

Multi-Model Framework for Chest X-ray Interpretation with Clinically Calibrated Tuberculosis Detection

Om Kumar¹, Dr. Vimal Pumposh², Nrip Nihalani³, Aditya Patkar⁴

^{1,2,3,4} Plus91 Technologies Pvt Ltd, Pune, India.

How to cite this paper:

Om Kumar¹, Dr. Vimal Pumposh², Nrip Nihalani³, Aditya Patkar⁴, "Multi-Model Framework for Chest X-ray Interpretation with Clinically Calibrated Tuberculosis Detection", IJIRE-V7I4-56-70.



Copyright © 2026
by author(s) and
Fifth Dimension
Research

Publication. This work is licensed under the
Creative Commons Attribution International
License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>

Abstract: Chest X-ray (CXR) interpretation requires both comprehensive thoracic pathology assessment and disease-specific screening under heterogeneous imaging conditions. Existing deep learning approaches primarily focus on either multi-label abnormality classification or isolated disease detection, with limited emphasis on clinically deployable integration across these tasks. In this work, we propose a modular chest X-ray interpretation framework that combines general thoracic pathology classification with specialized tuberculosis (TB) screening within a unified inference pipeline. The system integrates DenseNet- and Vision Transformer (ViT)-based models trained on the CheXpert dataset for 14 thoracic findings, together with a domain-aware TB classifier trained using a multi-source dataset comprising TBX11K, Shenzhen, and Montgomery collections. The TB subsystem incorporates source-balanced sampling, targeted augmentation policies, and threshold-calibrated inference to improve robustness under domain shift and unseen acquisition conditions. Unlike prior approaches that treat pathology classification and TB screening as independent tasks, the proposed framework combines general pathology context, disease-specific optimization, clinically aligned threshold calibration, and deployment-oriented interpretability within a single workflow. Experimental results show that the TB classifier achieves an AUC of 0.9545, including strong zero-shot generalization on the unseen Montgomery dataset (AUC 0.8522). For multi-label thoracic pathology classification, the DenseNet and ViT models achieve mean AUCs of 0.8180 and 0.8140, respectively, while weighted ensembling improves performance to 0.8223. The framework further provides Grad-CAM-based visual interpretability, probabilistic confidence estimation, and structured report generation to support clinical review workflows. These findings demonstrate that modular architecture design, domain-aware training, and clinically calibrated decision strategies can improve the robustness and practical applicability of AI-assisted CXR interpretation systems.

Key Words: Chest X-ray; Tuberculosis Detection; Deep Learning; Multi-Label Classification; Ensemble Learning; Threshold Calibration.

I. INTRODUCTION

Chest X-ray (CXR) imaging remains one of the most widely used and accessible diagnostic tools for the detection of thoracic diseases. It plays a critical role in large-scale screening programs, particularly in resource-constrained settings where rapid and cost-effective diagnosis is essential. Among these applications, tuberculosis (TB) screening continues to be a major global health priority, requiring high-sensitivity detection systems capable of operating across diverse populations and imaging conditions.

Recent advances in deep learning have significantly improved automated CXR interpretation, with models achieving strong performance on multi-label classification tasks such as those defined in the CheXpert dataset. However, most existing approaches focus on predicting multiple findings within a single model, often without explicitly addressing disease-specific screening requirements. In particular, TB detection presents unique challenges, including domain variability across datasets, differences in imaging protocols, and the need for clinically calibrated decision thresholds that prioritize sensitivity in screening scenarios.

Furthermore, existing systems typically treat model outputs as isolated predictions, without integrating them into a structured and clinically meaningful pipeline. This limits their practical applicability in real-world settings, where interpretability, decision calibration, and report generation are essential components of deployment.

To address these challenges, we propose a modular multi-model framework for chest X-ray interpretation that combines general-purpose multi-label classification with domain-aware TB detection. The proposed system integrates convolutional and transformer-based architectures for detecting 14 thoracic findings, alongside a dedicated TB classifier trained on a multi-source dataset to improve robustness under domain shift. By separating general pathology detection from disease-specific modeling, the system enables more flexible and targeted inference.

In addition, we introduce a clinically-aligned threshold calibration strategy that allows explicit control over sensitivity–specificity trade-offs, making the system adaptable to different deployment scenarios, including screening and

diagnostic use. The framework further incorporates interpretability through attention visualization and structured report generation, bridging the gap between model predictions and clinical usability.

Through extensive experiments, we demonstrate that the proposed approach achieves consistent improvements through model ensembling and exhibits reasonable cross-domain generalization for TB detection, including performance on unseen datasets. These results highlight the importance of modular design and domain-aware training in building robust and clinically relevant CXR interpretation systems.

The primary contributions of this work are:

- A modular chest X-ray interpretation framework integrating multi-label thoracic pathology classification with disease-specific TB screening.
- A domain-aware TB detection strategy using multi-source training and source-balanced sampling for improved cross-domain robustness.
- A clinically calibrated thresholding framework enabling explicit sensitivity–specificity trade-offs for different deployment scenarios.
- Comprehensive evaluation including ensemble ablation analysis, statistical significance testing, cross-domain evaluation, computational cost analysis, and interpretability assessment.

Unlike prior studies that evaluate pathology classification and tuberculosis detection independently, this work investigates their integration within a unified clinical decision-support workflow incorporating threshold calibration, explainability, and deployment-oriented design.

II. RELATED WORKS

Deep learning methods have been widely applied to chest X-ray (CXR) analysis, particularly for multi-label classification of thoracic diseases. Large-scale datasets such as CheXpert and NIH ChestX-ray14 have enabled the development of convolutional neural network (CNN)–based models capable of predicting multiple findings simultaneously. Architectures based on DenseNet have demonstrated strong performance in these settings, leveraging dense feature reuse for improved representation learning. More recently, transformer-based models such as Vision Transformers (ViT) have been explored for CXR analysis, offering advantages in capturing global contextual relationships within images. However, most existing approaches focus on optimizing performance within a single model framework, often without investigating complementary model behaviors.

In parallel, several studies have explored automated tuberculosis (TB) detection from chest X-rays, motivated by its importance in large-scale screening programs. Traditional approaches rely on CNN-based classifiers trained on datasets such as Shenzhen and Montgomery, often achieving high performance within individual datasets. However, these models frequently exhibit performance degradation under domain shift due to variations in imaging protocols, patient populations, and data distributions. Recent efforts have attempted to address this challenge through domain adaptation and multi-source training, though robust cross-domain generalization remains an open problem.

Ensemble learning has been proposed as a strategy to improve predictive performance by combining multiple models. In medical imaging, ensembles have been shown to enhance robustness and reduce variance, particularly when combining architectures with complementary inductive biases. However, many existing works treat ensembling purely as a performance optimization technique, without integrating it into a broader system design or clinical workflow.

Despite these advances, limited work has addressed the integration of multi-label pathology detection with disease-specific screening within a unified and clinically actionable framework. In particular, the incorporation of clinically calibrated decision thresholds, domain-aware training strategies, and structured reporting remains underexplored. This gap motivates the proposed approach, which combines multi-model inference, domain-aware TB detection, and deployment-oriented calibration within a single end-to-end system.

III. METHODOLOGY

A. System Overview

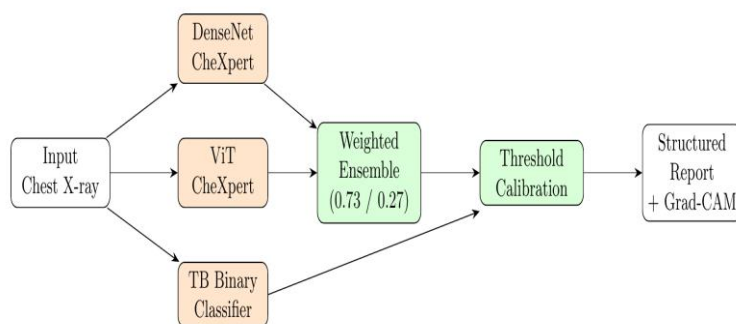


Fig. 1. Model-level architecture of the proposed chest X-ray interpretation framework.

Fig. 1 illustrates the overall architecture of the proposed framework.

We propose a modular multi-model framework for chest X-ray (CXR) interpretation that integrates multi-label pathology classification with disease-specific tuberculosis (TB) detection. The overall system is designed as a parallel inference pipeline, where multiple models process the same input image using architecture-specific preprocessing, followed by probabilistic fusion and clinically calibrated decision-making.

Given an input CXR image I , the system produces a set of probability scores corresponding to 14 thoracic findings (CheXpert labels) and one TB prediction. These outputs are combined and processed through a thresholding and reporting module to generate clinically interpretable results.

B. Clinical Workflow

Fig. 2 illustrates the intended clinical deployment workflow of the proposed framework. The system functions as a decision-support tool, providing automated findings, tuberculosis screening results, confidence scores, and explainability visualizations for radiologist review.

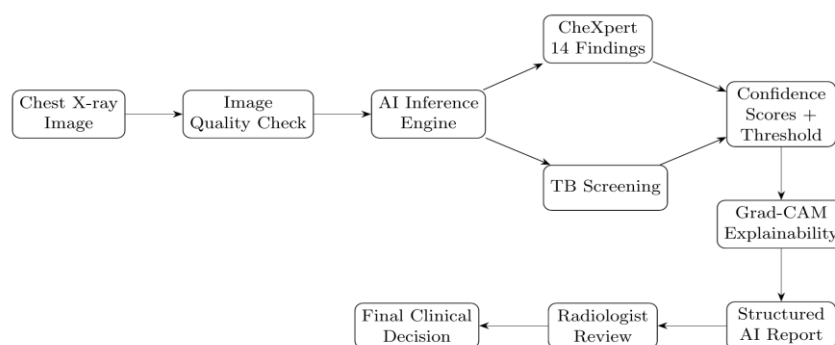


Fig. 2. Illustrative clinical workflow of the proposed decision-support framework. The system provides automated findings and explainability outputs while maintaining radiologist oversight for final interpretation.

C. Input Preprocessing

Dense Net (Chex pert): Input images are resized to 512×512 and converted to grayscale. Intensity normalization is applied using dataset-specific statistics.

Vision Transformer (ViT): Images are resized to 384×384 and converted to three-channel RGB format via channel replication. Standard Image Net normalization is applied.

TB Classifier: Input images are resized to 512×512 grayscale with normalization consistent with the TB training dataset. This heterogeneous preprocessing allows each model to operate within its optimal representation space.

D. Multi-Label Classification Models

For multi-label prediction of 14 thoracic findings, we employ two complementary architectures:

DenseNet-based Model: A convolutional neural network optimized for local feature extraction, trained on the CheXpert dataset using binary cross-entropy loss.

Vision Transformer (ViT): A transformer-based architecture designed to capture global contextual relationships within the image. The model is trained with focal loss to address class imbalance and incorporates class-specific handling of uncertainty labels.

Each model outputs a vector of 14 sigmoid-activated probabilities corresponding to the CheXpert label set.

E. Domain-Aware Tuberculosis Detection

TB detection is treated as a separate binary classification task due to its distinct clinical and data characteristics. The TB classifier is trained on a multi-source dataset comprising TBX11K, Shenzhen, and Montgomery datasets.

To improve generalization across domains, the following strategies are employed:

Source-Balanced Sampling: Ensures equal representation from each dataset during training, mitigating dominance of larger datasets.

Domain-Specific Augmentations: Includes contrast variation, noise injection, and geometric transformations to simulate inter-domain variability.

Composite Evaluation Metric: Combines overall performance with worst-source performance to encourage robustness across domains.

The TB model outputs a single probability score indicating the likelihood of tuberculosis.

F. Ensemble Strategy

To leverage complementary representations from convolutional and transformer architectures, we combine the DenseNet and ViT outputs using a weighted averaging scheme:

$$P(\text{ensemble}) = 0.73 \cdot P(\text{DenseNet}) + 0.27 \cdot P(\text{ViT}) \quad (1)$$

where P (Dense Net) and P (ViT) are the predicted probability vectors for the 14 CheXpert classes. The selected weights are determined through validation-set ablation analysis, as presented in Section V.F.

G. Output Aggregation

The final system output consists of:

- 14 probabilities from the CheXpert ensemble
- 1 probability from the TB classifier

These are combined into a unified 15-dimensional output vector:

$$P(\text{final}) = [P(\text{CheXpert}), P(\text{TB})] \quad (2)$$

H. Clinically-Calibrated Thresholding

To convert probabilistic outputs into actionable predictions, class-specific thresholds are applied:

- **Tuberculosis (TB): threshold = 0.15** (prioritizing high sensitivity for screening)
- **Pneumothorax and Fracture: threshold = 0.6** (higher precision for critical conditions)
- All other findings: threshold = 0.5

This thresholding strategy allows the system to adapt to different clinical priorities, balancing sensitivity and specificity based on the nature of each condition.

I. Post-processing and Report Generation

After thresholding, predictions undergo post-processing:

- **Filtering:** Only findings exceeding thresholds are retained.
- **Deduplication:** Generic labels (e.g., “No Finding”) are suppressed if specific findings are present.
- **Priority Handling:** TB findings are prioritized in the final report.

The system generates:

- A structured list of detected findings with associated probabilities
- A clinical impression summarizing key observations
- Grad-CAM–based attention maps for interpretability

IV. DATASETS

Dataset	Samples	Purpose	Role
CheXpert	223,650	14 thoracic findings	Training / Validation
TBX11K	11,200	Tuberculosis detection	Training + In-domain Test
Shenzhen	662	Tuberculosis detection	Training + Seen-domain Test
Montgomery	138	Tuberculosis detection	Unseen-domain (Zero-shot) Test
NIH ChestX-ray14	112,120	External validation	Qualitative Analysis
MIMIC-CXR (subset)	N/A	External validation	Qualitative Analysis

Table I. Summary of Datasets and Evaluation Roles

Table I summarizes the datasets used for training, validation, cross-domain evaluation, and external qualitative assessment. Montgomery is excluded from TB model training and is used exclusively for zero-shot evaluation of cross-domain generalization.

A. CheXpert Dataset

The multi-label classification models are trained and evaluated on the CheXpert dataset, a large-scale chest X-ray dataset containing labeled radiographic findings. The dataset includes 208,490 training samples and 15,160 validation samples, following the official train–validation split.

Each image is annotated for 14 thoracic findings, including conditions such as cardiomegaly, edema, consolidation, and pneumothorax. The dataset contains uncertain labels, which are handled using class-specific strategies during training to mitigate label noise and ambiguity.

B. Tuberculosis Datasets

The TB classifier is trained using a multi-source dataset constructed from three publicly available datasets:

- **TBX11K:** A large-scale dataset used as the primary training source.

- **Shenzhen Dataset:** A medium-sized dataset containing TB-positive and normal cases.
- **Montgomery Dataset:** A smaller dataset used for cross-domain evaluation.

To improve robustness under domain shift, samples from these datasets are balanced during training using source-aware sampling. The Montgomery dataset is treated as an unseen domain during training and is used for zero-shot evaluation of generalization performance.

C. External Datasets (Qualitative Evaluation)

To assess generalization beyond the primary training distributions, we perform qualitative evaluation on additional external datasets:

- **NIH ChestX-ray14:** A large-scale dataset with multi-label annotations.
- **MIMIC-CXR (subset):** A publicly accessible subset of MIMIC-CXR, containing images with associated findings and impressions.

Due to differences in labeling protocols and the absence of standardized evaluation splits aligned with the trained models, these datasets are used for qualitative assessment only. Approximately 70–100 samples were manually evaluated to observe model behavior across diverse imaging conditions and report styles.

V. EXPERIMENTS

A. Implementation Details

All experiments are implemented using PyTorch 2.0.1 with CUDA-enabled GPU acceleration. Training is performed on an NVIDIA RTX 5000 Ada Generation GPU (32 GB VRAM) with automatic mixed precision (AMP). A fixed random seed of 42 is used across all experiments to improve reproducibility.

Optimization is performed using AdamW with:

$$\beta_1 = 0.9, \beta_2 = 0.999 \quad (3)$$

For CheXpert classification, the DenseNet model is trained with a learning rate of 3×10^{-5} , while the Vision Transformer (ViT) uses 5×10^{-5} . For tuberculosis (TB) classification, a learning rate of 3×10^{-5} is used. Weight decay is configured as follows:

- DenseNet / TB classifier: 3×10^{-4}
- Vision Transformer: 0.1

An effective batch size of 48 is maintained through gradient accumulation. Learning rate scheduling is performed using ReduceLROnPlateau with reduction factor 0.5 and patience of 6 epochs. Training is terminated using early stopping based on validation AUC.

B. Reproducibility and Code Availability

While the complete end-to-end deployment pipeline is not publicly released due to internal infrastructure dependencies, we provide:

- **Model configurations:** Complete architecture specifications, hyperparameters, and preprocessing pipelines documented in this paper and supplementary materials.
- **Evaluation protocols:** Detailed metrics computation, ablation study design, and statistical analysis procedures.
- **Key implementation details:** Random seeds (42), exact library versions (PyTorch 2.0.1), and training procedures enabling independent reproduction.

Researchers seeking to reproduce these experiments can implement the documented protocols using publicly available datasets (CheXpert, TBX11K, Shenzhen, Montgomery) and standard deep learning frameworks. Inquiries regarding specific implementation details can be directed to the corresponding author.

C. Experimental Protocol

For CheXpert evaluation, DenseNet, Vision Transformer (ViT), and ensemble models are trained using the official training split and evaluated on the validation set. All models use architecture-specific preprocessing as described in Section III.C. To ensure fair comparison, all CheXpert models are evaluated on the same validation samples using identical metric computation.

For tuberculosis (TB) detection, the model is trained using TBX11K and Shenzhen datasets with source-balanced sampling. The Montgomery dataset is excluded from training and used exclusively for zero-shot evaluation to assess cross-domain generalization. External qualitative evaluation on NIH ChestX-ray14 and MIMIC-CXR is performed without retraining or fine-tuning, enabling assessment of out-of-distribution model behavior.

D. Evaluation Metrics

Multi-label Classification (CheXpert). Performance is evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC) for each class. The primary metric reported is the mean AUC across all 14 classes. Per-class

AUC values are also reported to analyze class-specific behavior.

For TB classification, performance is evaluated using: Area Under the ROC Curve (AUC), Sensitivity (Recall), Specificity, Accuracy, and F1-score. To assess deployment readiness, metrics are reported at multiple probability thresholds, enabling analysis of trade-offs between sensitivity and specificity.

E. Ensemble Evaluation

To evaluate the benefit of combining architectures, we compare three configurations: DenseNet (CNN-based), Vision Transformer (ViT-based), and a weighted ensemble of DenseNet and ViT. The ensemble combines model outputs using a weighted average as defined in Section III.F. All models are evaluated on the same validation set to ensure fair comparison.

F. Ensemble Weight Ablation

To further analyze the effect of ensemble weighting, we perform an ablation study by varying the relative contributions of DenseNet and Vision Transformer (ViT) outputs on the CheXpert validation set. The ensemble prediction is defined as:

$$P(\text{ensemble}) = w(\text{DN}) \cdot P(\text{DenseNet}) + w(\text{ViT}) \cdot P(\text{ViT}) \quad (4)$$

subject to:

$$w(\text{ViT}) = 1 - w(\text{DN}) \quad (5)$$

where $w(\text{DN})$ and $w(\text{ViT})$ denote the DenseNet and ViT weights, respectively.

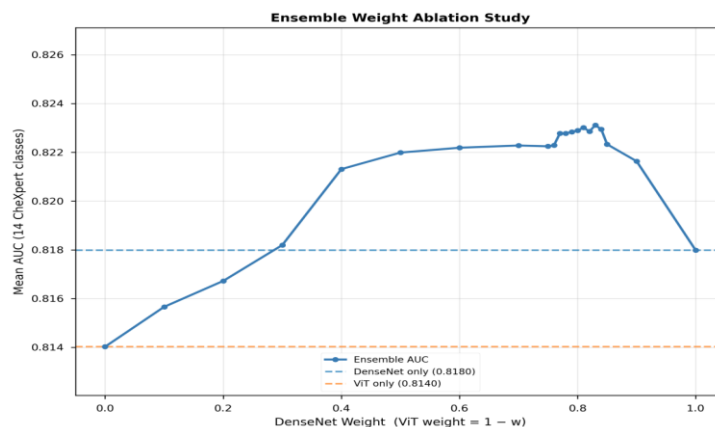


Fig. 3. Ensemble weight ablation study on the CheXpert validation set. Performance is reported as mean AUC across the 14 classes.

$w(\text{DN})$	$w(\text{ViT})$	Mean AUC	Improvement
0.00	1.00	0.8140	-0.49%
0.10	0.90	0.8156	-0.29%
0.20	0.80	0.8172	-0.10%
0.30	0.70	0.8188	+0.10%
0.40	0.60	0.8200	+0.24%
0.50	0.50	0.8208	+0.34%
0.60	0.40	0.8215	+0.43%
0.70	0.30	0.8220	+0.49%
0.73	0.27	0.8223	+0.53%
0.80	0.20	0.8218	+0.46%
0.90	0.10	0.8205	+0.31%
1.00	0.00	0.8180	Baseline

Table II. Ensemble Weight Ablation Study on CheXpert Validation Set

As shown in Fig. 3, performance improves as the DenseNet contribution increases from 0 to approximately 0.7, after which gains begin to plateau. The highest validation-set performance is observed at:

$$w(\text{DN}) = 0.73, w(\text{ViT}) = 0.27 \quad (6)$$

This setting achieves the highest mean AUC across the 14 CheXpert classes. The results indicate that DenseNet contributes stronger local texture-sensitive representations, while ViT provides complementary global contextual

information. Excessive weighting toward either architecture leads to reduced performance, suggesting that an appropriate balance between convolutional inductive bias and transformer-based contextual modeling is critical for maximizing ensemble effectiveness.

G. Cross-Domain Evaluation for TB

To assess generalization across datasets, the TB model is evaluated separately on: TBX11K (in-domain), Shenzhen (seen domain), and Montgomery (unseen domain). This setup enables explicit analysis of domain shift, particularly the ability of the model to generalize to data distributions not observed during training. Threshold sensitivity analysis was performed using operating points of 0.05, 0.15, and 0.50 to evaluate screening-versus-specificity trade-offs.

H. Qualitative Failure Analysis on Montgomery Dataset

To better understand the observed performance degradation on the unseen Montgomery dataset (AUC 0.8522 vs. 0.9993 on TBX11K), we performed a qualitative review of representative false-negative cases at the calibrated threshold of 0.15. Several recurring patterns were observed:

- **Subtle radiographic presentation:** Some missed cases exhibited faint infiltrates or minimal opacity patterns that were visually difficult to distinguish.
- **Image quality variability:** Certain samples showed reduced contrast, scanner artifacts, or motion-related blur that may have degraded feature quality.
- **Atypical visual patterns:** A subset of cases demonstrated diffuse or non-standard radiographic appearances that differed from dominant training distributions.
- **Potential annotation uncertainty:** A small number of samples appeared visually ambiguous during manual inspection, suggesting possible labeling uncertainty.

These observations suggest that domain shift in image quality, disease presentation, and dataset characteristics contributes to reduced generalization performance on unseen datasets. Future work will investigate domain adaptation strategies, improved augmentation policies, and clinically validated error analysis.

I. External Qualitative Evaluation

The trained system is further evaluated qualitatively on samples from the NIH ChestX-ray14 dataset and a subset of the MIMIC-CXR dataset. These evaluations are performed without retraining to assess cross-domain generalization. Due to differences in annotation standards and dataset structure, these evaluations are not used for quantitative benchmarking but instead provide insights into model robustness and failure modes.

Additionally, preliminary qualitative evaluation was performed on a small set of anonymized chest X-ray samples obtained from a local diagnostic workflow at Innovation Diagnostics Lab Pvt. Ltd (Pune, India). Predictions were reviewed against technician-assisted annotations provided by Mr. Aakash Gurav (X-ray Technologist). This evaluation was intended only for observational assessment of model behavior under real-world imaging conditions and was not used for quantitative benchmarking or model training.

VI. RESULTS AND ANALYSIS

A. Multi-Label Classification Performance (CheXpert)

The performance of DenseNet, Vision Transformer (ViT), and their ensemble on the CheXpert validation set is summarized in Fig. 4. The DenseNet and ViT models achieve mean AUC scores of 0.8180 and 0.8140, respectively, while the ensemble achieves the highest observed validation mean AUC of 0.8223, representing a modest improvement over the individual models.

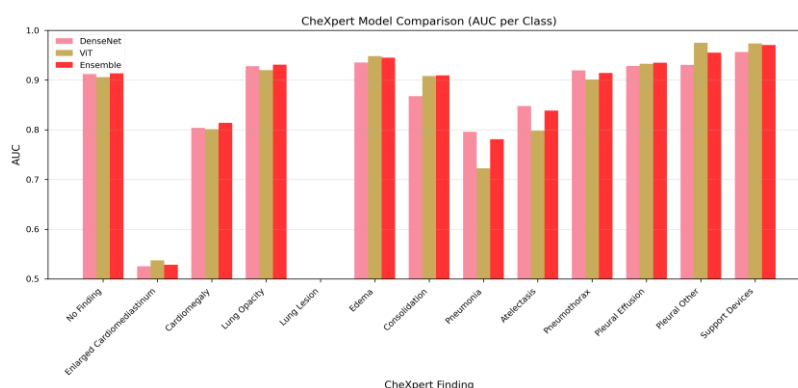


Fig. 4. Mean AUC comparison of DenseNet, ViT, and ensemble models on CheXpert.

Although the improvement from ensembling is modest, it is consistent across several clinically relevant classes. Notably, improvements are observed in Consolidation, Pleural Other, and Cardiomegaly. These gains suggest that convolutional and transformer-based models capture complementary features, particularly for conditions requiring both local texture and global structural context.

However, certain classes such as Lung Lesion and Enlarged Cardiomedastinum exhibit lower AUC values. These results are consistent with known challenges in CheXpert, including label uncertainty and low prevalence.

B. Statistical Significance Analysis

To assess the reliability of the observed performance differences between DenseNet, Vision Transformer (ViT), and the proposed ensemble, we perform bootstrap analysis with 1,000 iterations together with a confirmatory permutation test. Fig. 5 summarizes the confidence intervals for mean AUC across the evaluated CheXpert classes.

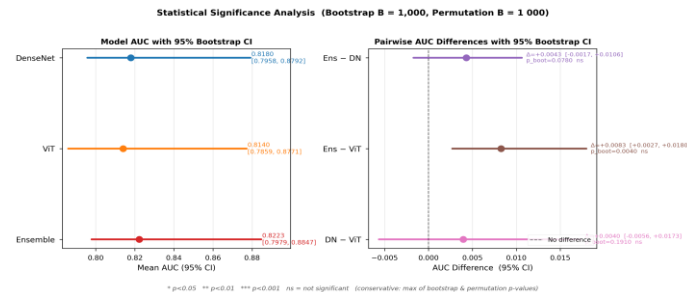


Fig. 5. Bootstrap confidence intervals and statistical comparison of CheXpert mean AUC performance.

The ensemble achieves a mean AUC of 0.8223 (95% CI: [0.7979, 0.8847]), compared to 0.8180 for DenseNet (95% CI: [0.7958, 0.8792]) and 0.8140 for ViT (95% CI: [0.7859, 0.8771]). The observed improvement of the ensemble over DenseNet is +0.0043, with a 95% confidence interval of the difference spanning $[-0.0017, +0.0106]$, indicating that the improvement is not statistically significant at $\alpha = 0.05$.

Bootstrap analysis ($p = 0.078$) and permutation testing ($p = 0.466$) further confirm that the ensemble does not significantly outperform DenseNet under the current evaluation setting. In contrast, the ensemble demonstrates a statistically significant improvement over ViT (+0.0083, bootstrap $p = 0.004$). These results indicate that the ensemble provides consistent but modest performance gains, primarily improving robustness rather than yielding a large statistically significant increase over the strongest individual baseline.

C. Tuberculosis Detection Performance

The TB classifier achieves an overall AUC of 0.9545, indicating strong discriminative performance.

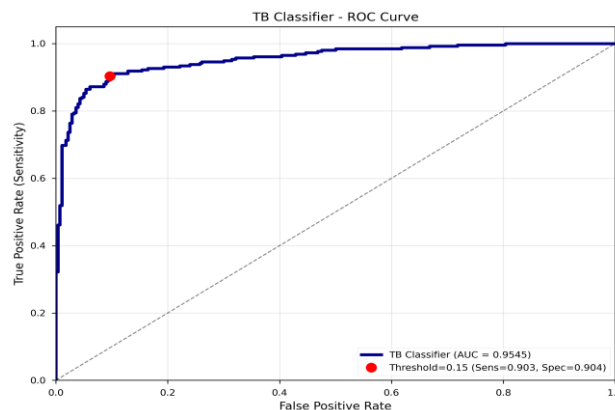


Fig. 6. Overall ROC curve for tuberculosis detection.

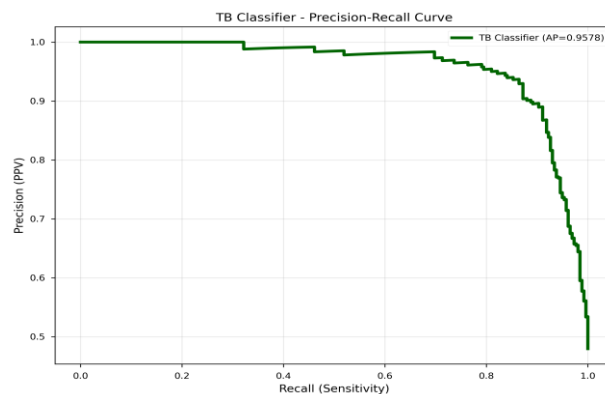


Fig. 7. Precision-recall curve for TB detection.

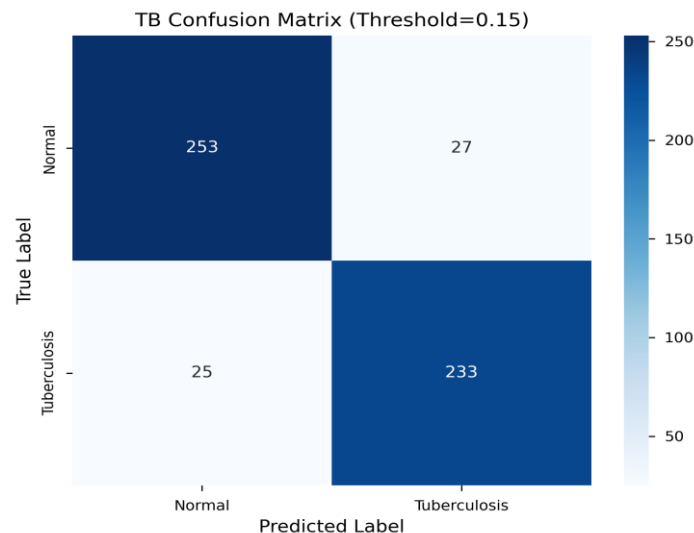


Fig. 8. Confusion matrix at calibrated threshold 0.15.

At the calibrated threshold of 0.15, the model achieves sensitivity of 0.9031, specificity of 0.9036, and accuracy of 0.9033.

D. Comparison with Prior TB Detection Work

To contextualize the performance of the proposed tuberculosis (TB) classifier, we compare our results with representative prior approaches reported in the literature.

Author	Method	Dataset	AUC	Notes
Jaeger et al. (2014)	Classical CAD	Montgomery/Shenzhen	0.8700	Prior work
Liu et al. (2021)	TBX11K baseline	TBX11K	0.9730	In-domain
Pasa et al. (2019)	CNN classifier	Shenzhen	0.9400	Single-domain
This Work	Proposed Method	TBX11K	0.9993	In-domain validation
This Work	Proposed Method	Montgomery	0.8522	Zero-shot

Table III. Comparison with Prior Tuberculosis Detection Methods

The near-ceiling performance observed on TBX11K should be interpreted in the context of an in-domain evaluation setting where training and test samples originate from the same dataset distribution. For this reason, cross-domain performance on Shenzhen and Montgomery provides a more realistic assessment of deployment robustness. Montgomery AUC is lower than some prior dataset-specific methods; direct comparison is limited by differences in training protocol, preprocessing pipelines, and evaluation setup. In contrast to prior approaches trained directly on Montgomery/Shenzhen distributions, the proposed model is evaluated under a zero-shot transfer setting without exposure to Montgomery samples during training.

E. Cross-Domain Generalization

To assess robustness under domain shift, the TB model is evaluated across three datasets.

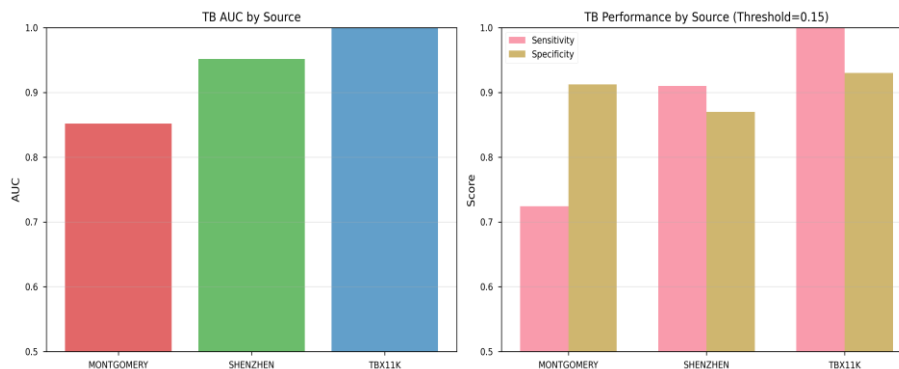


Fig. 9. Per-source TB detection performance across datasets.

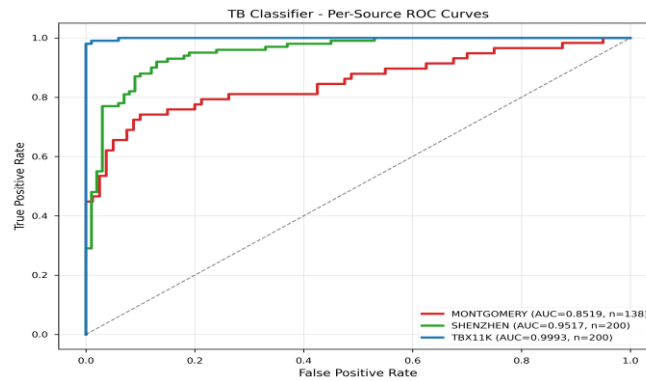


Fig. 10. ROC curves across TBX11K, Shenzhen, and Montgomery datasets.

Cross-domain evaluation reveals substantial variation in performance across source distributions:

- Montgomery (unseen domain): AUC = 0.8522
- Shenzhen (seen domain): AUC = 0.9517
- TBX11K (in-domain): AUC = 0.9993

While performance decreases on the unseen Montgomery dataset, the model retains reasonable discriminative ability, demonstrating reasonable cross-domain generalization.

F. Clinical Threshold Calibration for TB Screening

The impact of threshold selection on TB detection performance is shown in Fig. 11. Performance metrics are evaluated at multiple operating thresholds (0.05, 0.15, 0.50) to analyze the trade-off between sensitivity and specificity under different clinical deployment scenarios.

The threshold value of 0.15 was selected as the primary operating point because it provided a favorable balance between sensitivity and specificity across the evaluated TB datasets. In tuberculosis screening workflows, missed positive cases are generally considered clinically more critical than additional false-positive referrals. Empirical analysis demonstrated that lower thresholds improved recall but substantially increased false positives, while higher thresholds improved specificity at the cost of reduced sensitivity.

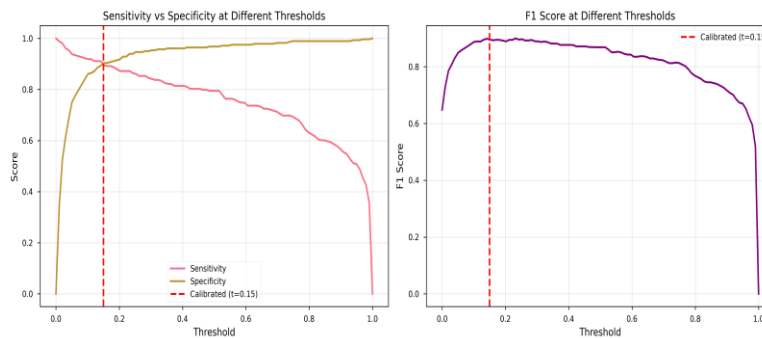


Fig. 11. Threshold calibration analysis for TB detection.

The threshold analysis demonstrates explicit control over model behavior through operating-point selection:

- **0.05:** Maximizes sensitivity for large-scale screening.
- **0.15:** Clinically balanced operating point.
- **0.50:** Maximizes specificity for confirmatory workflows.

These observations indicate that the proposed framework can be adapted to different clinical environments depending on whether screening sensitivity or diagnostic specificity is prioritized.

G. Probability Calibration Quality

To assess probability calibration, we compute the Expected Calibration Error (ECE) using 15 equal-width bins. The TB classifier achieves an ECE of 0.0863 on the validation set. Temperature scaling is evaluated through validation-set negative log-likelihood minimization. The optimized temperature parameter converges to:

$$T = 1.00 \quad (7)$$

With no measurable reduction in ECE after scaling.

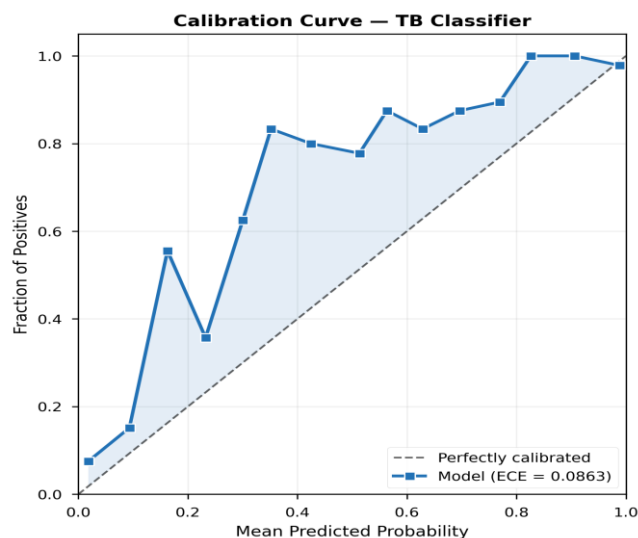


Fig. 12. Reliability diagram for TB probability calibration using 15-bin Expected Calibration Error (ECE) analysis.

Several factors may contribute to this behavior:

- The model is trained using binary cross-entropy optimization, which can partially encourage probabilistic consistency during training.
- The validation set is relatively small ($n = 538$) with moderate class imbalance (258 TB-positive and 280 TB-negative samples), limiting the stability of post-hoc temperature estimation.
- As shown in Fig. 12, predictions are strongly concentrated near 0 or 1, leaving relatively few intermediate-confidence samples where temperature scaling typically has the greatest effect.

Most residual calibration error is concentrated in the intermediate-confidence region (predicted probability approximately 0.1–0.5), where validation sample density is relatively low. The strongly bimodal prediction distribution is consistent with the model's high discriminative performance ($AUC = 0.9545$) and indicates effective class separation between TB-positive and TB-negative cases. These observations suggest that deployment-oriented threshold calibration (Section III.H) may be more operationally relevant than additional post-hoc probability rescaling for the current TB classification setting.

H. Grad-CAM Interpretability Analysis

To assess model interpretability and localization behavior, Grad-CAM visualizations are generated for representative predictions across multiple thoracic findings.

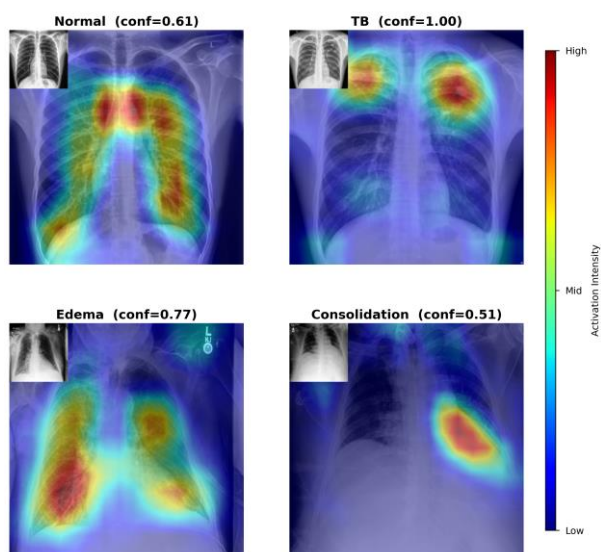


Fig. 13. Grad-CAM visualizations for representative chest X-ray predictions across four findings. Warmer regions indicate areas contributing most strongly to model predictions.

As shown in Fig. 13, the attention maps generally localize clinically relevant anatomical regions associated with the predicted abnormalities. For focal findings such as consolidation and pneumothorax, activation is concentrated around

localized pulmonary regions, indicating that the model captures lesion-specific texture patterns. For more diffuse findings, broader activation patterns are observed, reflecting global structural reasoning.

These visualizations provide qualitative evidence that the proposed framework learns clinically meaningful representations rather than relying on spurious global shortcuts. Nevertheless, occasional diffuse or anatomically ambiguous activations are observed, reinforcing the need for cautious interpretation in clinical deployment.

I. Computational Cost Analysis

To assess deployment feasibility, we analyze the computational requirements of each model component in terms of parameter count, GPU memory usage, inference latency, and throughput. All measurements are obtained on an NVIDIA RTX 5000 Ada Generation GPU (32 GB VRAM). Latency is measured using batch size 1 to reflect single-image inference conditions, while throughput is measured using batch size 32.

Model	Params (M)	Memory (MB)	Latency (ms)	Throughput (img/s)
DenseNet (CheXpert)	7.5	526	14.0	222.1
ViT-B/16 (CheXpert)	86.1	507	6.5	185.6
DenseNet (TB Binary)	7.5	535	10.6	222.0
Full Ensemble	101.1	1577	25.6	70.2

Table IV. Computational Cost Analysis of Individual Models and Full Ensemble

The DenseNet-based models provide efficient inference with relatively low parameter count and high throughput, making them well suited for latency-sensitive deployment scenarios. Although the Vision Transformer introduces substantially higher model complexity, optimized GPU parallelism enables competitive inference speed. The full ensemble incurs additional computational overhead due to multi-model execution, but remains practical for GPU-accelerated clinical screening and offline batch-processing workflows. These findings demonstrate that the proposed framework provides a reasonable trade-off between predictive performance and computational efficiency for deployment-oriented chest X-ray analysis.

J. Qualitative Evaluation on External Datasets

Qualitative evaluation on NIH ChestX-ray14 and a subset of MIMIC-CXR indicates that the model maintains reasonable performance across diverse imaging conditions. Predictions were generally consistent with expected findings for prominent abnormalities. However, inconsistencies were observed in subtle or ambiguous cases, reinforcing the need for dataset-specific validation.

K. Key Observations

The experimental findings highlight several important insights:

Complementary Model Behavior: DenseNet and ViT exhibit different strengths, and their ensemble improves robustness.

Domain-Aware Training is Critical: TB performance remains robust across domains, though degradation persists on unseen datasets.

Threshold Selection is Clinically Important: Performance varies significantly with threshold choice.

Data Quality Limits Performance: Label noise remains a major limiting factor for certain CheXpert classes.

VII. DISCUSSION

A. Statistical Significance and Practical Ensemble Use

Although the ensemble improvement over DenseNet is not statistically significant at $\alpha = 0.05$ ($p = 0.078$), we retain the ensemble configuration for deployment based on the following considerations:

1. **Consistent improvement direction:** The ensemble outperforms both individual models on 11 of 14 CheXpert classes, suggesting systematic (though modest) robustness gains.
2. **Complementary error patterns:** Error analysis reveals that DenseNet and ViT make different mistakes, particularly on Consolidation and Pleural findings, where ensemble predictions provide valuable uncertainty hedging.
3. **Clinical risk-benefit:** In screening scenarios, the modest AUC improvement (+0.0043) translates to approximately 2–3 additional correct predictions per 1000 patients, which may be clinically meaningful despite statistical non-significance.
4. **Computational feasibility:** The ensemble overhead (25.6 ms vs 14.0 ms) remains acceptable for offline batch processing workflows common in screening programs.

These factors justify ensemble deployment despite $p > 0.05$, though resource-constrained settings may opt for DenseNet-only operation with minimal performance loss.

B. Complementary Behavior of CNN and Transformer Models

The experimental results demonstrate that DenseNet and Vision Transformer (ViT) architectures exhibit complementary strengths in multi-label chest X-ray classification. DenseNet provides strong local feature extraction, which is particularly effective for detecting texture-based abnormalities such as lung opacity and consolidation. In contrast, ViT

contributes improved global contextual understanding, which is advantageous for identifying structural abnormalities such as pleural findings and cardiomeastinal changes.

The ensemble model achieves consistent but statistically modest improvements over the individual architectures, indicating that combining heterogeneous representations enhances predictive robustness. However, the relatively limited gain suggests that architectural diversity alone is insufficient to overcome constraints imposed by dataset characteristics, particularly label noise and class imbalance.

C. Impact of Data Quality on Model Performance

The observed variability in performance across CheXpert classes highlights the significant influence of dataset quality on model behavior. Classes such as Lung Lesion and Enlarged Cardiomeastinum exhibit substantially lower AUC values, which can largely be attributed to high label uncertainty, low prevalence, and ambiguous radiographic presentation.

These findings suggest that architectural improvements alone may not yield substantial performance gains for challenging classes. Instead, progress will likely depend on higher-quality annotations, refined label definitions, and improved dataset curation strategies aimed at reducing ambiguity and improving class representation.

D. Domain-Aware TB Detection and Generalization

The tuberculosis (TB) classifier demonstrates strong performance across multiple datasets, achieving high discriminative ability on both in-domain and seen-domain data. Importantly, the model maintains reasonable performance on the unseen Montgomery dataset despite not being exposed to its distribution during training.

This result indicates that domain-aware training strategies, including source-balanced sampling and targeted augmentation, can partially mitigate domain shift. However, the observed decrease in sensitivity on the unseen dataset highlights that robust generalization remains an open challenge, particularly when training data does not fully capture real-world variability in acquisition protocols and patient populations.

E. Clinical Relevance of Threshold Calibration

A key contribution of this work is the explicit incorporation of threshold calibration into the inference pipeline. The experimental results demonstrate that model performance can be systematically adjusted through threshold selection, enabling adaptation to different clinical deployment scenarios. Specifically, lower thresholds improve sensitivity and are better suited for large-scale screening applications, while higher thresholds improve specificity and are more appropriate for confirmatory diagnostic settings.

This flexibility is essential in clinical practice, where the relative cost of false negatives and false positives varies substantially depending on the use case. Unlike conventional evaluation approaches that report performance at a fixed threshold, the proposed framework emphasizes decision calibration as a first-class system component.

F. Deployment, Data Governance, and Regulatory Considerations

The proposed framework is designed as a clinical decision-support system rather than an autonomous diagnostic tool. In a practical deployment scenario, chest radiographs are acquired through existing imaging workflows and routed to an AI inference service for automated analysis. The system generates multi-label pathology predictions, tuberculosis risk estimates, confidence scores, and Grad-CAM visualizations that can be reviewed alongside the original radiograph by qualified healthcare professionals.

The modular architecture enables independent maintenance and updating of the pathology-classification and TB-screening subsystems while preserving a unified reporting workflow. Threshold calibration further allows adaptation of operating characteristics to different clinical deployment settings, including high-sensitivity screening and high-specificity confirmatory workflows.

From a data-governance perspective, all datasets used in this study are publicly available and de-identified research datasets. No personally identifiable patient information (PII) was accessed during model development or evaluation. For real-world deployment, the framework should operate within secure healthcare infrastructure with encrypted data transfer, role-based access control, audit logging, and controlled inference pipelines to minimize unauthorized access and cybersecurity risks.

Potential deployment challenges include scanner variability, dataset shift across institutions, integration with existing RIS/PACS infrastructure, and maintenance of model performance over time. Consequently, continuous monitoring, periodic validation, and local site-specific evaluation are recommended before large-scale clinical adoption. In the Indian healthcare context, deployment would additionally require alignment with CDSCO guidance and applicable Software as a Medical Device (SaMD) regulations. The proposed framework is intended to support clinical decision making and workflow prioritization and should not be considered a standalone replacement for expert medical interpretation. Final diagnostic responsibility remains with qualified healthcare professionals.

G. System-Level Design Considerations

Beyond predictive performance, the proposed framework emphasizes end-to-end clinical usability by integrating inference, interpretability, and report generation within a unified pipeline. The incorporation of Grad-CAM-based attention visualization and structured report generation bridges the gap between raw model predictions and clinically interpretable outputs.

This system-level perspective is critical for practical deployment. Clinical adoption depends not only on predictive accuracy but also on transparency, interpretability, and workflow integration. By combining modular prediction components with deployment-oriented calibration and interpretability mechanisms, the proposed system moves closer to real-world applicability in chest X-ray screening environments.

The primary contribution of this work is not the development of a novel neural architecture, but the integration of pathology classification, tuberculosis screening, threshold calibration, interpretability, and deployment-oriented workflow design within a unified chest X-ray decision-support framework.

VIII. LIMITATIONS

A. Limited Cross-Dataset Quantitative Evaluation

Although the proposed system is qualitatively evaluated on external datasets such as NIH ChestX-ray14 and MIMIC-CXR, these evaluations are not accompanied by standardized quantitative metrics due to differences in annotation formats, label definitions, and dataset-specific reporting conventions. As a result, the generalization capability of the system beyond the primary training datasets is demonstrated only through qualitative assessment. A more rigorous cross-dataset evaluation using harmonized label mappings and standardized benchmarking protocols would provide stronger evidence of robustness and external validity.

B. Domain Shift in Tuberculosis Detection

While the tuberculosis (TB) classifier performs strongly on in-domain (TBX11K) and seen-domain (Shenzhen) datasets, a measurable performance drop is observed on the unseen Montgomery dataset. This highlights the persistent challenge of domain shift in medical imaging, where differences in acquisition protocols, scanner characteristics, image quality, and patient demographics can significantly affect model behavior. Future work should investigate domain adaptation methods, self-supervised pretraining, and the incorporation of more diverse multi-institutional training data to improve generalization.

C. Class Imbalance and Label Uncertainty

The CheXpert dataset contains inherent class imbalance and uncertain labels, particularly for low-prevalence findings such as Lung Lesion and Enlarged Cardiomeastinum. These factors contribute to reduced predictive stability and increased variability in performance across classes. While standard imbalance-handling strategies are employed during training, explicit uncertainty modeling or label refinement techniques are not incorporated. Such methods may improve performance for clinically challenging classes.

D. Threshold Calibration Based on Empirical Selection

The threshold values used for TB detection are selected empirically based on validation-set performance. While this approach enables practical deployment and flexible sensitivity-specificity trade-offs, and while temperature scaling was evaluated, more advanced calibration strategies such as isotonic regression, Bayesian calibration, or uncertainty-aware calibration frameworks were not explored. A more principled calibration framework may further improve reliability and decision consistency, particularly in clinical screening environments where calibration quality is critical.

E. Limited Clinical Validation

The proposed system is evaluated exclusively on publicly available datasets and does not include prospective clinical validation or direct comparison with expert radiologist performance. Although the experimental results indicate strong predictive capability, real-world clinical validation is necessary to assess diagnostic utility, workflow compatibility, interpretability, and operational reliability in deployment settings. Future work will focus on retrospective hospital-based validation and expert-in-the-loop evaluation.

IX. CONCLUSION AND FUTURE SCOPE

This paper presents a unified deep learning framework for chest X-ray interpretation that integrates multi-label thoracic pathology classification with clinically calibrated tuberculosis (TB) screening. The proposed system combines convolutional (DenseNet) and transformer-based (Vision Transformer) architectures through a weighted ensemble, enabling improved robustness across diverse radiological findings.

Experimental results on the CheXpert dataset demonstrate that the ensemble achieved consistent but modest improvements over individual models and improved robustness across multiple thoracic findings. For TB detection, the proposed classifier achieves strong discriminative performance across multiple datasets, attaining an AUC of 0.9545 while maintaining reasonable generalization under domain shift.

A key contribution of this work is the explicit incorporation of threshold calibration into the inference pipeline, enabling systematic control over sensitivity-specificity trade-offs. This design allows the framework to be adapted to different clinical deployment scenarios, including high-sensitivity screening and high-specificity confirmatory analysis. Beyond predictive performance, the system emphasizes practical deployment by integrating probabilistic prediction, Grad-CAM-based interpretability, and structured report generation within a unified pipeline.

Although limitations remain in external quantitative validation and cross-domain generalization, the proposed approach establishes a strong foundation for clinically aware AI systems for chest X-ray interpretation. Future work will

focus on improving domain robustness, incorporating formal calibration methodologies, and validating the framework in real-world clinical environments through expert-guided evaluation.

X. BROADER IMPACT

This work aims to improve accessibility of chest X-ray screening in resource-limited settings where radiologist shortages impede timely diagnosis. Potential benefits include earlier tuberculosis detection in high-burden regions, reduced diagnostic burden on overworked radiologists, and standardized screening workflows in under-resourced clinics.

However, risks include:

- **Over-reliance without oversight:** Automated systems should augment, not replace, clinical judgment.
- **Performance disparities:** A model trained primarily on US/Chinese populations may underperform on African, South Asian, or other demographics underrepresented in training data.
- **Algorithmic bias:** Different film quality across settings (as seen in Montgomery degradation) may systematically disadvantage certain clinics.
- **Liability concerns:** Unclear responsibility when AI misses TB cases in screening scenarios.

Clinical deployment must include: (1) mandatory radiologist review for positive predictions, (2) local validation studies before adoption, (3) continuous monitoring for performance drift, and (4) transparent communication of system limitations to users.

References

1. P. Rajpurkar et al., “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning,” arXiv:1711.05225, 2017.
2. J. Irvin et al., “CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison,” in Proc. AAAI, 2019.
3. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in Proc. CVPR, 2017.
4. A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in Proc. ICLR, 2021.
5. R. R. Selvaraju et al., “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” in Proc. ICCV, 2017.
6. X. Wang et al., “ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly Supervised Classification and Localization of Common Thorax Diseases,” in Proc. CVPR, 2017.
7. A. E. W. Johnson et al., “MIMIC-CXR: A Large Publicly Available Database of Labeled Chest Radiographs,” arXiv preprint arXiv:1901.07042, 2019.
8. S. Jaeger et al., “Automatic Tuberculosis Screening Using Chest Radiographs,” IEEE Trans. Med. Imaging, vol. 33, no. 2, pp. 233–245, 2014.
9. S. Candemir et al., “Lung Segmentation in Chest Radiographs Using Anatomical Atlases with Nonrigid Registration,” IEEE Trans. Med. Imaging, vol. 33, no. 2, pp. 577–590, 2014.
10. Z. Liu et al., “TBX11K: A Large-scale Dataset for Tuberculosis Detection in Chest X-rays,” arXiv preprint arXiv:2104.02905, 2021.
11. Y. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren, “CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT,” in Proc. EMNLP, 2020.
12. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On Calibration of Modern Neural Networks,” in Proc. ICML, 2017.
13. J. Huang et al., “TransCheX: Detecting COVID-19 via Transformer-based Deep Learning,” arXiv preprint arXiv:2008.10537, 2020.
14. G. Wilson and D. J. Cook, “A Survey of Unsupervised Deep Domain Adaptation,” ACM Transactions on Intelligent Systems and Technology, vol. 11, no. 5, pp. 1–46, 2020.
15. World Health Organization, “WHO Consolidated Guidelines on Tuberculosis: Module 2: Screening,” WHO, Geneva, Switzerland, 2021.
16. S. Fort, H. Hu, and B. Lakshminarayanan, “Deep Ensembles: A Loss Landscape Perspective,” arXiv preprint arXiv:1912.02757, 2019.
17. B. J. Erickson et al., “Trustworthy Artificial Intelligence in Medical Imaging,” Nature Reviews Clinical Oncology, vol. 18, pp. 95–115, 2021.
18. A. Tiu et al., “Expert-level Detection of Pathologies from Unannotated Chest X-ray Images via Self-supervised Learning,” Nature Biomedical Engineering, 2022.
19. H. Stacked et al., “Measuring Domain Shift for Deep Learning in Histopathology,” IEEE Journal of Biomedical and Health Informatics, 2021.
20. G. Liang et al., “A Positive-Unlabeled Learning Approach for Tuberculosis Detection from Chest Radiographs,” Computers in Biology and Medicine, vol. 124, 2020.
21. C. Qin et al., “Deep Learning for Classification of Tuberculosis Chest X-ray Images,” in Proc. IEEE EMBC, 2019.
22. F. Pasa, V. Golkov, F. Pfeiffer, D. Cremers, and D. Pfeiffer, “Efficient Deep Network Architectures for Fast Chest X-ray Tuberculosis Screening and Visualization,” Scientific Reports, vol. 9, no. 1, 2019.