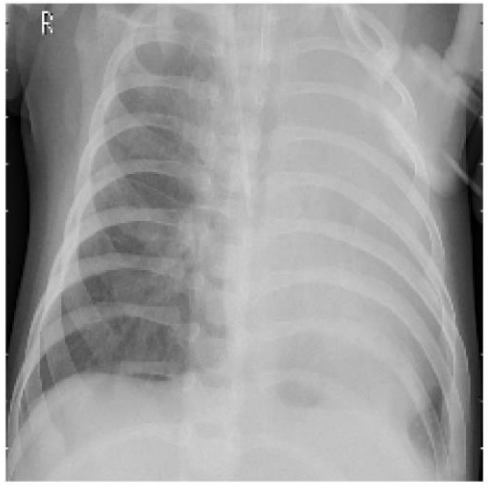
Original

Grad-CAM



True: Pneumonia | Pred: Pneumonia | Confidence: 1.000 | Explaining: Pneumonia (1.000)

Original

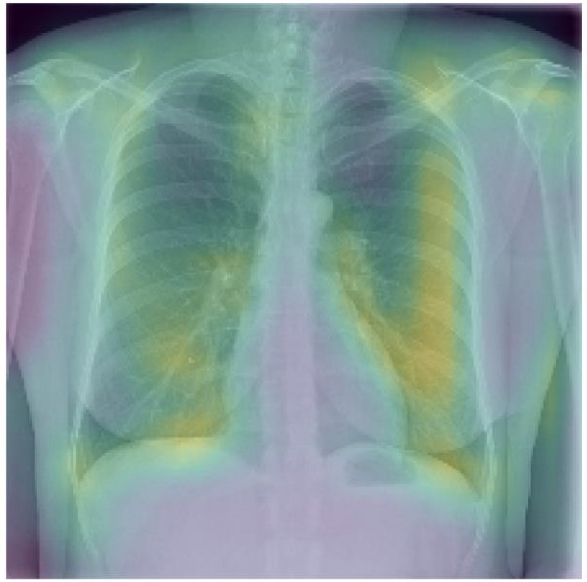


Grad-CAM

True: Normal | Pred: Normal | Confidence: 1.000 | Explaining: Normal (1.000)

Original

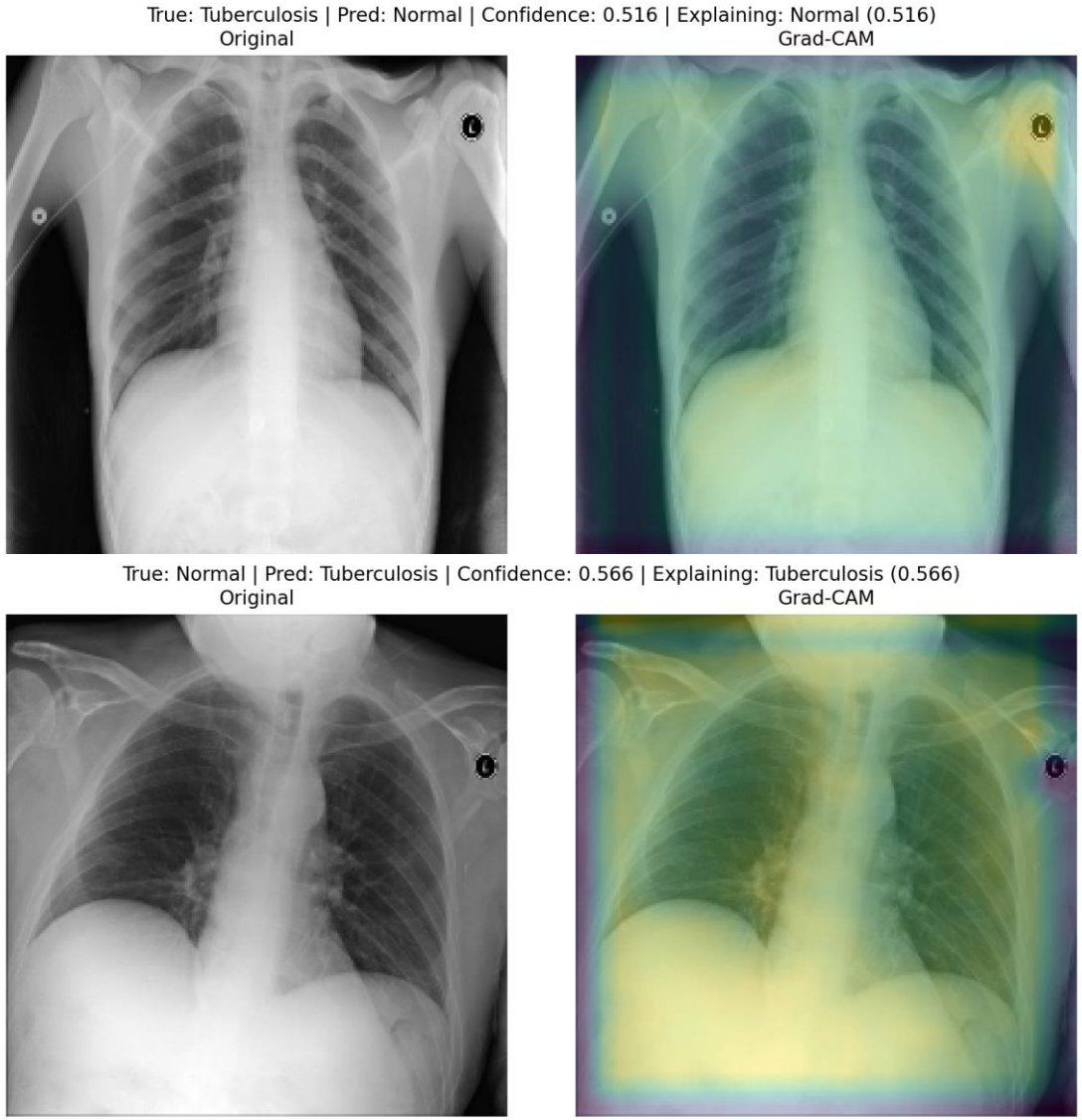
Grad-CAM

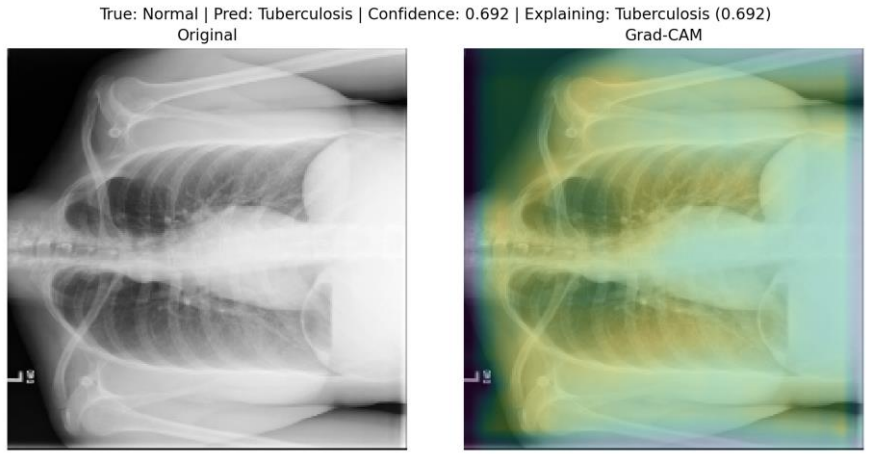


Observation:

- TB & Normal → broad bilateral focus
- Pneumonia → localized lung regions

Failure Modes & Safety Checks





Borderline

TB → Normal & Normal → TB

1. weak lung signal
2. Diffuse Attention

Shortcut learning: Occasional border/marker activation dataset bias

MLOps, Deployment and Demo

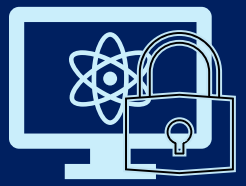
MLOps Pipeline

From Experiment to Containerized Artifact

Experimentation & Tracking

ML Flow Tracking.
Log metrics, params and artifacts for every run.

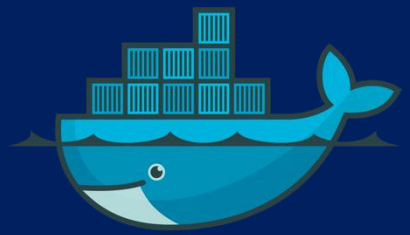
Model Bundling



Bundle-best for Locked artifact.
Weights | Config | Thresholds | Metadata

Stable versioning via bundles/latest

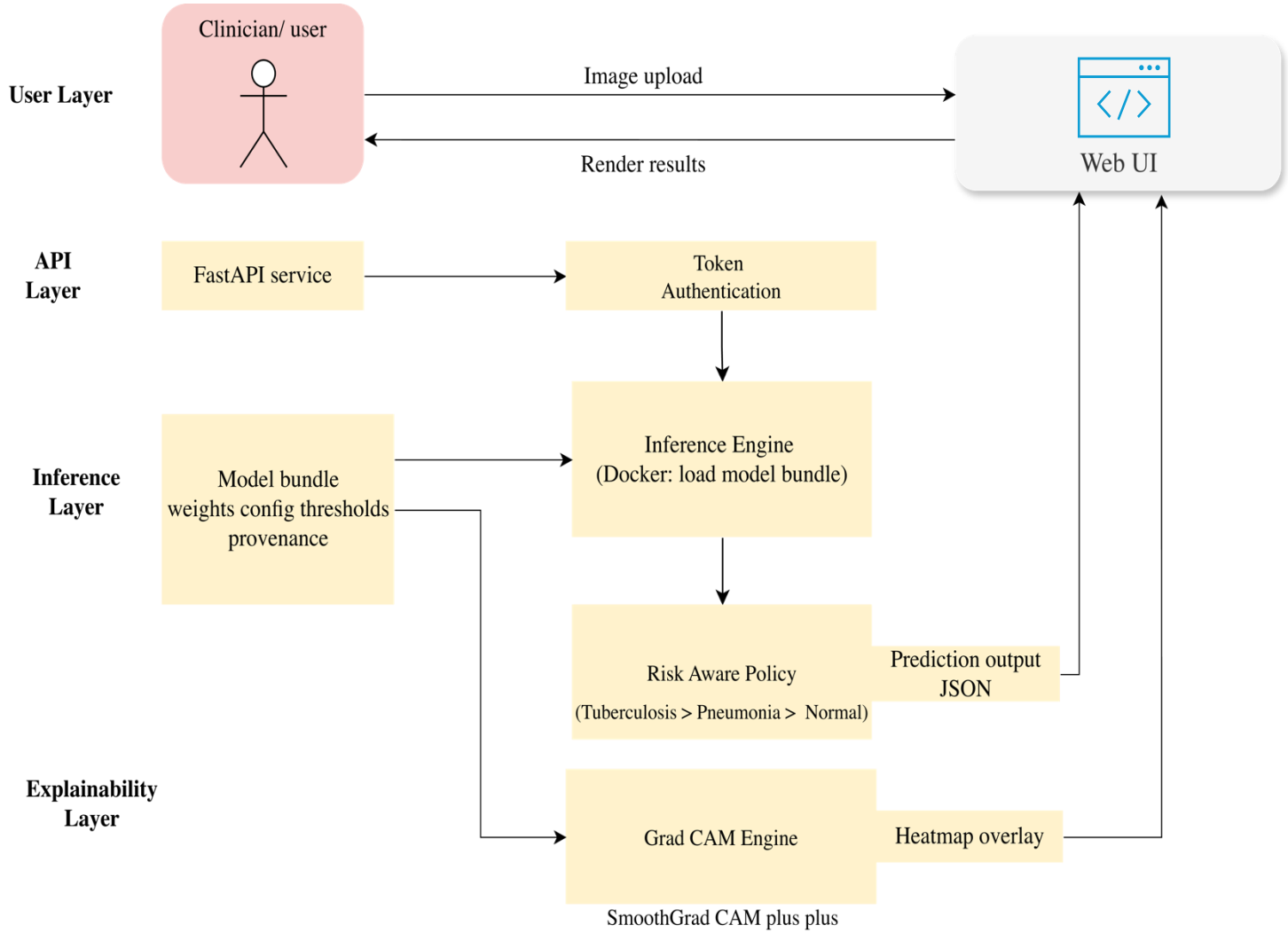
Containerization



Two Docker Images

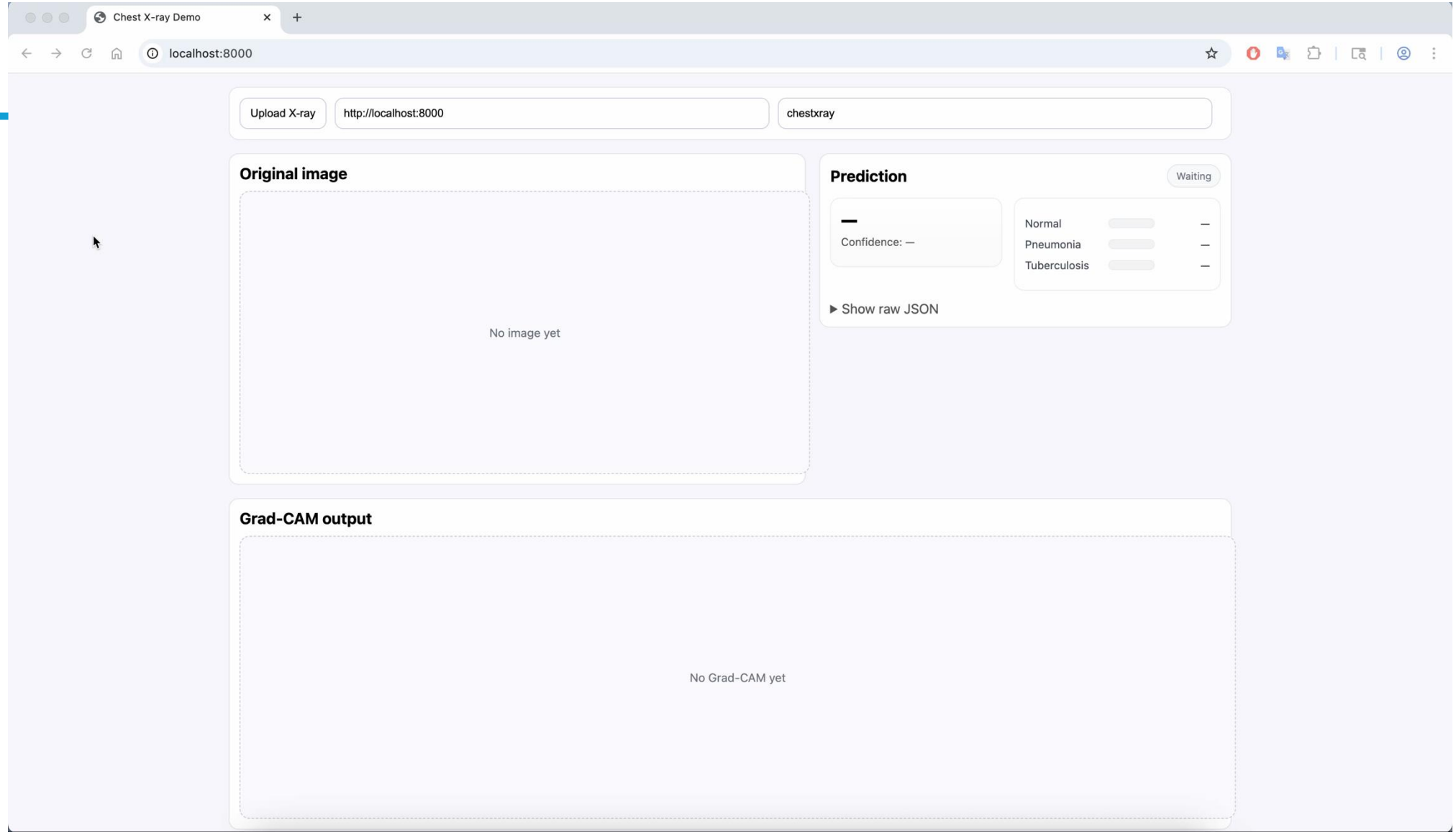
1. Inference CLI
2. FastAPI + Web UI

Deployment System Design



Demo

http://3.99.49.50/
labviewer:
labpassword



Discussion

Limitation & Future Improvements

Images Alone Are Not Diagnostic

Current Limitation

Compute: Model depth limited by hardware (Mac MPS).

Input: Single-image inference for UI only (no batch API yet).

Explainability: Grad-CAM is qualitative (a hint, not a proof).

...

Future Improvements

Mitigate shortcut learning

Monitoring System
Feature & Model drift detection.

Robustness Testing
Evaluation on external datasets to test for hospital-specific bias.

Self-Supervised Learning
Pre-training on unlabeled data to improve feature robustness.

...

Clinical Integration

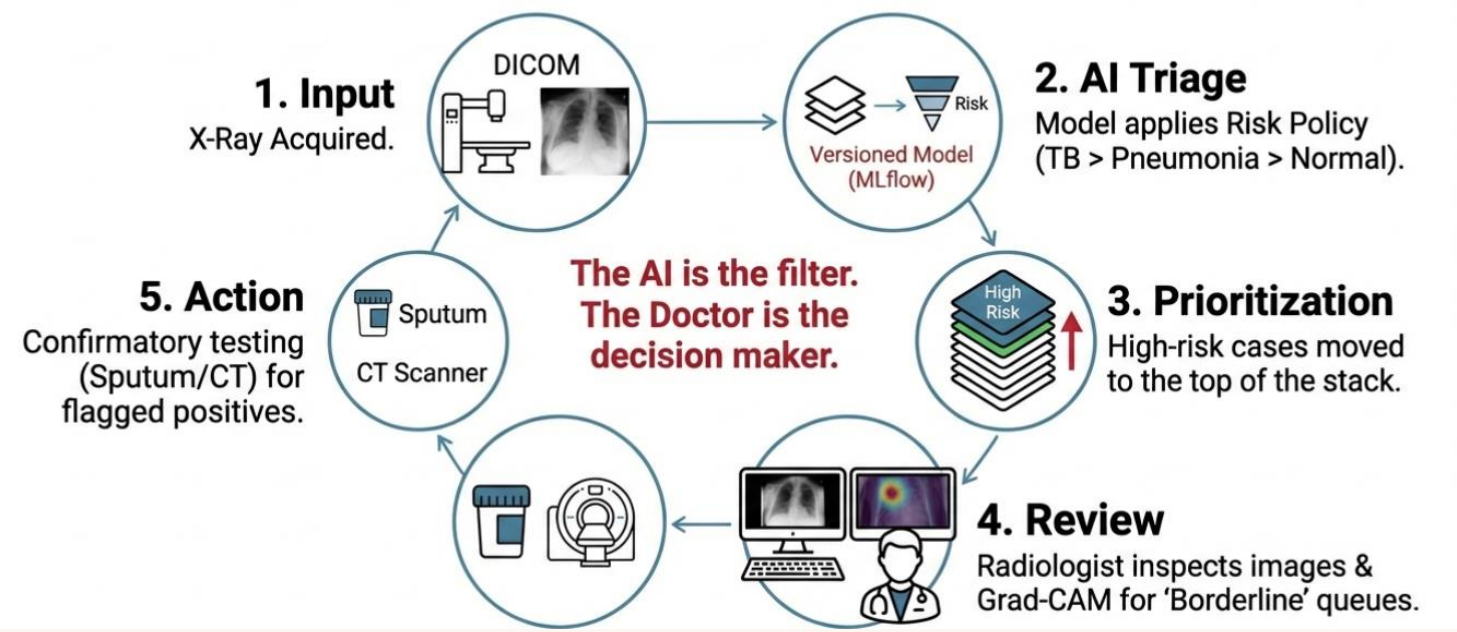
API deployment directly into PACS Viewers

Feedback Loops
Mechanism for radiologists to correct predictions and retrain the model.

Interoperability : Batch Triage
Extending API to handle full patient queues.

...

The Human in the loop workflow



Foundations of Responsible AI

Four Ethical Pillars

Compliance

- Store / Compute On-premise
- Compliant matters ~
- Software as Medical Devices ?

Software as Medical Devices Classification Level

	Treat or Diagnose	Drive clinical management	Inform clinical management
Critical	III	III	I or II
Serious	II or III	II or III	I or II
Non-serious	I or II	I or II	I or II

Conclusion

Responsible Al in healthcare isn’t just about higher accuracy numbers.

It is about choosing the right Operating Point and aligning technical parameters with clinical values, we created a viable safety net.

System bundled, containerized, and ready for deployment.

Thank You!