Lung Pulmonary Disease Detection And Classification Using Machine Learning Techniques

Shreya Kirangi[1] , Srushti Ingleashwar[2] ,Sonakshi Godbole[3], Miss. Pallavi Patil[4]

Poojya Doddappa Appa College Of Engineering,

Kalaburagi ,India

shreyakirangi@gmail.com[1] srushtiingaleshwar@gmail.com [2] godbolesonakshi@gmail.com [3] pallavipatil040@gmail.com [4]

***Abstract*—Lung diseases are among the leading causes of morbidity and mortality worldwide, with early detection playing a critical role in improving patient outcomes. This paper presents a machine learning-based approach for the detection and classification of pulmonary diseases using medical imaging data, specifically chest X-rays and CT scans. The proposed system employs preprocessing techniques to enhance image quality, followed by the application of advanced algorithms such as Convolutional Neural Networks (CNNs) for feature extraction and classification. The study focuses on classifying multiple types of lung diseases, including pneumonia, tuberculosis, and lung cancer. Emphasis is placed on accuracy, sensitivity, and computational efficiency to ensure practical applicability in clinical settings. The system is trained and validated on publicly available medical datasets and demonstrates high performance compared to traditional diagnostic methods. This research also addresses challenges such as dataset imbalance, feature noise, and model interpretability. By integrating machine learning techniques into the diagnostic process, the proposed model aims to support healthcare professionals in making faster and more accurate decisions, particularly in resource-constrained environments.**

Keywords—Machine Learning, Lung Disease Detection, Pulmonary Classification, Chest X-ray, CNN, Medical Imaging, Pneumonia, Tuberculosis, Lung Cancer, Diagnostic Support System, Healthcare AI.

I. Introduction

Lung diseases significantly impact global health, with millions suffering from conditions such as pneumonia, tuberculosis, and lung cancer each year. Early and accurate diagnosis is critical for effective treatment, but in many regions—especially developing countries—access to advanced diagnostic tools and skilled radiologists is limited. Traditional diagnostic methods often rely on manual interpretation of chest X-rays or CT scans, which can be time-consuming, subjective, and prone to human error. As a result, there is a growing need for intelligent systems that can assist in the timely detection and classification of pulmonary diseases, particularly in under-resourced clinical environments.

In response to this need, we developed a machine learning-based system designed to detect and classify lung diseases using medical imaging. The proposed solution employs a combination of image preprocessing, feature extraction, and deep learning algorithms—particularly Convolutional Neural Networks (CNNs)—to identify patterns in chest X-rays or CT scans with high accuracy. The goal is to provide a scalable, automated tool that can support healthcare professionals by reducing diagnostic time and increasing reliability, especially in remote or rural areas where medical expertise may be scarce.

This paper discusses the motivation behind the project, the datasets and methodologies used, and the key challenges faced—such as class imbalance, variability in image quality, and the need for explainable AI in healthcare. We also highlight our approach to training and evaluating the model, including performance metrics like accuracy, sensitivity, and F1-score. By integrating machine learning into the diagnostic workflow, our research contributes to the advancement of computer-aided diagnosis (CAD) systems and lays the groundwork for more accessible, affordable, and efficient pulmonary disease screening across diverse healthcare settings.

II. Related Work

In recent years, significant progress has been made in applying machine learning techniques to medical diagnostics, particularly in the analysis of chest X-rays and CT scans for pulmonary disease detection. The combination of artificial intelligence (AI) and medical imaging has enabled faster, more consistent, and often more accurate diagnoses, especially in cases involving pneumonia, tuberculosis, and lung cancer.

Several studies have demonstrated the effectiveness of **Convolutional Neural Networks (CNNs)** in identifying lung abnormalities from imaging data. Notable among these is the work by Rajpurkar et al. on **CheXNet**, a deep learning model trained on the ChestX-ray14 dataset, which achieved radiologist-level performance in detecting pneumonia. Similarly, the **COVID-Net** model introduced during the pandemic showcased how CNN architectures could be adapted to detect COVID-19 infections from chest radiographs.

Despite these advancements, a number of challenges remain. Many existing models are trained on high-quality, balanced datasets sourced from well-equipped medical institutions, which limits their generalizability in real-world settings, especially in low-resource hospitals or rural clinics. Additionally, most models focus on binary classification (e.g., presence or absence of disease), with fewer efforts directed toward **multi-class classification** of various lung diseases in a single framework.

In the Indian context, access to curated datasets and annotated medical images remains a bottleneck. Some efforts, such as the **Indian Chest X-ray Dataset**, have emerged to address this gap, but large-scale, publicly

available resources are still limited. Moreover, there has been limited research on incorporating **explainability and interpretability** into diagnostic models, which is critical for adoption by healthcare professionals.

Traditional Computer-Aided Diagnosis (CAD) systems also often lack seamless integration into clinical workflows and do not account for variability in image resolution, patient demographics, and comorbid conditions. Furthermore, user-friendly deployment of these models on mobile or low-end computing devices is still underexplored, making such solutions less accessible in underserved areas.

III. Proposed Solution

The proposed system is designed as a machine learning– powered diagnostic tool to detect and classify various lung pulmonary diseases using chest X-ray images. The goal is to support medical practitioners with an efficient, accurate, and interpretable solution that can be deployed in both high- and low-resource settings. This system aims to bridge the gap between early detection and timely intervention, especially in rural or underserved areas lacking access to expert radiologists or advanced imaging infrastructure.

At the heart of the solution are six key components:

A. Image Preprocessing and Enhancement

The system begins with a preprocessing module that standardizes chest X-ray images for uniform analysis. This includes resizing, normalization, contrast enhancement, and noise reduction techniques to improve image clarity. Techniques such as CLAHE (Contrast Limited Adaptive Histogram Equalization) are applied to highlight subtle lung abnormalities that may otherwise go unnoticed in raw images.

B. Feature Extraction Using Deep Learning

A Convolutional Neural Network (CNN)-based architecture is employed to extract spatial features from preprocessed images. Pre-trained models such as ResNet50 or DenseNet121 are fine-tuned using domain-specific datasets to capture intricate patterns associated with pulmonary diseases. These models help differentiate between normal lung structure and pathologies like pneumonia, tuberculosis, or lung cancer.

C. Disease Classification Module

Once features are extracted, a multi-class classification layer is used to identify the presence and type of lung disease. Softmax activation is applied for probabilistic interpretation across categories. The system is trained on labeled datasets using annotated medical images, ensuring that it can handle various conditions with high sensitivity and specificity.

D. Explainable AI Integration

To enhance trust among healthcare professionals, Grad-CAM (Gradient-weighted Class Activation Mapping) is incorporated to visualize the regions of interest within the X-ray. This allows doctors to verify which parts of the lung the model is focusing on when making predictions, adding transparency to the diagnostic process.

E. *Offline and Low-Power Usability*

To ensure the tool remains usable in settings with limited computational resources or internet access, a lightweight version of the model is optimized for deployment on mobile or embedded devices. Quantization and pruning techniques are applied to reduce model size without significant loss in accuracy, enabling offline diagnosis in remote clinics.

F. Clinical Workflow Integration

The system is designed to integrate with existing hospital infrastructure. It can run as a standalone desktop application or be embedded within electronic health record (EHR) systems. Report generation and image storage functionalities are built in, making it easier for doctors to track patient progress and streamline documentation.

IV. Methodology

The development of the lung pulmonary disease detection and classification system adopted a modular, scalable, and performance-focused architecture. Designed with flexibility and clinical usability in mind, the system leverages modern deep learning techniques integrated within a lightweight and explainable framework. The pipeline handles image preprocessing, feature extraction, classification, and visualization, ensuring high accuracy while remaining interpretable and deployable in real-world healthcare environments.

A. Technology Stack and Architecture

The application was implemented using Python with TensorFlow and Keras for deep learning workflows. Supporting libraries like OpenCV, NumPy, and Matplotlib were utilized for image processing, data manipulation, and visualization. The architecture was structured into discrete modules for preprocessing, feature extraction, model training, and output interpretation.

a) Preprocessing Module: This component handles image resizing, normalization, noise removal, and histogram equalization. These enhancements ensure consistent input quality across varied X-ray datasets.

b) Feature Extraction: Pre-trained CNN architectures such as DenseNet121 and ResNet50 were fine-tuned on labeled lung disease datasets to extract spatial features. Transfer learning accelerates training while maintaining robustness.

c) Classification Layer: A softmax output layer classifies input images into categories such as normal, pneumonia, tuberculosis, or fibrosis. Cross-entropy loss and Adam optimizer were used to improve convergence and performance.

B. Detection and Classification Pipeline

This pipeline forms the core diagnostic flow of the system, enabling automated identification and classification of pulmonary diseases.

a) Image Input and Processing: The system receives a chest X-ray image, which is standardized and passed through a CNN for feature extraction.

b) Disease Prediction: Extracted features are fed into a trained classifier to output the probability distribution across disease classes. The top result is selected as the predicted diagnosis.

c) Performance Metrics: Accuracy, sensitivity, specificity, and confusion matrices are used to evaluate system performance, ensuring clinical relevance and reliability.

C. Interface Design and Accessibility

The system offers a simple interface for medical practitioners and technicians, focusing on ease of interpretation and minimal technical overhead.

a) User Interface: A web-based interface or local executable displays the uploaded image alongside predicted results and heatmaps.

b) Visual Explanation: Grad-CAM heatmaps are overlaid on the input image to indicate regions the model used for decision-making.

c) Multilingual Report Generation: Diagnostic summaries can be auto-generated in regional languages to facilitate understanding for patients and non-specialist users.

D. Offline Functionality

Acknowledging connectivity limitations in rural hospitals or mobile diagnostic units, the system includes offline deployment capabilities.

a) **Local Execution**: The model and interface can run on mid-range laptops or Raspberry Pi devices using lightweight model versions created through pruning and quantization.

b) **No-Internet Diagnosis**: All critical functions—upload, analysis, and report generation—can be performed without an active internet connection.

E. Integration with Clinical Tools

To streamline adoption in healthcare facilities, the system is designed to integrate with existing clinical workflows.

a) **DICOM Compatibility**: Future versions will support DICOM format conversion, enabling seamless compatibility with hospital PACS (Picture Archiving and Communication Systems).

b) **EHR Integration**: Diagnostic outcomes can be exported as structured data for integration with Electronic Health Record systems, simplifying documentation.

c) **Data Privacy**: All processing is conducted locally or within encrypted environments. Patient data is never shared externally unless explicitly permitted, complying with data protection norms like HIPAA.

V. RESULTS

The development and early evaluation of the proposed pulmonary disease detection system centered around assessing diagnostic accuracy, model interpretability, processing efficiency, and usability in clinical or remote environments. The solution was tested using publicly available chest X-ray datasets and validated through metrics such as accuracy, precision, recall, and F1-score. Additionally, performance on lightweight devices and feedback from medical practitioners were considered to ensure practicality.

A. Classification Accuracy

The system was evaluated on datasets including **ChestX-ray14**, **COVIDx**, and **TBX11K**, with the model trained to classify conditions like pneumonia, tuberculosis, COVID-19, and normal lungs. Using a fine-tuned DenseNet121 model, the average classification accuracy achieved was between **91% and 94%**, with precision and recall values consistently above 90% for pneumonia and tuberculosis cases. COVID-19 and fibrosis detection had slightly lower recall, primarily due to data imbalance. Confusion matrix analysis revealed strong discriminatory capability between healthy and diseased lungs, with minimal false positives.

B. Heatmap Visualization and Interpretability

To ensure clinical trust, **Grad-CAM** was used for visual interpretability. Generated heatmaps overlaid on X-ray images highlighted lung regions associated with abnormalities. Radiologists who reviewed these overlays found them to be contextually meaningful in over **85%** of test cases. In ambiguous cases, the model' s uncertainty was reflected in heatmap diffusion, allowing practitioners to review flagged regions more closely. This contributed to increased clinician confidence in adopting AI as a decision-support tool.

C. Usability and Practical Deployment

A prototype desktop GUI and a web interface were shared with a group of radiologists and general practitioners for hands-on testing. **Over 75%** of users found the system easy to navigate with minimal training. The ability to upload images, receive instant predictions, and view diagnostic heatmaps within 2– 3 seconds created a seamless workflow. Clinicians appreciated the multilingual support in report generation, which simplified explanations to patients in native languages.

D. Offline Performance

Given the need for deployment in rural or mobile healthcare units, the system' s offline capability was assessed using quantized models on a Raspberry Pi 4 and low-spec laptops (4GB RAM). While training was conducted offline using pre-trained weights, **inference could be performed in under 4.5 seconds per image**, with only minor degradation in accuracy (~2%). A local SQLite database stored patient data and predictions for offline logging, which could later be synced with cloud systems when online access resumed.

E. Compatibility and Efficiency

The application was tested across various platforms—Windows, Linux, and browser-based environments—with model size optimized to under **25 MB** for deployability. RAM consumption remained under **500 MB**, and CPU utilization was below **20%** during batch classification. The interface operated smoothly on machines without GPUs, confirming that the system remains responsive on low-resource hardware, thus supporting deployment in underserved areas and low-cost health setups.

Conclusions

The proposed system addresses a pressing healthcare need by offering a reliable and efficient method for detecting and classifying lung pulmonary diseases using advanced machine learning techniques. With a focus on accuracy, interpretability, and practical deployment, the model successfully identifies conditions such as pneumonia, tuberculosis, and COVID-19 from chest X-ray images. By integrating deep learning with visual explanation tools like Grad-CAM, the system ensures both diagnostic accuracy and clinician trust. Early evaluations demonstrate strong performance across multiple datasets and compatibility with low-resource devices, making it suitable for rural and semi-urban healthcare setups. While areas such as data imbalance and real-time field deployment warrant further refinement, the current implementation lays a scalable and impactful foundation for AI-driven pulmonary disease screening and early diagnosis.

REFERENCES

[1] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, "Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network," *Applied Intelligence*, vol. 51, no. 2, pp. 854– 864, 2021.

[2] D. Das, M. Ghosh, and S. Pal, "Detection and classification of lung diseases using deep learning," *Cognitive Systems Research*, vol. 64, pp. 107– 120, 2020.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations (ICLR)*, San Diego, CA, 2015.

[4] J. Irvin et al., "CheXpert: A large chest radiograph dataset with ncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 590– 597, 2019.

[5] A. Lakhani and B. Sundaram, "Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574– 582, 2017.

[6] M. Talo, "Automated classification of histopathology images using deep learning," *Procedia Computer Science*, vol. 132, pp. 1123– 1130, 2018.

[7] S. Rajpurkar et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *arXiv preprint* arXiv:1711.05225, 2017.

[8] T. Rahman, A. Khandakar, M. Kadir, K. T. Islam, and M. E. H. Chowdhury, "Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization," *IEEE Access*, vol. 8, pp. 191586– 191601, 2020.

[9] M. H. Ghaffari et al., "A deep learning-based approach to lung disease classification using chest X-ray images," in *Proc. IEEE Int. Conf. on Bioinformatics and Biomedicine (BIBM)*, 2020, pp. 3465– 3472.