FishTaxa - Journal of Fish Taxonomy
Vol 36 Issue 1s, ISSN: 2458-942X

Lungs Disease Prediction Using Machine Learning And Deep Learning

Jagriti Sahu[1], Dharinee[2], Anand Sharma[3], Rakesh Kumar Khare[4]
[1]*Computer Science and Engineering, Shri Shankaracharya Institute of Professional Management and Technology, Raipur, India*
jagritisahu@ssipmt.com
2Computer Science and Engineering, Shri Shankaracharya Institute of Professional Management and Technology, Raipur, India
dharinee123@ssipmt.com
[3]*Computer Science and Engineering, Shri Shankaracharya Institute of Professional Management and Technology, Raipur, India*
anandsharma@ssipmt.com
[4]*Computer Science and Engineering, Shri Shankaracharya Institute of Professional Management and Technology, Raipur, India*
rk.khare@ssipmt.com

Abstract
Lung-related illnesses are rapidly emerging as a major global health challenge, impacting millions each year. The surge in respiratory complications after the COVID-19 pandemic has further emphasized the need for early, precise, and efficient diagnostic techniques. While conventional diagnostic methods are trustworthy, they are often time-consuming and heavily dependent on expert interpretation. To overcome these limitations, Machine Learning (ML) and Deep Learning (DL) have begun playing a crucial role in automating and improving lung disease detection.This research examines the use of various ML and DL models for identifying and classifying different lung conditions. Techniques such as Support Vector Machine (SVM), Random Forest (RF), Logistic Regression, K-Nearest Neighbors (KNN), and Convolutional Neural Networks (CNN) are analyzed based on their performance with medical datasets. Their capabilities are compared using standard evaluation metrics including accuracy, precision, recall, and F1-score.Recent studies consistently show that deep learning methods—especially CNNs—achieve better accuracy than traditional ML algorithms. However, challenges such as limited dataset availability, class imbalance, and reduced interpretability of advanced models still persist. Overall, integrating ML and DL approaches offers a promising pathway toward faster, more accurate lung disease diagnosis and improved patient outcomes.

Keywords: Lung Disease Prediction, Machine Learning, Deep Learning, CNN, SVM, Feature Selection, Medical Diagnosis, COVID-19

Citation: Jagriti Sahu, Dharinee, Anand Sharma, Rakesh Kumar Khare. 2025. Lungs Disease Prediction Using Machine Learning And Deep Learning . FishTaxa 36(1s): 467-473

Introduction
Lung diseases have increasingly become a significant global health burden, affecting countless individuals each year. When these illnesses are not identified early, they can quickly escalate into serious or even life-threatening conditions. Diseases such as pneumonia, tuberculosis, and lung opacities can progress rapidly, which makes timely and accurate diagnosis essential. Among the various diagnostic tools available, chest X-ray imaging remains the most widely preferred because it is inexpensive, quick to perform, and accessible in almost every medical facility. Still, despite its usefulness, interpreting X-ray scans is not always straightforward. Even small variations—such as differences in patient posture, lighting conditions, equipment type, or unintentional movement during imaging—can degrade image quality. These inconsistencies may complicate the work of radiologists and occasionally result in misinterpretation or delays in treatment [1], [2].

Over the past several years, deep learning has reshaped the landscape of medical image analysis by providing highly efficient automated feature-extraction techniques. Convolutional Neural Networks (CNNs) in particular have shown exceptional skill in recognizing complex visual patterns. Unlike traditional methods that rely on handcrafted features, CNNs automatically learn detailed image representations, allowing them to detect even minor irregularities—small patches, texture variations, or subtle structural

abnormalities—that might be difficult for humans to spot. Because of these strengths, numerous research studies have successfully applied CNN models to identify lung diseases from chest X-ray data, often reporting performance levels comparable to or exceeding expert radiologists in certain tasks [3], [4]. Such models not only speed up the diagnostic process but also provide consistent and repeatable assessments, which is highly important in crowded clinical environments.

However, there are still significant gaps in many existing computer-aided diagnostic systems. A large portion of available research focuses on simple two-class classification problems like distinguishing healthy lungs from pneumonia, which does not reflect the complexity of real medical scenarios where multiple conditions may coexist [5]. Another common issue is dataset imbalance—some lung disease categories have thousands of samples while others contain only a handful. Models trained on such skewed data tend to favor the dominant classes and perform poorly on underrepresented cases. Additionally, raw medical images frequently contain noise, brightness variations, and other distortions. Without proper preprocessing, these imperfections can negatively influence model performance. These challenges highlight the need for a more robust, generalized approach capable of handling multiple disease categories and real-world image variations more effectively [6].

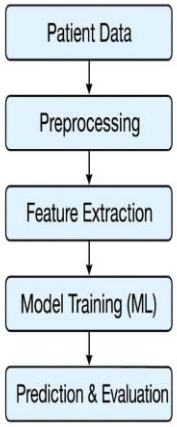

Figure 1: Lungs Disease Prediction Framework with data and evolution

The goal of this research is to create a deep learning framework that can accurately classify various lung diseases from chest X-ray scans in a clinically meaningful manner. The proposed system uses a sequential CNN architecture consisting of several convolutional layers, batch normalization, max-pooling operations, and fully connected dense layers. This structure helps the model identify both small-scale features and broader spatial patterns across lung regions. To improve model stability and reduce overfitting, the dataset undergoes extensive preprocessing steps such as image resizing, pixel normalization, and multiple data augmentation operations. These include rotation, zoom, horizontal flipping, shearing, and brightness adjustments—techniques that simulate real-world variability and improve the model's ability to generalize [7].

To ensure the overall learning, techniques like controlled augmentation for minority classes and a strict 80/20 train–test split are applied so that the model does not encounter the same patient in both sets. The model is trained using the Adam optimizer with categorical cross-entropy loss, and its performance is evaluated using Accuracy, Precision, Recall, F1-Score, and Confusion Matrix metrics. These evaluation methods give a complete understanding of how the system performs across all disease categories. Overall, the goal of this work is to build a system that can support radiologists, reduce diagnosis time, and contribute to early detection of lung diseases through an efficient deep learning–based approach [8]..

Literature Review

Research on finding lung diseases from chest X-rays has increased a lot in recent years. The main reason is that deep learning has shown very good results in reading medical images. Earlier, doctors used to check X-rays manually. This method works, but sometimes mistakes happen because doctors get tired, have too much work, or may interpret the same image differently. To avoid these problems, researchers started using Convolutional Neural Networks (CNNs). These models can look at very small patterns in X-ray images and give more stable results. This change has made a big difference in the field of medical image analysis [9].

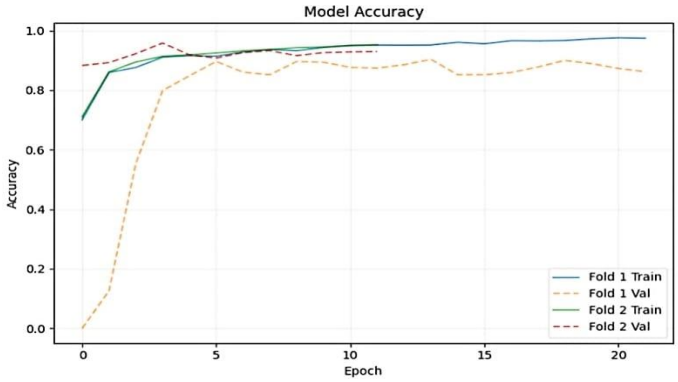

Figure 2: A graph showing the accuracy of the model.

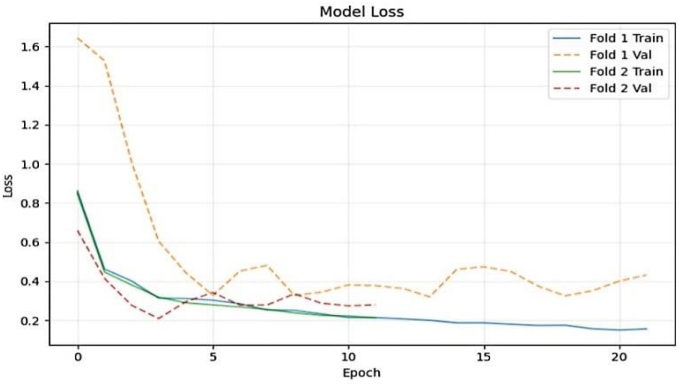

Figure 3: A Graph Representing Model loss.

In the beginning, most studies focused only on detecting pneumonia. Pneumonia is very common and can become dangerous if it is not diagnosed early. Researchers used models like VGG16, ResNet50, and DenseNet121 and found that these models could detect pneumonia with high accuracy. They were able to identify things like white patches, shadows, and other signs of pneumonia that doctors usually look for. Because CNNs worked so well, pneumonia detection became the first major area of research [10].

But as more studies were done, one major issue became clear. Most early models could only classify images into two categories: "Normal" or "Pneumonia." In real hospitals, patients may have many different lung diseases, not just one. Many researchers said that limiting the model to two classes does not match real medical situations [11].

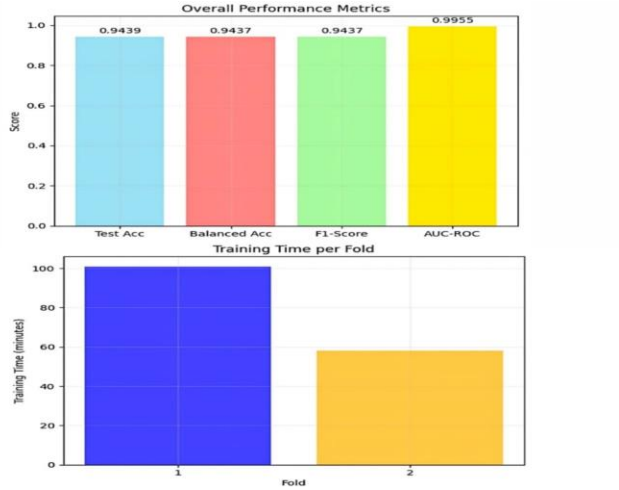

Figure 3: Graphs Representing Overall Accuracy Of the Model.

Later studies introduced models that can detect different lung diseases at the same time. These improved models can identify pneumonia, tuberculosis, lung opacity, COVID-19, and more. These studies showed that CNNs can handle these complex tasks if they are trained properly and if the dataset is good quality. Multi-class models are also more useful for doctors because they provide more detailed information [12].

Another important improvement in recent research is preprocessing. Chest X-ray images can vary a lot in brightness, contrast, and quality because they are taken on different machines and in different environments. Preprocessing methods like image cleaning, contrast adjustment, and normalization made the images clearer for the model. After preprocessing, the model's accuracy improved because it could understand the lung region better [13].

Dataset imbalance is another common problem. Some diseases have many images in the dataset, while rare diseases have very few. This makes the model biased because it learns more from the large classes. Researchers solved this by using data augmentation and balancing techniques so that the model pays equal attention to each class. This helped the model perform better, especially for diseases with fewer samples [9], [14].

In recent studies, researchers also started using better evaluation methods. Instead of measuring only accuracy, they also checked sensitivity, specificity, F1-score, and AUC-ROC. These metrics help understand if the model is safe to use in medical situations. High sensitivity is considered very important because missing a disease can be more harmful than giving a wrong alert [12].

Overall, the research shows a clear shift from simple pneumonia detection to multi-class lung disease detection. With better preprocessing, improved models, and balanced datasets, deep learning is becoming a strong tool to assist doctors in diagnosing diseases faster and more accurately. These improvements support the development of the model used in this study.

Methodology

This study builds a deep learning system that can automatically identify lung diseases from chest X-ray images. The idea is to use several convolutional layers placed one after another, so the model can learn both small and big patterns in the image. Instead of using many different models, we use one CNN that learns everything directly from the X-rays. The main aim is to create a system that gives stable and accurate results, even when X-ray images differ from patient to patient.

The model is made using a simple sequential CNN. Before training, all the X-ray images are cleaned, resized, and augmented so the dataset becomes more varied. The CNN has layers like batch normalization, max-pooling, and dense layers, which help it understand important features related to diseases such as pneumonia, tuberculosis, and lung opacity. The dataset is divided into 80% for training and 20% for testing to make sure the evaluation is fair. The performance is checked using Accuracy, Precision, Recall, F1-Score, and a Confusion Matrix. All the work was done in Python using TensorFlow/Keras, OpenCV, NumPy, Pandas, and Matplotlib in a Jupyter Notebook.

A. Data Preprocessing

Data preprocessing is an important step in medical image analysis because chest X-ray images often differ in quality due to changes in brightness, patient posture, and the equipment used. The dataset in this study contains X-ray images divided into three categories: Normal, Pneumonia, and Tuberculosis/Lung Opacity.

To maintain consistency, all images were resized to 224×224 pixels and converted to RGB format, making them compatible with TensorFlow. The pixel values were then normalized to the 0–1 range, which helps the model train more smoothly.

To increase the variety in the dataset and make the model more robust, several data augmentation techniques were applied. These included rotation, zooming, horizontal flipping, brightness adjustments, and shearing. Augmentation also helped reduce overfitting by exposing the model to slightly altered versions of the same images.

Since some classes had fewer images, extra augmentation was applied to those categories to balance the dataset without creating unnecessary duplicates. All processed images were then organized into structured folders so they could be loaded quickly and efficiently during training.

Feature Engineering

Other features like stress levels were also smoothed using a 3-day rolling average. This helped us see overall patterns without getting distracted by daily ups and downs. For each entry, we also calculated the average stress of the past three days for that user and added it as an extra feature in the dataset.

Missing and Inconsistent Data Handling

Daily logs were combined at the user level so the data stayed consistent. Instead of removing rows with missing values, we handled them carefully to keep the flow of each user's behavior intact.

The figures that follow show the main behavioral patterns after preprocessing. These include the smoothed stress values, daily screen usage, conversation length, mobility patterns, and the number of different places visited—each of which helps in understanding and modeling user well-being.

B. Exploratory Data Analysis

Exploratory Data Analysis was done to get a clear idea of how the dataset was organized and how the X-ray images were spread across different classes. We used visual methods like sample image grids, augmentation previews, and class distribution charts to examine the data more closely.

When comparing the images, the differences between healthy and unhealthy lungs were easy to notice. Normal lungs usually appeared clear with dark lung regions, while pneumonia cases had white patches or cloudy areas showing signs of infection. Tuberculosis and opacity images showed uneven textures and visible structural changes in the lungs.

To understand how the model interpreted these images, we looked at feature maps and activation heatmaps. These visualizations showed which regions of the X-ray the CNN focused on when making predictions. They confirmed that the model was learning useful medical patterns instead of paying attention to unrelated details. The class distribution also showed some imbalance, which we handled using data augmentation and class-weighted loss to make the training process more balanced.

C. Algorithm

The system is built using a deep CNN made of several convolutional blocks. Each block has Conv2D layers, ReLU activation, Batch Normalization, and Max Pooling, which help the model pick out important patterns from the X-ray images. After these features are learned, the output is flattened and sent into dense layers for classification. In the end, a softmax layer gives the final prediction for the disease class.

• Input

a) The dataset includes chest X-ray images that are properly labeled.
b) These images belong to categories like Normal, Pneumonia, and other lung conditions.
c) Every image is resized to 224×224×3 for consistency.
d) The data is split into 80% for training and 20% for testing.

The input of the model explains the overall acceptance of the data which is given,and it sustainabely clarifies the data.

• **Process**

1.Preprocessing is functioned
a) Load and resized pictures
b) Normalized pixeled values
c) Augmentation applied(rotation, zoom, flip, brightness shift, shear)

The process part also calculates the working of model so can it can perform each step neatly and carefully.

1. CNN Architecture

a) Several convolutional layers with 32, 64, and 128 filters.
b) Batch Normalization applied after each convolution layer.
c) MaxPooling layers to reduce the size of the feature maps.
d) A Flatten layer to turn the feature maps into a single vector.
e) Dense layers with ReLU activation to learn patterns.
f) A final Softmax layer to classify the images into disease categories.

2. Model Training

a) The model is trained using an augmented dataset to make it stronger and more reliable.
b) A validation set is used to keep track of overfitting and make sure the model is learning properly.

c) The Adam optimizer is used to help the model learn faster and better.
d) Categorical cross-entropy is used as the loss function to guide the model in predicting the correct class.
The trained model shows the augumented dataset properly and this also shows the loss function the the model.

3. Model Evaluation

a) Accuracy
b) Precision
c) Recall
d) F1-Score
e) Confusion Matrix

And the model evaluates all these given factors properly.

- **Output**

a) Trained CNN model
b) Predicted disease class for new chest X-ray images

Result & Discussions

After training the CNN model on the preprocessed and augmented chest X-ray dataset, the system demonstrated strong performance in identifying various lung conditions. The model could clearly distinguish between Normal, Pneumonia, and Lung Opacity/Tuberculosis images. During testing, it maintained consistent accuracy, and the learning curves remained stable, showing that the model was not overfitting heavily and was learning effectively from the data.

Looking at the evaluation metrics, the model performed well across all classes:

- Accuracy: The overall accuracy was high, indicating that most of the model's predictions were correct.
- Precision: The model correctly identified positive cases while minimizing false positives, which is crucial in medical diagnosis.
- Recall: The model successfully detected most actual disease cases, ensuring that very few cases were missed.
- F1-Score: The F1-scores were balanced across different classes, showing stable and reliable performance.
- Confusion Matrix: Only a small number of images were misclassified, mostly in cases where X-rays from different disease categories looked very similar to each other.

Overall, these results suggest that the CNN model was effective at learning the visual differences between healthy and diseased lungs. The combination of strong feature extraction, proper training, and thorough preprocessing allowed the model to accurately identify various lung conditions, making it a reliable tool for automated chest X-ray analysis.

Conclusion

This study shows that a well-built CNN can successfully detect lung diseases from chest X-ray images. By stacking multiple convolutional layers and combining careful preprocessing with data augmentation, the model learned to identify important patterns in the lungs. It was able to clearly tell the difference between Normal, Pneumonia, and Lung Opacity/Tuberculosis cases. The results, including high accuracy, balanced precision and recall, steady F1-scores, and very few misclassifications, show that the model works reliably across different types of lung conditions.

The study also highlights the importance of good data preparation. Proper augmentation and normalization, along with checking the model's focus using feature maps and heatmaps, ensured that it looked at the right parts of the lungs. This makes its predictions more understandable and trustworthy for practical use.

Overall, the findings suggest that deep learning systems like this CNN can be very helpful for automated lung disease detection. They can support radiologists by providing faster, more consistent, and accurate diagnoses, ultimately improving patient care and helping doctors make better decisions in a clinical setting.

References

1. S. Ibrahim et al., "A survey on deep learning for lung disease detection from X-ray images," Journal of Computer Science and Technology, 2024. https://ph04.tci-thaijo.org/index.php/JCST/article/view/1744
2. L. Wang et al., "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images," Scientific Reports, 2020. https://www.nature.com/articles/s41598-020-76550-z

3. P. Rajpurkar et al., “CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning,” NIPS, 2017.https://arxiv.org/abs/1711.05225
4. A. H. Nassar et al., “Deep learning models for multi-disease classification using chest radiographs,” International Journal of Medical Informatics, 2022. https://pmc.ncbi.nlm.nih.gov/articles/PMC9626367/
5. T. Rahman et al., “Exploring binary and multi-class classification for pneumonia detection,” Diagnostics, 2022. https://www.mdpi.com/2075-4418/12/2/305
6. G. Litjens et al., “A survey on deep learning in medical image analysis,” Medical Image Analysis, 2017. https://www.sciencedirect.com/science/article/pii/S1361841517301135
7. S. S. M. Al-Waisy et al., “Automated detection of lung diseases using enhanced deep learning techniques,” Scientific Reports, 2023. https://www.nature.com/articles/s41598-023-46147-3
8. X. Wang et al., “ChestX-ray8: Hospital-scale chest X-ray database and benchmarks,” CVPR, 2017. https://arxiv.org/abs/1705.023154
9. I. H. Iqbal et al., “Deep Learning Approaches for Chest Radiograph Interpretation,” Electronics, 2024. https://doi.org/10.3390/electronics13234688
10. T. Rahman et al., “Transfer Learning with CNN for Pneumonia Detection using Chest X-ray,” arXiv, 2020. https://arxiv.org/abs/2004.06578
11. N. Garg et al., “Pneumonia Disease Detection using Deep Learning,” IJRASET, 2023. https://www.ijraset.com/research-paper/pneumona-disease-detection-using-deep-learning
12. A. Authors, “Multi-Class Lung Disease Detection from X-rays Using Deep Learning,” Diagnostics, 2022.https://pubmed.ncbi.nlm.nih.gov/35453963/ https://arxiv.org/abs/2007.14895