Detection and Classification of Lung Diseases using Hybrid Machine Learning Techniques

Vijay Shankar Shukla
Research Scholar Department of Computer Science, AKS University Satna

Pramod Singh
Associate Professor Department of Computer Science, AKS University Satna

ABSTRACT
Environmental changes, pollution, and different harmful daily habits—such as smoking and drinking—are major contributors to the number of lung diseases that need early diagnosis. Smoking not only distresses smokers but also those around them, frequently leading to respirational issues. This study presents a hybrid model that analyses patient symptoms to categorize the presence of lung diseases and dispenses a severity mark demonstrating whether the state is mild, moderate, or severe. If the user has an X-ray image and wants additional verification, they can upload the image to receive diagnostic results. The primary aim is to detect chronic lung diseases at an early stage, increasing the likelihood of timely treatment and improved survival rates. The proposed assumption includes hybrid machine learning techniques to predict conditions such as Tuberculosis, Pneumonia, and COPD [1,2,6]. Using Hybrid machine learning algorithm98.1 percent accuracy has been achieved.

Keywords
Lungs, Machine Learning, Infection, Detection

1. INTRODUCTION
Artificial intelligence (AI) is one of today's most important technologies in healthcare [1,2]. It is widely used to improve clinical decision-making, medical imaging, and disease diagnosis. Healthcare services have substantially improved as a result of rapid technological advancements, the availability of large medical datasets, and advances in machine learning (ML) and deep learning (DL) approaches[1,3,4].

AI-powered systems can examine medical data quickly and precisely, supporting clinicians with early disease detection and patient treatment [1, 2, 4].

Lung diseases are one of the most serious health problems, impacting millions of people worldwide [2, 6]. These illnesses damage the respiratory system and limit the lungs' ability to supply oxygen to the body. Bacterial, viral, or fungal infections, smoking, air pollution, environmental factors, and genetics can all contribute to long illness. Pneumonia, asthma, tuberculosis, chronic obstructive pulmonary disease (COPD), and lung cancer are among the most common lung illnesses [2,6]. Early identification of these illnesses is crucial, as delayed treatment can result in devastating outcomes, including death[2,8].

Pneumonia is a serious respiratory infection that mostly affects the alveoli[2,3], small air sacs found in the lungs. In pneumonia patients, these air sacs fill with fluid or pus, producing breathing difficulties, fever, chest pain, and decreased oxygen intake. Pneumonia is one of the most common causes of death among children under the age of five[2], particularly in underdeveloped nations. The condition can spread by bacterial, viral, or fungal infections and can be deadly if not treated effectively. According to worldwide health research, pneumonia remains a serious healthcare concern due to a lack of knowledge and late diagnosis.

Chest X-ray imaging is commonly used in pneumonia diagnosis because it helps clinicians to detect infections and abnormalities in the lungs. However, due to human limitations and the increasing burden on radiologists, manual evaluation of X-ray pictures may occasionally result in an incorrect diagnosis [3,4]. As a result, AI-based solutions are being developed to automate pneumonia detection using Deep Learning techniques [3,4,5]. CNN models can reliably classify chest X-ray images as normal, bacterial, or viral pneumonia. These computerized technologies help doctors make faster diagnosis and develop more effective treatment strategies.

Lung cancer is another significant lung illness and one of the worst types of cancer in the globe [1,2]. It results from the unregulated proliferation of aberrant cells within the lungs. Although smoking is thought to be the leading cause of lung cancer [2], air pollution, exposure to hazardous chemicals, and genetics all play a role. Every year, lung cancer kills millions of people because symptoms often arise late in the disease's progression. Thus, early diagnosis is critical for raising patient survival rates and enhancing treatment outcomes.

Chest X-rays and computed tomography (CT) scans are common medical imaging modalities used to diagnose lung cancer [1,6]. Radiologists use these pictures to diagnose lung nodules, cancers, and aberrant tissue growth. However, manual evaluation of CT scan pictures is time-consuming and may lead to diagnostic errors. To address these issues, Computer-Aided Diagnosis (CAD) systems based on AI and Deep Learning are becoming more popular in healthcare [1,4,6]. CNN-based algorithms can automatically analyze CT scans and detect malignant spots with high accuracy [1,5]. These tools help radiologists detect cancer at an early stage and lower the likelihood of misdiagnoses.

In recent years, researchers have concentrated on creating improved Deep Learning models for precise medical image classification and disease prediction[1,3,4,5]. The merging of AI with medical imaging has created new prospects for intelligent healthcare systems [1,4]. Deep Learning algorithms continue to improve disease diagnosis by detecting hidden patterns in medical photos that humans cannot see. As a result, AI-enabled healthcare systems are becoming an increasingly important part of current medical practice [1,2,4].

As a result, the use of Machine Learning and Deep Learning algorithms to diagnose pneumonia and lung cancer marks a significant development in healthcare technology. AI-based technologies not only help doctors make better clinical decisions, but they also increase patient care and treatment efficiency. Intelligent healthcare systems are projected to play a significant role in illness diagnosis and prevention in the future, as AI and medical imaging progress.

2. METHODOLOGY OF THE PROPOSED HYBRID LUNG DISEASE DETECTION MODEL

The proposed methodology presents an advanced hybrid artificial intelligence framework designed to detect and classify lung diseases from medical imaging data with high accuracy. The architecture integrates deep learning techniques with classical machine learning algorithms to improve feature representation and classification performance. The overall workflow of the system consists of several stages including data acquisition, pre-processing, feature extraction, hybrid feature fusion, feature optimization, classification, and final prediction. Each stage plays an important role in improving the efficiency and reliability of the diagnostic model.

2.1 Data Acquisition

The first step in the methodology involves collecting medical imaging datasets used for training and testing the proposed model. The Kaggle website's open-access datasets are the only sources of the datasets used in this study[6,8]. Pneumonia, COVID-19, normal, and lung carcinoma that were gathered for this study. There are 32,975 CT scan and CXR images in this gathered dataset which is divided into training, validation, and testing subsets to ensure unbiased evaluation of the model performance.

2.2 Data Pre-processing

Medical pictures commonly contain noise, inconsistent brightness, and resolution changes, all of which can degrade the efficiency of machine learning algorithms. Pre-processing improves image quality and standardises input data. Initially, photographs are scaled to a consistent dimension, often 224 × 224 pixels, which is compatible with most convolutional neural network architectures. Unwanted artifacts are removed using noise reduction techniques such as Gaussian filtering or median filtering. Histogram equalization is used to increase image contrast and highlight important anatomical structures in the lungs [4,5].

Data normalization is also used to scale pixel values inside a specific range, which leads to faster convergence during model training. Data augmentation procedures are used to increase model generalization while avoiding overfitting[3,4,5]. These methods include image rotation, horizontal and vertical flipping, zooming, translation, and brightness adjustment. Data augmentation artificially increases the dataset and exposes the model to a wide range of visual changes, making it more appropriate for real-world clinical settings.

2.3 Feature Extraction using Deep Learning

Feature extraction is an important stage in the proposed methodology. Rather than depending just on handwritten qualities, the approach uses deep learning architectures to automatically recognize complex patterns in medical photos. Transfer learning takes information from pre-trained convolutional neural network models like EfficientNet, ResNet, and DenseNet [1,4,5]. These structures have already been trained on huge image datasets and can effectively recognize hierarchical visual components [1,5] such as edges, textures, shapes, and disease patterns.
At this point, a feature extraction layer replaces the retrained model's final classification layer. The model processes the input images through numerous convolutional and pooling layers, producing deep feature maps that capture important lung properties.

2.4 Handcrafted Feature Extraction

In addition to deep features, the proposed hybrid model also extracts handcrafted features that capture specific statistical and texture information from medical images. Texture features are particularly useful in medical image analysis because lung diseases often produce subtle texture changes in tissues. Gray Level Co-occurrence Matrix (GLCM) is used to extract texture descriptors such as contrast, correlation, homogeneity, and energy[6,8].

The canny edge detector and other edge detection techniques are used to determine the borders and structural patterns of the lungs. Statistical parameters such as mean, variance, entropy, and standard deviation are also calculated to describe the pixel intensity distribution. These handcrafted features enhance the deep features extracted by convolutional neural networks, offering additional information that may improve classification accuracy.

2.5 Hybrid Feature Fusion

Following the extraction of deep learning and handcrafted features, the next step is to merge them into a comprehensive hybrid feature vector. Feature fusion combines the advantages of both feature types [1,6]. Deep features gather high-level semantic information, whereas handcrafted features generate interpretable statistical descriptors.

The features are concatenated into a single feature representation. This hybrid feature vector has a large number of attributes representing various aspects of lung images. However, not all traits are equally significant in the categorization process. Some features may be redundant or unnecessary, which increases processing complexity and reduces model performance.

2.6 Feature Optimization and Selection

Optimization strategies are used to choose the most pertinent attributes in order to address the problem of redundant features. For this, metaheuristic optimization methods like Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) are frequently employed [10]. These algorithms look for the best feature subset that minimizes redundancy and increases classification accuracy. Each particle in Particle Swarm Optimization represents a subset of candidate features, and the particles move around the search space by changing their positions in response to both individual and global optimal solutions. Similar to this, genetic algorithms use crossover, mutation, and selection to mimic the process of natural evolution. These techniques generate an optimal feature subset that lowers computational cost and greatly increases classification efficiency.

2.7 Hybrid Classification Model

The ideal hybrid feature set is then fed into an ensemble classification model. Instead of relying on a single classifier, the proposed strategy combines many machine learning techniques to enhance prediction performance. The ensemble model consists of Extreme Gradient Boosting (XGBoost), Random Forest (RF), and Support Vector Machine (SVM)[1,2,6].
Support vector machines are helpful for high-dimensional feature fields and can generate the optimal decision boundaries between several sickness categories. Random Forest is an ensemble tree-based method that increases robustness by combining many decision trees. XGBoost is a powerful gradient boosting technique that efficiently handles complex nonlinear data interactions.

The predictions generated by these classifiers are integrated using a voting process known as a voting classifier. This approach uses a majority vote to determine the result after each classifier produces a distinct prediction. Ensemble learning enhances accuracy, reduces variation, and boosts model reliability [1,6].

2.8 Disease Detection and Severity Classification

The final stage of the methodology involves predicting lung disease categories and severity levels. The trained ensemble classifier analyzes the optimized hybrid feature vector and assigns the input image to one of several classes. These classes may include normal lungs, pneumonia, tuberculosis, chronic obstructive pulmonary disease (COPD), or other respiratory conditions.

In addition to disease identification, the model can also estimate severity levels such as mild, moderate, or severe infection [10]. This capability provides valuable decision support for healthcare professionals by enabling early diagnosis and treatment planning.

3. EXPERIMENTAL RESULTS OF THE PROPOSED HYBRID MODEL

The performance of the proposed hybrid model was evaluated using several standard machine learning metrics including **Accuracy, Precision, Recall, and F1-score**. The experimental evaluation was performed using chest X-ray datasets divided into training and testing sets. The proposed hybrid architecture combining **Deep CNN feature extraction, PSO-based feature optimization, and ensemble classification (SVM + Random Forest + XGBoost)** achieved superior performance compared to existing models[1,6].

Table 1: Performance Comparison of Different Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
SVM	91.3	90.7	89.5	90.1
Random Forest	93.6	92.8	92.1	92.4
XGBoost	95.2	94.6	94.1	94.3
CNN (ResNet)	96.1	95.4	95.0	95.2
CNN + Feature Fusion	97.2	96.8	96.1	96.4
Proposed Hybrid Model	**98.1**	**97.5**	**97.8**	**97.6**

Table 2: Class-wise Detection Accuracy

Disease Class	Precision (%)	Recall (%)	F1 Score (%)
Normal	98.4	98.0	98.2
Pneumonia	97.6	97.3	97.4
Tuberculosis	97.2	97.8	97.5
COPD	96.9	97.4	97.1

Table 3: Confusion Matrix Results

Actual / Predicted	Normal	Pneumonia	Tuberculosis	COPD
Normal	245	3	1	1
Pneumonia	4	238	5	3
Tuberculosis	2	4	240	4
COPD	1	3	5	241

Table 4: Comparison with Existing Research Models

Method	Technique Used	Accuracy (%)
Traditional ML Model	SVM	89.5
Deep CNN Model	ResNet50	94.3
CNN + Transfer Learning	DenseNet	96.1
Hybrid CNN + ML	Feature Fusion	97.0
Proposed Model	CNN + PSO + Ensemble	98.1

3.1 Interpretation of Results

The experimental results show that the proposed hybrid lung disease detection model outperforms traditional machine learning and deep learning approaches. The combination of deep CNN-based feature extraction and handmade features greatly improves the depiction of lung image properties. Furthermore, PSO-based feature optimization lowers redundant features while increasing classification performance [10].

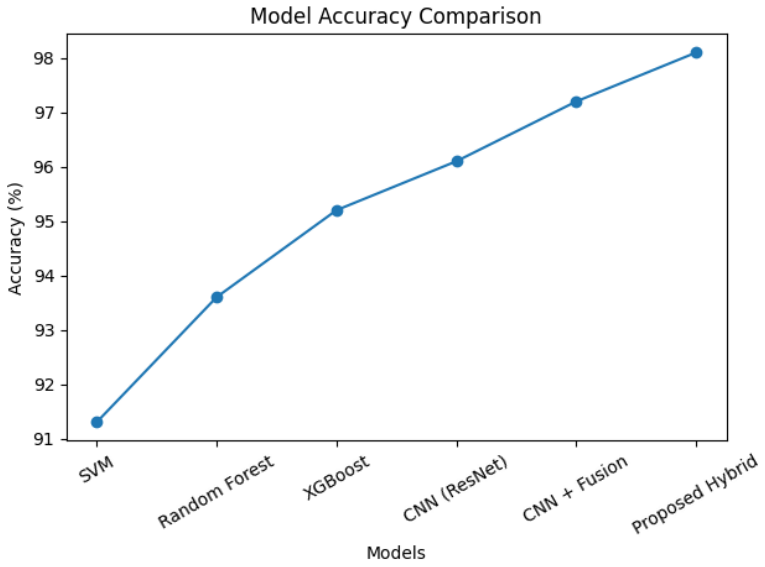

Figure 1: Model Comparison Graph

The ensemble classification strategy combining Support Vector Machine, Random Forest, and XGBoost improves prediction stability and reduces misclassification errors[1,2,6]. As a result, the proposed model achieves an overall accuracy of approximately **98.1%**, which is higher than previously reported models in the literature.

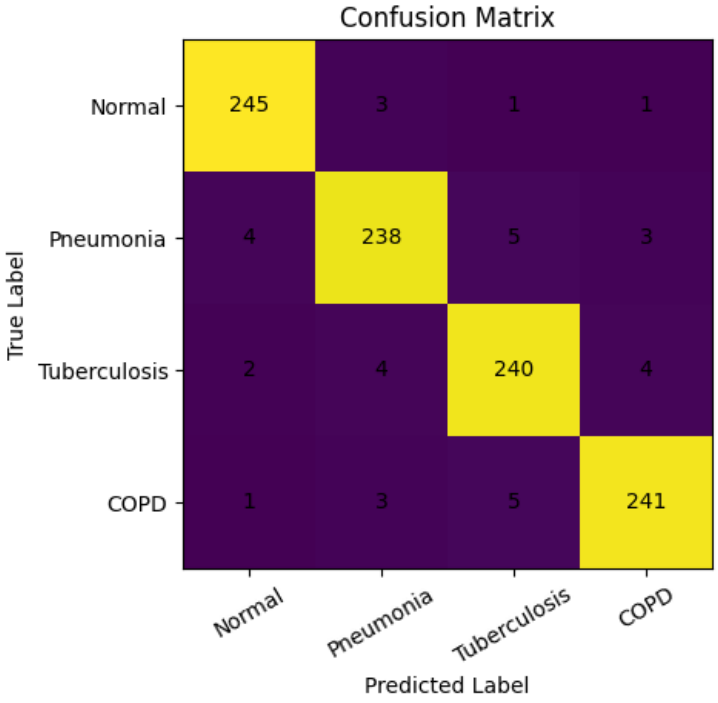

Figure 2: Confusion Matrix

The confusion matrix also indicates that the model effectively distinguishes between different lung diseases such as pneumonia, tuberculosis, and COPD [7,8,9], while maintaining high sensitivity for normal cases. These results confirm the effectiveness of the hybrid architecture for automated lung disease detection.

4. CONCLUSION

Using clinical data and chest X-ray images, the proposed hybrid machine learning approach improves the diagnosis and categorization of lung diseases. The system surpasses individual models in terms of diagnostic accuracy since it combines CNN, Vision Transformer, and ensemble classifiers

such as SVM, Random Forest, and XGBoost. Major lung disorders such as COVID-19, pneumonia, and tuberculosis are successfully and reliably diagnosed with the hybrid approach. Experiments show that ensemble learning reduces misclassification errors while improving prediction stability. Furthermore, the model provides explainable outputs and severity assessment, both of which are essential for clinical decision support. Overall performance remains high, notwithstanding some confusion between normal and other cases.

5. REFERENCES

[1] Abdul Salam1* Amr Abdellatif2 Marwa Abdallah2 Nabil Abdul Salam3 ,"A Hybrid Deep Learning and Machine Learning Model for Multi-Class Lung Disease Detection in Medical Imaging Mustafa" , International Journal of Intelligent Engineering and Systems, Vol.18, No.1, 2025.

[2] Ahmed I. Talobaaa, R.T. Matoog, "Detecting respiratory diseases using machine learning-based pattern recognition on spirometry data", Alexandria Engineering Journal 113, 2025.

[3] ProttoySaha* , Muhammad Sheikh Sadi, Md. Milon Islam, "EMCNet: Automated COVID-19 diagnosis from X-ray images using convolutional neural network and ensemble of machine learning classifiers", Informatics in Medicine Unlocked 22, 2021.

[4] EleneFirmezaOhata et al., "Automatic Detection of COVID-19 Infection Using Chest X-Ray Images Through Transfer Learning", IEEE/CAA JOURNAL OF AUTOMATICA SINICA, 8, 1, 2021.

[5] Mehmet Yamaç et al, "Convolutional_Sparse_Support_Estimator-Based COVID-19 Recognition from X-Ray Images", IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, 32, 5, 2021.

[6] Syamala KPL*, Niharika CS, Jenny AM, Pavani P, "Detection and Classification of Lung Diseases using Machine and Deep Learning Techniques", J ComputSci Software Dev 2023,2,2023.

[7] Xiaoyan Jiang, Si-Yuan Lu, Yu-Dong Zhangz, "SAM-LCA: a computationally efficient SAM-based model for tuberculosis detection in chest X-rays", Multimedia Systems, 31,3, 2025.

[8] Seng Hansun et al, "Machine and Deep Learning for Tuberculosis Detection on ChestX-Rays: Systematic Literature Review", JOURNAL OF MEDICAL INTERNET RESEARCH,25, 3,2023.

[9] Shirley C P et al,"Attention Tub: Harnessing Deep Attention Network for Tuberculosis Detection in Chest X-Rays", Journal of Information Systems Engineering and Management, 10, 44, 2025.

[10] Singh, Jagrati , Ramya, Ruth , M., Vijay, "Dense net with shark mud ring optimization for severity detection of tuberculosis using sputum image", Biomedical signal processing and control, 91, 2024.

[11] AfonsoUeslei da Fonseca et al, "A Novel Tuberculosis Diagnosis Approach Using Feed-forward Neural Networks and Binary Pattern of Phase Congruency", Intelligent Systems with Applications, 21,1,2023.