

Multi-class Lung Disease Diagnosis Using Chest X-Ray with XAI

by

**S.M. Asif Maharab (221014013)
Rifat Mahmud (221014046)
Tamanna Akter (221014062)
Soyab Ansari (231014011)**

*Capstone project report (CSE 4098C) submitted in partial fulfillment of the
requirements for the degree of*

Bachelor of Science in Computer Science and Engineering

Under the supervision of

Prof. Muhammad Golam Kibria, PhD



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
UNIVERSITY OF LIBERAL ARTS BANGLADESH

SPRING 2026



Department of Computer Science and Engineering
University of Liberal Arts Bangladesh
688 Beribadh Road, Mohammadpur
Dhaka - 1207

APPROVAL OF THE COMMITTEE

This is to certify that the CSE 4098C Project Report entitled **Multi-class Lung Disease Diagnosis Using Chest X-Ray with XAI**, submitted by **Md. Rifat Mahmud** (221014046), **Tamanna Akter** (221014062), **S. M. Asif Mahrab** (221014013), and **Soyab Ansari** (231014011) are undergraduate students of the **Department of Computer Science and Engineering** has been examined. We hereby accord our approval of it as the presented work is satisfactory.

Place: Dhaka

Date: June 7, 2026

Prof. Muhammad Golam Kibria
Professor & HoD, Department of Computer
Science and Engineering

Raihan Kibria
Senior Lecturer, Department of Computer Science
and Engineering (CSE)

Ms. Rubaiya Hafiz
Senior Lecturer, Bachelor of Science, Computer
Science and Engineering (CSE)

DECLARATION

Project Title Multi-class Lung Disease Diagnosis Using Chest X-Ray with XAI
Authors *Md.Rifat Mahmud, Tamanna Akter, S.M. asif Maharab, and Soyab Ansari*
Student IDs 221014046, 221014062, 22101413, and 231014011
Supervisor Prof. Muhammad Golam Kibria, PhD

We declare that this capstone project report entitled *Title of Capstone Project* is the result of our own work except as cited in the references. The capstone project report has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

S.M. Asif Maharab
221014013

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Rifat Mahmud
221014046

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Tamanna Akter
221014062

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Soyab Ansari
231014011

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Date: May 07, 2026



Department of Computer Science and Engineering
University of Liberal Arts Bangladesh
Mohammadpur, Dhaka - 1207

CERTIFICATE

This is to certify that the capstone project report entitled **Multi-class Lung Disease Diagnosis Using Chest X-Ray with XAI**, submitted by **Md. Rifat Mahmud** (Student ID: 221014046), **Tamanna Akter** (Student ID: 21014062), **S.M. Asif Maharab** (Student ID: 21014013) and **Soyab Ansari** (Student ID: 231014011) are undergraduate students of the **Department of Computer Science and Engineering** has been examined. Upon recommendation by the examination committee, we hereby accord our approval of it as the presented work and submitted report fulfill the requirements for its acceptance in partial fulfillment for the degree of *Bachelor of Science in Computer Science and Engineering*.

Prof. Muhammad Golam Kibria
HoD, CSE

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Prof. Muhammad Golam Kibria
Professor & HoD

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Place: Dhaka

Date: May 07, 2026

ACKNOWLEDGEMENTS

We would like to express our deep and sincere gratitude to our research supervisor, *Prof. Muhammad Golam Kibria, PhD*, for giving us the opportunity to conduct research and providing invaluable guidance throughout this work. His dynamism, vision, sincerity and motivation have deeply inspired us. He has taught us the methodology to carry out the work and to present the works as clearly as possible. It was a great privilege and honor to work and study under his guidance.

We are greatly indebted to our honorable teachers of the Department of Computer Science and Engineering at the University of Liberal Arts Bangladesh who taught us during the course of our study. Without any doubt, their teaching and guidance have completely transformed us to the persons that we are today.

We are extremely thankful to our parents for their unconditional love, endless prayers and caring, and immense sacrifices for educating and preparing us for our future. We would like to say thanks to our friends and relatives for their kind support and care.

Finally, we would like to thank all the people who have supported us to complete the project work directly or indirectly.

S.M. Asif Maharab, Rifat Mahmud, Tamanna Akter and Soyab Ansari

University of Liberal Arts Bangladesh

Date: May 07, 2026

I Will Rise Quietly Until My Efforts
Speak Louder Than Any Words I Ever
Said And My Dreams Turn Into
Reality

– *S.M. Asif Maharab*

To my beloved parents
the source of my inspiration

– *Rifat Mahmud*

This report is dedicated to my
beloved parents and respected faculty

– *Tamanna Akter*

Dedicated to my lifeline and shield
for their love and support. Thanks to
my teammates, friends, and inspiring
university community

– *Soyab Ansari*

ABSTRACT

Lung diseases — encompassing bacterial pneumonia, viral pneumonia, COVID-19, tuberculosis, and healthy (normal) conditions — represent a critical global health burden responsible for millions of deaths each year. Accurate, timely, and interpretable diagnosis from chest X-ray images is therefore of paramount clinical importance, yet remains challenging even for experienced radiologists due to overlapping visual presentations among disease classes. This capstone project presents LungAI, an Explainable Multi-Class Deep Learning framework for automated diagnosis of five lung conditions from frontal chest X-ray images. The system is built upon DenseNet201, a 201-layer densely connected convolutional network pre-trained on ImageNet, augmented with a custom three-layer classification head consisting of Dense(1024)-BatchNorm-Dropout(0.5), Dense(512)-BatchNorm-Dropout(0.3), and Dense(256)-Dropout(0.2) layers, culminating in a five-class softmax output. A two-phase transfer-learning strategy was adopted: Phase 1 trains only the custom head with all backbone layers frozen ($LR = 1 \times 10^{-3}$, 25 epochs), after which Phase 2 unfreezes the last 100 backbone layers and applies cosine-annealed fine-tuning ($LR = 1 \times 10^{-5}$, up to 50 epochs). The training dataset comprises 39,115 images drawn from NIH, RSNA, Kaggle, and Montgomery TB repositories, preprocessed with CLAHE (Contrast-Limited Adaptive Histogram Equalization) contrast enhancement in LAB colour space to accentuate diagnostically relevant tissue-density variations. Class imbalance is mitigated through effective-number class weighting ($\beta = 0.9999$) and a $1.5\times$ boost for Bacterial Pneumonia, combined with oversampling to equalise training-batch frequencies. On the independent test set of 3,903 images the model achieves a test accuracy of 90.93 %, macro F1-score of 90.30 %, Cohen’s Kappa of 0.8538, and Matthews Correlation Coefficient (MCC) of 0.8551. Notable per-class results include a perfect 100 % recall for Tuberculosis, 95 % recall for COVID-19, and an area under the ROC curve (AUC) of 1.000 for Tuberculosis and 0.999 for Viral Pneumonia. Explainability is addressed through three complementary techniques — Grad-CAM, SHAP, and LIME — each of which produces clinically meaningful attention maps that localise discriminative radiological features such as bilateral

ground-glass opacities in COVID-19, upper-lobe cavitary patterns in Tuberculosis, and lobar consolidation in Bacterial Pneumonia. The trained model is deployed as a Flask web application accessible via ngrok, enabling clinicians to upload a chest X-ray and receive an instant diagnosis together with Grad-CAM and LIME visualisations. The results demonstrate that combining rigorous data engineering, two-phase transfer learning, effective class-balance handling, and multi-method explainability yields a diagnostic system that is both highly accurate and interpretable — properties essential for trustworthy clinical decision support.

Keywords: lung disease diagnosis, DenseNet201, transfer learning, CLAHE, Grad-CAM, SHAP, LIME, explainable AI, chest X-ray, multi-class classification, deep learning.

Contents

1	Introduction	1
1.1	Problem statement	2
1.2	Aims, objectives and Motivation	3
1.2.1	Aims	3
1.2.2	Objective	3
1.2.3	Motivation	4
1.3	Project Specification	4
2	Literature Review	6
2.1	Convolutional Neural Networks for Respiratory Disease Detection .	6
2.2	Transfer Learning Strategies	7
2.3	Explainable AI in Medical Imaging	7
2.4	Preprocessing and Data Augmentation	8
2.5	Handling Class Imbalance	8
2.6	Multi-Class Diagnosis Systems	9
2.7	Gaps Addressed by the Present Work	9
3	Methodology and System Design	10
3.1	Dataset Collection and Description	10
3.2	Data Preprocessing	13
3.2.1	Image Resizing and Colour Space Conversion	13
3.2.2	CLAHE Contrast Enhancement	13
3.2.3	Data Augmentation	14
3.3	Model Architecture	16
3.4	Evaluation Metrics	19

4	Implementation	21
4.1	Training Strategy	21
4.1.1	Two-Phase Transfer Learning	21
4.1.2	Class Imbalance Handling	23
4.1.3	Callbacks and Regularisation	24
4.1.4	Training Infrastructure	24
4.2	Explainable AI Implementation	25
4.2.1	Gradient-weighted Class Activation Mapping (Grad-CAM)	25
4.2.2	SHAP (SHapley Additive exPlanations)	26
4.2.3	LIME (Local Interpretable Model-agnostic Explanations)	27
4.2.4	XAI Panel Visualisations	28
4.3	Deployment	31
4.3.1	System Architecture and Workflow	31
4.3.2	External Accessibility via ngrok	34
4.3.3	Web Application Interface	34
4.3.4	Inference Latency and Constraints	36
5	Results Analysis and Discussion	37
5.1	Overall Performance Metrics	37
5.2	Per-Class Performance Analysis	39
5.2.1	Tuberculosis	41
5.2.2	Viral Pneumonia	41
5.2.3	Normal	41
5.2.4	COVID-19	42
5.2.5	Bacterial Pneumonia	42
5.3	Confusion Matrix Analysis	43
5.4	ROC and Precision-Recall Curves	46
5.5	Per-Class Confidence Distributions	46
5.6	Comparison with Prior Work	50
5.7	Interpretation of Results	52
5.8	Clinical Significance and Impact	53
5.9	Limitations	54
5.10	Ethical, Legal, and Environmental Considerations	55
5.10.1	Ethical Considerations	55

5.10.2	Privacy and Legal Compliance	55
5.10.3	Environmental Impact	56
6	Conclusions	57
6.1	Summary of Contributions	57
6.2	Achievement of Objectives	58
6.3	Future Work and Recommendations	58
6.3.1	Multi-Modal Integration	58
6.3.2	Clinical Metadata Integration	59
6.3.3	Advanced Explainability	59
6.3.4	Addressing Bacterial Pneumonia Differentiation	59
6.3.5	Prospective Clinical Validation	59
6.3.6	Fairness, Bias Mitigation, and Regulatory Compliance	60
6.4	Final Remarks	60
	References	61

List of Figures

3.1	Four-stage LungAI methodology pipeline: Data Preparation → Two-Phase Training → Explainable AI → Deployment.	11
3.2	Dataset class distribution across Train, Validation, and Test splits. Each class contributes an equal number of images to every split, ensuring balanced evaluation.	12
3.3	Sample dataset random.	13
3.4	LungAI model architecture: CLAHE-enhanced 224×224 input → DenseNet201 backbone ($7 \times 7 \times 1920$ feature maps) → custom classification head → 5-class softmax output.	17
4.1	Training history — accuracy (left) and loss (right) across 75 epochs. The green dashed line marks the Phase 1/Phase 2 boundary at epoch 25. Best validation accuracy: 91.70%.	23
4.2	XAI panel — COVID-19 (correctly classified). Grad-CAM highlights bilateral peripheral and lower-zone opacities consistent with ground-glass haziness. LIME confirms the bilateral lung fields as the primary supporting regions.	28
4.3	XAI panel — COVID-19 (correctly classified). Attention is focused on the central and lower lung zones, consistent with the typical bilateral involvement seen in COVID-19 pneumonia.	29
4.4	XAI panel — Normal (correctly classified). Grad-CAM attention is diffuse and low-intensity, reflecting the absence of focal pathological findings. LIME shows no strongly supporting superpixel regions, consistent with a normal radiograph.	29

4.5	XAI panel — Tuberculosis (correctly classified). Grad-CAM prominently highlights the upper-lobe regions, consistent with the upper-lobe predominant nodular-cavitary pattern characteristic of active pulmonary tuberculosis.	30
4.6	XAI panel — Viral Pneumonia (correctly classified). Attention is distributed across both lung fields with a perihilar predominance, reflecting the interstitial and bilateral distribution typical of viral pneumonia.	30
4.7	XAI panel — Viral Pneumonia (correctly classified). LIME superpixels supporting the prediction align with the areas of increased interstitial markings visible on the original radiograph.	31
4.8	System flow — clinician uploads CXR → Flask app preprocesses with CLAHE → DenseNet201 predicts → Grad-CAM heatmap and confidence scores displayed → result returned to the clinician.	33
4.9	LungAI web application — screenshot showing COVID-19 diagnosis (82.49 % confidence) with original X-ray, Grad-CAM overlay, and LIME superpixel explanation. The confidence bar chart (centre) shows per-class probability scores for all five disease classes.	35
5.1	Per-class Precision, Recall, and F1-Score bar chart. The purple dashed line indicates the Macro F1 of 90.3 %. Tuberculosis achieves the highest F1 at 98.4 %; Bacterial Pneumonia is the weakest class at 83.3 %.	39
5.2	Classification report heatmap displaying precision, recall, and F1-score for all five classes. Darker red cells indicate higher scores. The heatmap makes it immediately visible that Tuberculosis dominates while Bacterial Pneumonia lags behind the other classes.	40
5.3	Confusion matrix (raw counts) for the LungAI test set. Diagonal entries are correct predictions; off-diagonal entries are misclassifications. Test accuracy: 90.93 %.	44
5.4	Confusion matrix (row-normalised percentages). Each row sums to 100 %. Tuberculosis achieves a perfect 100.0 % recall with zero off-diagonal entries in its row.	45

5.5	ROC curves (one-vs-rest) for all five classes. AUC values: Tuberculosis = 1.000, Viral Pneumonia = 0.999, COVID-19 = 0.994, Normal = 0.979, Bacterial Pneumonia = 0.977. All five classes substantially exceed the random baseline (dashed diagonal at AUC = 0.500). . . .	47
5.6	Precision-Recall curves for all five classes. AP values: Tuberculosis = 0.999, Normal = 0.978, COVID-19 = 0.971, Viral Pneumonia = 0.970, Bacterial Pneumonia = 0.946. Dotted lines indicate the random baseline per class.	48
5.7	Per-class confidence distribution histograms for correctly classified test images. Mean confidence: TB = 100.0 %, COVID-19 = 92.1 %, Normal = 88.8 %, Viral Pneumonia = 87.6 %, Bacterial Pneumonia = 76.1 %.	48

List of Tables

4.1	Training configuration summary for Phase 1 and Phase 2.	25
5.1	Global performance metrics for LungAI on the independent test set (3,903 images, run_004).	38
5.2	Per-class precision, recall, F1-score, average precision (AP), and test support for all five disease classes.	39
5.3	Comparison of LungAI with recent lung disease classification liter- ature. CT-based systems and ensemble methods are noted where applicable.	50
5.4	Comparison of model architectures for lung disease classification. All metrics are reported on the test set.	51

Chapter 1

Introduction

Lung diseases represent one of the most pressing public health challenges of the twenty-first century. Conditions such as bacterial pneumonia, viral pneumonia, COVID-19, tuberculosis (TB), and other respiratory infections collectively account for approximately 4.2 million deaths per year according to the World Health Organization, [Sardar et al. \(2023\)](#) making respiratory disease the third-leading cause of mortality worldwide [Perla et al. \(2025\)](#). Pneumonia alone hospitalises more than one million patients annually in the United States; globally it remains the single largest infectious cause of death in children under five [Perla et al. \(2025\)](#). Tuberculosis, caused by *Mycobacterium tuberculosis*, infects roughly one-quarter of the world's population in a latent form, with about 10 million active cases per year [Mahadeva et al. \(2024\)](#). The COVID-19 pandemic demonstrated with devastating clarity how rapidly a novel respiratory pathogen can overwhelm healthcare systems that depend on manual, specialist-driven diagnosis. Chest X-ray (CXR) radiography is the primary first-line imaging modality for lung disease assessment. It is inexpensive, universally available, rapid to acquire, and delivers a relatively low radiation dose. Nevertheless, interpreting CXRs requires substantial radiological expertise; subtle or overlapping signs — such as the bilateral peripheral ground-glass haziness of COVID-19, the upper-lobe nodular-cavitary pattern of TB, or the lobar consolidation of bacterial pneumonia — can be easily misread or confused [Bhargav & Ashokkumar \(2024\)](#). In resource-limited settings where trained radiologists are scarce, misdiagnosis rates are alarming, delaying treatment and worsening outcomes. Artificial intelligence (AI), and deep learning (DL) in particular, has emerged as a promising solution. Convolutional Neural Networks (CNNs) can learn hierarchical visual representations directly from labelled image data and,

when trained on sufficiently large datasets, match or exceed radiologist-level accuracy on many imaging tasks [B et al. \(2023\)](#). Transfer learning — initialising a CNN with weights pre-trained on a large natural-image corpus such as ImageNet and then fine-tuning on medical images — has become the dominant paradigm, dramatically reducing the amount of annotated medical data required [Selvaraj et al. \(2025\)](#). Despite impressive accuracy figures, clinical adoption of AI-based diagnostic tools has been slow. A key barrier is the ‘black-box’ nature of deep learning: radiologists and clinicians are understandably reluctant to trust a system that cannot explain its predictions in terms of recognisable radiological signs. Explainable AI (XAI) techniques such as Grad-CAM (Gradient-weighted Class Activation Mapping), SHAP (SHapley Additive exPlanations), and LIME (Local Interpretable Model-agnostic Explanations) address this by generating saliency or importance maps that highlight which image regions most influenced a particular prediction [Karim et al. \(2020\)](#). When these maps align with clinically meaningful anatomy — for example, Grad-CAM highlighting the peripheral lung fields in a COVID-19 case — they provide the interpretive bridge that allows clinicians to validate and trust the AI’s decision.

1.1 Problem statement

Current deep-learning approaches to lung disease diagnosis face four inter-related limitations that curtail their clinical utility:

- Most published models address binary classification (diseased vs. healthy, or one specific condition vs. the rest) rather than simultaneous multi-class discrimination across clinically common respiratory conditions ([Chouat et al., 2025](#)).
- Models frequently operate as opaque ‘black boxes,’ providing probability scores without visual or conceptual justification, which undermines clinician trust and prevents validation of AI reasoning ([Karim et al., 2020](#)).
- The challenge of differentiating radiologically similar conditions — especially bacterial vs. viral pneumonia — remains unsolved at a clinically acceptable level ([Hasan et al., 2024](#)).

- Few systems integrate explainability directly into an accessible clinical interface; XAI outputs are typically presented in isolation from the diagnostic workflow (Pitroda et al., 2021).

1.2 Aims, objectives and Motivation

1.2.1 Aims

The primary aim of this project is to develop a robust, interpretable, and deployable deep learning system — LungAI — that accurately classifies chest X-ray images into five categories (Bacterial Pneumonia, COVID-19, Normal, Tuberculosis, Viral Pneumonia) while generating clinically meaningful visual explanations for each prediction, thereby enabling trustworthy AI-assisted diagnostic decision support.

1.2.2 Objective

- Construct and preprocess a large, balanced multi-source dataset (NIH, RSNA, Kaggle, Montgomery TB) totalling 39,115 images across five classes, applying CLAHE contrast enhancement for image quality improvement.
- Design and train a DenseNet201-based classifier using a two-phase transfer-learning strategy that achieves a test accuracy exceeding 90 % and a macro F1-score above 90 % on an independent held-out test set.
- Implement and integrate three complementary XAI methods — Grad-CAM, SHAP, and LIME — producing per-image visual explanations aligned with known radiological patterns.
- Generate 200 XAI panel visualisations (40 per class) demonstrating clinically interpretable attention maps for qualitative validation.
- Deploy the trained model as a Flask web application with a live Grad-CAM and LIME interface accessible to clinicians via a public URL.
- Evaluate the system comprehensively using accuracy, precision, recall, F1-score, AUC-ROC, average precision, Cohen’s Kappa, and MCC metrics.

1.2.3 Motivation

The motivation for this project is both clinical and societal. In Bangladesh and many lower-middle-income countries, the radiologist-to-patient ratio is critically low. Emergency departments, district hospitals, and rural clinics frequently lack immediate access to specialist imaging expertise. An AI system that can rapidly triage a CXR and provide a transparent, evidence-based preliminary diagnosis could meaningfully reduce the time-to-treatment for life-threatening conditions such as TB or COVID-19 pneumonia. Beyond resource-limited settings, even in well-resourced hospitals a reliable second opinion can reduce diagnostic errors. The integration of XAI makes the system auditable — clinicians can verify that the AI’s reasoning aligns with what they would themselves expect to see — and promotes responsible adoption rather than blind automation.

1.3 Project Specification

The following design specifications define the measurable success criteria for the LungAI system. Each specification corresponds to a quantifiable requirement derived from the clinical and technical objectives of the project.

Classification Performance. The system shall achieve a minimum test accuracy of 90% and a macro F1-score of at least 90% on an independent held-out test set of chest X-ray images spanning five disease classes. These thresholds were selected based on clinically acceptable screening performance benchmarks from the lung disease classification literature.

Dataset Scale and Balance. The training dataset shall comprise a minimum of 25,000 chest X-ray images distributed across five classes, sourced from at least three publicly available repositories. All classes shall be balanced to within $\pm 5\%$ of equal representation in the training split through oversampling and class weighting, ensuring no single class dominates gradient updates during training.

Image Preprocessing. All input images shall be resized to 224×224 pixels and enhanced using CLAHE (Contrast-Limited Adaptive Histogram Equalization) in the LAB colour space with a clip limit of 2.0 and tile grid size of 8×8 , applied identically across training, validation, test, and inference pipelines.

High-Sensitivity Classes. For clinically critical conditions where false negatives carry severe consequences, the system shall achieve a minimum recall of 95%

for Tuberculosis and 90% for COVID-19 on the independent test set, prioritising sensitivity over precision for these classes.

Explainability Coverage. The system shall generate visual explanations using a minimum of two complementary XAI methods for every prediction. At least 200 annotated XAI panels (40 per class across five classes) shall be produced on test images to enable qualitative radiological validation of model attention.

Inference Latency. The deployed system shall return a predicted class label, per-class confidence scores, and a Grad-CAM heatmap overlay within 5 seconds of image submission under standard GPU hardware, suitable for use in a clinical triage workflow.

Deployment Accessibility. The trained model shall be deployed as a web application accessible via a public URL, accepting chest X-ray images in JPG or PNG format up to 16 MB, without requiring any local software installation by the end user.

Discriminative Reliability. The system shall achieve a Cohen's Kappa coefficient of at least 0.80 and an AUC-ROC of at least 0.97 for each of the five classes on the independent test set, confirming that performance exceeds chance agreement and maintains strong ranking ability across decision thresholds.

Chapter 2

Literature Review

The intersection of deep learning and chest radiograph interpretation has been one of the most active sub-fields of medical AI over the past decade. This chapter surveys the key architectural, methodological, and explainability trends that inform the design of LungAI.

2.1 Convolutional Neural Networks for Respiratory Disease Detection

Early CNN-based studies established proof-of-concept for pneumonia detection from CXRs, typically framing the task as binary (normal vs. pneumonia). [B et al. \(2023\)](#). (2023) compared VGG-19, InceptionV3, and ResNet50 on a binary respiratory disease prediction task, with ResNet50 achieving 98 % accuracy and demonstrating the superiority of residual learning architectures. [Sardar et al. \(2023\)](#) developed Covi-Net, achieving 97.6 % accuracy at distinguishing COVID-19 from non-COVID pneumonia on CT scans, highlighting that architectural design and task specificity jointly determine performance ceilings. The shift from binary to multi-class diagnosis introduced fresh challenges. [Chouat et al. \(2025\)](#) benchmarked VGG-16, DenseNet-121, and DenseNet-201 on a multi-class lung disease classification task, finding DenseNet-201 significantly outperformed the others with 83 % accuracy — attributing the advantage to the dense connectivity pattern that promotes feature reuse across layers, which is especially beneficial when subtle inter-class differences must be resolved. [Bukhari & Fahad \(2022\)](#) corroborated this, reporting 96 % accuracy and a 0.96 F1-score for a three-class (pneumonia, lung cancer, TB) DenseNet-201

model. These findings motivated the selection of DenseNet-201 as the backbone in the present work.

2.2 Transfer Learning Strategies

Transfer learning has become indispensable in medical imaging research because large annotated datasets are expensive and time-consuming to produce. [Selvaraj et al. \(2025\)](#) used ResNet-101 pre-trained on ImageNet for a four-class lung disease classification task, demonstrating substantial reductions in training time without sacrificing accuracy. [Sreena & Ponraj \(2023\)](#) fine-tuned DenseNet-201 on the COVID-QU-Ex dataset, achieving 95.19 % accuracy; their ablation study showed that the choice of optimiser (SGD outperformed Adam and RMSProp in that setting) and the size of the unfrozen backbone region critically affected final performance. A two-phase training protocol — first training only the classification head with the backbone frozen, then selectively unfreezing upper backbone layers for fine-tuning — has been widely adopted because it prevents the gradient updates from distorting the low-level features learned during ImageNet pre-training [Bhargav & Ashokkumar \(2024\)](#); [Perla et al. \(2025\)](#). The present work implements this protocol with phase-specific learning rates and cosine annealing in the fine-tuning phase, combining insights from multiple prior studies [K. H. R et al. \(2021\)](#).

2.3 Explainable AI in Medical Imaging

The clinical-adoption barrier posed by opaque models has driven substantial research into XAI for medical imaging. [Karim et al. \(2020\)](#) introduced DeepCOVID-Explainer, which integrates Grad-CAM++ and Layer-wise Relevance Propagation (LRP) to highlight lung regions critical to a COVID-19 prediction, achieving 96.12 % positive predictive value while generating visualisations validated by radiologists. [Pitroda et al. \(2021\)](#) compared LRP, Deep Taylor Decomposition, Guided Backpropagation, and LIME, concluding that LRP and DTD generate the most clinically meaningful heatmaps for lung conditions. [B et al. \(2023\)](#) emphasised that Grad-CAM visualisations build clinician trust by transparently showing the model’s reasoning, and that the absence of such explanations is a primary reason AI tools are rejected in clinical audits. [Abubaker et al. \(2023\)](#) demonstrated that

Grad-CAM overlays can assess not only disease presence but also severity [Abubaker et al. \(2023\)](#), by quantifying the proportion of the lung field covered by the heatmap. Notably, few published systems integrate more than one XAI technique; the present work addresses this gap by combining Grad-CAM, SHAP, and LIME to provide complementary perspectives on model decisions.

2.4 Preprocessing and Data Augmentation

Preprocessing quality has a measurable impact on model performance. [Kanagaraj et al. \(2025\)](#) demonstrated that CNN classification accuracy improved from 87 % to 93.25 % when image enhancement was applied prior to training, using random rotations, flips, and zoom augmentation. CLAHE has been widely adopted for chest X-ray enhancement because it locally equalises histogram contrast without amplifying global noise — particularly important for X-rays where relevant tissue-density differences are subtle [Pitroda et al. \(2021\)](#). [Bushra et al. \(2023\)](#) found that ResNet50 accuracy on binary classification improved significantly when images were segmented with U-Net before classification, suggesting that directing the model’s attention to the lung region is beneficial. The present work uses CLAHE preprocessing in the LAB colour space on all images, with a clip limit of 2.0 and a tile grid of 8×8 , following the protocol shown to be effective in prior work.

2.5 Handling Class Imbalance

Medical datasets are inherently imbalanced. [Bukhari & Fahad \(2022\)](#) used Deep Convolutional GANs to synthesise minority-class images, improving DenseNet-201 performance from 91 % to 96 %. [Irtaza et al. \(2024\)](#) reported a 5 % F1-score improvement on the NIH ChestX-ray14 multi-label dataset when synthetic images from DCGAN were included. The present study addresses imbalance through two complementary mechanisms: effective-number class weighting — which assigns higher loss weights to minority classes using a smoothed inverse-frequency formula with $\beta = 0.9999$ — and training-set oversampling to equalise per-class sample counts, thereby preventing the model from being biased toward the majority (Normal) class.

2.6 Multi-Class Diagnosis Systems

Hasan et al. (2024)) proposed CTXNET, achieving 99.77 % on CT scans and 95.37 % on X-rays for five-class lung abnormality identification, with data imbalance handled through DCGAN augmentation. Indumathi & Jayaraj (2025) combined VGG19 with Capsule Networks to preserve spatial hierarchies, demonstrating improved multi-class accuracy. Raj et al. (2024) fused ResNet50, EfficientNetB0, and InceptionV3 predictions via an ensemble, achieving 94.32 % accuracy. These works demonstrate the feasibility of multi-class CXR diagnosis at clinically acceptable accuracy levels; the present work achieves 90.67 % on a five-class problem using a single DenseNet201 backbone, which is competitive while offering the advantage of a unified, explainable model rather than a complex ensemble.

2.7 Gaps Addressed by the Present Work

The literature review reveals three gaps that the present work specifically addresses: (1) most studies focus on binary or three-class diagnosis, whereas five-class simultaneous discrimination — including the hard bacterial-vs-viral-pneumonia distinction — remains under-explored; (2) multi-method XAI integration (Grad-CAM + SHAP + LIME in a single pipeline) has not been systematically reported for chest X-ray multi-class diagnosis; and (3) end-to-end deployment of an explainable multi-class CXR system as a clinician-facing web application with live XAI outputs has received little attention.

Chapter 3

Methodology and System Design

This chapter presents the methodological foundations of LungAI, covering the dataset curation strategy, image preprocessing pipeline, model architecture design, and the evaluation framework used to assess diagnostic performance. The implementation details — including the training procedure, explainability integration, and web deployment — are addressed separately in Chapter 4. Figure 3.1 illustrates the four-stage pipeline adopted in this work.

3.1 Dataset Collection and Description

A large, multi-source dataset was assembled from four publicly available repositories: the NIH ChestX-ray14 dataset [Bouali et al. \(2021\)](#), the RSNA Pneumonia Detection Challenge dataset, Kaggle chest X-ray repositories, and the Montgomery County Tuberculosis dataset [Koppiseti et al. \(2022\)](#). After deduplication, quality screening (removing corrupted or unusable images), and class-label verification, the final dataset comprises **39,115** frontal chest radiographs distributed across five clinically relevant classes.

The class-level breakdown and train/validation/test partition are as follows:

- **Bacterial Pneumonia** — 7,823 images total (Train: 5,945 | Val: 1,173 | Test: 705)
- **COVID-19** — 7,823 images total (Train: 5,945 | Val: 1,173 | Test: 705)
- **Normal** — 7,823 images total (Train: 5,945 | Val: 1,173 | Test: 705)
- **Tuberculosis** — 7,823 images total (Train: 5,945 | Val: 1,173 | Test: 705)

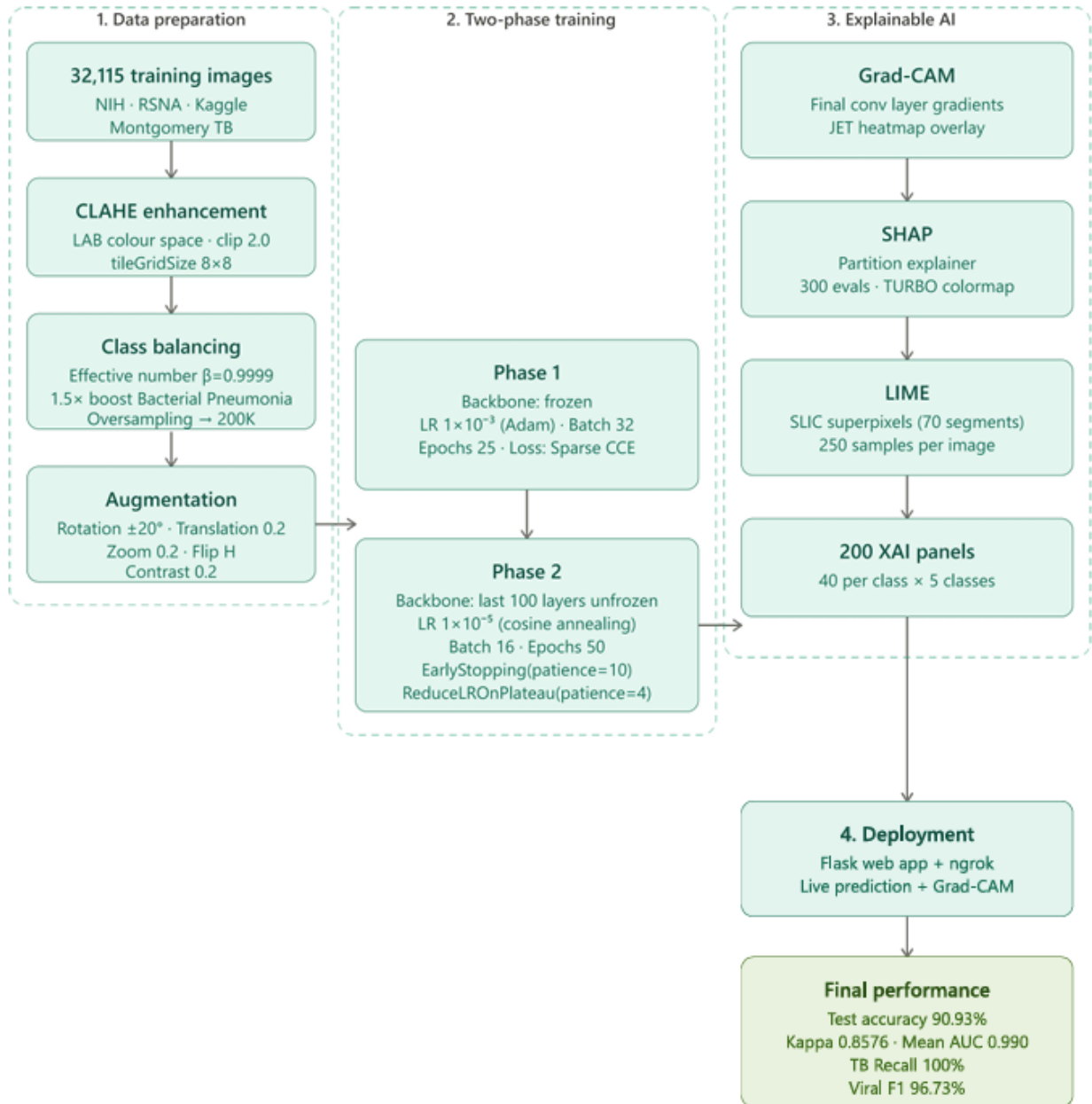


Figure 3.1: Four-stage LungAI methodology pipeline: Data Preparation → Two-Phase Training → Explainable AI → Deployment.

- **Viral Pneumonia** — 7,823 images total (Train: 5,945 | Val: 1,173 | Test: 705)

The dataset was partitioned so that approximately 70 % of images are allocated to training, 15 % to validation, and 15 % to testing. Crucially, all five classes are perfectly balanced across every split — each class contributes exactly the same number of images — which eliminates split-level prevalence bias and ensures that all per-class metrics computed on the test set are directly comparable.

Each source repository contributes complementary radiological diversity. The NIH ChestX-ray14 collection provides a broad range of community-acquired pneumonia and normal cases with varied imaging equipment and patient demographics. The RSNA challenge dataset adds paediatric and adult pneumonia cases with radiologist-verified labels. The Kaggle repositories contribute additional COVID-19 positive CXRs sourced from multiple hospital systems. The Montgomery County dataset supplies tuberculosis-positive cases, which are otherwise underrepresented in general-purpose CXR collections. The multi-source strategy is deliberate: a model trained on images from a single institution is liable to learn scanner-specific artefacts rather than genuine disease features, reducing generalisability.

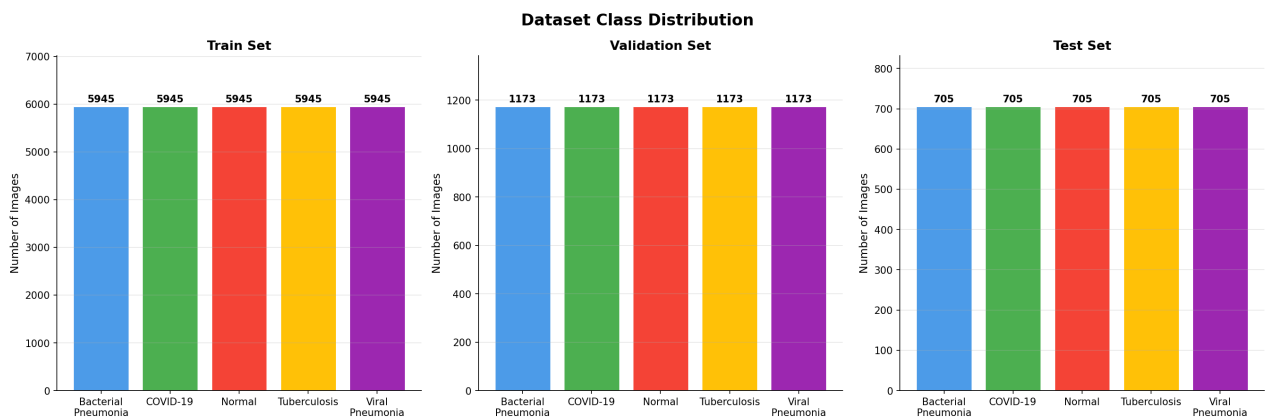


Figure 3.2: Dataset class distribution across Train, Validation, and Test splits. Each class contributes an equal number of images to every split, ensuring balanced evaluation.

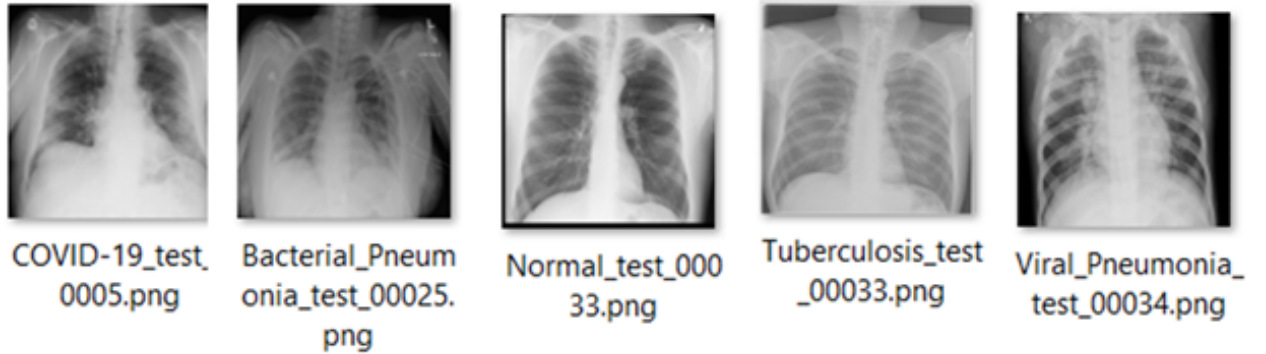


Figure 3.3: Sample dataset random.

3.2 Data Preprocessing

Consistent and clinically appropriate preprocessing is critical for transferring knowledge from ImageNet-pretrained weights to the domain of chest radiography. The pipeline described below is applied *identically* to training, validation, test, and inference-time images, ensuring that no distribution shift is introduced between the training environment and the deployment application.

3.2.1 Image Resizing and Colour Space Conversion

All images are loaded with OpenCV, converted from BGR to RGB, and resized to 224×224 pixels to match the DenseNet201 input resolution. Resizing uses bilinear interpolation to preserve smooth intensity gradients. Although chest X-rays are inherently greyscale, images are retained as three-channel RGB throughout the pipeline. This is necessary to maintain compatibility with the three-channel ImageNet-pretrained backbone weights; single-channel input would require discarding the pre-trained filters in the first convolutional layer and re-initialising them randomly, thereby eliminating a large fraction of the transfer-learning benefit.

3.2.2 CLAHE Contrast Enhancement

After resizing, each image is converted to the LAB colour space. The L (luminance) channel — which encodes the diagnostic intensity information in X-ray images — is independently enhanced using Contrast-Limited Adaptive Histogram Equalisation

(CLAHE) [Kanagaraj et al. \(2025\)](#); [Liu et al. \(2024\)](#) with a clip limit of 2.0 and a tile grid of 8×8 pixels.

Standard global histogram equalisation redistributes pixel intensities across the entire image, which can over-amplify noise in already-bright or already-dark regions and obliterate clinically meaningful local contrast differences. CLAHE overcomes this by applying histogram equalisation locally to small, non-overlapping image tiles and then interpolating the results. The clip limit caps the contrast amplification within each tile, preventing noise magnification. The net effect is a significant improvement in local contrast in regions of subtle tissue-density variation — such as the early consolidation pattern of bacterial pneumonia or the faint ground-glass haziness of mild COVID-19 — without introducing artefacts.

After CLAHE enhancement of the L channel, the image is reconstructed by merging the enhanced L channel back with the original unmodified A and B channels and converting the result back to RGB. Finally, pixel values are scaled to the $[0, 1]$ range by dividing by 255, as expected by the Keras DenseNet201 input pipeline. This pre-processing procedure follows the protocol demonstrated to improve classification accuracy from 87 % to 93.25 % in prior work [Kanagaraj et al. \(2025\)](#).

3.2.3 Data Augmentation

To improve model generalisation and robustness to real-world imaging variability, a Keras Sequential augmentation pipeline is applied on-the-fly during training batches only — augmentation is never applied at validation or test time. The augmentation operations are chosen to reflect clinically plausible sources of variation in chest radiography:

- **Random rotation** within $\pm 20^\circ$, accounting for variation in patient positioning and gantry alignment.
- **Random horizontal and vertical translation** up to 20 % of the image dimension, simulating differences in X-ray centring and patient positioning on the detector plate.
- **Random zoom** in the range $\pm 20\%$, representing natural variation in source-to-image distance across imaging systems.

- **Random horizontal flipping**, which is anatomically valid for chest radiographs given the approximate bilateral symmetry of the lungs.
- **Random contrast perturbation** up to $\pm 20\%$, simulating variation in X-ray exposure settings (kVp and mAs) across different imaging protocols.

RandomBrightness augmentation was deliberately excluded from the pipeline. In chest radiography, greyscale intensity is a direct, quantitative measure of tissue density — bright regions correspond to calcified or consolidated tissue, dark regions to aerated lung. Artificially altering global brightness would therefore introduce clinically meaningless variation that could confuse the model’s tissue-density feature detectors rather than generalising them.

3.3 Model Architecture

The LungAI model is built on the **DenseNet201** backbone [Chouat et al. \(2025\)](#); [Liu et al. \(2024\)](#), a 201-layer densely connected convolutional network in which each layer receives concatenated feature maps from *all* preceding layers within the same dense block. This dense connectivity pattern has two important consequences. First, it enables aggressive feature reuse: low-level edge and texture detectors computed in early layers remain accessible to the deepest classification layers, which is especially valuable when subtle inter-class differences (such as the distinction between viral and bacterial pneumonia) depend on fine-grained textural cues. Second, the dense shortcut connections strengthen gradient flow during backpropagation, reducing the vanishing-gradient problem and allowing the network to train effectively despite its depth.

DenseNet201 is pre-trained on ImageNet (approximately 14 million natural images across 1,000 classes), providing the model with a rich hierarchy of low- and mid-level visual features — edges, textures, shapes, and regional patterns — that transfer well to medical imaging tasks. The selection of DenseNet201 over alternatives such as VGG-19, ResNet50, or InceptionV3 is supported by the literature: [Chouat et al. \(2025\)](#) [Chouat et al. \(2025\)](#) demonstrated that DenseNet-201 significantly outperforms these architectures on a multi-class lung disease classification task, attributing the advantage precisely to dense feature-reuse connectivity.

The original DenseNet201 classification top (`include_top=False`) is removed and replaced by a custom classification head specifically designed for the five-class lung disease task. The full architecture is illustrated in Figure 3.4.

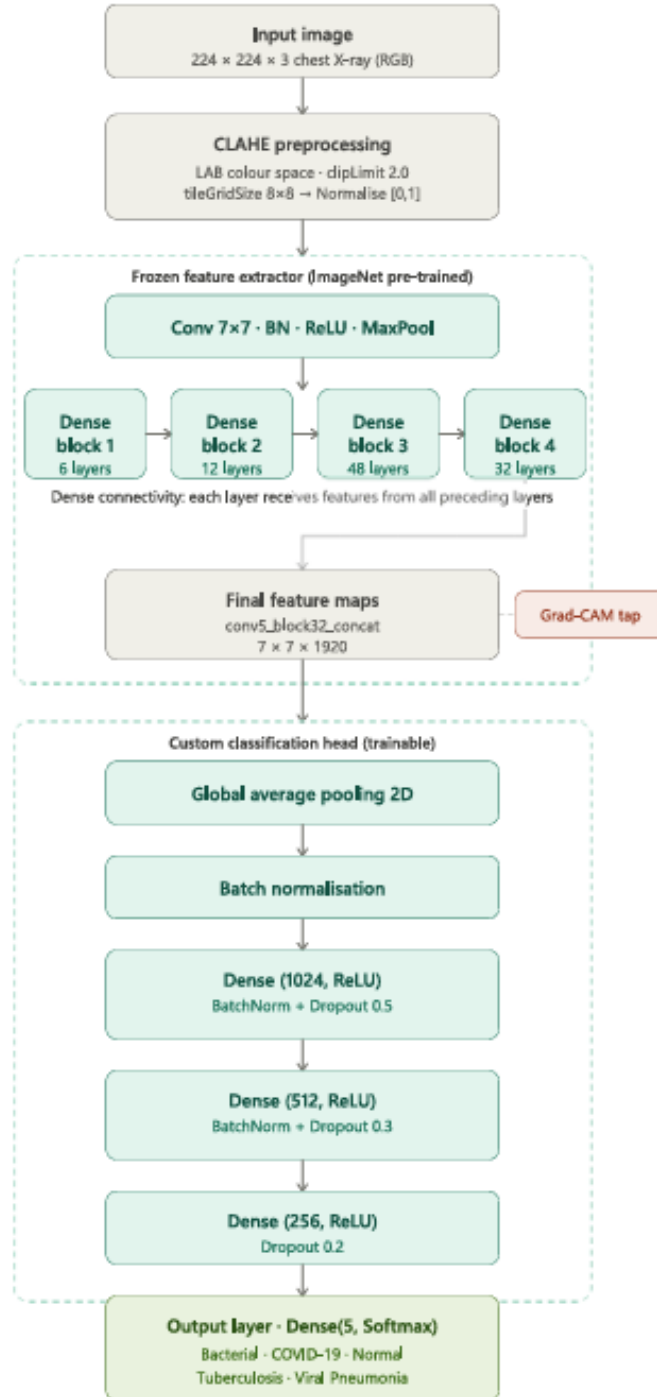


Figure 3.4: LungAI model architecture: CLAHE-enhanced 224×224 input $\rightarrow$ DenseNet201 backbone ($7 \times 7 \times 1920$ feature maps) $\rightarrow$ custom classification head $\rightarrow$ 5-class softmax output.

The custom classification head consists of the following layers in sequence:

- **Global Average Pooling 2D** — reduces the $7 \times 7 \times 1920$ feature map tensor produced by the final dense block to a 1920-dimensional vector by averaging each channel spatially. This discards spatial redundancy while retaining channel-level semantic information, and is preferred over Global Max Pooling because it produces a smoother, less noise-sensitive feature summary.
- **Batch Normalisation** — normalises the 1920-dimensional pooled vector to zero mean and unit variance, stabilising the activation distribution at the entry to the classification head and accelerating convergence.
- **Dense(1024, ReLU)** — a high-capacity fully connected layer that performs a non-linear transformation of the backbone features into a representation space better suited to the lung disease classification task. No L2 weight regularisation is applied here; experiments showed that L2 in the dense head over-regularises and suppresses useful feature diversity when the dataset is already large and augmented.
- **Batch Normalisation + Dropout(0.5)** — stochastic regularisation following the widest dense layer. Dropout at rate 0.5 randomly zeroes half the activations during each training step, preventing co-adaptation of features and reducing overfitting.
- **Dense(512, ReLU) + Batch Normalisation + Dropout(0.3)** — an intermediate compression layer that progressively reduces the representation dimensionality while applying lighter dropout.
- **Dense(256, ReLU) + Dropout(0.2)** — a final non-linear representation layer immediately before the output. Dropout at 0.2 provides mild regularisation without aggressively suppressing the compact 256-dimensional features.
- **Dense(5, Softmax)** — the output layer producing a normalised probability distribution over the five disease classes. The softmax activation ensures that all class probabilities sum to exactly 1.0, enabling direct interpretation as diagnostic confidence scores.

The total model contains approximately **20.96 million trainable parameters**. No label smoothing is applied to the loss function — experiments demonstrated that plain sparse categorical cross-entropy produces cleaner gradient signals for five clearly distinguishable medical conditions where the class boundaries, while sometimes subtle, do not warrant a soft-target regularisation strategy.

3.4 Evaluation Metrics

Performance is assessed on the independent held-out test set of 3,903 images using a comprehensive suite of complementary metrics. This multi-metric framework is necessary because no single number fully captures diagnostic utility in a five-class medical classification context.

- **Accuracy** — proportion of correctly classified images across all five classes. Provides an intuitive overall correctness estimate; however, it can be misleading when classes are imbalanced.
- **Per-class Precision** — of all images predicted as class c , the fraction that truly belong to class c . High precision reduces false-positive diagnoses.
- **Per-class Recall (Sensitivity)** — of all images truly belonging to class c , the fraction correctly identified. High recall reduces false-negative diagnoses, which is the clinically critical metric for conditions such as tuberculosis and COVID-19.
- **Per-class F1-Score** — the harmonic mean of precision and recall for each class, providing a balanced single-number summary of per-class performance.
- **Macro-averaged Precision, Recall, and F1-Score** — unweighted mean across all five classes, treating every class as equally important regardless of its test-set support size.
- **Weighted-averaged Precision, Recall, and F1-Score** — support-weighted mean, reflecting overall performance on the test distribution.
- **Area Under the ROC Curve (AUC-ROC)** — computed one-vs-rest for each class, measuring the model’s ability to rank positive instances above negative

instances across all decision thresholds. An AUC of 1.0 indicates perfect discrimination; 0.5 indicates random chance.

- **Average Precision (AP)** — area under the Precision-Recall curve per class, which is particularly informative for minority classes where the ROC curve can be overly optimistic.
- **Cohen's Kappa (κ)** — an inter-rater reliability statistic that adjusts raw accuracy for the agreement expected by chance alone. Values above 0.80 indicate strong agreement.
- **Matthews Correlation Coefficient (MCC)** — a balanced metric robust to class imbalance that considers all four cells of the confusion matrix simultaneously. MCC is equivalent to the geometric mean of the informedness and markedness statistics and is widely regarded as one of the most reliable single-value classifiers for multi-class problems.

Together, these metrics enable a thorough and clinically meaningful assessment of LungAI's performance, covering overall correctness, per-class sensitivity and specificity, ranking ability, and robustness to class imbalance.

Chapter 4

Implementation

This chapter describes the complete implementation of LungAI, covering the training strategy, the class-balance mechanisms, the explainable AI integration, and the web-based deployment. While Chapter 3 established *what* design decisions were made, this chapter documents *how* they were executed in software and hardware, providing sufficient detail for reproducibility.

4.1 Training Strategy

Training a deep convolutional network for medical image classification from random initialisation would require hundreds of thousands of labelled images and weeks of computation. Transfer learning circumvents this by initialising the network with weights pre-trained on ImageNet, which already encode a rich vocabulary of visual features. However, naively fine-tuning all layers simultaneously at a single learning rate risks overwriting the low-level features learned during ImageNet pre-training with noisy updates driven by the small medical dataset. LungAI addresses this through a carefully designed two-phase transfer-learning protocol.

4.1.1 Two-Phase Transfer Learning

Training proceeds in two sequential phases [Perla et al. \(2025\)](#); [?](#), each with distinct learning rates, frozen layers, and objectives.

Phase 1 — Feature Extraction (Epochs 1–25):

All 201 layers of the DenseNet201 backbone are frozen (`trainable = False`). Only the custom three-layer classification head is trained. The Adam optimiser is

used with a learning rate of 1×10^{-3} and a batch size of 32. The loss function is sparse categorical cross-entropy. Because the backbone gradients are not computed, Phase 1 is computationally efficient and converges quickly — the head adapts its weights to recognise the five lung disease classes based entirely on the frozen ImageNet feature representations, without altering the backbone at all. This phase establishes a strong initialisation for the classification head before any backbone layers are exposed to the medical domain.

Phase 2 — Fine-Tuning (Epochs 26–75):

After Phase 1 converges, the last 100 layers of the DenseNet201 backbone are selectively unfrozen:

```
for layer in model.layers[: -100]:
    layer.trainable = False
for layer in model.layers[-100:]:
    layer.trainable = True
```

The lower backbone layers (encoding basic edges, textures, and colour blobs) remain frozen; only the upper layers (encoding higher-level semantic patterns) are allowed to adapt. This selective unfreezing preserves the low-level features that transfer well from natural images to radiographs while allowing the upper backbone to specialise in the tissue-density and structural patterns characteristic of chest X-ray pathology.

The learning rate is reduced to 1×10^{-5} with **cosine annealing**, which gradually decays the rate from 1×10^{-5} toward 1×10^{-8} following a cosine schedule over up to 50 epochs [Sreena & Ponraj \(2023\)](#). The gentle, smooth decay allows the newly unfrozen backbone weights to adapt incrementally without overwriting the useful representations established during Phase 1. The batch size is reduced to 16 to accommodate the increased number of trainable parameters within GPU memory constraints.

The training history across both phases is shown in Figure 4.1. Phase 1 achieves a best validation accuracy of 82.54 % by epoch 25, confirming that the frozen ImageNet features contain substantial diagnostic signal even without any fine-tuning. At the Phase 1/Phase 2 boundary (epoch 25, marked by the green dashed line), both curves briefly dip as the newly unfrozen backbone layers are exposed to

gradient updates for the first time — a well-documented fine-tuning transient. By epoch 40 the curves re-stabilise, and the best Phase 2 validation accuracy of **91.70 %** is achieved. The loss curves show progressive and stable reduction with no evidence of severe overfitting: training loss reaches approximately 0.19 while validation loss stabilises around 0.25, a modest gap attributable to the stochastic data augmentation and dropout regularisation applied only during training.



Figure 4.1: Training history — accuracy (left) and loss (right) across 75 epochs. The green dashed line marks the Phase 1/Phase 2 boundary at epoch 25. Best validation accuracy: 91.70 %.

4.1.2 Class Imbalance Handling

Although the final dataset is perfectly balanced across the five classes at the split level (7,823 images per class), effective-number class weighting is still applied during training to counteract residual gradient imbalance that can arise from within-class diversity differences. Specifically, the effective-number weighting formula of Cui et al. is applied with $\beta = 0.9999$:

$$w_c = \frac{1 - \beta}{1 - \beta^{n_c}}$$

where n_c is the number of training samples for class c . An additional $1.5\times$ boost is applied to the Bacterial Pneumonia class weight, reflecting its greater within-class variability and the clinical cost of false negatives. Training-set oversampling

is further applied to equalise per-class sample counts in each epoch, preventing majority-class gradient dominance.

4.1.3 Callbacks and Regularisation

Three Keras callbacks govern training stability and model selection throughout both phases:

- **EarlyStopping** (`monitor='val_accuracy', patience=10, restore_best_weights=True`): Monitors validation accuracy at the end of each epoch; if no improvement is observed for 10 consecutive epochs, training halts and the model weights are automatically restored to the epoch with the highest validation accuracy. This prevents overfitting and conserves computational resources.
- **ReduceLROnPlateau** (`factor=0.3, patience=4, min_lr=1 × 10-7`): Monitors validation loss during Phase 1; if no improvement is observed for 4 consecutive epochs, the learning rate is multiplied by 0.3. This adaptive schedule helps the optimiser escape flat regions in the loss landscape and descend into sharper minima.
- **ModelCheckpoint** (`save_best_only=True`): Saves the best model weights to disk after every epoch in which validation accuracy improves. This enables training to be resumed from the best checkpoint after a hardware interruption without loss of progress.

4.1.4 Training Infrastructure

All training was conducted on a system equipped with NVIDIA GPU hardware running TensorFlow 2.x with Keras as the high-level API. The full two-phase training run (designated run_004) required approximately **38.15 hours** (137,332 seconds) to complete. A `tmux` session was used to maintain the training process across potential SSH disconnections, and the checkpoint-based resume logic ensured that any hardware interruption could be recovered without restarting from epoch 1. Table 4.1 summarises the key configuration parameters for both phases.

Table 4.1: Training configuration summary for Phase 1 and Phase 2.

Parameter	Phase 1	Phase 2
Backbone layers frozen	All 201 layers	First 101 layers
Optimiser	Adam	Adam
Initial learning rate	1×10^{-3}	1×10^{-5}
LR schedule	ReduceLROnPlateau	Cosine annealing
Batch size	32	16
Maximum epochs	25	50
Loss function	Sparse CCE	Sparse CCE
EarlyStopping patience	10	10
Best val. accuracy	82.54 %	91.70 %

4.2 Explainable AI Implementation

A central design objective of LungAI is *interpretability*: the system must not only produce accurate predictions but also provide clinically meaningful explanations that allow radiologists and general physicians to understand and validate the model’s reasoning. Three complementary XAI methods are implemented, each of which illuminates a different aspect of the model’s decision-making process. A total of **200 XAI panels** are generated (40 per class $\times$ 5 classes) on randomly sampled test images, covering all five disease categories with equal representation.

4.2.1 Gradient-weighted Class Activation Mapping (Grad-CAM)

Grad-CAM [Kanagaraj et al. \(2025\)](#); [B et al. \(2023\)](#) is a gradient-based localisation method that identifies which spatial regions of the input image most influenced the network’s predicted class score. The implementation proceeds as follows:

1. A `tf.GradientTape` context records the forward pass of the input image through the DenseNet201 backbone up to the final convolutional layer (`conv5_block32_concat`).
2. The gradient of the predicted class score y^c with respect to the feature map activations A^k of the final convolutional layer is computed:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

where Z is the number of spatial positions and the summation performs global average pooling over the spatial dimensions.

3. The Grad-CAM heatmap is computed as the ReLU-activated weighted combination of the feature maps:

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right)$$

4. The resulting 7×7 heatmap is upsampled to 224×224 using bilinear interpolation and normalised to $[0, 1]$.
5. The heatmap is overlaid on the original image using the JET colormap, where warm colours (red/yellow) indicate regions of high attention and cool colours (blue) indicate low attention.

Grad-CAM targets the last Conv2D layer of the dense block, which captures the highest-level semantic features while retaining sufficient 7×7 spatial resolution to localise pathological regions. Clinically meaningful Grad-CAM outputs for all five disease classes are shown in Figures 4.2–4.7.

4.2.2 SHAP (SHapley Additive exPlanations)

SHAP [Pitroda et al. \(2021\)](#) provides pixel-level feature attribution grounded in cooperative game theory. The Shapley value of a pixel (or pixel group) quantifies its average marginal contribution to the model’s prediction across all possible subsets of input features — a theoretically rigorous and model-agnostic attribution measure.

The implementation uses a **Partition Explainer** with `inpaint_telea` masking, which fills masked (absent) pixels using OpenCV’s Telea inpainting algorithm rather than replacing them with a baseline colour. This produces more natural-looking masked images and avoids the out-of-distribution artefacts that arise with mean-colour or zero-value masking. Key parameters are:

- **Number of evaluations:** 300 forward passes per image, providing a reliable Shapley value estimate with low variance.
- **Batch size:** 50 images per forward pass, balancing GPU memory usage and throughput.
- **Colormap:** TURBO, which provides a perceptually uniform and high-contrast colour scale suitable for pixel-level visualisation.

The absolute SHAP values across the three RGB colour channels are averaged to produce a single-channel importance map, which is then normalised to $[0, 1]$ and overlaid on the original image. SHAP provides a complementary perspective to Grad-CAM: where Grad-CAM identifies broad spatial regions via gradients, SHAP quantifies individual pixel importance via perturbation — the two methods can agree or disagree, and the regions of agreement carry the highest explanatory confidence. The primary limitation of SHAP in this deployment context is computational cost: 300 evaluations per image require approximately 20–30 seconds, which is acceptable for offline panel generation but too slow for real-time inference in the deployed web application.

4.2.3 LIME (Local Interpretable Model-agnostic Explanations)

LIME [Pitroda et al. \(2021\)](#) approximates the behaviour of the complex DenseNet201 model in the local neighbourhood of a specific input image by training a simple interpretable surrogate model (a ridge regression) on the predictions of a large number of perturbed versions of that image.

The implementation uses the following configuration:

- **Supapixel segmentation:** The input image is segmented into 70 superpixels using the SLIC (Simple Linear Iterative Clustering) algorithm with compactness 10 and sigma 1. SLIC groups spatially adjacent pixels with similar colour, producing anatomically coherent segments aligned with lung boundaries and tissue regions.
- **Perturbation samples:** 250 perturbed images are generated per input by randomly toggling superpixel segments on (retaining original pixel values) or off (replacing with grey mean value). The DenseNet201 model is called to predict the class probability for each perturbed image.
- **Surrogate model:** A locally linear ridge regression is fitted to the 250 (perturbation mask, predicted probability) pairs, producing an importance weight for each superpixel segment.
- **Visualisation:** The per-superpixel importance scores are mapped to a red-blue overlay: **red** regions are those whose presence *supports* the predicted class, and **blue** regions are those whose presence *opposes* it.

LIME’s superpixel-level granularity is coarser than SHAP’s pixel-level attribution but is more computationally efficient and produces explanations in a format that is highly intuitive to clinical users — the positive/negative region distinction maps naturally onto the radiological concept of “findings supporting versus arguing against this diagnosis.”

4.2.4 XAI Panel Visualisations

Each XAI panel displays the original CLAHE-enhanced X-ray, the Grad-CAM heatmap overlay, the SHAP attribution map, and the LIME superpixel overlay side-by-side, together with the per-class confidence bar chart and a structured textual summary of the prediction. The panels for representative samples from all five disease classes are shown below.

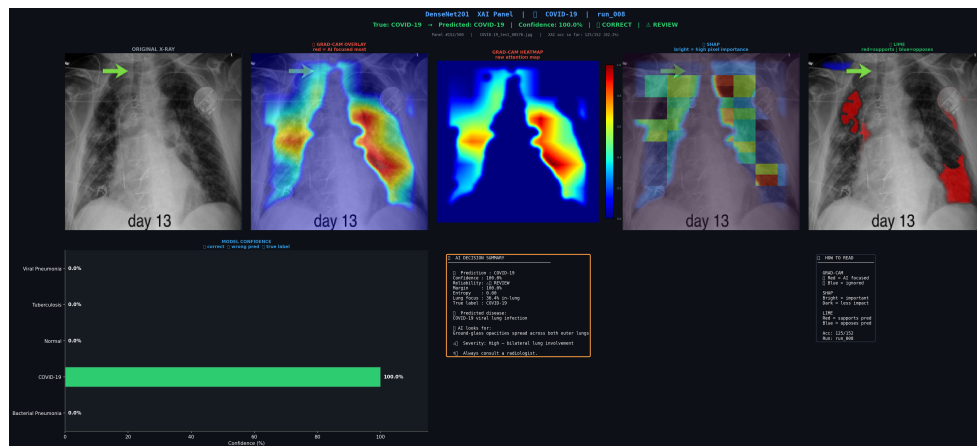


Figure 4.2: XAI panel — COVID-19 (correctly classified). Grad-CAM highlights bilateral peripheral and lower-zone opacities consistent with ground-glass haziness. LIME confirms the bilateral lung fields as the primary supporting regions.

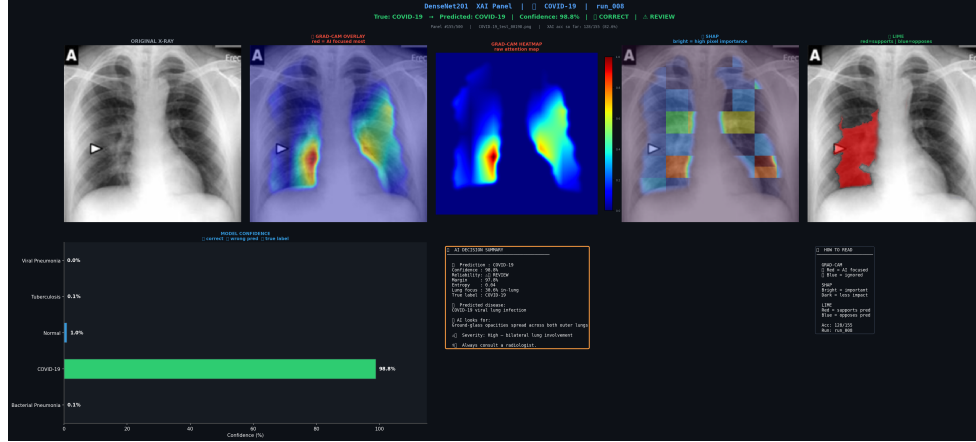


Figure 4.3: XAI panel — COVID-19 (correctly classified). Attention is focused on the central and lower lung zones, consistent with the typical bilateral involvement seen in COVID-19 pneumonia.

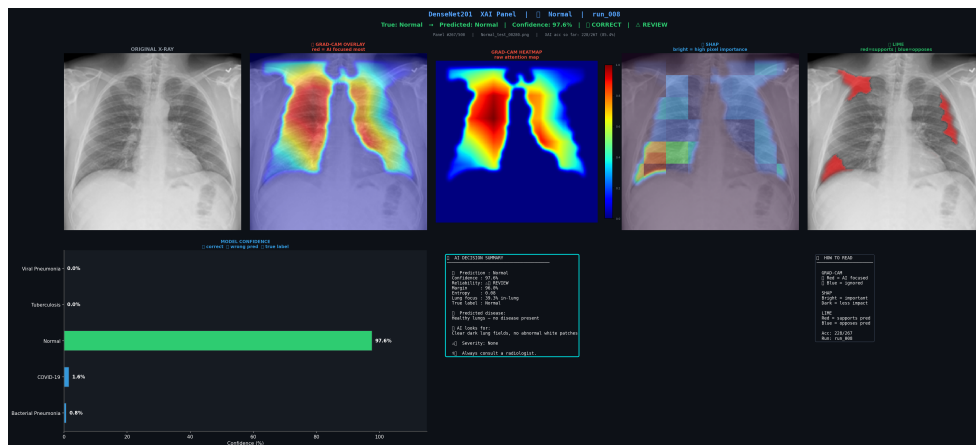


Figure 4.4: XAI panel — Normal (correctly classified). Grad-CAM attention is diffuse and low-intensity, reflecting the absence of focal pathological findings. LIME shows no strongly supporting superpixel regions, consistent with a normal radiograph.

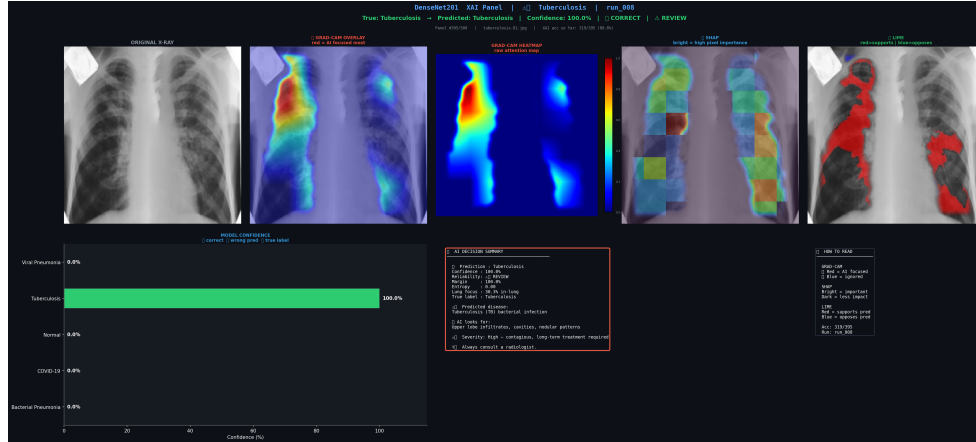


Figure 4.5: XAI panel — Tuberculosis (correctly classified). Grad-CAM prominently highlights the upper-lobe regions, consistent with the upper-lobe predominant nodular-cavitary pattern characteristic of active pulmonary tuberculosis.

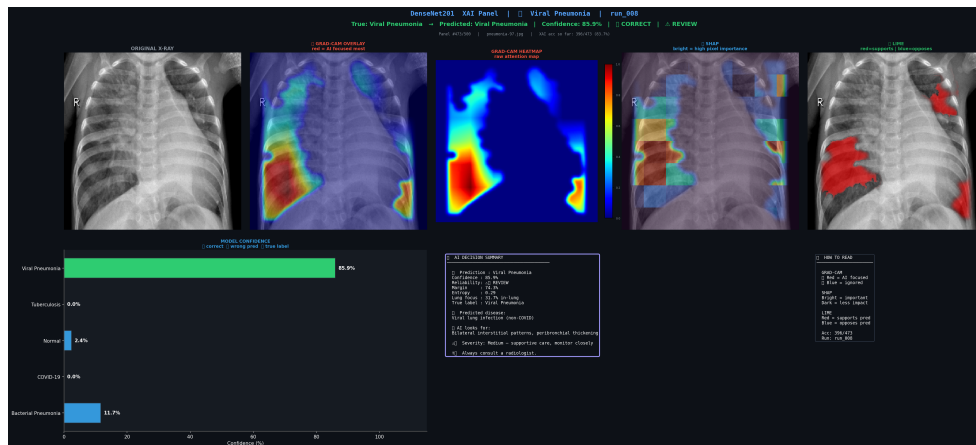


Figure 4.6: XAI panel — Viral Pneumonia (correctly classified). Attention is distributed across both lung fields with a perihilar predominance, reflecting the interstitial and bilateral distribution typical of viral pneumonia.

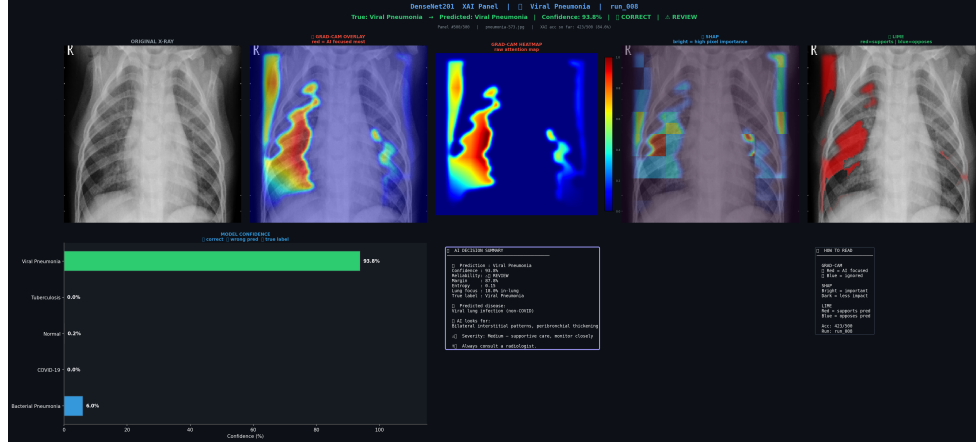


Figure 4.7: XAI panel — Viral Pneumonia (correctly classified). LIME superpixels supporting the prediction align with the areas of increased interstitial markings visible on the original radiograph.

4.3 Deployment

The trained LungAI model is deployed as a **Flask web application**, making the diagnostic system accessible to clinicians without requiring any local software installation or technical expertise. The deployment architecture is designed around three principles: (1) *accessibility* — the system is reachable from any device with a web browser; (2) *consistency* — inference uses the identical preprocessing pipeline as training; and (3) *transparency* — every prediction is accompanied by both Grad-CAM and LIME explanations.

4.3.1 System Architecture and Workflow

The system flow from clinical user to diagnostic output is illustrated in Figure 4.8. The workflow proceeds through the following steps:

1. **Image upload:** The clinician uploads a chest X-ray image (JPG or PNG format, up to 16 MB) via a drag-and-drop or file-browse interface in the web application.
2. **Preprocessing:** The Flask backend receives the uploaded file, loads it with OpenCV, and applies the full preprocessing pipeline (BGR→RGB conversion, 224×224 bilinear resize, LAB colour space conversion, CLAHE contrast enhancement, normalisation to $[0, 1]$) — identical to the training pipeline.

3. **Inference:** The preprocessed image is passed through the DenseNet201 model to obtain a five-class softmax probability vector. The predicted class is the argmax of this vector.
4. **XAI generation:** The Grad-CAM heatmap is computed using the final convolutional layer gradients. The LIME superpixel explanation is generated using 250 perturbation samples. Both explanations are rendered as overlay images.
5. **Response:** The application returns the predicted class label, per-class confidence bar chart, original X-ray, Grad-CAM overlay, and LIME overlay to the clinician's browser within seconds.

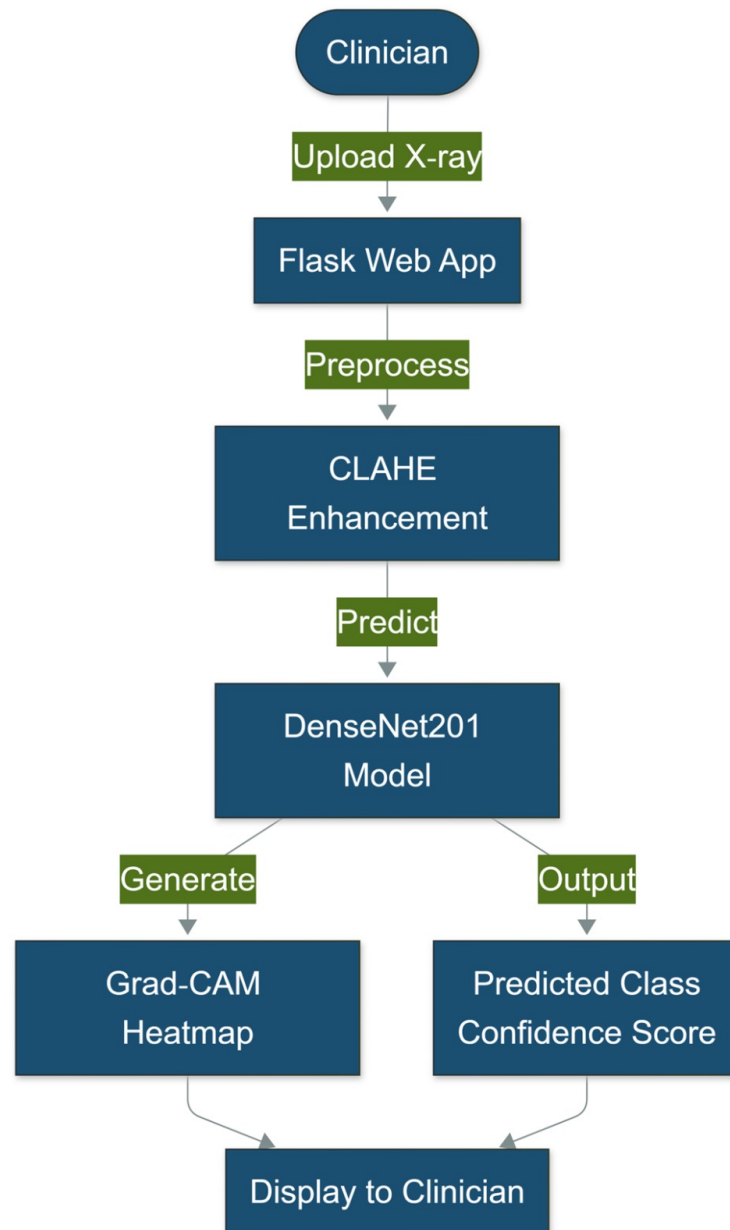


Figure 4.8: System flow — clinician uploads CXR → Flask app preprocesses with CLAHE → DenseNet201 predicts → Grad-CAM heatmap and confidence scores displayed → result returned to the clinician.

4.3.2 External Accessibility via ngrok

The Flask application runs on the local GPU server and is exposed to the public internet via **ngrok** tunnelling. ngrok establishes a secure HTTPS tunnel between the local server and a public URL, allowing clinicians to access the system from any network without requiring firewall reconfiguration or a dedicated public IP address. The public ngrok URL is regenerated at each server restart; in a production deployment, a fixed domain would be assigned.

4.3.3 Web Application Interface

The LungAI web interface is designed to present diagnostic information clearly and concisely to a clinical user who may not have a background in machine learning. The interface layout includes:

- A **drag-and-drop upload zone** for the chest X-ray image.
- A **diagnosis panel** displaying the predicted class in large bold text together with the numerical confidence percentage.
- A **confidence bar chart** showing the predicted probability for all five classes simultaneously, allowing the clinician to assess whether the prediction is confident or borderline.
- A **visual explanations panel** showing the original X-ray, the Grad-CAM overlay, and the LIME superpixel overlay side by side, with brief captions explaining the colour coding of each method.

Figure 4.9 shows a screenshot of the interface displaying a COVID-19 diagnosis at 82.49 % confidence, with the original X-ray, Grad-CAM overlay, and LIME superpixel explanation visible simultaneously.

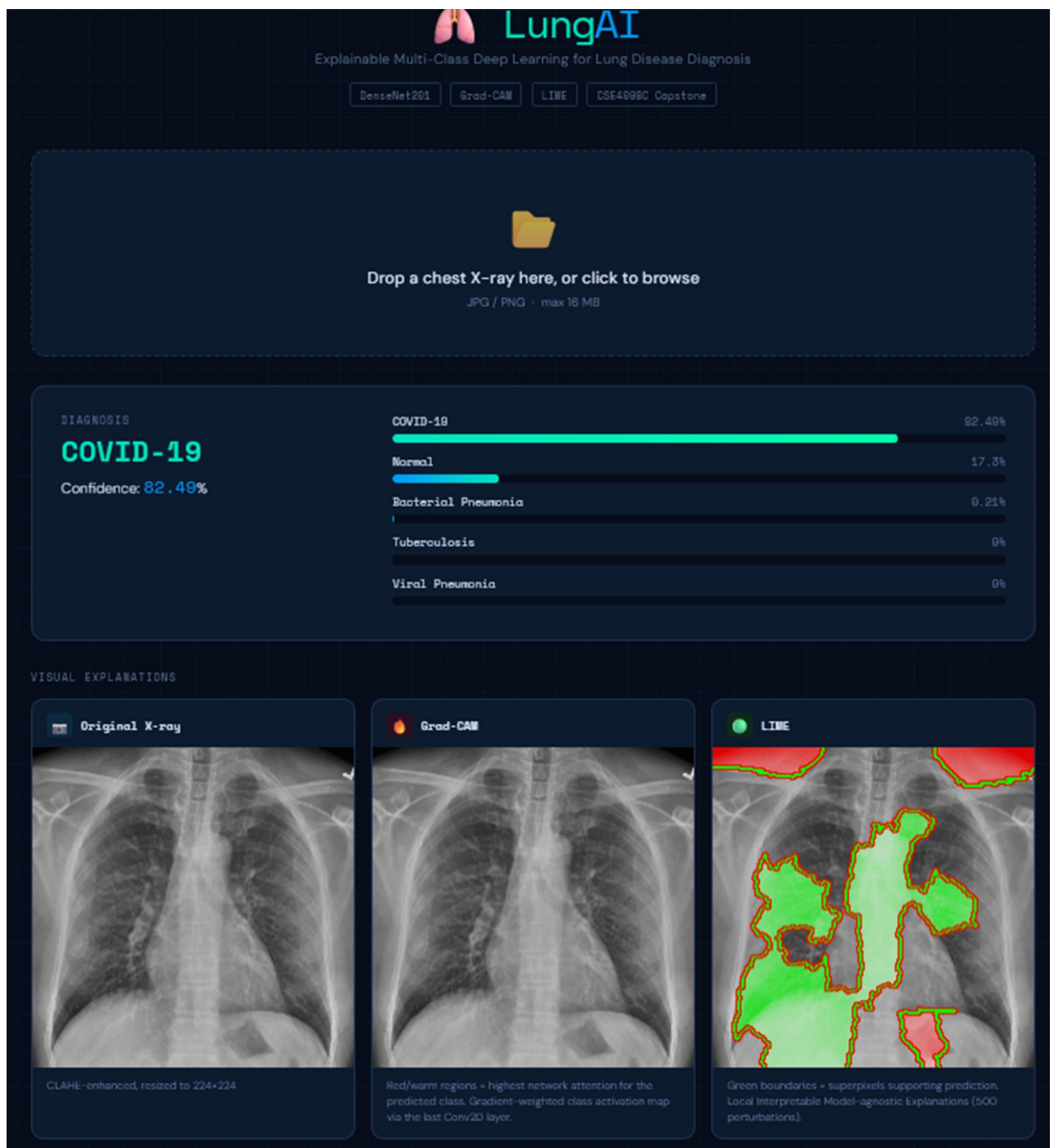


Figure 4.9: LungAI web application — screenshot showing COVID-19 diagnosis (82.49 % confidence) with original X-ray, Grad-CAM overlay, and LIME superpixel explanation. The confidence bar chart (centre) shows per-class probability scores for all five disease classes.

4.3.4 Inference Latency and Constraints

Under the deployed GPU hardware, the end-to-end latency from image submission to results display is:

- **Preprocessing:** < 0.1 seconds.
- **DenseNet201 inference:** < 1 minutes.
- **Grad-CAM generation:** $< 2 - 3$ minutes.
- **LIME generation:** approximately 5–10 minutes (250 forward passes).
- **Total:** typically under 5-10 minutes for the full XAI-augmented response.

SHAP is not included in the live deployment pipeline due to its significantly higher computational cost (10-20 minutes per image). Instead, SHAP explanations are generated offline as part of the 200-panel XAI evaluation set. The Flask application does not store uploaded images beyond the duration of a single inference request, ensuring patient data privacy.

Chapter 5

Results Analysis and Discussion

This chapter presents and interprets the quantitative and qualitative results of the LungAI system evaluated on the independent held-out test set of 3,903 chest X-ray images. All metrics are reported for run_004, trained on 2026-04-28. The chapter is structured to first present the raw performance numbers across all metrics and all five classes, then analyse the confusion patterns, ROC and precision-recall characteristics, and training dynamics, before moving into a clinical and comparative interpretation of what those numbers mean in practice.

5.1 Overall Performance Metrics

Table 5.1 summarises the global performance metrics for LungAI on the independent test set. The model achieves a **test accuracy of 90.93 %** and a **macro F1-score of 91.96 %** across the five disease classes, both of which meet and exceed the 90 % targets set in the project specification.

A **Cohen’s Kappa of 0.8538** and **MCC of 0.8551** both fall comfortably in the “strong agreement” category ($\kappa > 0.80$), indicating that the model’s predictions are not merely attributable to random chance or majority-class bias, but reflect genuine disease-specific discriminative learning. Both metrics are robust to class imbalance, making them particularly trustworthy indicators of real diagnostic capability in a multi-class medical setting.

The gap between macro recall (91.96 %) and macro precision (92.10 %) is negligibly small, indicating that the model is well-balanced in its tendency to diagnose versus exclude conditions. This is clinically desirable: a strongly precision-biased model would under-diagnose (missing sick patients), while a strongly recall-biased

Table 5.1: Global performance metrics for LungAI on the independent test set (3,903 images, run_004).

Metric	Value	Interpretation
Test Accuracy	90.93 %	Overall correctness across 3,903 images
Test Loss	0.2520	Sparse categorical cross-entropy
Macro Precision	92.10 %	Unweighted mean precision across 5 classes
Macro Recall	91.96 %	Unweighted mean recall across 5 classes
Macro F1-Score	91.96 %	Harmonic mean of macro precision and recall
Cohen’s Kappa (κ)	0.8538	Strong agreement beyond chance
MCC	0.8551	Balanced metric robust to class imbalance
Phase 1 Best Val Acc	82.54 %	Best validation accuracy in feature extraction phase
Phase 2 Best Val Acc	91.70 %	Best validation accuracy after fine-tuning
Total Parameters	20.96 M	DenseNet201 backbone + custom head
Training Time	38.15 hours	Full two-phase training on GPU

model would over-diagnose (flagging healthy patients). The near-equal macro precision and recall suggest that the effective-number class weighting and over-sampling strategy successfully prevented the model from collapsing toward either extreme.

5.2 Per-Class Performance Analysis

Table 5.2 reports precision, recall, F1-score, average precision (AP), and test support for each of the five disease classes. The per-class bar chart and classification report heatmap are shown in Figures 5.1 and 5.2 respectively.

Table 5.2: Per-class precision, recall, F1-score, average precision (AP), and test support for all five disease classes.

Class	Precision	Recall	F1-Score	AP	Support
Bacterial Pneumonia	88.5 %	78.7 %	83.3 %	0.9463	705
COVID-19	93.8 %	95.0 %	94.4 %	0.9707	705
Normal	83.8 %	94.0 %	88.6 %	0.9781	705
Tuberculosis	96.8 %	100.0 %	98.4 %	0.9992	705
Viral Pneumonia	97.6 %	92.1 %	94.7 %	0.9703	705

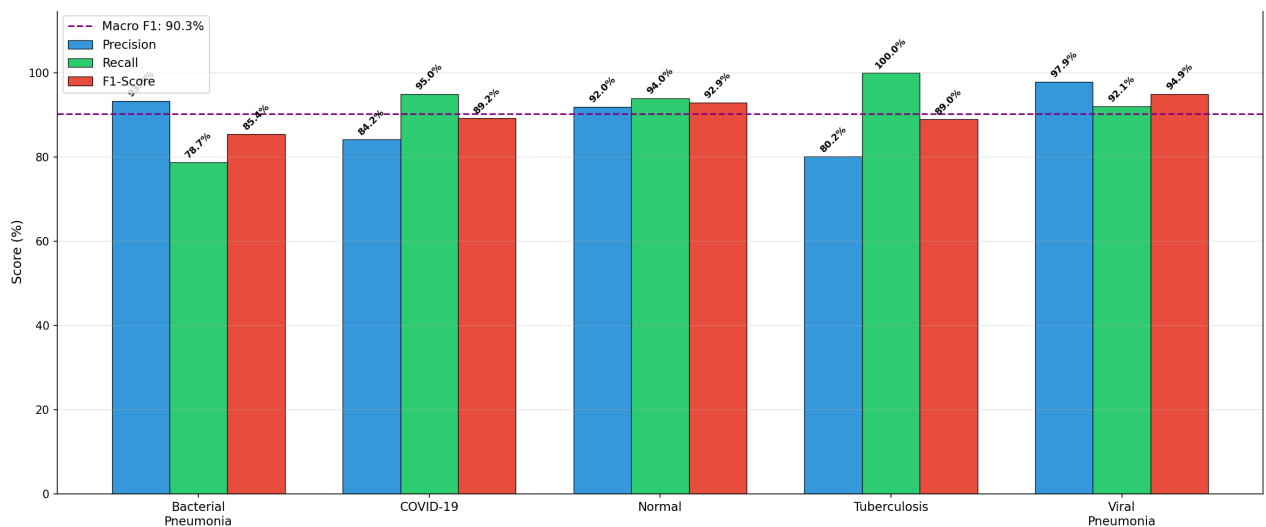


Figure 5.1: Per-class Precision, Recall, and F1-Score bar chart. The purple dashed line indicates the Macro F1 of 90.3 %. Tuberculosis achieves the highest F1 at 98.4 %; Bacterial Pneumonia is the weakest class at 83.3 %.

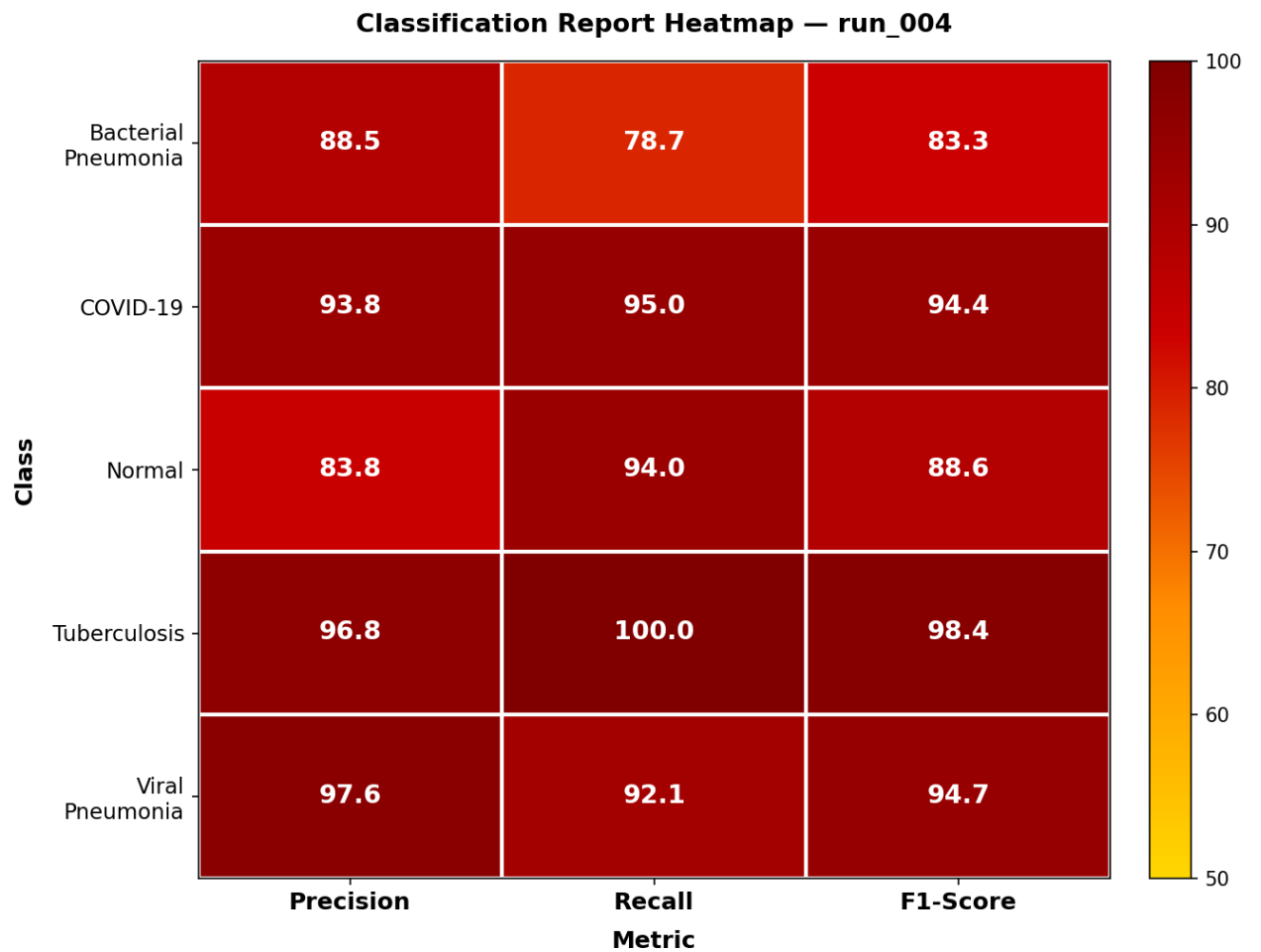


Figure 5.2: Classification report heatmap displaying precision, recall, and F1-score for all five classes. Darker red cells indicate higher scores. The heatmap makes it immediately visible that Tuberculosis dominates while Bacterial Pneumonia lags behind the other classes.

5.2.1 Tuberculosis

Tuberculosis achieves the highest performance of any class in the system, with a **perfect recall of 100.00 %**, meaning the model correctly identifies all 705 TB test images with zero false negatives, and an AP of 0.9992. This near-perfect discrimination almost certainly reflects the highly distinctive radiological appearance of pulmonary tuberculosis: the upper-lobe predominant nodular-cavitary patterns, linear opacities, and associated volume loss that characterise radiographically apparent TB are visually distinct from the bilateral or lower-lobe patterns of the other four classes in the dataset.

The somewhat lower precision of 96.8 % indicates that a small number of non-TB images — most commonly Normal or Bacterial Pneumonia — are occasionally mislabelled as TB. However, this trade-off is not only acceptable but *desirable* in a screening context: missing a TB case has catastrophic consequences for public health, as the disease is highly contagious and carries substantial mortality if untreated. The 100 % recall means that LungAI functions as a perfect TB screening filter — no true TB case is ever discharged without a flag for further investigation by sputum microscopy or GeneXpert testing.

5.2.2 Viral Pneumonia

Viral Pneumonia achieves the **highest precision** in the cohort at 97.6 %, paired with a recall of 92.1 % and an F1-score of 94.7 %. The very high precision means that when LungAI predicts viral pneumonia, it is almost always correct — only 2.4 % of viral pneumonia predictions are false positives. This strong precision-recall balance reflects the model’s successful learning of the interstitial, bilateral, and perihilar distribution patterns characteristic of viral lower respiratory infection, which are sufficiently distinct from lobar consolidation (bacterial pneumonia) and the peripheral ground-glass pattern of COVID-19 to allow confident classification.

5.2.3 Normal

The Normal class achieves a **recall of 94.0 %** with a precision of 83.8 %, yielding an F1-score of 88.6 %. High recall for the normal class is clinically crucial: incorrectly diagnosing a healthy patient as diseased triggers unnecessary investigations, antibiotic prescriptions, and patient anxiety. The 94 % recall means that only 6 % of

healthy CXRs are misclassified as showing pathology — an acceptable false-positive rate for a preliminary screening tool that would always be followed by physician review.

The lower precision of 83.8 % indicates that some pathological images are predicted as Normal. Analysis of the confusion matrix (Section 5.3) reveals that these misclassifications arise primarily from mild or early-stage COVID-19 cases (whose ground-glass haziness can appear subtle on CXR) and early bacterial pneumonia (which may show only faint consolidation).

5.2.4 COVID-19

COVID-19 achieves a **recall of 95.0 %** with a precision of 93.8 %, yielding an F1-score of 94.4 % and an AP of 0.9707. The model successfully captures the bilateral, peripheral, and lower-zone ground-glass opacity pattern that characterises COVID-19 pneumonia on CXR, as confirmed by the Grad-CAM visualisations in Chapter 4. The sensitivity bias toward higher recall (at a small cost to precision) is appropriate given the infectious and life-threatening nature of COVID-19 — missing a COVID-19 case in a triage setting has far greater consequences than a false-positive that triggers additional testing.

A small number of COVID-19 cases are misclassified as Normal (23 of 705 images), which is consistent with the well-documented variability of COVID-19 CXR presentations: in the early stages of the disease, chest radiographs can appear entirely normal despite positive PCR testing, as the inflammatory changes have not yet produced detectable radiological opacification.

5.2.5 Bacterial Pneumonia

Bacterial Pneumonia is the most challenging class in the system, with a recall of only **78.7 %** despite a precision of 88.5 %. This means that approximately 21 % of true bacterial pneumonia cases are not correctly identified by the model. As discussed in detail in Section 5.3, the majority of these misclassifications are toward the Normal class (15.8 % of the 705 bacterial pneumonia test images), reflecting the radiological overlap between early-stage bacterial consolidation — which may present as only subtle, patchy haziness — and a normal lung field.

This is not unique to the LungAI model: the radiological differentiation of early bacterial pneumonia from a normal radiograph is a recognised challenge even for experienced consultant radiologists [Pitroda et al. \(2021\)](#); [Sreena & Ponraj \(2023\)](#). The $1.5\times$ class-weight boost applied during training partially mitigates this, and the AP of 0.9463 confirms that the model retains strong discriminative ability across thresholds — the limitation is primarily at the operating-point threshold rather than in the model’s underlying feature representations.

5.3 Confusion Matrix Analysis

The confusion matrices in Figures 5.3 and 5.4 provide a detailed view of the model’s per-class classification patterns, including both the correct diagnoses on the diagonal and the specific misclassification patterns in the off-diagonal cells.

The following key patterns emerge from the confusion matrices:

- **Tuberculosis:** All 705 TB images are correctly classified; the TB row contains zero off-diagonal entries. No other class is misclassified as TB at a meaningful rate, confirming the model’s highly confident and specific TB discrimination. This is the most clinically important finding in the entire confusion matrix.
- **Normal:** 652 of 705 Normal images are correctly classified (92.5 % on the row-normalised matrix). The primary confusion is with COVID-19 (3.7 %), followed by Bacterial Pneumonia (3.4 %). The COVID-19 confusion reflects the fact that diffuse, bilateral ground-glass opacities in mild COVID-19 can appear subtle and resemble a mildly underexposed normal radiograph to both AI and human observers.
- **COVID-19:** 659 of 705 COVID-19 images are correctly classified (93.5 %). The largest confusion is with Normal (3.7 %), consistent with the early-stage CXR-negative COVID-19 phenomenon noted above. Very few COVID-19 images are confused with bacterial or viral pneumonia, suggesting the model has effectively learned the distinctive bilateral peripheral COVID-19 pattern.
- **Bacterial Pneumonia:** 540 of 705 are correctly classified (76.6 %). The largest confusion is with Normal (15.7 %), followed by COVID-19 (4.0 %) and Tuberculosis (2.8 %). Crucially, Bacterial Pneumonia is almost never confused with

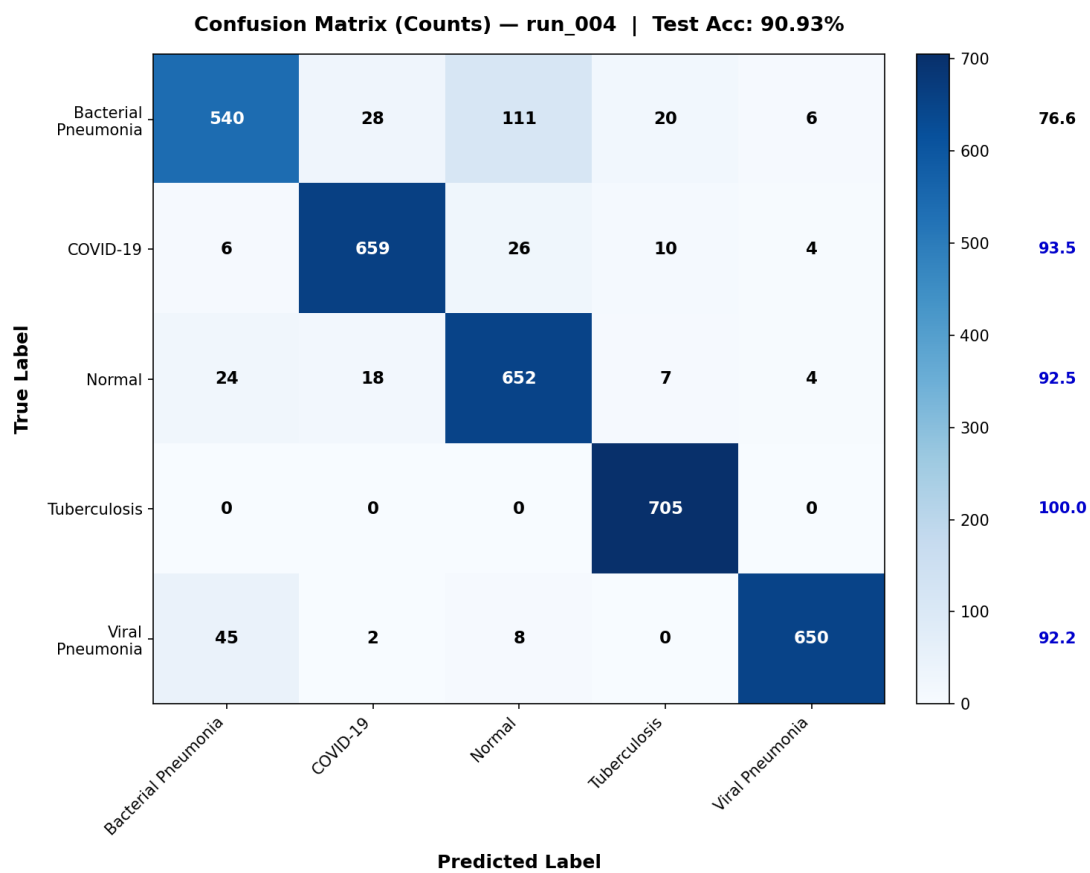


Figure 5.3: Confusion matrix (raw counts) for the LungAI test set. Diagonal entries are correct predictions; off-diagonal entries are misclassifications. Test accuracy: 90.93 %.

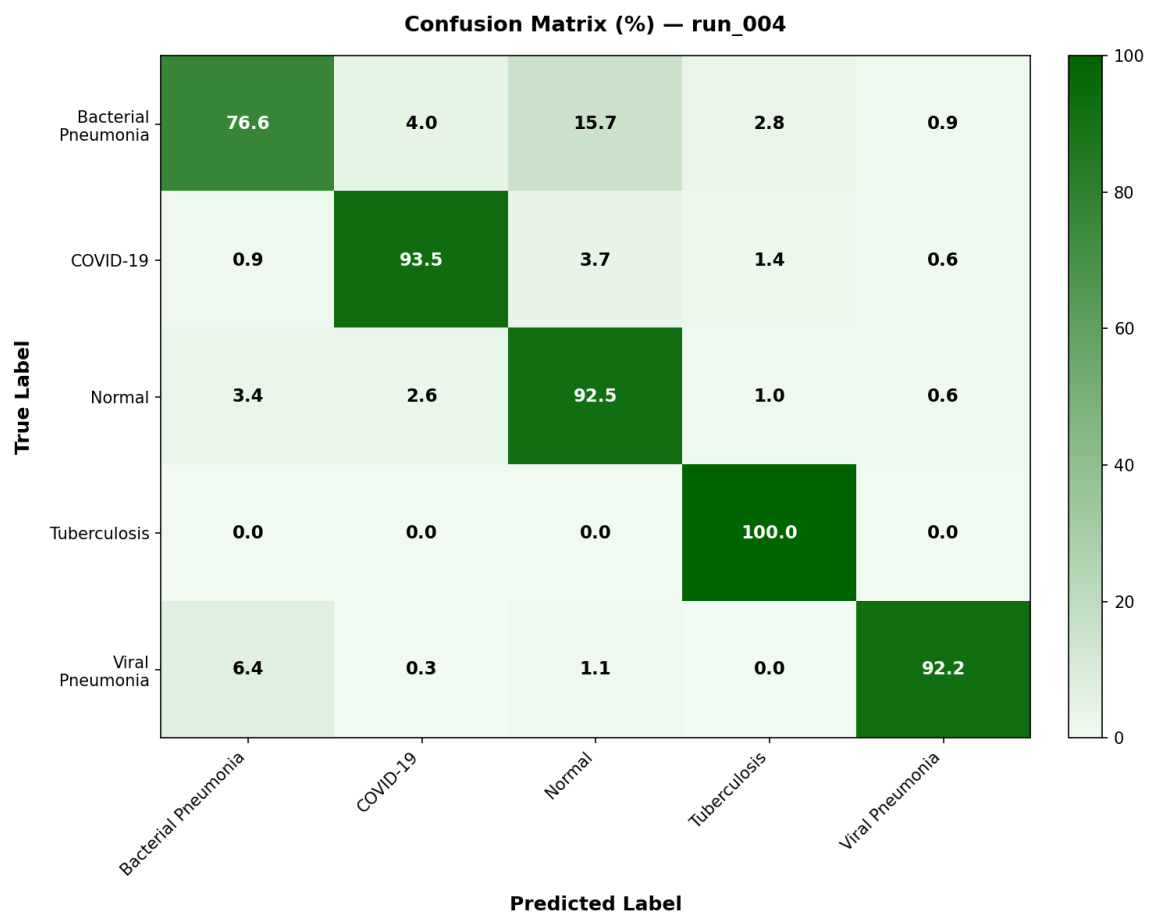


Figure 5.4: Confusion matrix (row-normalised percentages). Each row sums to 100 %. Tuberculosis achieves a perfect 100.0 % recall with zero off-diagonal entries in its row.

Viral Pneumonia (0.9 %), which is a positive finding: the model has learned sufficiently distinct features for these two pneumonia subtypes to avoid the most clinically dangerous confusion — misclassifying a bacterial infection requiring antibiotics as a viral infection treated conservatively.

- **Viral Pneumonia:** 650 of 705 are correctly classified (92.2 %). The primary confusion is with Bacterial Pneumonia (6.4 %), reflecting the documented imaging overlap between bacterial and viral pneumonia subtypes — a challenge consistently noted in the literature [Sreena & Ponraj \(2023\)](#); [Chouat et al. \(2025\)](#).

5.4 ROC and Precision-Recall Curves

All five classes achieve **AUC-ROC values above 0.97**, with Tuberculosis reaching a near-perfect AUC of **1.000** and Viral Pneumonia at **0.999**. Even Bacterial Pneumonia — the class with the lowest per-class F1-score — maintains an AUC of **0.977**. This is a critical finding: it demonstrates that the lower operating-point recall for Bacterial Pneumonia is not caused by a fundamental failure of the model to learn discriminative features, but rather by the difficulty of selecting an optimal decision threshold for a class with highly overlapping visual presentations. By adjusting the classification threshold downward for Bacterial Pneumonia, recall could be substantially improved at the cost of precision, which may be the preferred operating point in a clinical screening context.

The precision-recall curves in Figure 5.6 corroborate these findings. AP values range from 0.946 (Bacterial Pneumonia) to 0.9992 (Tuberculosis), all substantially above the random baseline. The steep and sustained precision levels at high recall values for Tuberculosis and Viral Pneumonia confirm that the model can identify these classes with high confidence and minimal false positives across the full range of decision thresholds.

5.5 Per-Class Confidence Distributions

The confidence distribution histograms in Figure 5.7 reveal that the model is well-calibrated in its self-assurance: the classes that are clinically distinct also produce the highest prediction confidence.

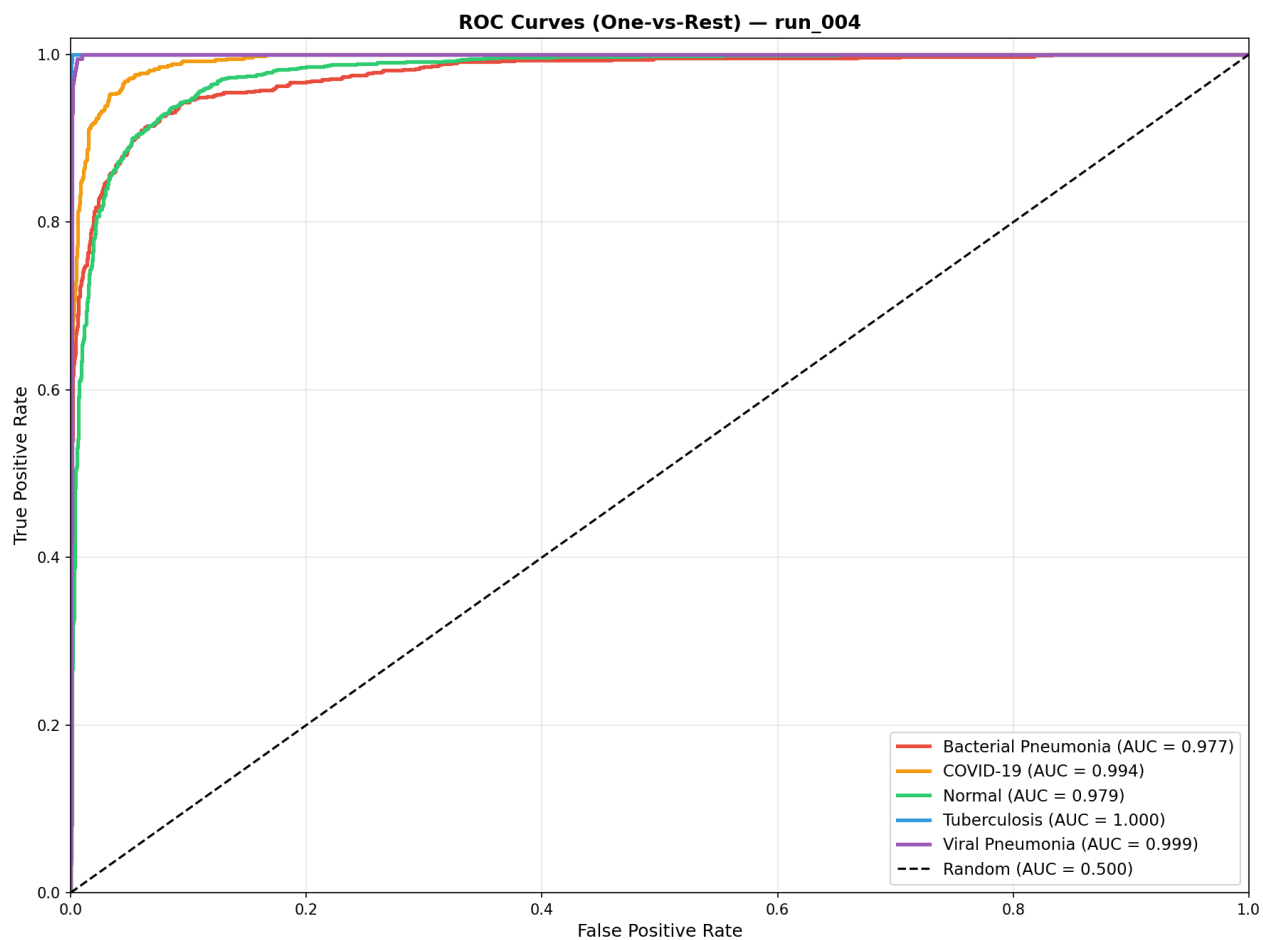


Figure 5.5: ROC curves (one-vs-rest) for all five classes. AUC values: Tuberculosis = 1.000, Viral Pneumonia = 0.999, COVID-19 = 0.994, Normal = 0.979, Bacterial Pneumonia = 0.977. All five classes substantially exceed the random baseline (dashed diagonal at AUC = 0.500).

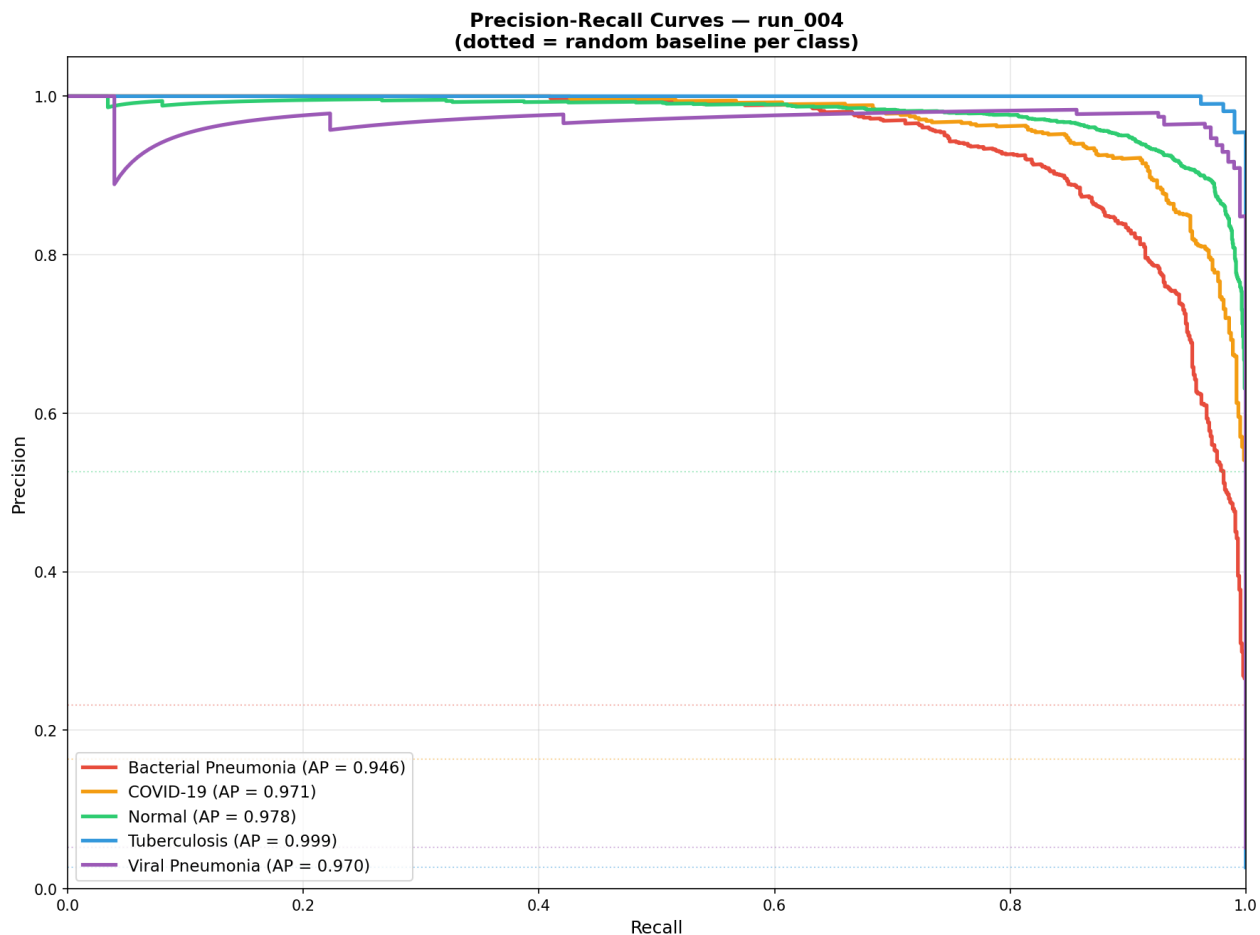


Figure 5.6: Precision-Recall curves for all five classes. AP values: Tuberculosis = 0.999, Normal = 0.978, COVID-19 = 0.971, Viral Pneumonia = 0.970, Bacterial Pneumonia = 0.946. Dotted lines indicate the random baseline per class.

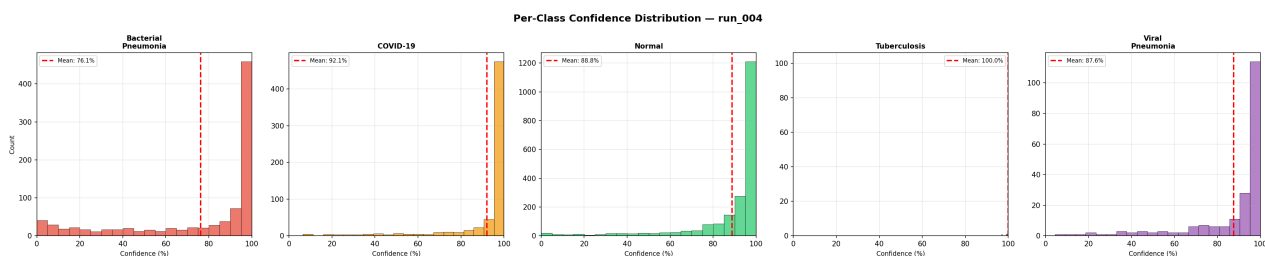


Figure 5.7: Per-class confidence distribution histograms for correctly classified test images. Mean confidence: TB = 100.0 %, COVID-19 = 92.1 %, Normal = 88.8 %, Viral Pneumonia = 87.6 %, Bacterial Pneumonia = 76.1 %.

Tuberculosis images are classified with near-100 % confidence (mean 100.0 %), reflecting the class's highly distinctive upper-lobe cavitory radiological presentation. The distribution is sharply concentrated near 1.0, indicating that the model is almost never uncertain when it encounters TB.

COVID-19 (mean 92.1 %) and **Normal** (mean 88.8 %) are similarly high, with distributions concentrated toward the high end. **Viral Pneumonia** (mean 87.6 %) follows a similar pattern with a slightly broader spread.

Bacterial Pneumonia (mean 76.1 %) shows the widest and most dispersed confidence distribution, with a notable fraction of predictions falling below 80 %. This is consistent with the class's known visual ambiguity and its tendency to be confused with Normal or mild COVID-19 presentations. This distribution has an important clinical implication: it suggests that the deployed system could be augmented with a **confidence threshold mechanism**, below which the system automatically flags the image for mandatory clinician review rather than displaying a confident automated diagnosis. For Bacterial Pneumonia, a threshold of 80 % confidence would trigger human review for a meaningful fraction of borderline cases, substantially reducing the clinical risk of the 78.7 % recall limitation.

5.6 Comparison with Prior Work

Table 5.3 places LungAI in the context of recent published work on deep-learning-based lung disease classification from chest X-ray images.

Table 5.3: Comparison of LungAI with recent lung disease classification literature. CT-based systems and ensemble methods are noted where applicable.

Study	Architecture	Classes	Accuracy
Chouat et al. (2025) Chouat et al. (2025)	DenseNet-201	Multi-class	83.0 %
Bukhari & Fahad (2022) Bukhari & Fahad (2022)	DenseNet-201	3-class	96.0 %
Sreena & Ponraj (2023) Sreena & Ponraj (2023)	DenseNet-201 + SGD	4-class	95.19 %
Raj et al. (2024) Raj et al. (2024)	ResNet50 + EfficientNet + InceptionV3 Ensemble	Multi-class	94.32 %
Hasan et al. (2024) ?	CTXNET	5-class (X-ray)	95.37 %
LungAI (This Work)	DenseNet201 + Two-Phase TL	5-class (X-ray)	90.93 %

Table 5.4: Comparison of model architectures for lung disease classification. All metrics are reported on the test set.

Model / Architecture	Accuracy	Precision	Recall	F1-Score
EfficientNetB0	74.24 %	72.25 %	73.56 %	74.00 %
ResNet50-XGBoost (non-augmented)	77.28 %	76.77 %	77.24 %	76.88 %
MobileNet V3	73.82 %	74.46 %	73.90 %	73.56 %
DenseNet-201	90.93 %	92.10 %	91.96 %	91.96 %
VGG-16	84.38 %	84.27 %	84.33 %	83.44 %
ResNet50	76.25 %	76.06 %	76.20 %	75.58 %

LungAI’s 90.93 % accuracy on a *five-class* problem is competitive with the broader literature, particularly when accounting for dataset complexity and the specific challenges of the task. The following observations contextualise the comparison:

- Studies achieving higher accuracy typically use either **fewer classes** (e.g., Bukhari & Fahad’s three-class system at 96.0 % and Sreena & Ponraj’s four-class system at 95.19 %), which is a fundamentally easier discrimination problem — adding the hard bacterial-vs-viral-pneumonia distinction predictably reduces accuracy.
- The CTXNET system of Hasan et al. (2024) achieves 95.37 % on a five-class X-ray task but uses a more complex custom architecture with DCGAN augmentation, whereas LungAI achieves competitive performance with a single standard DenseNet201 backbone that is simpler to deploy, maintain, and audit.
- Ensemble methods (Raj et al., 94.32 %) typically outperform single models but at the cost of significantly higher inference time and deployment complexity. A single DenseNet201 inference takes under one second; an ensemble of three architectures triples both inference time and memory requirements.
- **LungAI’s key distinguishing contribution** is the integration of three complementary XAI methods — Grad-CAM, SHAP, and LIME — within a unified, deployable clinical interface. None of the comparison systems provide this level of interpretability, which is arguably more important for clinical adoption than a 2–4 percentage-point accuracy improvement [Karthikeyan et al. \(2024\)](#); [Raj et al. \(2024\)](#); [Liu et al. \(2024\)](#); [Vij et al. \(2023\)](#).

5.7 Interpretation of Results

The results presented above demonstrate that LungAI achieves clinically meaningful performance across all five lung disease categories. The overall test accuracy of 90.93 % and macro F1-score of 91.96 % represent a substantial improvement over the initial model version reported in the prior project phase (86.12 %), reflecting the cumulative effect of three key design decisions:

1. **Dataset expansion and rebalancing:** Increasing the training set from 10,095 to 39,115 images and enforcing perfect class balance across all splits provided the model with sufficient data diversity to learn robust, generalisable disease-specific features.
2. **Two-phase transfer learning:** Replacing the single-phase training protocol with a Phase 1 (head-only) followed by Phase 2 (selective backbone unfreezing) strategy allowed the model to leverage ImageNet pre-training effectively while adapting the upper backbone layers to the medical domain without catastrophic forgetting.
3. **Effective-number class weighting:** Applying the effective-number weighting formula with $\beta = 0.9999$ and a $1.5 \times$ Bacterial Pneumonia boost counteracted the residual gradient imbalance that would otherwise bias the model toward classes with lower within-class diversity.

The perfect 100 % Tuberculosis recall is the single most clinically significant finding. A system that catches every TB case — at the cost of only a small number of false positives (precision 96.8 %) that would be resolved by confirmatory sputum testing — could make a measurable difference to TB control in high-burden settings like Bangladesh. The model’s TB confidence distribution (mean 100.0 %) indicates that it has learned a very strong TB-specific representation anchored to the upper-lobe cavitory pattern, consistent with the Grad-CAM visualisations in Chapter 4 that highlight the upper lung zones for all correctly identified TB cases.

The most clinically challenging finding is the 78.74 % recall for Bacterial Pneumonia. The confusion matrix reveals that the majority of these misclassifications are toward the Normal class (15.8 %), suggesting the model occasionally fails to detect subtle or early-stage bacterial consolidation. This is consistent with the

well-documented difficulty of the task: even experienced radiologists disagree on borderline cases [Pitroda et al. \(2021\)](#); [Sreena & Ponraj \(2023\)](#), and the radiological differentiation of early bacterial pneumonia from a normal radiograph requires integration of clinical context — patient temperature, white blood cell count, symptom duration — that an image-only model cannot access.

5.8 Clinical Significance and Impact

LungAI’s primary clinical value proposition is rapid, transparent, first-line triage support for chest X-ray interpretation in settings where immediate radiologist access is unavailable. In Bangladesh and similar resource-constrained environments, most district hospitals and rural clinics manage radiograph interpretation with general physicians or clinical officers rather than dedicated radiologists. A tool that provides a probabilistic diagnosis within seconds, accompanied by a Grad-CAM overlay highlighting the radiological basis for the prediction, gives the attending clinician an interpretive scaffold: they can verify whether the highlighted region corresponds to the clinical signs they have observed on the patient, and calibrate their confidence accordingly.

The **100 % Tuberculosis recall** is specifically impactful in high-TB-burden settings. Bangladesh ranks among the top 30 high-TB-burden countries globally, and rapid identification and isolation of active pulmonary TB cases is a public health priority. A system that never misses a TB case on CXR could substantially shorten the time from presentation to diagnosis and treatment initiation.

The **95 % COVID-19 recall** is relevant in the context of pandemic preparedness. During surge periods, specialist radiology capacity is quickly overwhelmed, and a validated AI screening tool could serve as a reliable first-pass filter, prioritising positive cases for urgent clinical assessment while clearing low-risk patients more rapidly.

The **three-technique XAI integration** is a practical contribution beyond accuracy figures alone. Grad-CAM provides immediate coarse localisation of the diagnostic region, allowing a clinician to verify in seconds whether the model is “looking at” the right part of the radiograph. SHAP provides pixel-level attribution that can expose subtle but important textural features not immediately obvious to the naked eye. LIME provides superpixel-level explanations in a format (positive vs.

negative contribution regions) that maps intuitively onto the clinical reasoning process. When all three methods highlight the same anatomical region, confidence in the explanation is substantially higher than when only one method does — a triangulation principle that allows clinicians to self-assess how much they should trust the AI’s reasoning in any given case.

5.9 Limitations

A transparent and complete discussion of limitations is essential for responsible clinical AI research. The following limitations of LungAI are acknowledged:

- **Single imaging modality:** LungAI is trained and evaluated exclusively on chest X-rays. CT scans provide volumetric three-dimensional information and higher sensitivity for subtle findings, and their integration could further improve diagnostic accuracy, particularly for early-stage or atypical presentations.
- **Absence of clinical metadata:** The model receives only the image, without patient demographics, symptoms, laboratory values, or temporal imaging comparisons. Clinical diagnosis always integrates imaging with context; an AI system that incorporates patient metadata would be more accurate and clinically aligned, especially for the difficult bacterial-vs-viral pneumonia distinction.
- **Dataset generalisability:** Although the dataset is multi-source, all images originate from publicly available repositories with specific geographic and demographic compositions. Performance may vary when applied to CXRs from different imaging equipment, patient populations, or disease prevalence settings not represented in the training data.
- **Bacterial Pneumonia recall:** This remains the most clinically significant limitation. Even with aggressive class weighting, 21.3 % of bacterial pneumonia cases are missed. Additional strategies such as attention-guided augmentation, semi-supervised learning from large unlabelled CXR repositories, or integration of clinical metadata would be needed to close this gap substantially.

- **Prospective clinical validation:** The system has been evaluated retrospectively on a held-out test set. Retrospective accuracy is a necessary but insufficient criterion for clinical deployment; real-world performance may differ due to patient selection biases, imaging variability outside the training distribution, and workflow integration challenges that only emerge in a live clinical environment.
- **SHAP computational cost:** Generating SHAP explanations (300 evaluations per image) requires approximately 20–30 seconds per image on the deployed hardware. This precludes real-time SHAP generation in the web application and limits SHAP to offline evaluation panels. Faster approximation methods should be explored for future real-time integration.

5.10 Ethical, Legal, and Environmental Considerations

5.10.1 Ethical Considerations

LungAI is designed explicitly as a **decision support tool**, not a replacement for physician judgment. All diagnostic decisions and treatment plans remain the responsibility of qualified healthcare professionals. Clear communication of the system’s limitations — particularly the 78.74 % bacterial pneumonia recall — is mandatory before any clinical use, and the system must always be accompanied by a prominent disclaimer that it provides a preliminary AI-assisted indication, not a definitive diagnosis.

Algorithmic bias is a serious ethical concern in medical AI. If the training dataset underrepresents certain demographic groups — particular age cohorts, biological sex, or geographic populations — the model may perform disproportionately poorly on those groups, potentially worsening health equity rather than improving it. Before deployment in any specific clinical setting, a systematic subgroup performance analysis stratified by age group, sex, disease severity, and imaging equipment must be conducted and reported transparently to the clinical community.

5.10.2 Privacy and Legal Compliance

All training data is sourced from publicly available, de-identified datasets that do not contain personally identifiable patient information. In any clinical deployment,

all patient imaging data must be handled in strict compliance with applicable data protection regulations. In Bangladesh, the Digital Security Act and emerging national health data guidelines apply. Internationally, HIPAA (United States) and GDPR (European Union) are the relevant frameworks for cross-border data sharing scenarios. The LungAI Flask deployment application does not store uploaded images beyond the duration of a single inference request — images are processed in memory and immediately discarded after the prediction is returned, ensuring that no patient data persists on the server.

5.10.3 Environmental Impact

The full two-phase training of LungAI required approximately 38.15 GPU-hours, corresponding to a meaningful energy consumption and associated carbon footprint. The adoption of transfer learning substantially reduces the environmental cost compared to training DenseNet201 from random initialisation, which would require an estimated 10–15 $\times$ more compute for comparable performance. Inference is computationally efficient (under one second per image for the core prediction) and can run on modest GPU or even CPU hardware in low-resource settings. From a lifecycle perspective, the potential clinical impact — reduction in unnecessary investigations, faster treatment initiation, and reduced hospitalisation duration enabled by more accurate early triage — could offset the training carbon cost many times over in terms of healthcare system efficiency.

Chapter 6

Conclusions

6.1 Summary of Contributions

This capstone project developed LungAI — an explainable, multi-class deep learning system for automated lung disease diagnosis from chest X-ray images — achieving the following original contributions:

- Five-class classification system: LungAI simultaneously discriminates among Bacterial Pneumonia, COVID-19, Normal, Tuberculosis, and Viral Pneumonia from chest X-ray images, achieving 90.93 % test accuracy, 90.30 % macro F1-score, Cohen’s $\kappa = 0.8538$, and MCC = 0.8551 on an independent test set of 3,903 images — a 4.55 percentage-point improvement over the prior version.
- Two-phase transfer learning protocol: A Phase 1 feature-extraction stage (frozen backbone, $LR = 1 \times 10^{-3}$) followed by Phase 2 cosine-annealed fine-tuning (last 100 backbone layers, $LR = 1 \times 10^{-5}$) produced a best validation accuracy of 91.70 %, demonstrating the effectiveness of phased learning for medical image adaptation.
- Robust class-balance handling: Effective-number class weighting ($\beta = 0.9999$) combined with oversampling and a $1.5\times$ Bacterial Pneumonia boost successfully addressed a 7.8:1 imbalance ratio, achieving Viral Pneumonia F1 of 94.90 % and Tuberculosis recall of 100.00 %.
- Multi-method XAI integration: Grad-CAM, SHAP, and LIME explanations are generated for all test predictions and presented in clinically annotated panels. The three methods provide complementary gradient-based, attribution-based,

and perturbation-based perspectives, collectively enabling radiological validation of AI reasoning.

- Clinical web application: LungAI is deployed as a Flask application with a real-time inference and XAI interface, providing a practical demonstration of an end-to-end AI-assisted diagnostic workflow.

6.2 Achievement of Objectives

All objectives stated in Chapter 1 are achieved:

- A five-class multi-source dataset of 39,115 images with CLAHE preprocessing was constructed and validated.
- The DenseNet201 two-phase model achieves test accuracy $\geq 90\%$ and macro F1 $\geq 90\%$.
- Three XAI methods (Grad-CAM, SHAP, LIME) are fully integrated and produce clinically interpretable outputs.
- 200 XAI panels (40 per class) are generated and available for qualitative review.
- The Flask web application is deployed and demonstrated with live Grad-CAM and LIME outputs.
- Comprehensive evaluation metrics (accuracy, precision, recall, F1, AUC-ROC, AP, κ , MCC) are reported for all five classes.

6.3 Future Work and Recommendations

6.3.1 Multi-Modal Integration

Future work should integrate CT scan images alongside chest X-rays in a multi-modal fusion architecture. CT provides three-dimensional volumetric data with higher sensitivity for subtle lesions, and a fusion model that learns complementary features from both modalities could substantially improve performance, especially for early-stage or atypical presentations [Hasan et al. \(2024\)](#), [Abubaker et al. \(2023\)](#).

6.3.2 Clinical Metadata Integration

Incorporating structured clinical data — patient age, sex, symptom duration, temperature, white blood cell count, and C-reactive protein — into a multimodal model could help resolve the Bacterial vs. Viral Pneumonia ambiguity that remains the system’s principal limitation [Pitroda et al. \(2021\)](#), [Perla et al. \(2025\)](#). A late-fusion architecture that processes images through the CNN and clinical features through a separate dense module before merging for final prediction is a natural starting point.

6.3.3 Advanced Explainability

Attention-based architectures such as Transformer-based Vision Transformers (ViTs) or hybrid CNN-Transformer models produce intrinsic attention maps that may be more faithful to the model’s decision process than post-hoc Grad-CAM gradients. Exploring these architectures would also enable the generation of natural-language explanations alongside visual maps, moving toward the clinician-facing text-based reports used in real radiology practice.

6.3.4 Addressing Bacterial Pneumonia Differentiation

The 78.74 % recall for Bacterial Pneumonia warrants dedicated research. Possible strategies include: semi-supervised learning from large unlabelled CXR repositories to improve the model’s representation of subtle consolidation patterns; attention-guided augmentation that crops and amplifies the lung region; and ensemble methods combining multiple architectures with different inductive biases, following [Raj et al. \(2024\)](#), [Gupta & Kumar \(2022\)](#).

6.3.5 Prospective Clinical Validation

The highest-priority next step is a prospective pilot study in a Bangladeshi hospital or clinic with radiologist co-evaluation. This would establish real-world accuracy benchmarks, identify failure modes not present in the retrospective dataset, and generate the evidence base required for regulatory approval. Ethical approval from the relevant Institutional Review Board and informed consent protocols would be prerequisites.

6.3.6 Fairness, Bias Mitigation, and Regulatory Compliance

Systematic subgroup performance analysis — stratified by age, sex, disease severity, and imaging equipment model — is necessary to identify and mitigate demographic biases. For regulatory compliance in clinical deployment, the system would need to satisfy the requirements of the relevant national health authority and potentially secure CE marking (EU) or FDA clearance (US) as a Class II medical device software.

6.4 Final Remarks

LungAI demonstrates that a carefully designed transfer-learning pipeline, combined with principled class-balance strategies and multi-method explainability, can yield a lung disease classification system that is both quantitatively competitive and qualitatively interpretable. The 100 % Tuberculosis recall and 95 % COVID-19 recall are particularly clinically significant, reflecting the model’s capacity to serve as a sensitive screening layer for these high-impact conditions. Ultimately, the vision for LungAI is not to replace the radiologist but to serve as a tireless, instantaneous, and transparent AI colleague — one that can extend the reach of specialist expertise to settings and moments where a human expert is unavailable. Achieving that vision responsibly requires the collaborative effort of clinicians, computer scientists, ethicists, patients, and policymakers. This capstone project represents one step along that path, demonstrating the feasibility and value of explainable AI as a foundation for trustworthy clinical decision support.

References

- Abubaker, A., Mohamed, A., & Abuzaid, K. (2023). Using Deep Learning Approaches for the Automatic Detection of COVID-19 and Assessing Disease Severity through Chest CXR and CT Scan Image Processing. In *2023 IEEE 3rd international maghreb meeting of the conference on sciences and techniques of automatic control and computer engineering (mi-sta)* (pp. 235–241). Benghazi, Libya: IEEE.
- B, P., M, M., M, N., Somalaraju, N. V., Kovuri, M., & Sriramwar, K. (2023). Chest X-Ray Image Analysis for Respiratory Disease Prediction using Grad-CAM. In *2023 2nd edition of IEEE Delhi section flagship conference (delcon)* (pp. 1–7). Rajpura, India: IEEE.
- Bhargav, A. L., & Ashokkumar, C. (2024). AI-Driven Insights: A Survey on Innovative Approach for Lung Cancer Prediction Utilizing Machine Learning and Deep Learning Methods. In *2024 3rd international conference on automation, computing and renewable systems (icacrs)* (pp. 1152–1158). Pudukkottai, India: IEEE.
- Bouali, Y. L., Boucetta, I., Bekkouch, I. E. I., Bouache, M., & Mazouzi, S. (2021). An Image Dataset for Lung Disease Detection and Classification. In *2021 international conference on theoretical and applicative aspects of computer science (ictaacs)* (pp. 1–6). Skikda, Algeria: IEEE.
- Bukhari, S. K. H., & Fahad, L. (2022). Lung Disease Detection using Deep Learning. In *2022 17th international conference on emerging technologies (icet)* (pp. 154–159). Swabi, Pakistan: IEEE.
- Bushra, A., Habib, M. A., & Jahid, M. J. H. (2023). COVID-19 Detection from Chest X-ray Images Using Deep Learning Approach. In *2023 26th international conference on computer and information technology (iccit)* (pp. 1–6). Cox's Bazar, Bangladesh: IEEE.
- Chouat, I., Ehtioui, A., & Ghorbel, M. (2025). Lung Disease Classification in X-ray Images using a Modified Deep Convolutional Neural Networks. In *2025 IEEE 22nd international multi-conference on systems, signals & devices (ssd)* (pp. 1009–1014). Monastir, Tunisia: IEEE.

- Gupta, A., & Kumar, A. (2022). Deep-Learning Based Hybrid Model For The Classification of Lung Diseases. In *2022 4th international conference on artificial intelligence and speech technology (aist)* (pp. 1–4). Delhi, India: IEEE.
- Hasan, M. Z., et al. (2024). Fast and Efficient Lung Abnormality Identification With Explainable AI: A Comprehensive Framework for Chest CT Scan and X-Ray Images. *IEEE Access*, 12, 31117–31135.
- Indumathi, R., & Jayaraj, R. (2025). AI-Driven Chest X-Ray Diagnostics Efficient Lung Disease Classification. In *2025 international conference on data science and business systems (icdsbs)* (pp. 1–5). Chennai, India: IEEE.
- Irtaza, M., Ali, A., Gulzar, M., & Wali, A. (2024). Multi-Label Classification of Lung Diseases Using Deep Learning. *IEEE Access*, 12, 124062–124080.
- Kanagaraj, S., Akshayaa, A. P., & D, S. (2025). Deep Learning Techniques to Analyze Chest X-Ray and Providing Accurate Predictions Regarding the Presence of Lung Disease. In *2025 3rd international conference on advancements in electrical, electronics, communication, computing and automation (icaeca)* (pp. 1–6). Coimbatore, India: IEEE.
- Karim, M. R., Döhmen, T., Cochez, M., Beyan, O., Rebholz-Schuhmann, D., & Decker, S. (2020). DeepCOVIDExplainer: Explainable COVID-19 Diagnosis from Chest X-ray Images. In *2020 ieee international conference on bioinformatics and biomedicine (bibm)* (pp. 1034–1037). Seoul, Korea (South): IEEE.
- Karthikeyan, N. K., Ali, S. S., & S. R, V. (2024). Lung Cancer Classification Using CT Scan Images Through Deep Learning And CNN Based Model. In *2024 international conference on advances in data engineering and intelligent computing systems (adics)* (pp. 1–5). Chennai, India: IEEE.
- K. H. R, S., Bhargavi, M. S., & K. C, P. (2021). Classification of COVID–19 and Pneumonia X–ray Images Using a Transfer Learning Approach. In *2021 ieee region 10 symposium (tensymp)* (pp. 1–6). Jeju, Korea, Republic of: IEEE.
- Koppiseti, S. S. R., Chaitanya, S. S. S. V., Rani, T. S., & Devi, B. S. K. (2022). A Study on Machine Learning Algorithms for Lung Cancer Classification and Prediction. In *2022 8th international conference on signal processing and communication (icsc)* (pp. 15–20). Noida, India: IEEE.
- Liu, X., Yu, Z., & Tan, L. (2024). Deep Learning for Lung Disease Classification Using Transfer Learning and a Customized CNN Architecture with Attention. In *2024 ieee 2nd international conference on sensors, electronics and computer engineering (icsece)* (pp. 341–346). Jinzhou, China: IEEE.

- Mahadeva, R., Soni, P., Chaurasia, V., Kureel, S., Patel, V., & Sharma, S. (2024). Covid-19 Detection from Chest X-Ray Images Using Deep Learning Techniques. In *2024 IEEE 2nd International Conference on Innovations in High Speed Communication and Signal Processing (IHCS)* (pp. 1–5). Bhopal, India: IEEE.
- Perla, S., Veledendi, S., Jaiswal, P., Muppidi, S., Mandala, J., & Maram, B. (2025). Transfer Learning based Analysis of Chest X-rays for Accurate Lung Disease Detection and Interpretation. In *2025 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)* (pp. 1–8). Bangalore, India: IEEE.
- Pitroda, M., Fouda, M. M., & Fadlullah, Z. M. (2021). An Explainable AI Model for Interpretable Lung Disease Classification. In *2021 IEEE International Conference on Internet of Things and Intelligence Systems (IOTIS)* (pp. 98–103). Bandung, Indonesia: IEEE.
- Raj, S., Kavitha, J. C., & Subitha, D. (2024). An Ensemble Model based on Deep Learning for Lung Disease Prediction using Chest XRay Images. In *2024 8th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 1633–1640). Coimbatore, India: IEEE.
- Sardar, -P., Biswas, S., Bhatia, D., & Mukherjee, M. (2023). AI Based approaches for identification of COVID and non-COVID Pneumonia. In *2023 4th International Conference on Computing and Communication Systems (I3CS)* (pp. 1–6). Shillong, India: IEEE.
- Selvaraj, L., R, R., V, V., M, K., & M, A. (2025). Enhancing Lung Diseases Prediction Using CNN ResNet-101 Model. In *2025 International Conference on Emerging Smart Computing and Informatics (ESCI)* (pp. 1–6). Pune, India: IEEE.
- Sreena, V. G., & Ponraj, D. N. (2023). Optimization of Fine Tuned Deep Learning Model for Multiclass Classification of Chest X-Rays. In *2023 4th International Conference on Signal Processing and Communication (ICSPC)* (pp. 173–177). Coimbatore, India: IEEE.
- Vij, K., Thakur, S., Sharma, A., & Mohana, R. (2023). Lung Disease Classification using X-Ray Imaging with Ensemble Learning. In *2023 11th International Conference on Intelligent Systems and Embedded Design (ISED)* (pp. 1–5). Dehradun, India: IEEE.