ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

PulmoScan AI: An Explainable Deep Learning-Based Clinical Decision Support System for Multi-Class Lung Disease Detection Using Chest X-Ray Images

Joel Jacob Varghese[1], Manchit Choudhary[2], Manna Sara Bilu[3], Navin Cholangi[4], Dr. Yogesh Golhar[5]

[1]Department of Computer Engineering St. Vincent Pallotti College of Engineering and Technology, Nagpur

[2]Department of Computer Engineering St. Vincent Pallotti College of Engineering and Technology, Nagpur

[3]Department of Computer Engineering St. Vincent Pallotti College of Engineering and Technology, Nagpur

[4]Department of Computer Engineering St. Vincent Pallotti College of Engineering and Technology, Nagpur

[5]Assistant Professor, Department of Computer Engineering St. Vincent Pallotti College of Engineering and Technology, Nagpur

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150700139

Received: 12 August 2026; Accepted: 17 August 2026; Published: 24 August 2026

ABSTRACT

Lung diseases such as pneumonia and tuberculosis (TB) represent a major global health burden, particularly in low- and middle-income regions with limited access to expert radiological interpretation. Early and accurate diagnosis through chest X-ray (CXR) imaging is critical, yet conventional radiological interpretation suffers from inter-observer variability and a shortage of expert radiologists in resource-limited settings. This paper proposes PulmoScan AI, a deep learning-based full-stack clinical decision support system for automated detection and classification of lung diseases from CXR images. The system employs EfficientNetB0 with transfer learning for multi-class classification, trained on a curated dataset of 11,910 images, and detects three categories: Normal, Pneumonia, and Tuberculosis. The model achieves a test accuracy of 97.9%, AUC-ROC of 0.9982, macro precision of 98.28%, macro recall of 97.70%, and macro F1-score of 97.96%, outperforming four baseline architectures (a shallow CNN, a custom CNN, MobileNetV2, and ResNet50) trained under an identical protocol; five-fold cross-validation confirms stable performance across data partitions (97.20 ± 0.39% mean accuracy). A web-based clinical decision support interface integrates real-time prediction with confidence thresholding, Grad-CAM explainability, and an occupational risk assessment module. Experimental results demonstrate the feasibility and clinical potential of the approach for deployment in health screening programs.

Keywords—Deep learning, chest X-ray, lung disease detection, EfficientNetB0, transfer learning, Grad-CAM, explainable AI, computer-aided diagnosis, pneumonia, tuberculosis, clinical decision support

INTRODUCTION

Lung diseases represent one of the most pressing global health challenges. Pneumonia and tuberculosis (TB) alone account for millions of preventable deaths annually, with a disproportionate impact on low- and middle-income countries [1], [2]. The burden is compounded in occupational settings, where prolonged exposure to crystalline silica dust, coal particulates, and infectious aerosols additionally gives rise to conditions such as silicosis and coal workers' pneumoconiosis (CWP).

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

Chest X-ray (CXR) imaging remains the primary first-line screening tool owing to its wide availability, low cost, and minimal radiation exposure. However, CXR-based diagnosis carries inherent limitations in sensitivity and reproducibility; for early-stage disease in particular, reported sensitivity can be low—for example, as low as 0.76 for silicosis screening in a recent systematic review of 20 studies covering 2,156 participants [3]—implying a substantial number of missed cases and a clinically significant diagnostic gap.

The rapid advancement of deep learning and convolutional neural networks (CNNs) has opened new frontiers in medical image analysis. Transfer learning approaches have demonstrated competitive performance even when labelled clinical data is scarce [4], [13], [19]. Prior work has validated architectures including VGG19, ResNet, DenseNet121, and MobileNetV2 for multi-class CXR classification.

This paper presents PulmoScan AI, a full-stack deep learning system using EfficientNetB0 transfer learning. While the broader motivation of this work is the under-diagnosis of occupational lung disease, the present system targets three classes for which large, well-curated public CXR datasets exist—Normal, Pneumonia, and Tuberculosis—and is architected for future extension to silicosis and pneumoconiosis as suitable datasets become available. The principal contributions are: (1) an EfficientNetB0 classifier trained on 11,910 images with two-stage warm-up and fine-tuning; (2) Grad-CAM explainability with test-time augmentation (TTA) and confidence thresholding; (3) a Flask REST API and React frontend with a clinical dashboard; (4) an occupational risk assessment module for contextual screening support; and (5) a controlled benchmark against four baseline architectures trained on the same dataset and protocol.

Problem Statement and Objectives

Problem Statement

Conventional CXR-based diagnosis faces several compounding challenges. Radiological interpretation is inherently subjective, producing significant intra- and inter-observer variability [3]. Early-stage disease produces subtle radiographic changes that are frequently missed, and resource-limited settings lack sufficient expertise, leaving many patient populations under-screened [14].

These issues create a critical need for an accurate, accessible, and automated AI-assisted screening system that can function as a first-pass triage tool and reduce the burden on specialist radiologists.

Objectives

1) Analyse existing deep learning approaches for CXR-based lung disease classification to identify architectural best practices.

2) Design and implement a multi-class classification system using EfficientNetB0 transfer learning on a curated 11,910-image dataset.

3) Address training stability through label smoothing, data augmentation, and a two-phase training strategy.

4) Integrate confidence thresholding and Grad-CAM explainability to minimise high-confidence misclassifications.

5) Evaluate performance using accuracy, AUC-ROC, per-class precision/recall/F1, and confusion matrix, and benchmark the proposed backbone against four baseline architectures trained on the same dataset.

6) Deploy a full-stack Flask+React interface with an occupational risk assessment module.

Related Work

The use of AI and deep learning in medical imaging has grown significantly. CNN-based models and transfer learning have clearly improved detection accuracy, and the addition of explainability tools like Grad-CAM supports clinical trust [8].

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

Alshmrani et al. [4] proposed a multi-class lung disease classification system using VGG19+CNN achieving 96.48% accuracy and AUC of 0.998 on 80,000 images with six classes. Sharma et al. [5] combined UNet and Xception for TB detection achieving 99.29% accuracy. Priego-Torres et al. [6] proposed a deep learning pipeline for engineered-stone silicosis screening and staging from CXR, reporting near-perfect screening discrimination and approximately 84% staging accuracy. Zhang et al. [7] applied DenseNet121 for pneumoconiosis staging (81%). Tan and Le [9] introduced EfficientNet, demonstrating superior accuracy-to-parameter efficiency through compound scaling of depth, width, and resolution. For pneumonia specifically, Hashmi et al. [20] showed that an optimally weighted ensemble of transfer-learned CNNs achieves high detection accuracy on CXR. More recently, Veeramani et al. [18] proposed an explainable transfer learning framework (XAI-TRANS) for multi-class CXR classification using LIME and Grad-CAM.

Despite these advances, most models are never deployed in a practical platform. PulmoScan AI directly targets this gap by providing a full-stack deployable system with explainability and occupational risk assessment.

METHODOLOGY

The methodology is organised into four phases: Dataset Preparation, Model Architecture, Training Strategy, and System Deployment.

Dataset Preparation

The dataset comprises 11,910 chest X-ray images from publicly available repositories. Normal (4,841) and Pneumonia (3,875) images were sourced from the RSNA Pneumonia Detection Challenge dataset (Kaggle) [16]. Tuberculosis (3,194) images were obtained from the NIAID TB Portals Program dataset [17].

A stratified 75/15/10 split was applied, yielding 8,932 training, 1,787 validation, and 1,191 test images. The per-class split is summarised in Table I and visualised in Fig. 1.

Table I Dataset Class Distribution Across Splits

Class	Total	Train	Val	Test
Normal	4,841	3,630	727	484
Pneumonia	3,875	2,906	581	388
Tuberculosis	3,194	2,396	479	319
Total	11,910	8,932	1,787	1,191

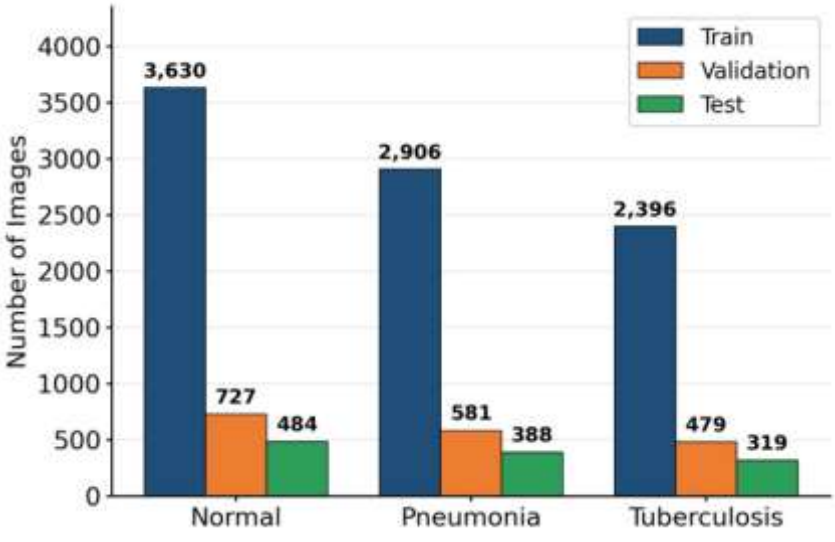

Fig. 1. Dataset distribution: number of images per class across training, validation, and test splits.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

Preprocessing and Data Augmentation

All images were resized to 224×224 pixels. Pixel values were normalised using the EfficientNet preprocessing function. During training, the following stochastic augmentation operations were applied: random rotation, zoom variations (±10%), width and height shifts (±10%), brightness adjustment, shear, and horizontal flipping. Label smoothing of 0.04 was applied to reduce overconfidence. No augmentation was applied during validation or testing.

Model Architecture

The architecture centres on EfficientNetB0 [9], a CNN that uniformly scales depth, width, and resolution using a compound coefficient, achieving superior accuracy-to-parameter efficiency. The end-to-end system architecture—from image ingestion through classification, explainability, and web-based delivery—is shown in Fig. 2. The classification pipeline consists of: (1) a 224×224×3 input; (2) EfficientNet preprocessing normalisation; (3) the EfficientNetB0 backbone (ImageNet weights, include_top=False); (4) GlobalAveragePooling2D; (5) BatchNormalization; (6) Dropout (rate=0.40); (7) a Dense layer (192 units, ReLU, L2 regularisation); (8) Dropout (rate=0.35); and (9) a final Dense(3, Softmax) output over Normal, Pneumonia, and Tuberculosis.

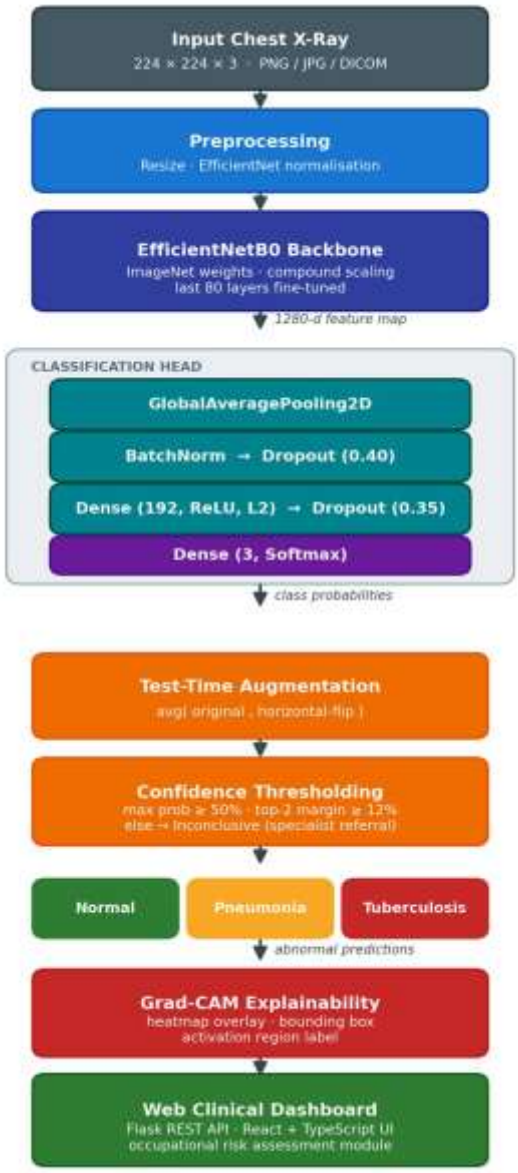

Fig. 2. End-to-end system architecture of PulmoScan AI: preprocessing, EfficientNetB0 feature extraction, classification head, inference-time safeguards (TTA and confidence thresholding), Grad-CAM explainability, and the web clinical dashboard.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

Training Strategy

Training was conducted in two stages. Stage 1 — Warm-up (frozen backbone): the EfficientNetB0 backbone is frozen; only the classification head is trained using Adam with learning rate 4×10^{-4}. Stage 2 — Fine-tuning (partial unfreeze): the last 80 layers of the backbone are unfrozen and fine-tuned using Adam with reduced learning rate 1×10^{-5} to prevent catastrophic forgetting.

Training callbacks included ModelCheckpoint (best validation AUC), EarlyStopping (patience=5), and ReduceLROnPlateau (factor=0.5, patience=3). The loss function was CategoricalCrossentropy with label smoothing of 0.04. The technology stack is summarised in Table II.

Table II Technologies And Components Used In System Implementation

Component	Technology / Detail
Base Model	EfficientNetB0 (ImageNet weights)
Input Size	224×224×3 pixels
Classification Head	GAP → BN → Dropout → Dense(Softmax)
Loss Function	CategoricalCrossentropy (label smooth 0.04)
Optimizer	Adam (lr: 4×10^{-4} warm-up; 10^{-5} fine-tune)
Saved Model	model_3class.keras / .h5
Backend / API	Flask REST API (POST /predict)
Frontend	React + TypeScript + Tailwind CSS + Vite
Dev Tools	Python 3, TensorFlow/Keras, Google Colab Pro

Confidence Thresholding and Test-Time Augmentation

At inference time, test-time augmentation (TTA) is applied: the original image and a horizontally flipped version are passed through the model and their softmax probability vectors are averaged, reducing variance and improving calibration. Predictions are then evaluated against a minimum confidence threshold of 50% and a minimum margin of 12% between the top-2 predictions. Predictions below threshold are flagged as Inconclusive, prompting referral for specialist review.

Grad-CAM Explainability

Gradient-weighted Class Activation Mapping (Grad-CAM) [8] is implemented to provide visual explanations for model predictions. For abnormal predictions (Pneumonia or Tuberculosis), the backend returns: a heatmap overlay on the original image, a bounding box around the high-activation region, a peak activation score, and a region label (e.g., upper/lower left/right lung). For Normal predictions, the heatmap is suppressed to avoid misleading activations. Fig. 3 illustrates the Grad-CAM output for a Pneumonia prediction.

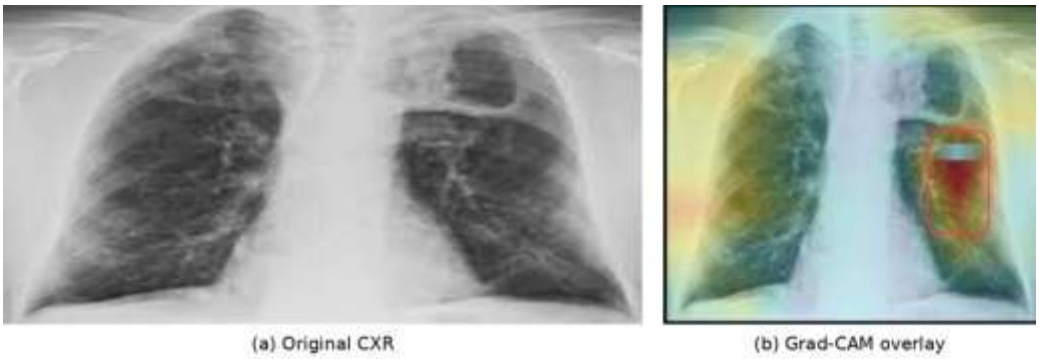

Fig.3. Grad-CAM explainability output: (a) original CXR, (b) heatmap overlay with bounding box identifying the high-activation lung region for a Pneumonia prediction.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

System Deployment

The trained model is served via a Flask REST API with endpoints /predict, /model-metrics, /health, and authentication routes. The backend supports PNG, JPG, JPEG, WebP, BMP, and optional DICOM. An SQLite authentication database with password hashing and session token generation is integrated.

The React/TypeScript frontend is built with Vite, Tailwind CSS, Zustand (state management), Recharts (visualisation), Framer Motion (animation), and Three.js/React Three Fiber (3D lung model). Key features include scan upload and preview, Grad-CAM overlay with zoom/contrast controls, confidence bar charts, JSON export, text report download, image histogram, and a model metrics dashboard. Representative views of the deployed web interface are presented in Section V-J (Figs. 9 and 10).

An additional occupational risk assessment module computes a tiered risk score (Low / Moderate / High / Critical) based on occupation type, exposure years, dust/silica/asbestos/fumes/coal exposure, smoking status, pack years, age, BMI, and symptom severity. This module is rule-based and is distinct from the trained deep learning classifier.

RESULTS AND EXPERIMENTAL ANALYSIS

Experimental Setup

Test environment: Python 3 with TensorFlow/Keras; model trained on Google Colab Pro (NVIDIA A100 GPU). Total dataset: 11,910 CXR images; held-out test set: 1,191 images (484 Normal, 388 Pneumonia, 319 Tuberculosis).

Overall Performance Results

Table III summarises the overall test performance. The model achieves 97.9% accuracy and a macro AUC-ROC of 0.9982, indicating exceptional class discrimination. Fig. 4 visualises these metrics.

Table III Overall Test Performance Metrics

Metric	Value
Accuracy	97.90%
Test Loss	0.2787
AUC-ROC (Macro)	0.9982
Macro Precision	98.28%
Macro Recall	97.70%
Macro F1-Score	97.96%

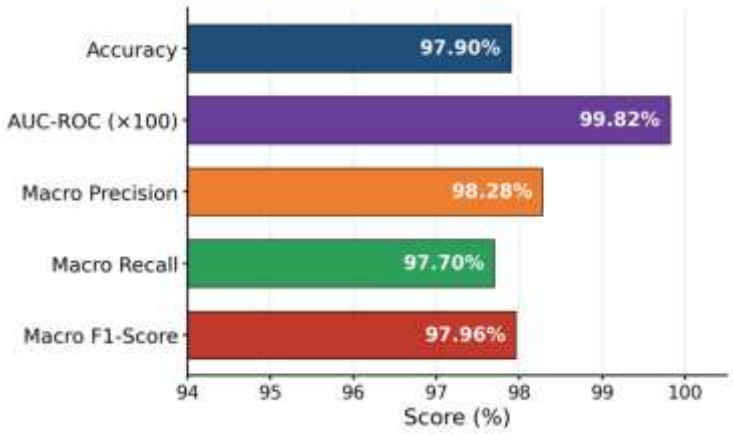

Fig. 4. Overall model performance metrics on the held-out test set.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

Per-Class Performance Analysis

Table IV presents the per-class metrics on the 1,191-image test set. Key observations: (i) Tuberculosis achieves perfect precision (100.00%); (ii) Normal achieves the highest recall (99.59%), ensuring almost all normal cases are correctly identified; (iii) the lowest individual metric is Tuberculosis recall at 96.87%, still a strong clinical result. Fig. 5 shows the grouped bar chart.

Table IV Per-Class Classification Metrics (Test Set, N = 1,191)

Class	Prec.	Recall	F1	Supp.
Normal	95.63%	99.59%	97.57%	484
Pneumonia	99.21%	96.65%	97.91%	388
Tuberculosis	100.00%	96.87%	98.41%	319
Macro Avg	98.28%	97.70%	97.96%	1,191

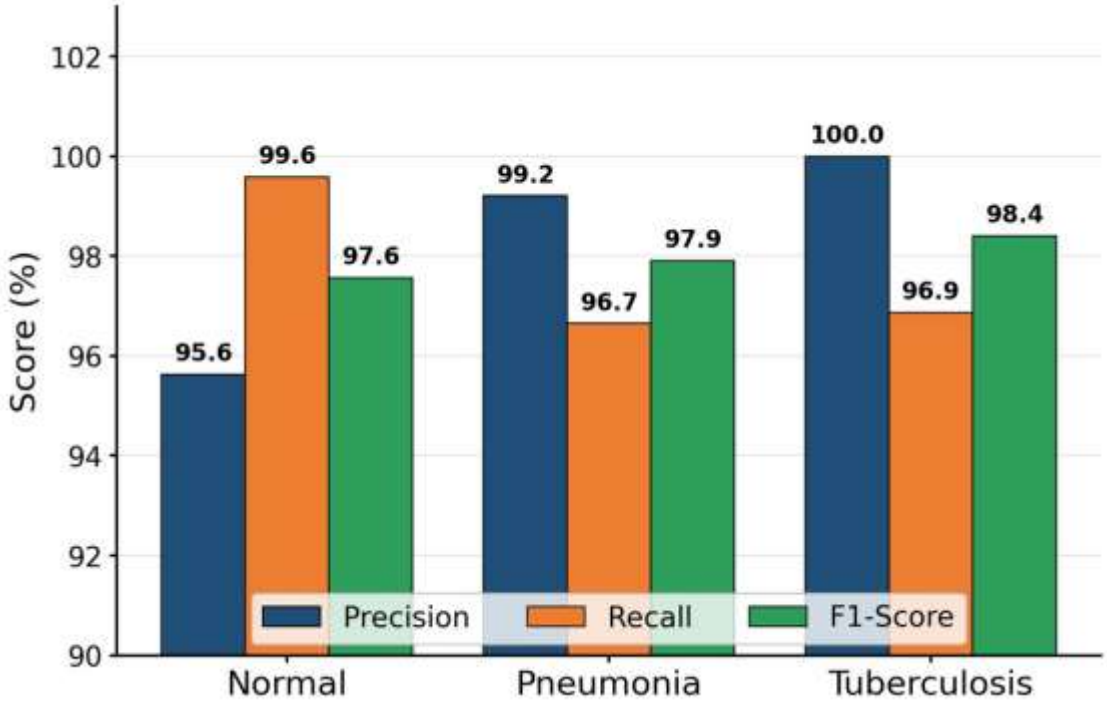

Fig. 5. Per-class Precision, Recall and F1-Score on the test set (n = 1,191).

Confusion Matrix Analysis

Table V presents the confusion matrix. Of 1,191 test samples, 1,166 are correctly classified (97.9% accuracy). The dominant error is the misclassification of 13 Pneumonia images as Normal (3.35% of Pneumonia cases), reflecting the radiological similarity between early consolidation patterns and normal parenchyma. No Normal or Pneumonia images were misclassified as Tuberculosis. Fig. 6 provides a colour-coded visualisation.

Table V Confusion Matrix On Held-Out Test Set (N = 1,191)

Actual \ Predicted	Normal	Pneumonia	TB
Normal	482	2	0
Pneumonia	13	375	0
Tuberculosis	9	1	309

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

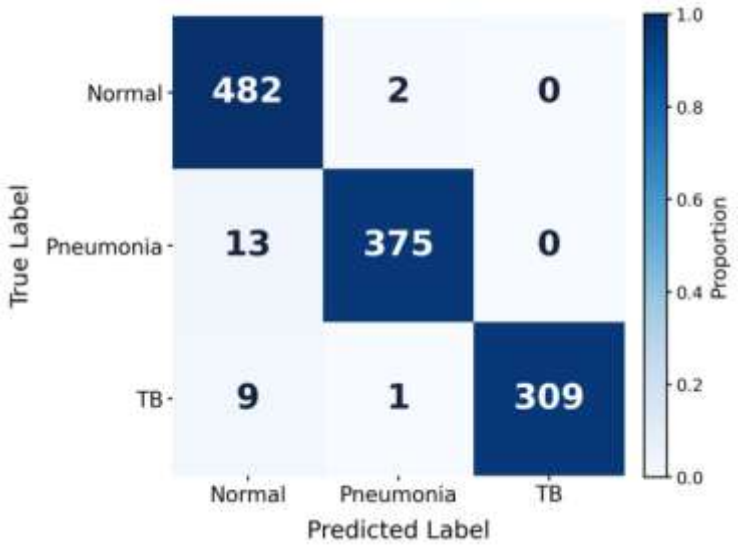

Fig. 6. Colour-coded confusion matrix on the held-out test set (n = 1,191). Diagonal cells show correct classifications; off-diagonal cells show misclassifications. Values are raw counts; colour intensity represents per-row proportion.

ROC-AUC Analysis

Aggregate metrics such as accuracy and the confusion matrix summarise predictions at a single fixed decision point and are not sufficient to reconstruct a ROC curve, which requires the predicted probability score for every individual test sample across varying decision thresholds. To plot the true empirical ROC curves rather than estimate them from aggregate statistics, the trained model was evaluated again on the independent 1,191-image test set and the per-sample softmax probabilities were exported. For each test image, the filename, ground-truth label, predicted label, and the softmax scores for Normal, Pneumonia, and Tuberculosis were recorded.

Because this is a three-class task, ROC curves were computed using a one-versus-rest (OvR) scheme: each class is treated as the positive class in turn while the remaining two are treated as negative, and the true and false positive rates are evaluated across all probability thresholds. Fig. 7 presents the resulting curves. The per-class AUC values obtained from the exported per-sample scores were 0.9969 (Normal), 0.9976 (Pneumonia), and 0.9999 (Tuberculosis), giving a macro-averaged AUC-ROC of 0.9982 and confirming exceptional discriminative capability across all three classes. The confusion matrix in Section V-D, by contrast, summarises only the final class predictions at the fixed operating threshold.

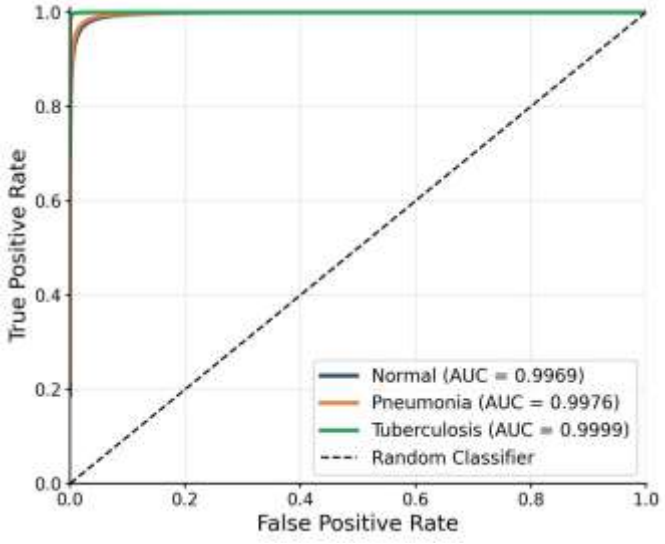

Fig. 7. Multi-class ROC curves (OvR) computed from per-sample softmax scores on the held-out test set. Macro AUC-ROC = 0.9982.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

Backbone Comparison on the Same Dataset

To isolate the contribution of the EfficientNetB0 backbone, four additional models were trained and evaluated on the identical dataset, stratified train/validation/test split (8,932 / 1,787 / 1,191 images), preprocessing pipeline, and evaluation protocol. The baselines span a range of capacities: a shallow CNN and a deeper custom CNN, both trained from scratch, and two ImageNet-pretrained transfer-learning backbones, MobileNetV2 [10] and ResNet50. All models were evaluated on the same 1,191-image held-out test set. Table VI reports the results, and Fig. 8 visualises accuracy and macro F1 across models.

Table VI Backbone Comparison On The Same Dataset And Protocol

Model	Acc.	Prec.	Recall	Auc
Shallow CNN	64.48%	69.28%	67.71%	0.8568
Custom CNN	86.23%	89.08%	83.83%	0.9884
MobileNetV2 [10]	96.31%	96.75%	96.17%	0.9982
ResNet50	96.47%	97.24%	96.13%	0.9995
EfficientNetB0 (ours)	97.90%	98.28%	97.70%	0.9982

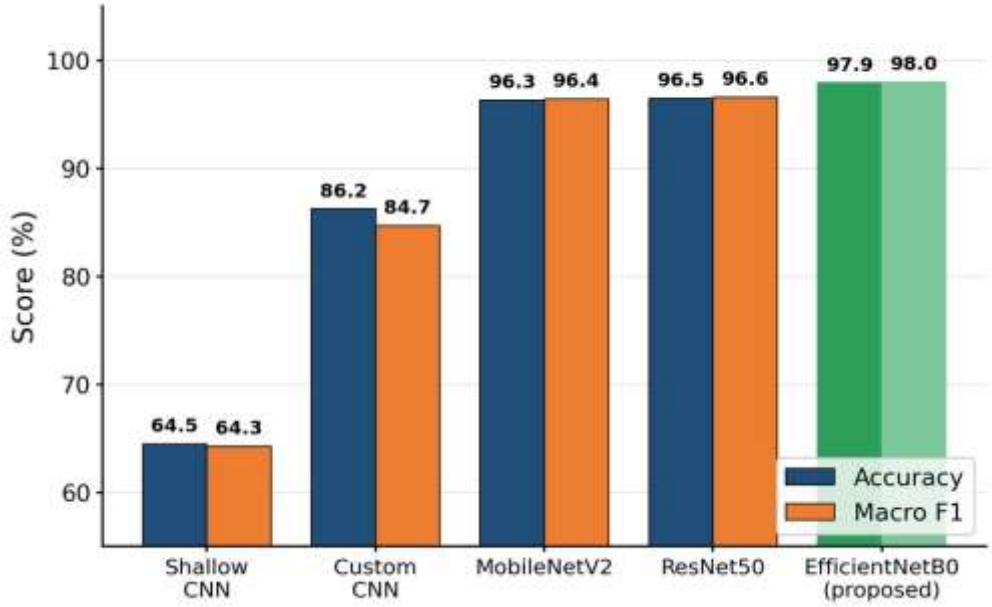

Fig. 8. Accuracy and macro F1-score of the proposed EfficientNetB0 model against four baselines trained under the identical dataset split and protocol.

The from-scratch models lag substantially: the shallow CNN reaches only 64.48% accuracy and the deeper custom CNN 86.23%, confirming that limited medical datasets are insufficient to train competitive networks without transfer learning. The two pretrained backbones close most of this gap, with MobileNetV2 and ResNet50 reaching 96.31% and 96.47% accuracy respectively. The proposed EfficientNetB0 model attains the highest accuracy (97.90%) and macro F1 (97.96%) of all five models. Notably, ResNet50 records a marginally higher AUC (0.9995), reflecting strong ranking ability, but EfficientNetB0 achieves better accuracy and F1 at a fraction of ResNet50's parameter count (approximately 4.30M versus 23.99M for the deployed models; Section V-G), underlining its favourable accuracy-to-efficiency trade-off for deployment.

Model Complexity and Inference Efficiency

Beyond classification quality, deployment feasibility in screening environments depends on model size and prediction latency. Table VII reports the total parameter count of each evaluated model together with its

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

average per-image inference time. Model complexity and inference latency were measured on an Apple M4 CPU system (TensorFlow, GPU unavailable); inference time was averaged over the 1,191-image held-out test set with batch size 1, TTA disabled, and timing recorded around single-image forward passes using Python's time.perf_counter(). Parameter counts correspond to the models as actually constructed and saved (ImageNet backbone with include_top=False plus the custom classification head), and therefore differ slightly from canonical full-ImageNet figures. The proposed EfficientNetB0 model attains the highest accuracy of all five models while requiring approximately one-fifth of the parameters of ResNet50 (4.30M versus 23.99M) at comparable CPU latency (82.66 ms versus 86.39 ms per image), confirming its favourable accuracy-to-efficiency trade-off for clinical decision support and for prospective deployment on resource-constrained hardware. MobileNetV2 remains the fastest pretrained backbone (37.96 ms per image), but at a cost of 1.59 percentage points of accuracy relative to the proposed model.

Table VII Model Complexity And Average Inference Time Per Image

Model	Params (M)	Avg. Inference Time (ms/image)
Shallow CNN	0.028	3.89
Custom CNN	0.424	9.14
MobileNetV2 [10]	2.51	37.96
ResNet50	23.99	86.39
EfficientNetB0 (ours)	4.3	82.66

Cross-Validation Analysis

To quantify the variance of the reported metrics across data partitions, stratified five-fold cross-validation was performed with the EfficientNetB0 architecture, training for five epochs per fold under the same preprocessing and augmentation pipeline; the abbreviated per-fold schedule was adopted for computational tractability, whereas the final deployed model uses the full two-stage warm-up and fine-tuning schedule of Section IV-D. Table VIII reports the mean and standard deviation across the five folds. The model achieved an average accuracy of 97.20 ± 0.39%, macro F1-score of 97.33 ± 0.39%, and macro AUC-ROC of 0.9984 ± 0.0005 (per-fold AUCs: 0.9977, 0.9983, 0.9983, 0.9988, 0.9988), demonstrating stable performance across different data partitions. The low standard deviations indicate that the single held-out split result of Section V-B is not an artefact of a favourable partition.

Table VIII Five-Fold Cross-Validation Results (Mean ± Sd Across Folds)

Metric	Mean ± SD
Accuracy	97.20 ± 0.39%
Macro Precision	97.58 ± 0.25%
Macro Recall	97.13 ± 0.50%
Macro F1-Score	97.33 ± 0.39%
AUC-ROC (Macro)	0.9984 ± 0.0005

Comparative Analysis with Related Work

Table IX contextualises PulmoScan AI against the most closely related published studies. Direct comparison should be interpreted with caution, since these studies use different datasets, class counts, and evaluation protocols. Within the body of recently reported multi-class chest X-ray classification systems, PulmoScan AI demonstrates competitive accuracy and AUC-ROC while using a comparatively lightweight backbone.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

Table IX Comparative Analysis With Related Work

Ref.	Model	Cls.	Acc.	AUC
[4]	VGG19+CNN	6	96.48%	0.998
[5]	UNet+Xception	2 (TB)	99.29%	0.999
[6]	YOLOv8+MNV2	3	84%	0.94
[7]	DenseNet121	3	81%	0.92
Ours	EffNetB0 TL	3	97.90%	0.998

Deployed Clinical Interface

Figs. 9 and 10 show the deployed PulmoScan AI web application. Fig. 9 presents the imaging intake and preprocessing page: the user selects the modality, uploads a chest image (DICOM, PNG, JPG, or NIfTI), previews it in-browser, and triggers backend inference against the trained three-class Keras model. Fig. 10 presents the AI results dashboard for a Tuberculosis prediction: the central viewer renders the Grad-CAM attention overlay with a bounding box around the high-activation region (mid left lung), accompanied by per-class confidence bars (actual model softmax outputs), a clinical probability ring, case metadata, an in-browser image histogram, the AI pipeline log, and a downloadable image report. Zoom, contrast, overlay toggling, and export controls support radiologist-style review of each prediction.

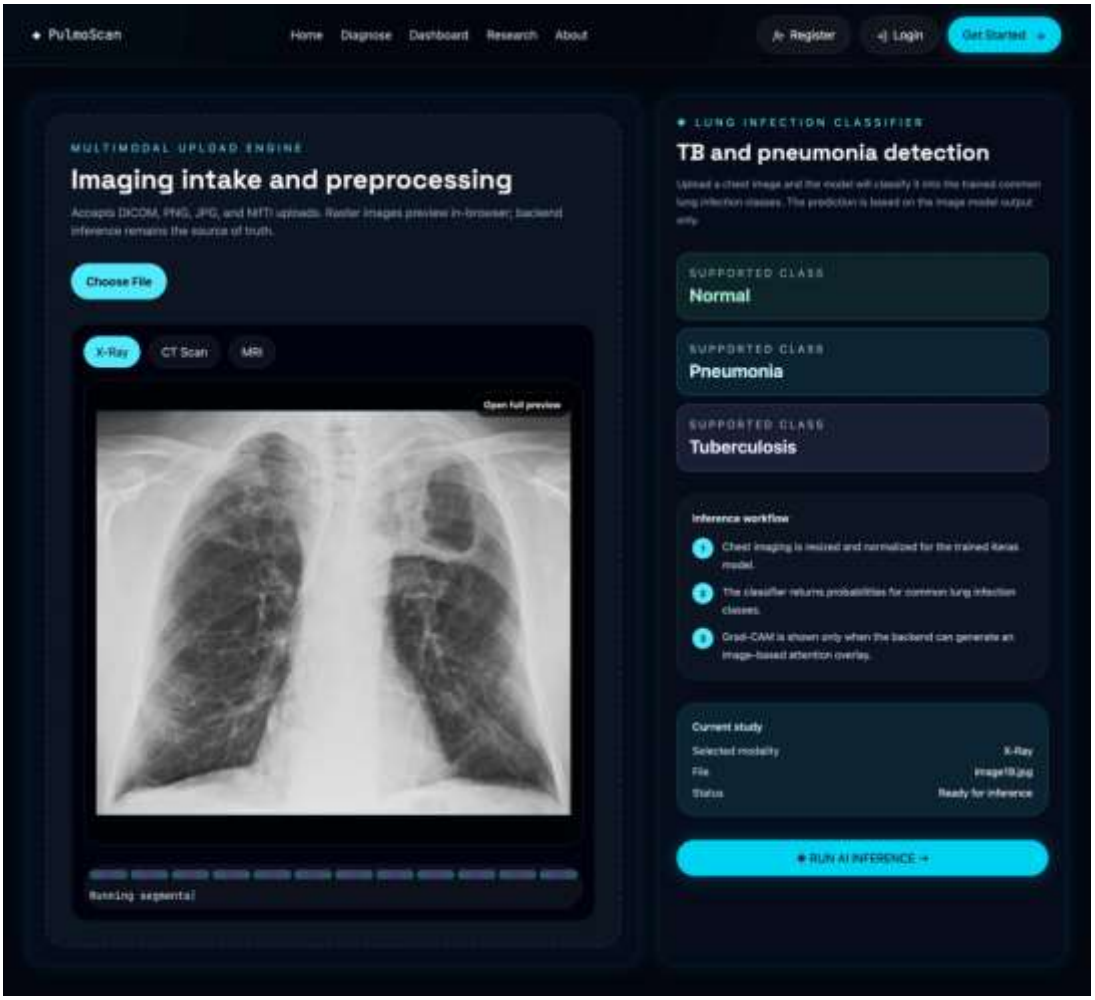

Fig. 9. Deployed web interface: imaging intake and preprocessing page with modality selection, in-browser preview, supported classes, inference workflow, and current study status.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

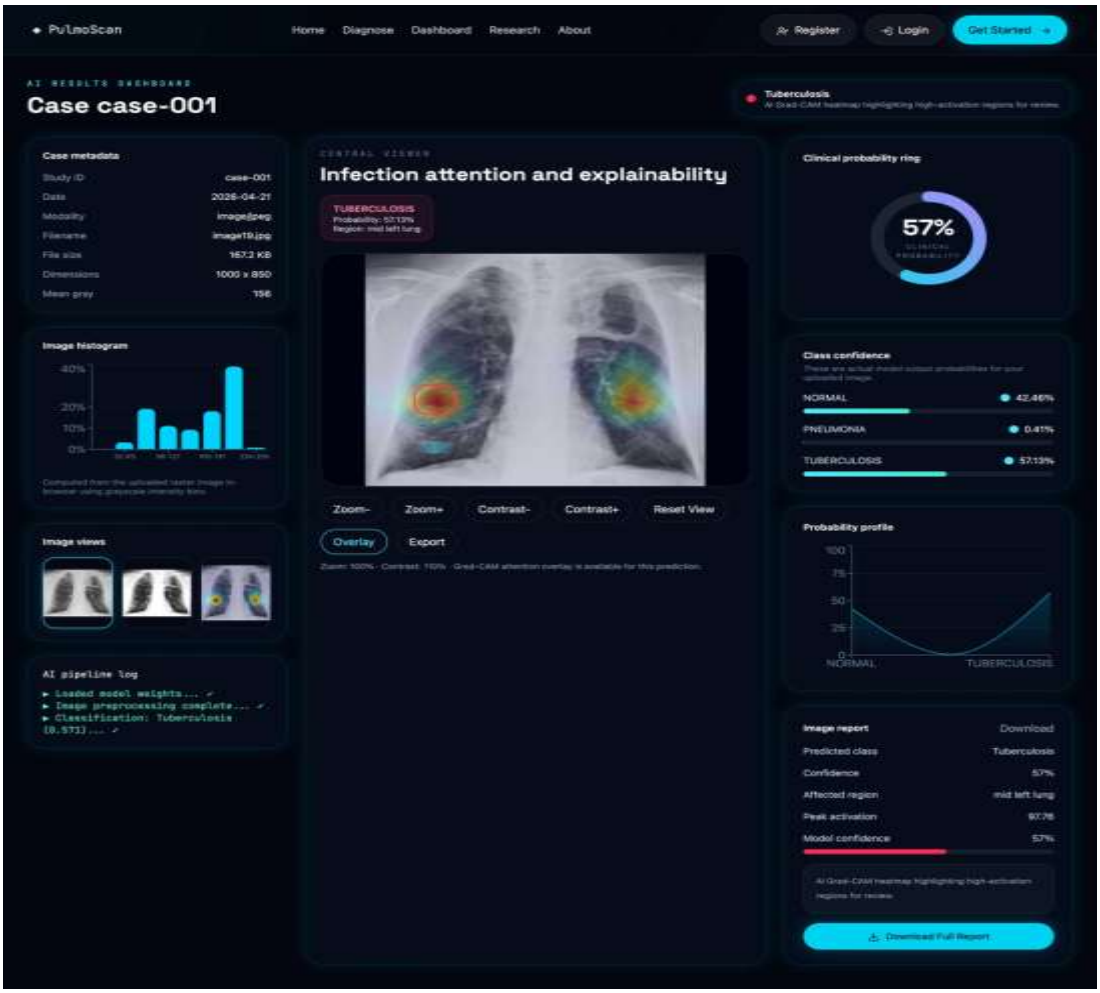

Fig. 10. AI results dashboard for a Tuberculosis prediction: Grad-CAM attention overlay with bounding box and region label, class confidence bars, clinical probability ring, image histogram, pipeline log, and downloadable image report.

DISCUSSION

Analysis of Results

PulmoScan AI demonstrates competitive performance for three-class CXR-based lung disease detection compared with recently reported systems. The macro AUC-ROC of 0.9982 indicates strong class discrimination. The controlled comparison in Section V-F, in which four baseline architectures were trained under an identical protocol, supports the choice of EfficientNetB0: it outperformed both from-scratch CNNs and the MobileNetV2 and ResNet50 transfer-learning baselines on accuracy and macro F1, while remaining markedly more parameter-efficient than ResNet50. The choice of EfficientNetB0 is therefore justified: it balances model size, depth, width, and resolution scaling efficiently, performing well on image classification even with relatively limited medical datasets compared to large natural image collections [9].

The two-phase training strategy was critical: the warm-up phase allowed the classification head to converge before backbone weights were disturbed, while partial fine-tuning of the last 80 layers enabled task-specific feature adaptation without catastrophic forgetting. TTA further improved calibration by reducing prediction variance.

The confidence thresholding mechanism—reporting Inconclusive when confidence is below threshold or the margin between the top-2 predictions is below 12%—provides an important clinical safeguard against overconfident misclassifications. Grad-CAM explainability supports radiologist trust by visually highlighting the anatomical regions driving each prediction.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

LIMITATIONS

Despite strong empirical results, this study has several limitations that bound the scope of its claims:

1) Absence of silicosis-specific data: although occupational lung disease motivates this work, the current model does not detect silicosis or pneumoconiosis, as no sufficiently large public CXR dataset for these conditions was available. The occupational risk module is rule-based and independent of the classifier.

2) No radiologist validation: Grad-CAM heatmaps have not been validated by expert radiological assessment, so the clinical plausibility of the highlighted regions is not yet established.

3) Potential dataset bias: the Normal and Pneumonia images originate from the RSNA repository while the Tuberculosis images come from the NIAID TB Portals dataset. Because the classes are drawn from different repositories, the model may in principle learn repository-specific acquisition characteristics (scanner type, exposure settings, post-processing) rather than disease-specific patterns. Dataset bias therefore cannot be completely ruled out, and the perfect inter-class separation observed for Tuberculosis should be interpreted with this caveat in mind.

4) No external validation: all images derive from two public repositories; performance on multi-institution data with different scanners, populations, and acquisition protocols is unknown and remains to be established through prospective evaluation.

Despite strong performance, the study is therefore limited principally by the absence of external validation and the possibility of dataset-source bias. Accordingly, the model should be interpreted as an experimental decision-support prototype rather than a clinically validated diagnostic system.

Future Work

Several directions follow naturally from the limitations above and the system's extensible architecture:

1) Silicosis and pneumoconiosis detection: incorporate silicosis-specific CXR and HRCT datasets (e.g., occupational cohorts as in [6], [7]) to extend the classifier to the occupational diseases that motivate this work.

2) CT/HRCT integration: extend the pipeline beyond CXR to high-resolution CT, which offers superior sensitivity for early-stage interstitial disease.

3) Expert validation of explainability: conduct radiologist assessment of Grad-CAM heatmaps following the protocol of [6] to quantify localisation agreement.

4) Multi-hospital external validation: evaluate on independent multi-institution datasets to establish generalisability beyond the two source repositories.

5) Federated learning: enable privacy-preserving training across hospitals without centralising patient data.

6) Conversational clinical assistant: integrate a medical chatbot into the dashboard to explain predictions, answer screening-protocol questions, and guide referral decisions.

CONCLUSION

This paper presented PulmoScan AI, a deep learning-based clinical decision support system for automated multi-class detection of lung diseases from chest X-ray images. The EfficientNetB0 transfer learning framework, trained on 11,910 CXR images with two-stage fine-tuning, label smoothing, TTA, and confidence thresholding, achieved 97.9% test accuracy, 0.9982 macro AUC-ROC, 98.28% macro precision, and 97.96% macro F1-score. Five-fold cross-validation further confirmed the stability of these results across data

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

partitions (97.20 ± 0.39% mean accuracy; macro AUC-ROC 0.9984 ± 0.0005). Perfect Tuberculosis precision (100%) and near-perfect Normal recall (99.59%) are clinically significant results.

The full-stack deployment with Flask REST API, React frontend, Grad-CAM explainability, and occupational risk assessment module provides a deployable screening tool for real-world clinical environments. As outlined in Section VII, future work will extend the system to silicosis and pneumoconiosis, incorporate CT imaging, and pursue expert and multi-institution validation.

ACKNOWLEDGMENTS

The authors sincerely thank Dr. Yogesh Golhar for his guidance and supervision throughout this work, and the faculty of the Department of Computer Engineering, St. Vincent Pallotti College of Engineering and Technology, Nagpur, for their support.

REFERENCES

1. P. Cullinan et al., “Occupational lung diseases: from old and novel exposures to effective preventive strategies,” Lancet Respir. Med., vol. 5, no. 5, pp. 445–455, 2017.
2. R. F. Hoy et al., “Current global perspectives on silicosis—convergence of old and newly emergent hazards,” Respirology, vol. 27, no. 6, pp. 387–398, 2022.
3. P. Howlett, A. Durairaj, M. Lesosky, and J. Feary, “Diagnostic accuracy of chest X-ray screening for silicosis: a systematic review, meta-analysis and modelling study,” Occup. Environ. Med., 2025, doi: 10.1136/oemed-2024-109985.
4. G. M. M. Alshmrani, Q. Ni, R. Jiang, H. Pervaiz, and N. M. Elshennawy, “A deep learning architecture for multi-class lung diseases classification using chest X-ray (CXR) images,” Alexandria Eng. J., vol. 64, pp. 923–935, 2023.
5. V. Sharma, Nillmani, S. K. Gupta, and K. K. Shukla, “Deep learning models for tuberculosis detection and infected region visualization in chest X-ray images,” Intell. Med., vol. 4, no. 2, pp. 104–113, 2024, doi: 10.1016/j.imed.2023.06.001.
6. B. Priego-Torres, D. Sanchez-Morillo, E. Khalili, M. Á. Conde-Sánchez, A. García-Gámez, and A. León-Jiménez, “Automated engineered-stone silicosis screening and staging using deep learning with X-rays,” Comput. Biol. Med., vol. 191, art. 110153, 2025, doi: 10.1016/j.compbiomed.2025.110153.
7. L. Zhang et al., “A deep learning-based model for screening and staging pneumoconiosis,” Sci. Rep., vol. 11, art. 2201, 2021.
8. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626.
9. M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105–6114.
10. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 4510–4520.
11. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.
12. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 4700–4708.
13. C. Gu and M. Lee, “Deep transfer learning using real-world image features for medical image classification, with a case study on pneumonia X-ray images,” Bioengineering, vol. 11, no. 4, art. 406, 2024.
14. L. Devnath, P. Summons, S. Luo, D. Wang, K. Shaukat, I. A. Hameed, and H. Aljuaid, “Computer-aided diagnosis of coal workers’ pneumoconiosis in chest X-ray radiographs using machine learning: a systematic literature review,” Int. J. Environ. Res. Public Health, vol. 19, no. 11, art. 6439, 2022, doi: 10.3390/ijerph19116439.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VII, July 2026

15. W. Sun et al., “A fully deep learning paradigm for pneumoconiosis staging on chest radiographs,” IEEE J. Biomed. Health Inform., vol. 26, no. 10, pp. 5154–5164, 2022.
16. Radiological Society of North America, “RSNA Pneumonia Detection Challenge,” Kaggle, 2018. [Online]. Available: https://www.kaggle.com/c/rsna-pneumonia-detection-challenge
17. National Institute of Allergy and Infectious Diseases (NIAID), “TB Portals Program,” U.S. Dept. of Health and Human Services. [Online]. Available: https://tbportals.niaid.nih.gov
18. N. Veeramani et al., “NextGen lung disease diagnosis with explainable artificial intelligence,” Sci. Rep., vol. 15, art. 33052, 2025, doi: 10.1038/s41598-025-07603-4.
19. D. S. Kermany et al., “Identifying medical diagnoses and treatable diseases by image-based deep learning,” Cell, vol. 172, no. 5, pp. 1122–1131.e9, 2018, doi: 10.1016/j.cell.2018.02.010.
20. M. F. Hashmi, S. Katiyar, A. G. Keskar, N. D. Bokde, and Z. W. Geem, “Efficient pneumonia detection in chest X-ray images using deep transfer learning,” Diagnostics, vol. 10, no. 6, art. 417, 2020, doi: 10.3390/diagnostics10060417.