Augmenting Chest X-Ray Screening with Risk-Prioritized Deep Learning and Explainable AI

Nway Nway Aung

1. Introduction

Background

Chest X-ray imaging is widely used to assess disease due to its low cost, accessibility, and clinical relevance. Automated analysis of chest radiographs has gained increasing attention as a means of supporting clinicians, particularly in large-scale screening and resource-limited settings. Diseases such as pneumonia and tuberculosis remain major global health concerns, and timely identification is critical for patient outcomes.

Despite advances in deep learning, chest X-ray classification is challenging due to variation in image acquisition, exposure, and resolution, as well as the presence of non-anatomical artifacts (e.g., black borders and text overlays). In addition, clinically meaningful features are often subtle and spatially diffuse, increasing the risk that models learn spurious correlations rather than pathology-relevant representations.

Beyond predictive accuracy, clinical ML systems require reproducibility, interpretability, and deployment readiness. Models must be robust to minor distribution shifts, calibrated for clinical risk tolerance, and interpretable enough to support validation and trust. These considerations motivate an end-to-end pipeline that integrates data processing, model training, evaluation, explainability, and deployment.

Related work

Early chest X-ray systems relied on handcrafted features (e.g., texture descriptors and edge-based representations) combined with classical machine learning classifiers. While relatively interpretable, these approaches generally lack the representational capacity to capture complex and variable pathology.

More recent work demonstrates the effectiveness of convolutional neural networks (CNNs), especially when pretrained on large-scale datasets and fine-tuned for medical imaging tasks. Architectures such as ResNet, DenseNet, and EfficientNet have shown strong performance, benefiting from transfer learning when labeled medical data are limited. Dense connectivity and multi-scale feature reuse are particularly helpful for radiographic texture analysis.

Transformer-based models have also been explored, offering global context modeling through self-attention. However, data efficiency and interpretability remain open issues, especially for high-resolution grayscale imagery such as X-rays.

Recent studies emphasize that performance metrics alone are insufficient for clinical deployment. Calibration, robustness, and failure-mode analysis often supported by post-hoc explainability methods such as Grad-CAM are increasingly considered necessary for safe real-world use.

2. Technical Tasks

Task 1. Data Processing & EDA

1. Exploratory Data Analysis (EDA)

Dataset Overview

The dataset used in this study was obtained from the *Chest X-Ray Dataset* [1] . It contains 25,553 chest X-ray images labeled as Normal, Pneumonia, or Tuberculosis, and includes predefined training, validation, and test splits. These splits were used without modification.

All exploratory analysis was conducted using: src/notebooks/01.EDA.ipynb.

Class Distribution and Imbalance

Figure 1: Class distribution across splits

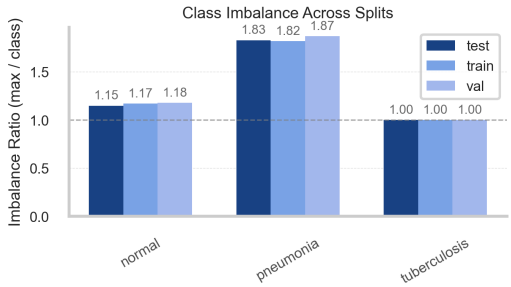


Figure 2: Class Imbalance and Per-Split Distributions

The dataset exhibits moderate class imbalance, with tuberculosis as the majority class and pneumonia as the most underrepresented. The maximum-to-minimum class ratio is approximately 1.8. Class proportions remain consistent across splits, suggesting effective stratification and minimizing split-induced bias.

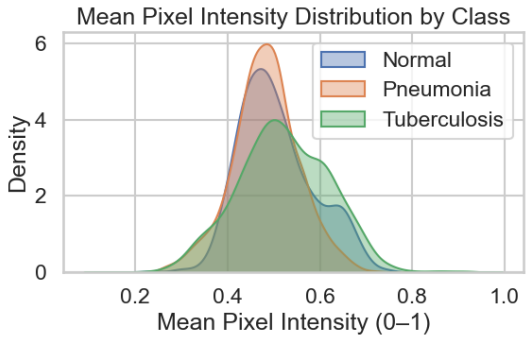


Figure 3: Mean Pixel Intensity Distribution by Class

Pixel intensity histograms show wide variation across images, reflecting differences in exposure and acquisition conditions. Texture inspection indicates that diagnostically relevant information is dominated by subtle lung parenchymal patterns rather than strong edges. Visual inspection shows significant variability in exposure (under/over), contrast, resolution, black border artifacts, Pixel intensity histograms confirm wide exposure variation. Texture inspection indicates that discriminative signal is dominated by subtle lung parenchymal patterns rather than strong edges.

2. Preprocessing Pipeline

Image Size and Aspect Ratio Standardization

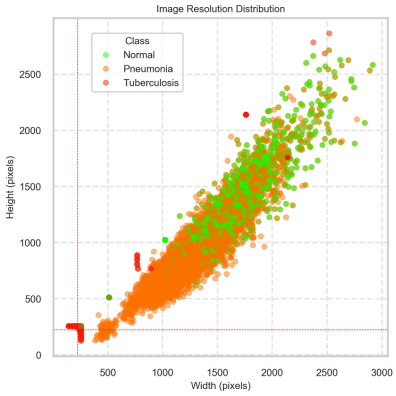


Figure 4: Image Resolution Distribution

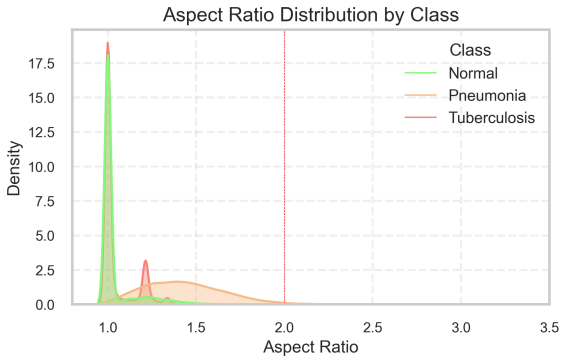


Figure 5: Aspect Ratio

Image geometry analysis depicts a strong linear correlation between image width and height, indicating largely consistent acquisition protocols across the dataset. Most images exceed the 224 × 224 resolution threshold, with only a small fraction falling below this minimum. These low-resolution samples are disproportionately concentrated in the tuberculosis class, whereas normal and pneumonia images are predominantly well-sized. From aspect-ratio analysis, it shows that most images are near-square (AR ≈ 1), with pneumonia images exhibiting slightly greater variability and extreme aspect ratios occurring rarely. Removing low-resolution or extreme-aspect-ratio images would disproportionately affect specific classes and introduce bias. Class imbalance was addressed during training through loss weighting rather than data resampling.

Images were resized to a fixed input resolution of *224 × 224* to match all deep learning architectures used in this study. Aspect-ratio-preserving cropping was intentionally avoided, as cropping may remove peripheral lung regions that could contain diagnostically relevant information. EDA confirmed that resolution variability was moderate and class-consistent, supporting this design choice.

Intensity Normalization and Contrast Enhancement

Under- and over-exposed images were identified using a combined detection strategy based on pixel saturation metrics (near-black and near-white fractions) and robust brightness statistics derived from median/MAD z-scores. Using the union of these methods, 474 images were flagged as under-exposed and 248 as over-exposed, with tuberculosis images exhibiting the highest exposure variability; overall, approximately 2% of images were identified as exposure outliers.

Qualitative inspection indicated that exposure artifacts were generally subtle and did not result in severe loss of anatomical detail. This was supported by quantitative analysis using mean intensity (μ), contrast (standard deviation), and structural similarity measured via pixel-wise correlation (ρ) between original and enhanced images.

Three contrast adjustment methods, global normalization, gamma correction, and Contrast-Limited Adaptive Histogram Equalization (CLAHE) were evaluated on flagged images. CLAHE consistently improved local contrast while preserving anatomical structure, achieving high correlation with original images ($\rho \approx 0.88$-0.99), whereas gamma correction and global normalization provided weaker enhancement or introduced contrast compression.
Based on these findings, CLAHE was selected and applied conditionally, only to images exhibiting low contrast without severe exposure degradation, rather than uniformly across the dataset.

Black Border Inspection and Removal

Black borders were detected using edge-region intensity ratios computed from fixed-width regions along all four image boundaries (Top, Left, Right, Bottom). Images exceeding a predefined black-border threshold were selectively cropped prior to resizing, while unaffected images were left unchanged.
The resulting border reduction distribution was bimodal, with the majority of images showing negligible change and a smaller subset exhibiting substantial border removal (often >60%), confirming that non-informative padding was present only in a limited fraction of samples.
All exploratory data analysis in this section was conducted using the notebook src/notebooks/ 02.Preprocessing_Blackborder.ipynb

Visual inspection confirmed no adverse impact on image geometry, with changes limited to revealing the true effective resolution of affected images, primarily within the tuberculosis class.

All subsequent experiments used border-cleaned images. Following border removal, image resolution and aspect ratio summaries were recomputed to verify that no substantial changes occurred.

label	small_image	extreme_aspect_ratio
normal	0.00	0.02
pneumonia	0.70	2.16
tuberculosis	18.66	0.01

label	small_image	extreme_aspect_ratio
normal	0.00	0.02
pneumonia	0.70	2.16
tuberculosis	18.84	0.01

Figure 6: Before and After Border Removal : Image Resolution (Small) & Aspect Ratio (Small)

Decision Made after Exploratory Data Analysis: Human chest anatomy is asymmetric (e.g., cardiac silhouette and mediastinal shift); however, radiographic diagnostic criteria for tuberculosis and pneumonia are largely side-invariant, and pathology is frequently bilateral or multifocal. Consequently, preprocessing steps such as border removal and horizontal flipping were deemed anatomically safe and unlikely to distort clinically relevant cues. Accordingly, CLAHE was applied only to images exhibiting low contrast without severe exposure degradation, ensuring contrast enhancement did not compromise image quality as see table below.

Train, validation, and test data loaders were constructed directly from the dataset's predefined splits, with fixed random seeds and deterministic shuffling applied where appropriate to ensure reproducibility.

3. Data Augmentation

To improve generalization while preserving clinical validity, augmentations were restricted to transformations that reflect realistic acquisition variability. Applied augmentations included horizontal flips, small rotations (≤10°), minor translations, and with gaussian noise. Contrast enhancement was performed only on over and under exposure images using CLAHE. All augmentations were implemented conservatively to avoid distortion of anatomical structure.

These operations simulate patient positioning differences and moderate scanner noise without altering pathology-relevant anatomy. Unrealistic transformations that could distort diagnostic cues were avoided, including vertical flips, large rotations, and aggressive geometric warps/elastic deformations, as these can invert anatomical orientation or introduce non-physiologic structures. Augmentation impact was quantified through an ablation comparison (augmentation vs. none) using validation performance and training curves.

4. Feature Engineering

Histogram of Oriented Gradients (HOG) was explored as a classical feature extractor due to its suitability for grayscale, low-contrast imagery and its focus on local gradient orientation and edge structure, which are salient cues in radiographic images. Chest X-rays lack color information and often rely on subtle texture and shape patterns, making HOG a reasonable baseline for assessing how much discriminative signal can be captured using handcrafted features alone.

HOG features were combined with standard classifiers (MLP, Random Forest, and XGBoost) to establish interpretable, non-deep-learning baselines. While these approaches captured coarse texture information, they consistently underperformed convolutional neural networks, indicating that handcrafted features are insufficient to model the subtle, spatially distributed patterns characteristic of chest X-ray pathology.

Self-supervised contrastive learning methods (e.g., SimCLR with linear probing[2]) were considered due to their suitability for limited-label medical imaging settings. However, such methods typically require large batch sizes and substantial computational resources to learn stable representations. Given the hardware constraints of the local training environment, contrastive pretraining was not feasible in this study. Nonetheless, self-supervised representation learning remains a promising direction for future work.

5. Missing or Corrupt Data Handling

The dataset consists solely of JPEG images with pixel data. All 25,553 images across training, validation, and test splits were inspected for EXIF metadata, and none was present; consequently, modeling relied exclusively on image-based features.
Prior study suggested that, auxiliary metadata (e.g., age or sex) when available can provide coarse supervisory signals that help impose global structure during training and improve representation learning, particularly in medical imaging tasks with limited labeled data.[3]

Data integrity checks identified no corrupt, unreadable, or duplicate images, confirming the dataset's suitability for downstream analysis and modeling.

Task 2 - Model Training and Fine-Tuning

1. Model Architecture Choice

Given limited computational resources, this study prioritizes lightweight and computationally efficient architectures while maintaining sufficient representational capacity for chest X-ray classification. The modeling strategy includes

(i) a custom convolutional neural network (CNN) [4] trained from scratch,
(ii) transfer-learning-based models using frozen pretrained backbones with linear probing techniques and
(iii) classical feature-based approaches using Histogram of Oriented Gradients (HOG) combined with standard classifiers (MLP, Random Forest, XGBoost).

Xray is grayscale image and it has identical values across all three channels, we use a single channel when training a CNN, whereas for linear probing we use three channels.

CNN Architecture Details

The baseline CNN consists of three sequential convolutional blocks operating on single-channel (grayscale) input images. Each block progressively increases channel depth to capture hierarchical texture representations:

- Block 1: Conv(1→32) → BatchNorm → ReLU → MaxPool
- Block 2: Conv(32→64) → BatchNorm → ReLU → MaxPool
- Block 3: Conv(64→128) → BatchNorm → ReLU

The final convolutional layer (Conv(64→128)) serves as the target layer for Grad-CAM analysis, as it provides the highest-level spatial feature maps while retaining sufficient resolution for meaningful localization. This design balances representational depth with interpretability, which is particularly important for explainability in medical imaging.

CNN Ablation Study

To isolate the impact of preprocessing choices, a CNN ablation study was conducted using a fixed architecture. Three configurations were evaluated:
(i) Base CNN without augmentation,
(ii) CNN with horizontal flipping, and
(iii) CNN with horizontal flipping combined with conditional CLAHE contrast enhancement.

All other training settings were held constant. Horizontal flipping improved generalization over the base model, while the Flip + CLAHE configuration achieved the best performance. Consequently, all subsequent CNN experiments used the Flip + CLAHE setting.

Pretrained Backbones (Linear Probing)

In addition to the CNN baseline, pretrained backbone architectures were evaluated using a linear probing strategy, in which the feature extractor is frozen and only a lightweight classifier head is trained. Backbones were initialized with ImageNet pretrained weights and sourced from the timm library, including ResNet-50 [5], DenseNet-121 [6], EfficientNet-B0 [7], and Swin-Tiny (patch4 window7 224)[8]. ImageNet [9] pretraining provides robust low-level representations (edges, gradients, textures) that transfer effectively to radiographic images despite domain differences. These architectures were selected for their computational efficiency, multi-scale texture modeling capability, and suitability for moderate-scale medical imaging datasets. Linear probing further reduces training cost and overfitting risk under limited compute.

2. Training Procedure

All models were implemented in PyTorch. Experiments were conducted with a fixed random seed (42) to ensure reproducibility. Training workflows were standardized using Makefile-based execution, with configurations defined in cnn.yaml: make train --cfg configs/cnn.yaml

Optimization was performed using the *AdamW* optimizer with weight decay fixed at 1×10^{-4}. Learning rates were selected according to the training regime:

- CNN trained from scratch: 3×10^{-4}
- Linear probing of pretrained backbones: 1×10^{-3}

A cosine annealing learning rate scheduler was applied over the full training horizon, decaying the learning rate toward 1×10^{-6}, with one scheduler step per epoch.

Batch Size and Hardware Constraints
All experiments were executed *on a Mac notebook* using the *Metal Performance Shaders (MPS)* backend. Due to limited GPU memory and throughput, batch sizes were constrained as follows:

- CNN baseline: *batch size = 8*
- Pretrained backbones (linear probing): *batch size = 4*

These values represent the largest stable batch sizes at an input resolution of *224 × 224* without memory exhaustion. Learning rates were adjusted accordingly to maintain optimization stability.

LossFunction : Training employed weighted cross-entropy loss, with class weights computed from the training set to address class imbalance and align optimization with macro-averaged metrics. Gradient clipping was applied at each training step using a global ℓ_2 norm of 1.0 to improve numerical stability.

Early Stopping and Checkpointing : Early stopping was enabled for pretrained backbone experiments, monitoring *validation macro-F1 score* with Patience: *5 epochs* and Minimum improvement (min_delta): *0.0.*

3. Hyperparameter Tuning

Hyperparameters were selected through structured experimentation rather than exhaustive search, reflecting hardware constraints. The primary hyperparameters explored included:

- Learning rate: {3×10^{-4} (CNN), 1×10^{-3} (linear probing)}
- Weight decay: 1×10^{-4}
- Batch size: {8 (CNN), 4 (pretrained)}
- Number of epochs: {30 (CNN), 20 (pretrained)}
- Input resolution: 224 × 224
- Training regime: linear probing vs. training from scratch
- Data augmentation: flip {on/off}, CLAHE {on/off}, Gaussian blur {on/off}

Multiple mechanisms were employed to control model capacity and generalization such as overfitting or underfittiong. For Regularization and Stability we use AdamW weight decay: 1×10^{-4} and Gradient clipping: ℓ_2 norm capped at 1.0 and Conservative data augmentation (flip, CLAHE, mild blur).

4. Model generalization

Model generalization is assessed by tracking training and validation loss, validation macro-F1, and per-class precision, recall, and F1 scores. All configurations were evaluated using validation macro-F1 as the primary criterion. Experiments were logged using MLflow [10] and Weights & Biases locally, enabling systematic comparison and controlled ablation.

Observed generalization gaps guided adjustments to regularization strength, training duration, and learning rates. The best-performing model was automatically restored prior to final validation and test evaluation.

5. Evaluation Metrics and Reporting

At this stage, multiple modeling paradigms were evaluated: (i) end-to-end CNN-based deep learning models, (ii) linear probing with frozen pretrained backbones (ResNet, DenseNet, EfficientNet, Swin-Tiny), and (iii) handcrafted feature pipelines using HOG combined with MLP, XGBoost, and Random Forest classifiers.

Three candidates are selected to undergo this analysis.

For each paradigm, the best-performing model was selected based on validation performance and *evaluated on a shared held-out test set,* while preserving each model's native preprocessing pipeline. Model selection was driven entirely by MLflow experiment tracking, without manual intervention or subjective tuning. CNN models were ranked by validation accuracy, while other approaches were selected based on best macro-averaged performance

Following model selection, detailed analyses were conducted for each winning model, including threshold optimization to determine operating points aligned with clinical risk tolerance. Evaluation included AUC-ROC, F1 score, precision, recall, sensitivity, specificity, confusion matrices, and calibration analysis via reliability diagrams, with per-class as well as macro- and micro-averaged metrics reported.

We decided to perform evaluation that was aligned with a screening-oriented risk posture: operating thresholds were chosen to maximize sensitivity for Tuberculosis and Pneumonia (i.e., minimize false negatives), while accepting an increased false-positive rate and lower recall for the

Normal class as an intentional trade-off to reduce the likelihood of missing clinically significant disease.

Risk-Prioritized Decision Policy and Threshold Selection

To reflect realistic clinical screening priorities, model predictions were evaluated using a policy-aware decision framework rather than naïve argmax classification. Instead of assigning the class with the highest posterior probability, a risk-prioritized decision policy was applied to minimize false negatives for high-risk conditions.

Let $f(x) = [p_{TB}(x), p_{PN}(x), p_N(x)]$ denote the predicted class probabilities for Tuberculosis *(TB),* Pneumonia *(PN)*, and Normal *(N).*
Class-specific operating thresholds τ_c, $c \in \{TB, PN\}$, were determined using *one-vs-rest (OvR) threshold sweeps* on the validation set. Thresholds were selected to satisfy minimum sensitivity constraints for clinically critical classes while maximizing specificity subject to these constraints:

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

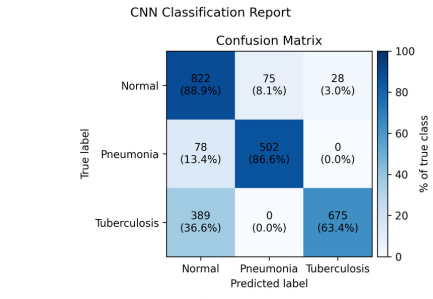
Final predictions were generated using a hierarchical, priority-ordered decision rule:

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

Under this policy, Tuberculosis predictions are first assigned using a sensitivity-prioritized threshold, followed by Pneumonia predictions for samples not already classified as Tuberculosis. All remaining samples are labeled Normal. This ordering mirrors real-world screening workflows, where failure to detect severe disease is substantially more costly than over-triage.
Model performance under this policy-aware framework was compared against a standard argmax baseline to quantify the impact of thresholding on sensitivity, specificity, and overall diagnostic behavior, with particular emphasis on Tuberculosis detection.

Performance Analysis and Clinical Implications

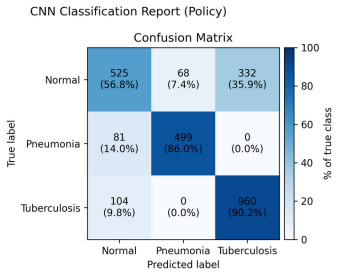
<u>CNN</u>



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

Figure 7: Confusion Matrix and Classification Report

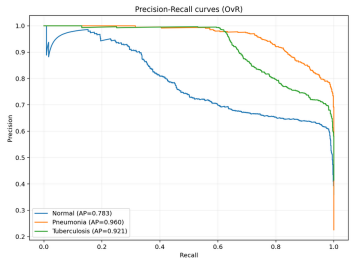


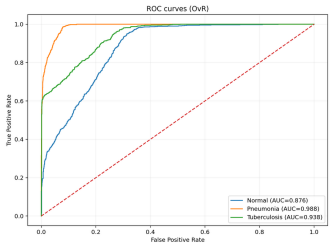
Classification Report

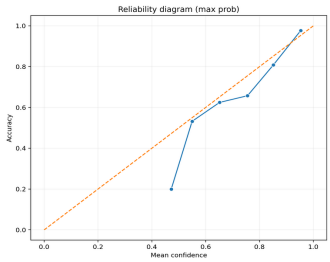
	precision	recall	f1-score	specificity	support
Normal	0.74	0.57	0.64	0.89	925
Pneumonia	0.88	0.86	0.87	0.97	580
Tuberculosis	0.74	0.9	0.81	0.78	1064
accuracy			0.77		
macro avg	0.79	0.78	0.78		2569
weighted avg	0.77	0.77	0.77		2569

Figure 8: Confusion Matrix and Classification Report

Comparison of CNN Model Performance Before and After Policy-Based Threshold Selection







Under a standard argmax decision rule, the CNN exhibits strong overall discriminative performance but shows clinically problematic behavior for Tuberculosis (TB). Normal cases are classified correctly in 88.9% of instances (822/925), with most errors arising from confusion with mild disease: *Normal → Pneumonia (8.1%) and Normal → TB (3.0%)* which is expected given overlapping radiographic appearances. Pneumonia demonstrates robust detection, with stable recall and limited confusion with Normal.

The most critical issue arises for Tuberculosis. More than one-third of TB cases are misclassified as Normal, yielding a false-negative rate exceeding 36% and a sensitivity of only 63.4% under argmax classification. This reflects a strong bias toward predicting "Normal" in ambiguous TB cases, which is unacceptable in a screening context where missed TB diagnoses carry substantial risk.

Importantly, this limitation is not due to insufficient model capacity. Precision-recall analysis shows high TB average precision (AP = 0.921), with the precision-recall curve remaining strong until recall approaches approximately 0.6.

ROC analysis supports this interpretation, showing strong discrimination for TB and near-perfect separability for Pneumonia, with Normal remaining the most challenging class due to its heterogeneity.

Calibration analysis with reliability diagrams shows that the CNN's probability estimates are trustworthy at high confidence levels, supporting the use of probability-based, risk-aware decision rules rather than naive argmax classification.

These findings motivate the adoption of class-conditional thresholding aligned with clinical risk tolerance.

Accordingly, a risk-prioritized decision policy was introduced, assigning predictions hierarchically: Tuberculosis first, followed by Pneumonia, with Normal as the fallback class. Thresholds were selected using one-vs-rest sweeps on the validation set, enforcing high sensitivity for critical diseases:

- TB threshold: $p(\text{TB}) \geq 0.18$
- Pneumonia threshold: $p(\text{PN}) \geq 0.51$

This policy increased TB sensitivity from 63.4% to 90.2%, yielding a +26.8 percentage point gain, corresponding to 285 fewer missed TB cases and an approximate 73% reduction in TB false negatives.

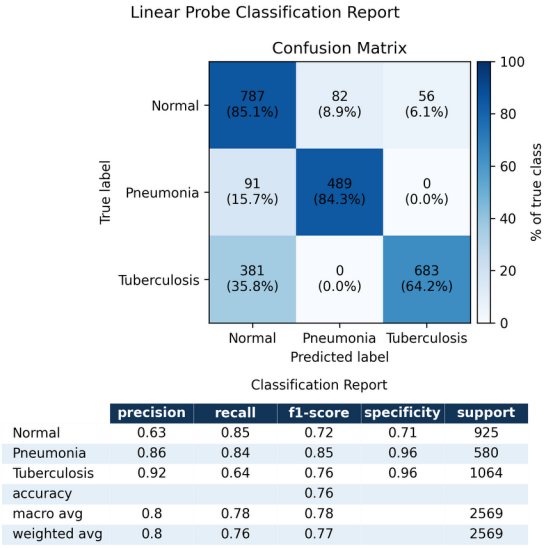
Pneumonia performance remained essentially unchanged (-0.6% recall), indicating robustness under the new policy. As expected, Normal recall decreased from 88.9% to 56.8%, reflecting an intentional trade-off.

Crucially, this trade-off is can be beneficial in a screening setting. Patients misclassified as diseased undergo confirmatory testing rather than immediate treatment, whereas missed TB cases

may lead to delayed diagnosis and continued transmission. Thus, accepting increased false positives among Normal cases is far safer than maintaining high false-negative rates for TB.

Overall, the CNN demonstrates strong discriminative performance across all classes (AUC: 0.88-0.99; AP: 0.78-0.96).
The primary limitation lies not in representation learning but in operating point selection. By replacing argmax classification with a policy-aware, risk-prioritized decision framework, the model's behavior is aligned with clinical priorities, substantially improving TB detection while preserving stable Pneumonia performance.

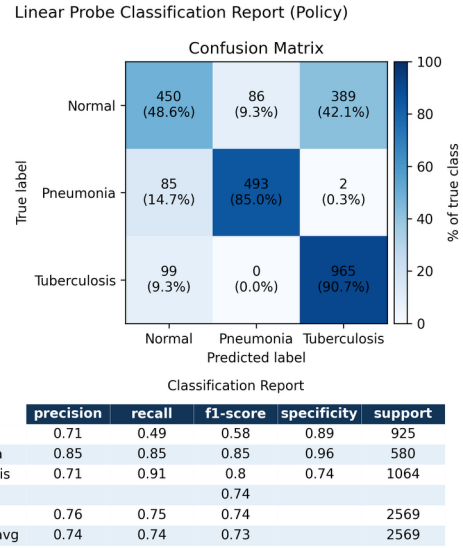
DenseNet 121 with Linear Probe Fine Tunning



Classification Report

	precision	recall	f1-score	specificity	support
Normal	0.63	0.85	0.72	0.71	925
Pneumonia	0.86	0.84	0.85	0.96	580
Tuberculosis	0.92	0.64	0.76	0.96	1064
accuracy			0.76		
macro avg	0.8	0.78	0.78		2569
weighted avg	0.8	0.76	0.77		2569

Figure 9: Confusion Matrix and Classification Report

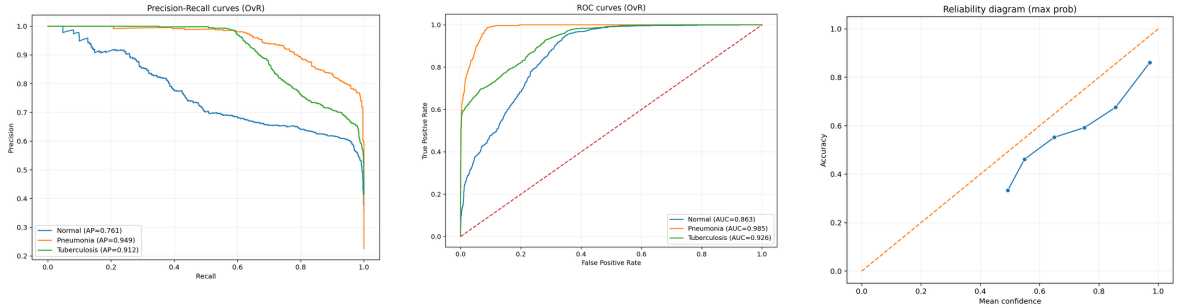


Classification Report

	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

Figure 10: Confusion Matrix and Classification Report

Comparison of Best Linear Probe Densenet 121 Model Performance Before and After Policy-Based Threshold Selection



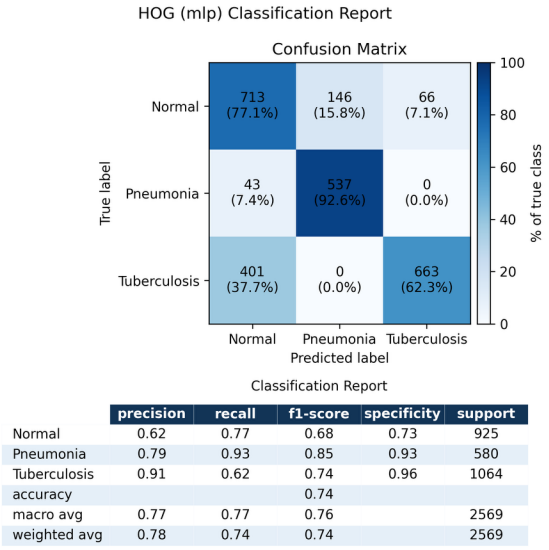
The DenseNet-121 linear probe exhibits slightly weaker performance than the end-to-end CNN, in Tuberculosis detection. While DenseNet produces high-confidence predictions when classifying TB correctly, it misses a substantial fraction of TB cases under a standard argmax decision rule. This behavior reflects a classic linear probing failure mode: although the frozen backbone encodes TB-relevant features, the linear classifier lacks sufficient capacity to separate subtle or atypical TB presentations from Normal cases.

Importantly, this limitation is not due to poor feature quality. Precision-recall analysis shows that TB cases are well ranked, indicating that discriminative information is present in the frozen representations. Rather, the issue lies in the operating point imposed by argmax classification, which favors precision over sensitivity and leads to excessive TB false negatives.

After applying class-specific thresholding within a risk-prioritized decision policy, TB false negatives were reduced from approximately 36% to 9%, substantially improving sensitivity at the expected cost of reduced precision. Under this calibrated policy, the DenseNet linear probe achieves clinically acceptable TB sensitivity, approaching the performance of the end-to-end CNN despite its limited classifier capacity.

These findings suggest that, while linear probes may underperform CNNs in capturing fine-grained decision boundaries, calibrated thresholding can partially compensate for this limitation, making frozen-backbone models viable for screening-oriented applications under appropriate decision policies.

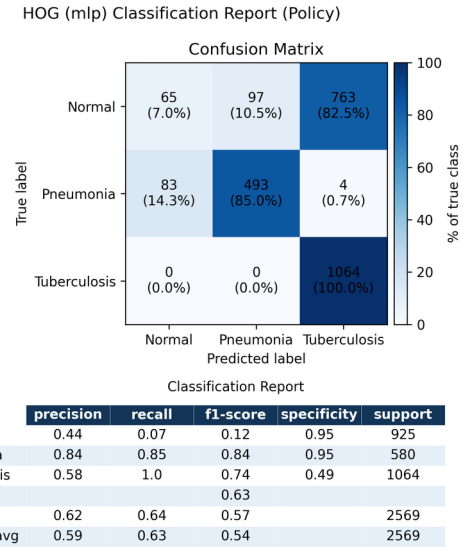
Histograms of Oriented Gradients + MLP



Classification Report

	precision	recall	f1-score	specificity	support
Normal	0.62	0.77	0.68	0.73	925
Pneumonia	0.79	0.93	0.85	0.93	580
Tuberculosis	0.91	0.62	0.74	0.96	1064
accuracy			0.74		
macro avg	0.77	0.77	0.76		2569
weighted avg	0.78	0.74	0.74		2569

Figure 11: Confusion Matrix and Classification Report

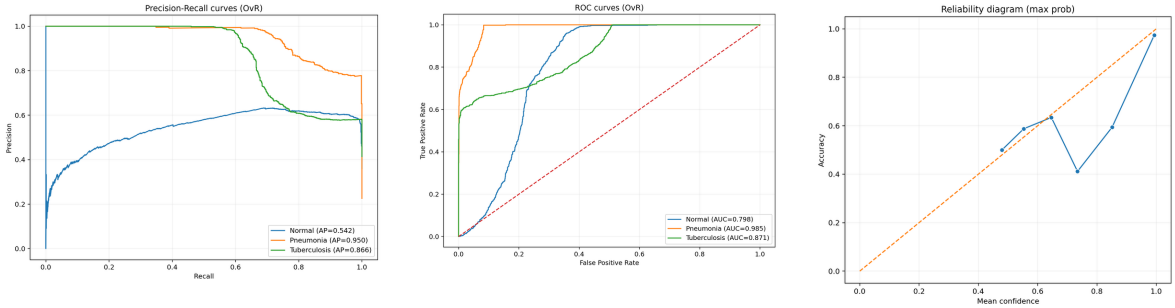


Classification Report

	precision	recall	f1-score	specificity	support
Normal	0.44	0.07	0.12	0.95	925
Pneumonia	0.84	0.85	0.84	0.95	580
Tuberculosis	0.58	1.0	0.74	0.49	1064
accuracy			0.63		
macro avg	0.62	0.64	0.57		2569
weighted avg	0.59	0.63	0.54		2569

Figure 12: Confusion Matrix and Classification Report

Comparison of HOG Model Performance Before and After Policy-Based Threshold Selection



From confusion matrix, we found HOG features capture coarse textural patterns that are sufficient for distinguishing Pneumonia but struggle to consistently separate Tuberculosis from Normal cases, where visual differences are more subtle and spatially diffuse.
Precision-recall analysis shows that the model ranks Pneumonia cases effectively, whereas Tuberculosis detection involves precision-recall trade-off and Normal-class confidence remains limited. Although HOG features provide useful ranking information, their predicted probabilities are poorly calibrated and should not be interpreted as reliable confidence estimates.
Under a Tuberculosis-prioritized decision policy, the HOG-based model achieves improved TB sensitivity at the expected cost of degraded Normal-class performance and overall accuracy.

Clinical Interpretation and Deployment Decision: *Tuberculosis-prioritized decision policy*

In summary, all models were evaluated using a policy-aware decision framework rather than naive argmax classification. Class-specific operating thresholds were selected via one-vs-rest (OvR) sweeps to enforce minimum sensitivity constraints for high-risk conditions, with predictions assigned hierarchically in the following priority order: Tuberculosis -> Pneumonia -> Normal. The objective was to maximize specificity subject to clinically defined sensitivity constraints, reflecting real-world screening pipelines where missed disease is substantially costlier than over-triage. Clinical operating constraints were defined as:

Tuberculosis: sensitivity ≥ 90% , **Pneumonia:** sensitivity ≥ 85%

Under the policy-aware framework, the CNN achieves the best overall balance between sensitivity, specificity, and calibration. It provides reliable probability estimates that support threshold-based decision-making, maintaining high sensitivity for Tuberculosis and Pneumonia while limiting unnecessary over-flagging of Normal cases. As a result, the CNN is the most deployment-ready model among those evaluated.

Linear probe models achieve clinically acceptable sensitivity for Tuberculosis and Pneumonia after threshold adjustment but at a higher operational cost, producing more false positives among Normal cases and increasing downstream workload. While suitable for screening in compute-constrained settings, they present less favorable trade-offs than the end-to-end CNN.

HOG-based models illustrate the limitations of classical pipelines under safety-critical constraints. Although Tuberculosis sensitivity can be enforced through aggressive thresholding, this leads to poor calibration, extremely low Normal recall, and excessive over-triage, making these models suitable primarily as baselines rather than for clinical deployment.

After reviewing results across all paradigms, the final deployed model was selected automatically based on policy-aware macro-F1 performance, with all subsequent analyses traced to the corresponding MLflow run identifier.
At this step, we automatically generate curated case lists that capture clinically relevant scenarios, including *high-risk errors*, *borderline cases near decision thresholds*, and *correct high-confidence* predictions. These structured outputs are then used as inputs for batch Grad-CAM visualization, enabling targeted inspection of model behavior on cases most relevant for safety, failure analysis, and clinical review.

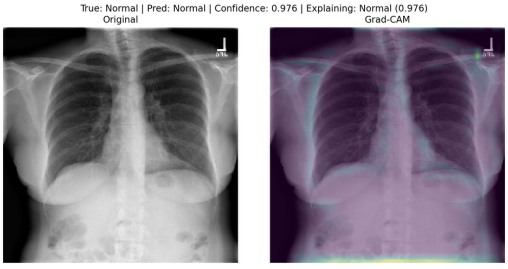
Task 4 — Model Explainability & Safety

For clinical adoption of AI systems, interpretability is essential to establish trust with clinicians and patients. In high-risk applications such as chest X-ray screening for tuberculosis and pneumonia, models must not only perform well numerically but also provide insight into why a prediction is made. Without transparency, even accurate models are difficult to validate, audit, or safely integrate into clinical workflows. Accordingly, this work treats explainability not as a tool to confirm correctness, but as an analytical mechanism to expose model limitations and inform safety-oriented system design.
We employ SmoothGrad-CAM++ to visualize spatial regions contributing to CNN predictions. Heatmaps are generated with respect to the target class at the final convolutional layer and overlaid on the original chest X-ray images [11]. These visualizations are interpreted as qualitative indicators of model attention rather than causal explanations.
The Grad-CAM pipeline additionally produces prioritized case lists for downstream radiologist review, with tuberculosis-related cases assigned the highest priority, followed by pneumonia. Explainability analysis surfaces several potential sources of issues, including subtle disease presentation, label noise, exposure variability, and preprocessing-related effects.

Grad-CAM heatmaps are generated in batch using curated case lists produced during error analysis. These case lists are automatically exported and organized by clinical risk and diagnostic ambiguity under:
reports/best_model_b01/cnn__cnn-30-flip-clahe/analysis/viz_cases/

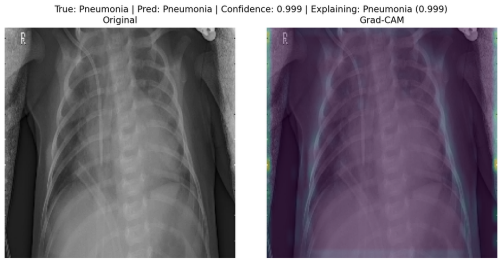
Observed Model Behavior on Correct Predictions

Normal Cases



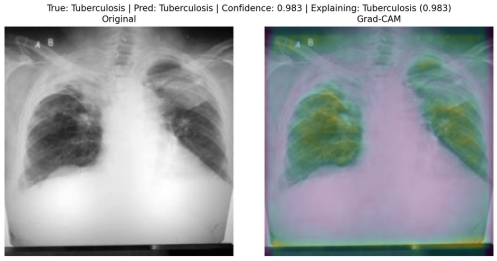
In correctly classified Normal images, Grad-CAM visualizations typically exhibit diffuse activation across the lung fields and strong attention to background regions and image borders is observed, suggesting partial reliance on global image characteristics or acquisition artifacts rather than exclusive focus on lung anatomy.

Pneumonia Cases



Correct pneumonia predictions generally produce more localized attention within the lungs, often concentrated in mid-to-lower zones. These regions correspond to plausible areas of consolidation. Nonetheless, activation is not always tightly confined to visually apparent opacities, indicating that the model may rely on coarse texture cues rather than precise lesion localization.

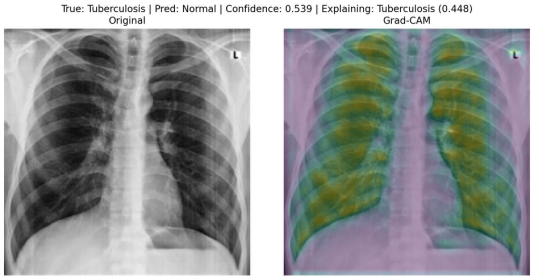
Tuberculosis Cases



Correct tuberculosis predictions show broad, bilateral lung activation, frequently involving upper lung regions. While this pattern is consistent with known TB radiographic presentations, the resulting Grad-CAM maps are often diffuse and low resolution. This limits direct association with specific pathological findings such as cavitation or nodularity, suggesting that the CNN predominantly relies on global texture changes rather than explicit structural abnormalities.

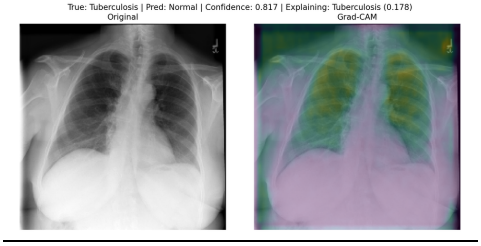
Failure Mode Analysis

Subtle Tuberculosis False Negatives (Borderline)



In tuberculosis cases misclassified as Normal with borderline prediction probabilities (gradcam_tb_fn_true), These errors occur when Tuberculosis presents with weak, diffuse, or low-contrast patterns that overlap with normal lung texture. Grad-CAM shows broad, non-focal attention across the lungs rather than a strong localized signal, indicating insufficient salience under argmax classification. Such cases fall near the decision threshold and are easily misclassified as Normal, reflecting the known difficulty of detecting early or subtle TB on chest X-rays and motivating sensitivity-prioritized thresholding.

High-Confidence Tuberculosis Misses



In this case, the model predicts Normal with high confidence, yet Grad-CAM for Tuberculosis shows distributed activation over both upper lung fields, a region commonly associated with TB involvement. The TB probability lies just below the decision threshold, indicating that TB-relevant features are detected but not strong enough to override the dominant Normal signal. This reflects a failure mode where subtle, bilateral TB patterns are underweighted relative to global normal appearance, reinforcing the need for sensitivity-prioritized thresholds and human review for borderline cases.

Evidence of Dataset Bias and Shortcut Learning

Grad-CAM analysis occasionally reveals activation outside anatomical lung regions, including image borders, film edges, and acquisition markers. While lung parenchyma remains the primary focus in most cases, these patterns indicate potential shortcut learning, where the model may partially rely on non-biological cues such as radiographic markers, borders, or hardware artifacts rather than pathology alone. These artifacts may influence model confidence, particularly in borderline cases, and may reduce generalizability to images acquired under different conditions.

Rather than undermining the system, these findings motivate a cautious*, policy-driven deployment strategy* and highlight opportunities for improvement, including enhanced data quality, higher-resolution modeling, and more explicit uncertainty estimation.

Task 5 — Deployment Strategy (Production-ready)

Packaging, Versioning, and Reproducible Deployment

Details workflow can be found at : docs/workflow-guide.md
This project uses a traceable packaging chain from experiment tracking to deployment:
MLflow -> model bundle -> Docker images (CLI inference + FastAPI + web UI).
All training and evaluation runs are logged in MLflow, enabling reproducible selection of the best model based on validation metrics and providing a stable run identifier that links artifacts, configuration, and results.
After selecting the best run, the model is exported as a self-contained bundle (make bundle-best). The bundle locks the deployment contract by packaging the model weights, model/config metadata, class names, preprocessing metadata, policy thresholds, and run provenance into a single directory. A stable symlink (bundles/latest) is maintained to avoid timestamp-dependent paths and to provide a consistent reference point for downstream builds. This ensures that deployment always uses an explicitly versioned artifact rather than an ad-hoc checkpoint.
Deployment is containerized to guarantee environment reproducibility and portability. Two Docker deliverables are produced:

1. Inference-only container (CLI): built with a lightweight dependency set (requirements.infer.txt) and a bundled model artifact (make docker-build-infer). This produces a slim, self-contained image that can run inference without access to MLflow or local Python environments.
   The baked-image smoke test (make docker-smoke-baked) validates bundle loading and JSON output inside Docker.
2. FastAPI + Web UI container: built with the same bundled model artifact and exposes prediction and Grad-CAM endpoints alongside a browser-based UI (make docker-build-api, make docker-run-api).

This single-image deployment provides an end-to-end workflow:
upload an image -> run inference, -> visualize both predictions and Grad-CAM without requiring additional web servers.

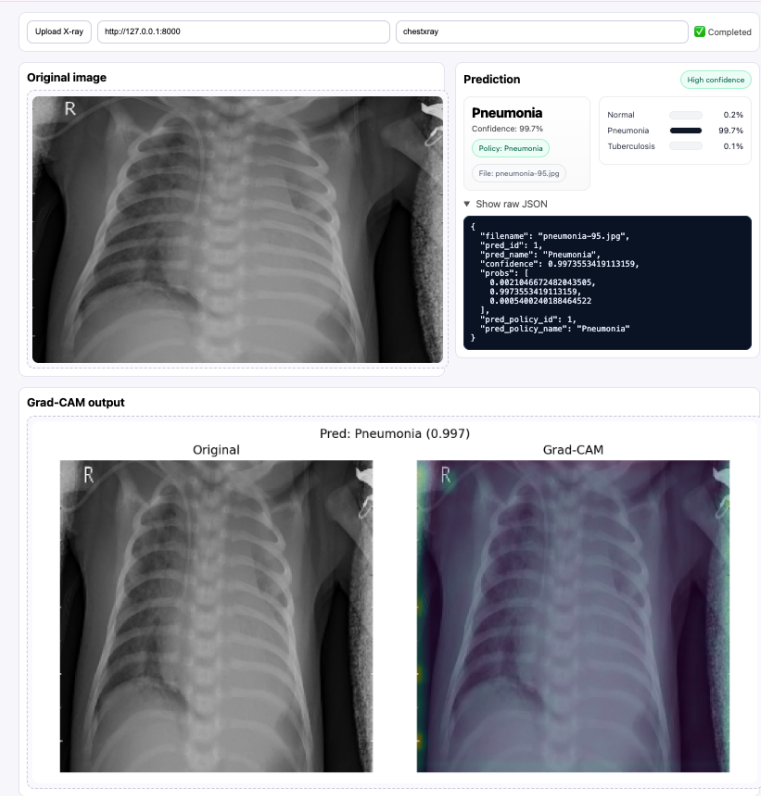


Figure 13: UserInterface for Inference System

All operational workflows are standardized through Makefile entry points, which act as a single interface for training, evaluation, bundling, Docker builds, and API deployment. This reduces configuration drift and ensures that the same commands can be executed consistently across machines and environments. A lightweight CI gate (make ci-smoke) provides a deployment readiness check by validating that the pipeline is reproducible, fully bundled, and Docker-compatible.
For distribution and remote deployment, the FastAPI image can be exported as a portable archive (make docker-api-save) and loaded on a target machine using standard Docker tooling. The resulting deployment is self-contained, does not depend on MLflow at runtime, and can be started with a single entry command (docker run ... --restart unless-stopped ...), making it suitable for controlled server environments. All development was performed on a Mac PC. The Docker image has been tested on macOS, Windows, and cloud environments, where it runs successfully.

Monitoring and Model Oversight

Model monitoring is supported through the existing training-evaluation-analysis pipeline, which is designed to surface performance drift, calibration issues, and clinically relevant failure modes prior to deployment. Model selection and ranking are performed automatically using MLflow, ensuring that changes in model performance across retraining cycles are traceable and reproducible.
Post-training analysis explicitly generates *policy-aware evaluation* artifacts, including operating-point thresholds, calibration diagnostics, and structured case lists. These case lists categorize

predictions into clinically meaningful groups such as high-risk error cases (e.g., Tuberculosis false negatives), borderline cases near decision thresholds, and high-confidence correct predictions. This enables targeted inspection of edge cases and supports periodic review of model behavior as new data become available.

While no continuous online monitoring is implemented in the deployed API, the workflow supports offline monitoring and re-validation through repeated execution of eval-best and analyze-best, allowing updated models to be compared against prior versions under identical decision policies.

The deployment strategy explicitly assumes human-in-the-loop oversight for safety-critical decisions. High-risk and ambiguous cases identified during analysis are exported as structured artifacts and prioritized for radiologist review, with Tuberculosis-related cases assigned highest priority. In the deployed system, all positive or borderline predictions are intended for confirmatory clinical assessment rather than autonomous diagnosis, aligning with real-world screening workflows.

Risk and Compliance Considerations

The system is designed as a screening decision-support tool, not a diagnostic device. Predictions are probabilistic, policy-aware, and conservative by design, prioritizing reduced false negatives for severe disease at the expense of increased false positives. This trade-off is explicitly documented and clinically motivated.

The deployment pipeline avoids handling patient-identifiable metadata; all inputs are image files without embedded EXIF or DICOM headers. Access to the inference API is restricted via a simple token-based authentication mechanism, providing basic access control suitable for controlled environments. Model artifacts are versioned and traceable through MLflow run identifiers and bundled metadata, supporting auditability and reproducibility.

Overall, the existing pipeline provides structured safeguards for monitoring, review, and traceability, while intentionally relying on human oversight for edge cases and final clinical decision-making, consistent with responsible deployment of machine learning in medical imaging.

3. Conclusion

Impact and Implications

The proposed system demonstrates how machine learning can augment clinical screening workflows to support radiologists and improve patient outcomes. Rather than operating as an autonomous diagnostic tool, the model is designed to augment clinician decision making by prioritizing high-risk cases and highlighting uncertain findings that require human review.

Through calibrated probability estimates, risk-aware decision policies, and visual explanations, the system helps clinicians focus attention on cases most likely to require urgent assessment, particularly for Tuberculosis and Pneumonia where delayed or missed diagnoses carry significant risk. In high-throughput screening environments, this approach has the potential to accelerate triage, improve consistency in case prioritization, and reduce time to specialist evaluation.

By reducing missed cases of severe disease and explicitly surfacing borderline predictions for targeted review, the workflow supports earlier confirmatory testing and intervention. The emphasis

on reproducibility, transparent decision logic, and human-in-the-loop oversight reinforces AI's role in augmenting clinical practice while preserving clinician autonomy and accountability.

Strengths

This work presents a complete and reproducible end-to-end pipeline for chest X-ray classification, spanning exploratory data analysis, model training, policy-aware evaluation, explainability, and deployment. A key strength lies in the explicit alignment of model evaluation with clinical risk priorities through class-specific thresholding and a hierarchical decision policy, which substantially reduces false negatives for Tuberculosis and Pneumonia. The systematic comparison of deep learning models, linear probing strategies, and classical feature-based approaches provides insight into performance trade-offs under realistic computational constraints. In addition, the use of MLflow-based experiment tracking, versioned model bundling, Dockerized inference, and a FastAPI-based web interface demonstrates strong attention to reproducibility and deployment readiness with human-in-the-loop review.

Limitations

Several limitations should be acknowledged. Model development and experimentation were constrained by limited computational resources, which restricted batch sizes, training duration, and the feasibility of more advanced approaches such as large-scale self-supervised pretraining or full fine-tuning of deep backbones. Time constraints also limited the depth of robustness testing and long-term evaluation under distribution shift. The deployed API currently supports single-image inference only, which may limit throughput in real-world screening environments. Additionally, explainability analysis using Grad-CAM is qualitative and exploratory in nature and does not provide formal guarantees of causal attribution or clinical correctness.

Future Improvements

Future work could leverage larger computational budgets to enable higher-resolution training, full fine-tuning of pretrained models, and self-supervised or semi-supervised learning to improve sensitivity to subtle disease patterns. Robustness could be further evaluated using external datasets or simulated acquisition shifts to better characterize performance under distribution change. From a deployment perspective, extending the API and user interface to support batch inference, asynchronous processing, and structured clinician feedback would improve usability. Incorporating radiologist feedback into iterative retraining and recalibration cycles would further strengthen the human-in-the-loop framework.

Discussion

Recent advances in large foundation models have demonstrated significant promise in medical imaging, particularly through improved representation learning and robustness across diverse datasets. However, such models typically require substantial computational resources and large-scale annotated data.

In contrast, the present work was conducted under practical constraints of limited computational budget and dataset size, reflecting conditions commonly encountered in applied clinical research. While the resulting model does not achieve state-of-the-art performance, the study provides a transparent characterization of its strengths and limitations. More importantly, it demonstrates how

explainability-driven analysis can be used to identify high-risk failure modes and to design safety-oriented deployment strategies.
This work therefore contributes not by claiming maximal performance, but by illustrating a pragmatic and responsible approach to developing and deploying deep learning models for high-risk clinical tasks, where understanding uncertainty and failure is as critical as achieving accuracy.

Conflict of Interest

The author declares no conflicts of interest related to this work.

References

1. Muhammad Rehan & Kaggle. Chest X Ray Dataset. *https://www.kaggle.com/datasets/muhammadrehan00/chest-xray-dataset*.
2. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. in *37th International Conference on Machine Learning, ICML 2020* vols PartF168147-3 (2020).
3. Drexlin, D. J. *et al.* MeDi: Metadata-Guided Diffusion Models for Mitigating Biases in Tumor Classification. in *Lecture Notes in Computer Science* vol. 15973 LNCS (2026).
4. Saxena, A. An Introduction to Convolutional Neural Networks. *Int. J. Res. Appl. Sci. Eng. Technol.* 10, (2022).
5. Bohlol, P., Hosseinpour, S. & Soltani Firouz, M. Improved food recognition using a refined ResNet50 architecture with improved fully connected layers. *Curr. Res. Food Sci.* 10, (2025).
6. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* vols 2017-January (2017).
7. Tan, M. & Le, Q. V. EfficientNet: Rethinking model scaling for convolutional neural networks. in *36th International Conference on Machine Learning, ICML 2019* vols 2019-June (2019).
8. Liu, Z. *et al.* Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. in *Proceedings of the IEEE International Conference on Computer Vision* (2021). doi:10.1109/ICCV48922.2021.00986.
9. Deng, J. *et al.* ImageNet: A Large-Scale Hierarchical Image Database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009* (2009). doi:10.1109/CVPR.2009.5206848.
10. MLFLow.
11. Azfar, M., Bharadwaj, S. & Sasikumar, A. Improving Smooth GradCAM++ with Gradient Weighting Techniques. in *2024 IEEE 21st India Council International Conference, INDICON 2024* (2024). doi:10.1109/INDICON63790.2024.10958549.