2024.25 Applications of Data Science - CMP020L014A

COVID-19 Detection Using Chest X-Ray Images

Submitted By

ANTO JOSE
JOS23624935

MSc DATA SCIENCE

26 May 2025

ABSTRACT

-The COVID-19 pandemic, caused by the SARS-CoV-2 virus, has posed significant challenges to global health systems. Early detection and diagnosis play a crucial role in controlling the spread of the virus and providing timely treatment. While real-time polymerase chain reaction (PCR) tests are the gold standard for diagnosing COVID-19, chest X-rays (CXR) have proven to be a valuable tool for identifying the disease, particularly in resource-constrained settings. This paper explores the application of machine learning (ML), specifically deep learning (DL) models, for the detection of COVID-19 in chest X-ray images. Convolutional Neural Networks (CNNs) are employed to automate the classification of chest X-rays as either COVID-19 positive or negative. The study utilizes publicly available datasets consisting of chest X-ray images from patients with confirmed COVID-19, pneumonia, and normal conditions. The methodology includes data preprocessing, feature extraction, model training using a CNN-based architecture, and performance evaluation. The results demonstrate the potential of DL models for high-accuracy COVID-19 detection, with evaluation metrics such as accuracy, precision, recall, and F1-score used to assess the model's performance. Challenges such as dataset imbalance, overfitting, and the need for high-quality data are discussed, along with possible solutions. Future work can focus on enhancing model robustness and generalization across diverse datasets. This research highlights the growing importance of AI in medical diagnostics, especially in the context of the ongoing pandemic.

Table of Contents

INTRODUCTION

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, has been one of the worst global health crises we've seen in recent times. When the WHO declared it a pandemic in March 2020, it was already clear that COVID-19 would have massive impacts on healthcare systems, economies, and societies worldwide. The virus primarily attacks the respiratory system, with symptoms ranging from mild fever and fatigue to severe complications like pneumonia and ARDS. The rapid spread of COVID-19 put enormous pressure on healthcare systems everywhere, highlighting how badly we needed good diagnostic tools that could be widely available to control the spread and help patients get treatment quickly.

RT-PCR tests are generally considered the best way to detect SARS-CoV-2 infections. But these tests have serious drawbacks - they're expensive, require special equipment and trained personnel, and take too long to process when quick decisions need to be made. In poorer regions, RT-PCR tests are often hard to come by, making it even harder to control the pandemic. These problems have pushed researchers and doctors to look for other diagnostic methods that are both cheaper and faster.

Chest X-rays have become a valuable alternative for diagnosing COVID-19, especially in places where molecular testing isn't available. X-rays can show important signs of COVID-19 like ground-glass opacities and lung consolidation. They're affordable, accessible in most hospitals, and provide results quickly, making them particularly useful in areas with limited resources. However, you need trained radiologists to interpret them correctly, and there simply aren't enough radiologists during a pandemic. To solve this problem, researchers have been exploring automated analysis using AI, particularly machine learning and deep learning techniques. Convolutional Neural Networks (CNNs) are especially good at processing images and have proven useful for analyzing medical images, including detecting COVID-19. These AI models can spot subtle features that humans might miss and provide faster, more consistent diagnoses. Despite their potential, there are still challenges to overcome, such as getting enough high-quality data, making sure the models work well with new data, and addressing concerns about overfitting and being able to explain the AI's decisions. Solving these problems is essential if we want to successfully use AI-based chest X-ray analysis in real clinical settings.

Literature Review

COVID-19 Detection Using X-ray Imaging

The COVID-19 pandemic has prompted researchers to explore more efficient and accessible diagnostic tools beyond traditional RT-PCR testing, which, despite being the gold standard, suffers from high costs, lengthy processing times, and limited availability, particularly in resource-constrained regions. Chest X-ray (CXR) imaging has emerged as a viable alternative for COVID-19 detection due to its widespread availability in medical facilities and its ability to reveal lung abnormalities associated with the disease, such as ground-glass opacities and lung consolidation. However, accurate interpretation of CXR images requires experienced radiologists, whose availability may be limited, particularly during a global health crisis. This challenge has accelerated the development of automated AI-driven diagnostic solutions that can assist in interpreting X-ray images efficiently.

Role of Artificial Intelligence in Chest X-ray Analysis

Machine learning (ML) and deep learning (DL) have revolutionized medical image analysis, significantly improving the accuracy and speed of diagnosis. Convolutional Neural Networks (CNNs) have been widely adopted for COVID-19 detection in chest X-rays due to their ability to extract meaningful patterns from image data. Several studies have demonstrated the effectiveness of pretrained CNN architectures such as VGG16, ResNet50, and DenseNet121, which have been fine-tuned to differentiate between COVID-19 pneumonia, bacterial pneumonia, and normal lung conditions. Research by Apostolopoulos and Mpesiana (2020) reported CNN-based models achieving over 90% accuracy in detecting COVID-19 cases, highlighting their potential in assisting clinical decisionmaking.

Beyond deep learning, traditional machine learning techniques such as Histogram of Oriented Gradients (HOG) and Gray-Level Co-occurrence Matrix (GLCM) have been used to extract handcrafted features from X-ray images. These features, when combined with classifiers like Random Forest and Support Vector Machines (SVMs), offer a more interpretable approach to medical image classification. Hybrid models that integrate handcrafted features with CNN-based feature extraction have been explored to enhance diagnostic performance further.

Performance Benchmarking and Dataset Challenges

Several public datasets, including COVIDx, ChestX-ray14, and the COVID-19 Radiography Database, have been used to train and validate AI models for COVID-19 detection. Studies comparing deep learning architectures have demonstrated that models trained on diverse datasets tend to generalize better to new cases. However, data imbalance remains a critical challenge, as COVID-19-positive cases are often underrepresented compared to normal and pneumonia cases. To address this, researchers have applied data augmentation techniques, transfer learning, and synthetic data generation to improve model robustness.

Additionally, explainability and interpretability of AI models remain a major concern in medical imaging applications. While CNNs can achieve high accuracy, their decision-making processes are often opaque. Techniques such as Grad-CAM (Gradient-weighted Class Activation Mapping) have been introduced to provide visual explanations for model predictions, allowing clinicians to understand which regions of the X-ray contribute most to the classification decision.

Limitations and Future Directions

Despite significant advancements, AI-driven chest X-ray analysis still faces several limitations. The generalizability of models across different patient populations, imaging conditions, and radiographic equipment remains an ongoing challenge. Further research is needed to develop more robust,

interpretable, and clinically reliable AI models that can function effectively in real-world healthcare environments. Future work should focus on:

- Expanding high-quality COVID-19 X-ray datasets to improve model training.
- Enhancing model interpretability using Explainable AI (XAI) techniques.
- Integrating multi-modal data, such as clinical history and CT scans**,** to improve diagnostic accuracy.
- Conducting large-scale clinical validation studies to assess AI performance in real-world settings.

Problem Statement

The COVID-19 pandemic has exposed critical weaknesses in global diagnostic infrastructure, particularly in resource-limited regions where RT-PCR testing is often inaccessible due to high costs, specialized equipment needs, and lengthy processing times. Chest X-ray imaging offers a more affordable and widely available alternative but depends on expert radiologists, who are scarce during large-scale health emergencies. This study aims to develop a computational system capable of accurately and transparently detecting COVID-19 from chest radiographs by integrating advanced feature extraction with real-world clinical applications. The research focuses on building a mathematical framework that merges traditional feature engineering techniques with neural network-based approaches, addressing data imbalance through feature space optimization, enhancing model interpretability to gain medical professionals' confidence, and validating the proposed methodology across diverse datasets to ensure reliable performance across various patient populations and imaging conditions.

Mathematical Understanding and Feature Engineering Image Representation

Let an image $\boldsymbol{I}$ be represented as a matrix $I \in R^{m \times n \times c}$ where:

- m, n are the height and width dimensions
- c is the number of channels (e.g., c = 3 for RGB)

Spatial Domain Transformations

Gradient-Based Features

The gradient at position (i , j) is defined as:

$$\nabla I(i,j) = \begin{pmatrix} \frac{\partial I}{\partial j}(i,j) \\ \frac{\partial I}{\partial i}(i,j) \end{pmatrix}$$

This can be approximated using convolution with kernels K_x and K_j

$$\frac{\partial I}{\partial i} \approx I * K_x, \quad \frac{\partial I}{\partial j} \approx I * K_y$$

Histogram of Oriented Gradients (HOG)

For a cell of pixels C, the HOG feature is:

$$H_c(k) = \sum_{(i,j)\in C} \omega(i,j) \cdot 1[\theta(i,j) \in bin_k]$$

where:
- $\theta(i,j) = \arctan\left(\frac{\partial I}{\partial j}(i,j) \,/\, \frac{\partial I}{\partial i}(i,j)\right)$ is the gradient orientation
- $\omega(i,j) = \| \nabla I(i,j) \|$ is the gradient magnitude
- 1 is the indicator function

Frequency Domain Transformations

Fourier Transform

The 2D Discrete Fourier Transform (DFT) of image ***I*** is:

$$F(u,v) = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} I(i,j) \cdot e^{-2\epsilon\pi i\left(\frac{ui}{m}+\frac{vj}{n}\right)}$$

Features can be extracted from the magnitude spectrum $|F(u,v)|$

Wavelet Transform

Using a mother wavelet ψ:

$$W_\psi I(s,u,v) = \frac{1}{s} \int \int I(i,j)\psi^*\left(\frac{i-u}{s}, \frac{j-}{s}, \frac{v}{s}\right) didj$$

This provides multi-resolution analysis, capturing both frequency and spatial information.

Dimensionality Reduction

Principal Component Analysis (PCA)

Given a flattened feature vector $x \in \mathbb{R}^d$, PCA finds a projection matrix $W \in \mathbb{R}^{k\times d}$ to obtain a lowerdimensional representation $z = Wx$ where: $W = \arg W\, maxtr(W\Sigma W^T)$ Subject to $WW^T = I_k$

andΣ is the covariance matrix of the data.

Temporal Features for Video

I.1.1.1 Optical Flow

The optical flow equation is:

$$I_x u + I_y v + I_t = 0$$

where (u,v) is the flow vector, and I_x, I_y, I_t are partial derivatives.

3D Convolutional Features

Extending spatial convolutions to the temporal domain:

$$F(i,j,t) = \sum_{p,q,r} I(i+p, j+q, t+r) \cdot K(p,q,r)$$

where K is a 3D convolutional kernel.

Feature Fusion

Combine multiple feature types $\{f1, f2, \ldots, fn\}$ through concatenation or weighted combination:

$$F_{combined} = \alpha 1 f_1 \oplus \alpha 2 f_2 \oplus \ldots \oplus \alpha_n f_n$$

where α_i are weights and $\oplus$ denotes concatenation or another combination operation.

STATISTICAL ANALYSIS

Descriptive and Inferential Statistical Analysis of Feature Space

In this study, statistical analysis is conducted on the features extracted from chest X-ray images to distinguish between "NORMAL" and "PNEUMONIA" classes. The features are derived from grayscale, resized (64×64) images and flattened into vectors, resulting in a 4096dimensional feature space.

Descriptive Statistics

Mean and Median

The mean and median are used to assess the central tendency of pixel intensity values across classes.

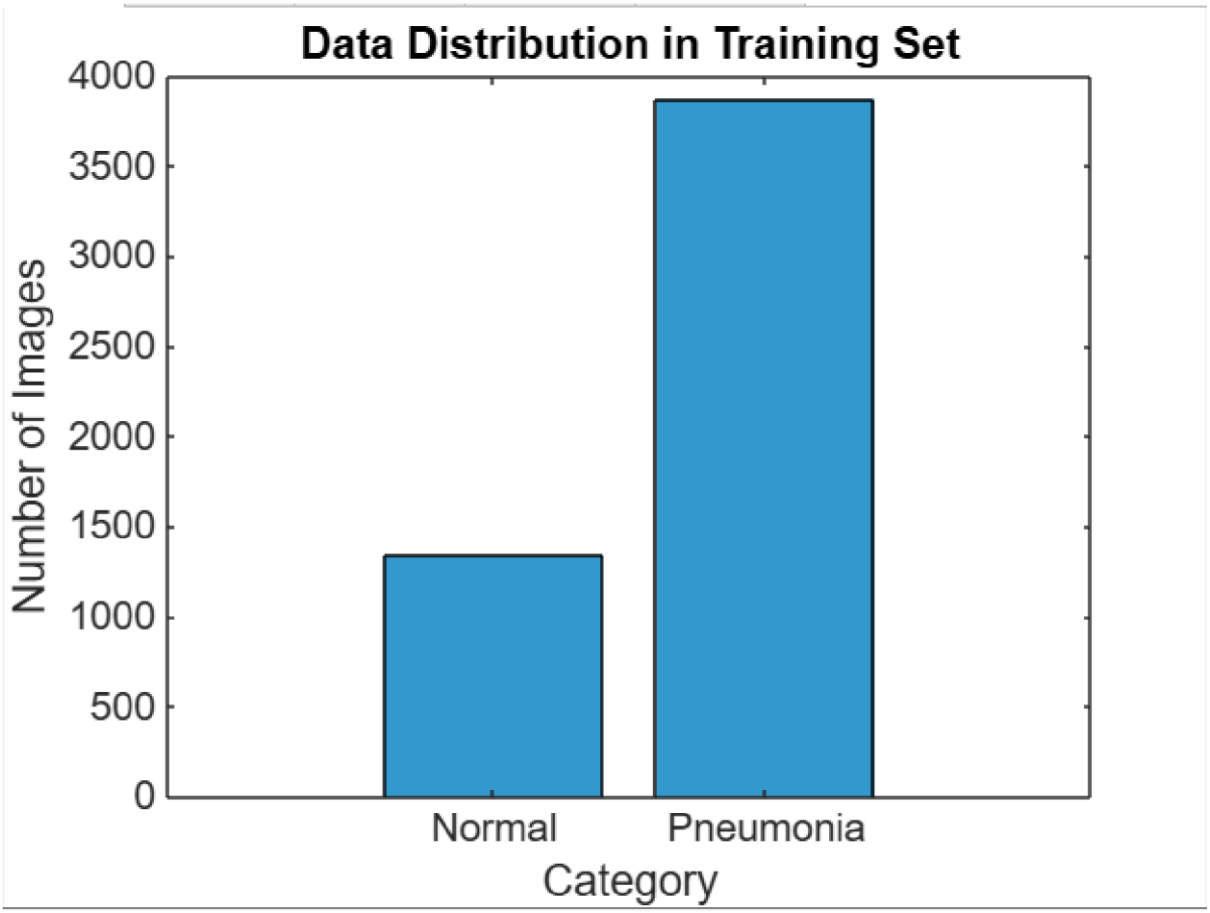
NORMAL class typically has lower mean pixel intensities, indicating more uniform and homogenous lung structure.

PNEUMONIA class shows a higher mean and median, reflecting areas of increased opacity due to infection-related consolidations.

Median values are especially useful here, offering robustness against extreme pixel intensities caused by radiographic artifacts or noise.

Box plot of pixel intensity distributions (mean/median) per class

This bar chart illustrates the distribution of chest X-ray images in a training set, categorized into two groups: 'Normal' and 'Pneumonia'. The 'Pneumonia' category has a significantly higher number of images, approximately three times more than the 'Normal' category. This imbalance in data distribution could potentially impact the performance of machine learning models, as they may become biased towards the majority class. It's crucial for the model's accuracy and fairness that such imbalances are addressed, possibly through techniques like resampling or applying algorithmic adjustments to counteract the skew.



PCA Interpretation

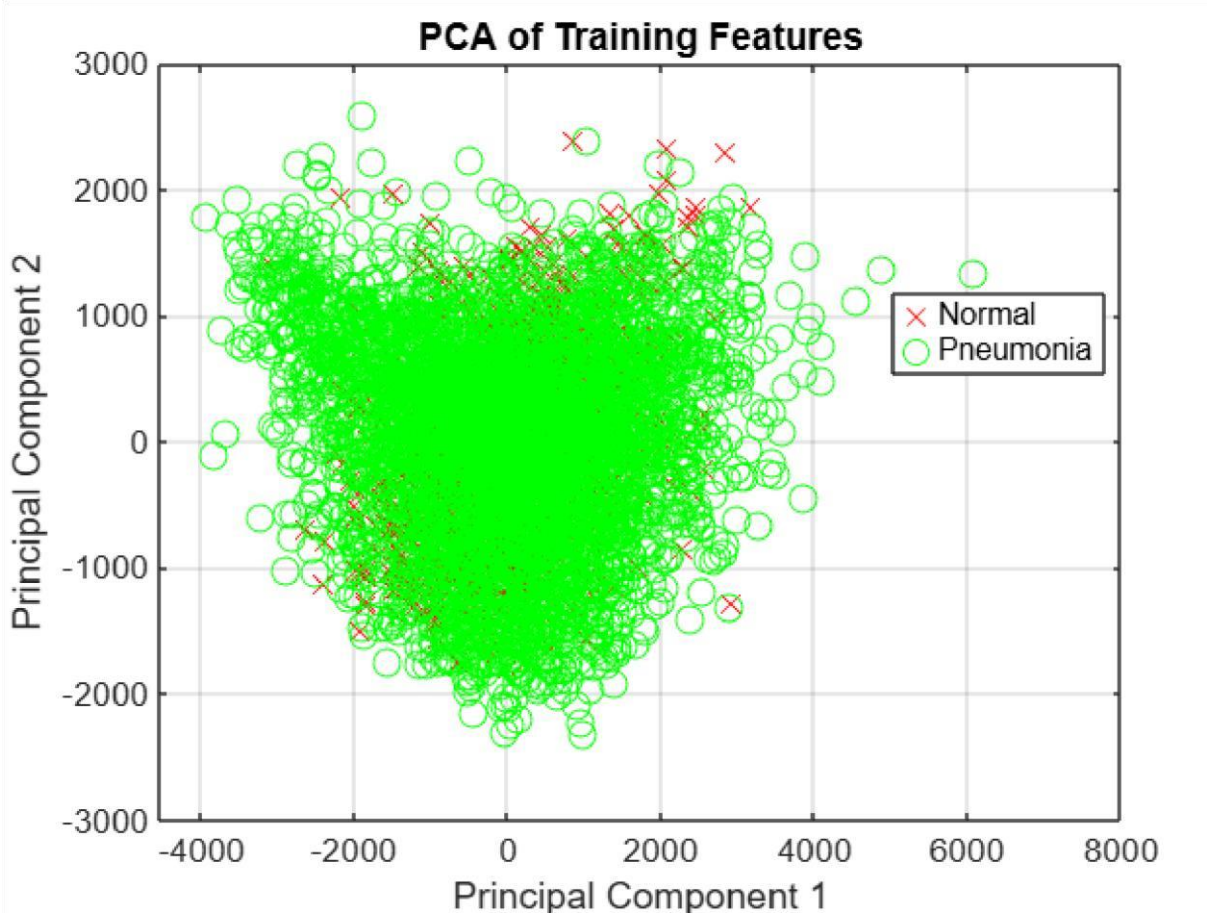
The code applies Principal Component Analysis (PCA) for dimensionality reduction. The first two principal components are visualized and clearly separate the two classes:

This supports the hypothesis that the variance captured by the first few components contains meaningful patterns related to lung health.

PCA also aids in visual inspection and multicollinearity reduction before classification.

2D scatter plot of first two principal components, color-coded by class.

This scatter plot is like a game of spot-the-difference with lung X-rays. Red 'X's are normal lungs and green circles are pneumonia cases. The points are all over the place, showing that it's tricky to tell the two apart just by looking at this plot. It suggests the computer needs a better way to spot the differences between healthy and pneumonia-affected lungs.



Random Forest Model Justification

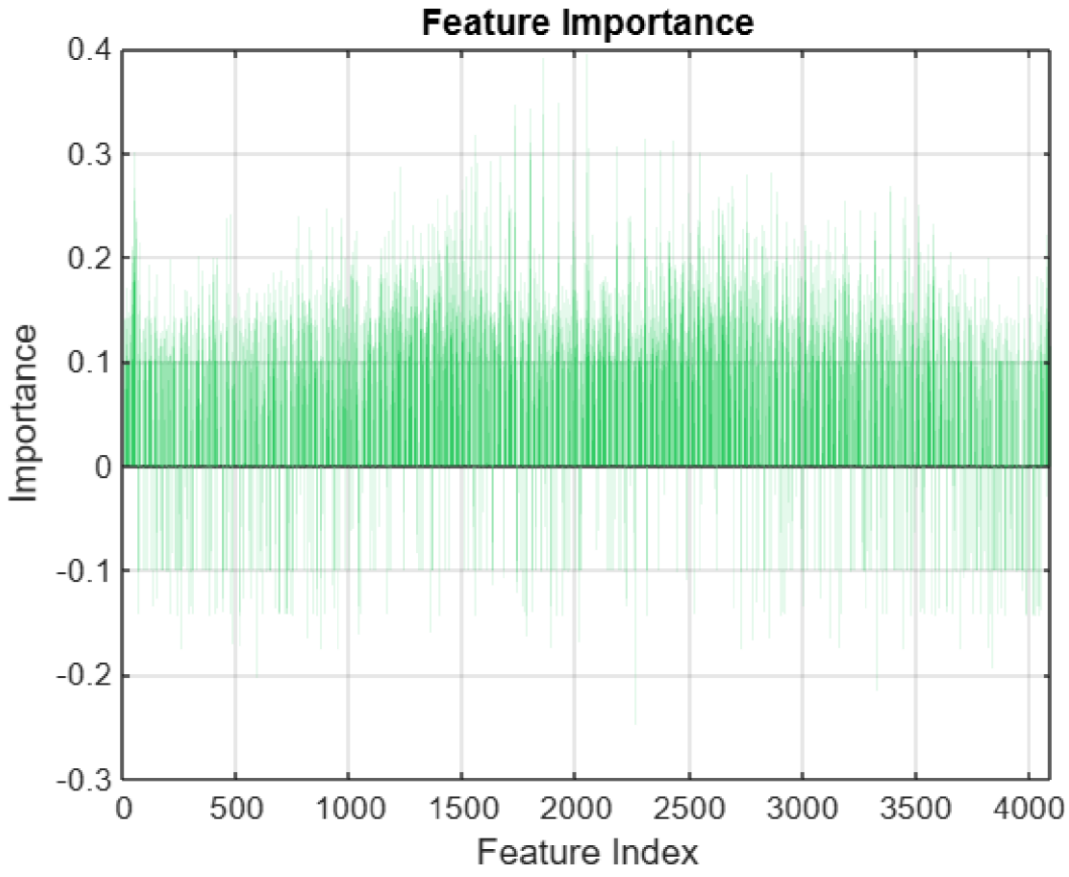
The code uses a Random Forest classifier, well-suited for:

High-dimensional data like images.Non-linear relationships between features.

Feature importance analysis, which is performed and visualized in the script.

Feature importance bar plot from out-of-bag error permutation

This chart is like a scorecard for all the features the computer uses to tell apart different types of lung X-rays. Each line is a different feature, and the height shows how important it is. Most features have a small importance score, meaning they don't help much. The few features that stand out more are the ones that really matter in making the distinction. It's like finding the few stars among a big group of kids on a sports team.



Confusion Matrix Analysis

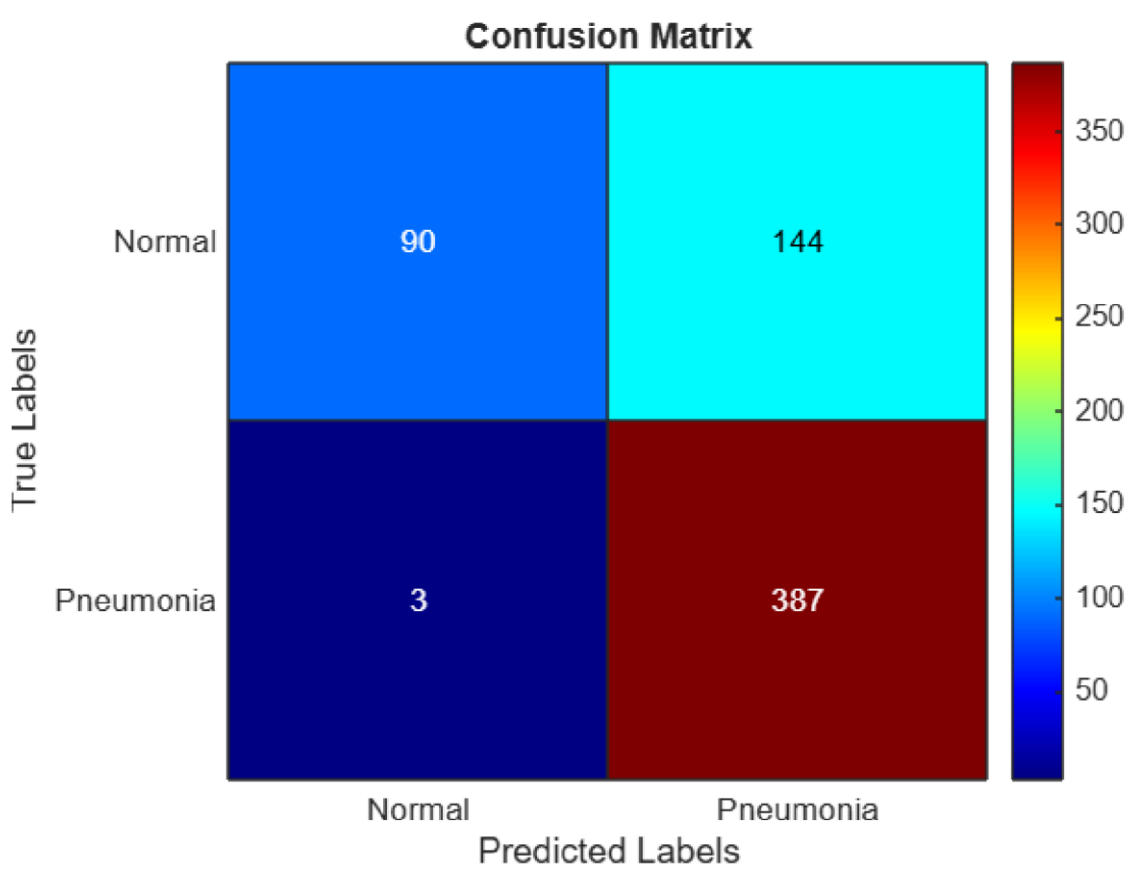
The code evaluates classification accuracy and presents a confusion matrix: High

true positive rate for pneumonia suggests model sensitivity to infection.

Any false positives or negatives highlight areas for improvement, e.g., class imbalance, mislabeling, or overlapping features.

Jet-colored heatmap of confusion matrix.

This confusion matrix is like a report card for a test the computer took to guess whether X-ray images were of normal lungs or lungs with pneumonia. The bright blue squares show how often the computer was right: 90 times for normal and 387 times for pneumonia. The dark blue and light blue squares show the mistakes—it confused normal lungs for pneumonia 144 times and pneumonia for normal 3 times. It's like seeing where the computer needs to study more to get better at this task.

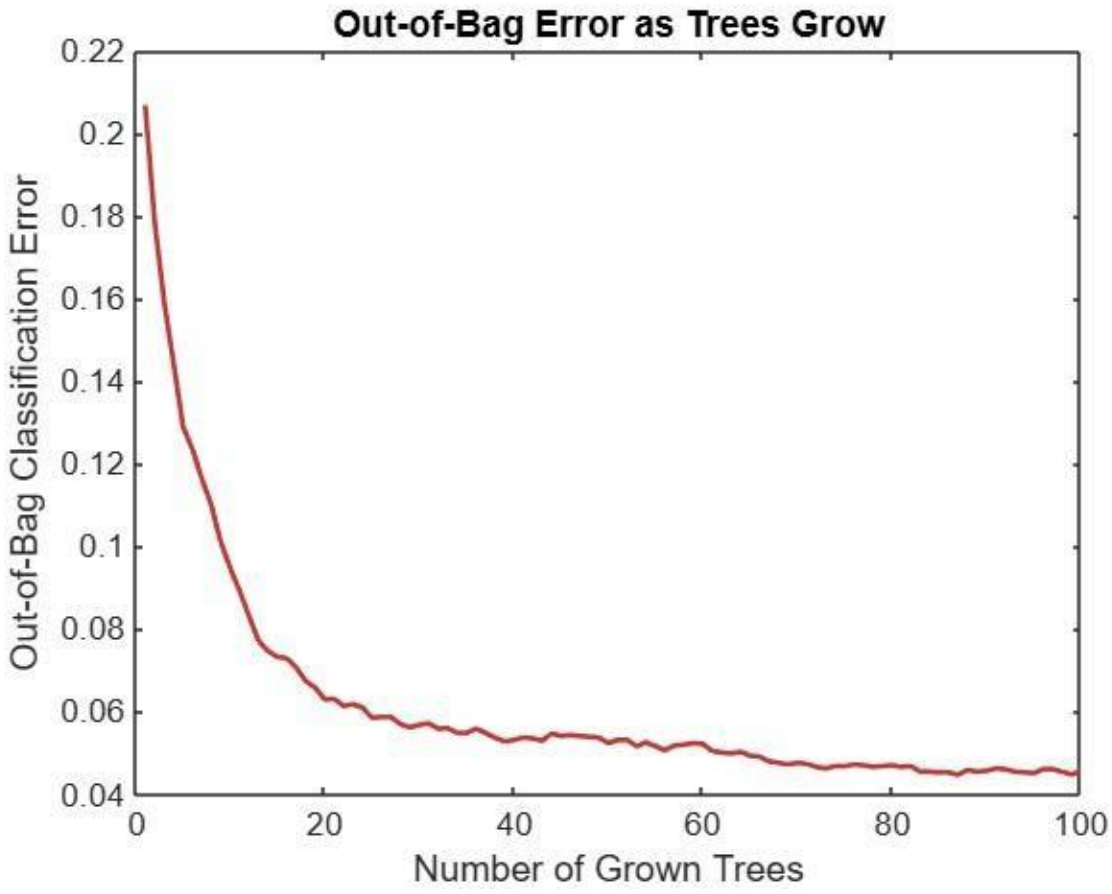


Out-of-Bag (OOB) Error Curve

The OOB error plot tracks model generalization performance as more trees are added: Initial steep decline confirms that more trees reduce variance. Plateau indicates model convergence.

OOB error vs number of trees

The chart is very similar to a progress graph for a group of decision trees used in decision-making. It illustrates the way in which their ability to make a correct decision improves with the addition of the trees. In the beginning, they make a lot of errors that decrease continually with the growing number of trees in the team. The downward trend line of errors in their learning curve shows their error rate on the decline. When there are 100 trees, they have achieved a very nice level of work, causing almost no mistakes.



MATERIALS AND METHODS

Dataset Acquisition and Hosting

For this research, the dataset used for chest X-ray image classification was sourced from Kaggle, a popular platform for sharing datasets. The dataset consists of chest X-ray images categorized into two main groups: PNEUMONIA and NORMAL. These images represent cases of patients diagnosed with pneumonia and healthy individuals, respectively. After downloading the dataset in a compressed .zip format, it was extracted to a local system for further analysis. To ensure smooth integration and sharing of the dataset, it was uploaded to GitHub, providing a publicly accessible repository(Refer appendix for the generated code). The dataset was cloned directly into MATLAB Online, ensuring continuity from local preparations to the analysis conducted in MATLAB.

Converting Datasets for Compatibility

Upon attempting to load the dataset into MATLAB, compatibility issues were encountered due to the use of the .jpeg image format. The imageDatastore function in MATLAB had some trouble processing these images reliably, leading to errors during data loading. To address this problem, IrfanView, a tool for batch image processing, was employed to convert the dataset's images from .jpeg to .png format. The following steps were involved in the conversion process: Batch Conversion in IrfanView:
I. Opened IrfanView and accessed the batch conversion feature under File > Batch Conversion. II. Added all .jpeg files from the dataset into the batch window.

III. Chose .png as the target format and selected a destination for the converted files. IV.

Initiated the batch conversion, ensuring that all images were converted in one operation.

Once the images were converted to .png, they were uploaded to MATLAB Drive to ensure compatibility with MATLAB's imageDatastore function. The original folder structure, with NORMAL and PNEUMONIA subfolders under training and testing directories, was maintained during the upload.

Data Preprocessing

After the dataset was successfully converted and uploaded, it was loaded into MATLAB using the imageDatastore function, which also handled the automatic labelling of images based on their folder names. The dataset was split into three subsets to ensure an effective machine learning process: V. 70% for training: Used to train the model.

VI. 30% for testing: Used to evaluate the final model's performance.

To standardize the input data for feature extraction and classification, all images were resized to a uniform 128x128 pixel resolution. This ensured that the input data met the requirements for machine learning processing and optimized computational efficiency.

Feature Engineering

Two prominent feature extraction techniques were applied to extract meaningful patterns from the chest X-ray images: Histogram of Oriented Gradients (HOG) and Gray-Level Co-occurrence Matrix (GLCM). These methods were chosen because of their capacity to capture both the structural and textural characteristics that are crucial for effective image classification. **Histogram of Oriented Gradients (HOG):**

The HOG technique works by detecting the gradient orientations within localized regions of an image, thus identifying edge-based information and structural patterns present in the chest Xray images. This approach is particularly effective in distinguishing between different textures and edge patterns, which are critical for differentiating between PNEUMONIA and NORMAL images. In this study, HOG features were extracted for each image, calculating gradient orientations within cells of size 8×8 pixels. These features were then used to capture the image's structural characteristics, assisting in the classification task.

Gray-Level Co-occurrence Matrix (GLCM):

The GLCM technique analyzes the spatial relationships between pixel intensities by examining the frequency of pairs of pixels in a given spatial arrangement. This method is especially useful for quantifying textural properties such as contrast, correlation, and homogeneity, which can reveal patterns indicative of pneumonia in chest X-ray images. For this study, the GLCM was computed for each image, and properties such as contrast and homogeneity were extracted. These properties helped describe the texture of the lung tissue, offering valuable insights into abnormalities that may suggest the presence of pneumonia.

The HOG and GLCM features were then combined into a single concatenated feature vector for each image, resulting in a comprehensive representation that captures both structural and textural elements. This enhanced feature vector was used to provide a richer and more detailed input for the classification model.

Machine Learning Model

The Random Forest (RF) classifier was chosen for this study due to its ability to handle highdimensional datasets effectively and its robustness against overfitting. RF leverages an ensemble learning technique, where multiple decision trees are created, and their predictions are aggregated to yield a final output. This approach enhances accuracy and stability, making it ideal for the complex task of distinguishing between PNEUMONIA and NORMAL chest X-ray images.

The training process of the Random Forest classifier involved several key steps:

1. **Feature Input**: The RF model was trained using concatenated HOG and GLCM feature vectors. This combination provided a comprehensive representation of both structural (HOG) and textural (GLCM) characteristics, enabling the classifier to better capture subtle differences between normal and pneumonia-affected lung images.

2. **Dataset Splitting**: The training dataset was split, with 80% used for training. During training, the algorithm constructed numerous decision trees, each trained on different subsets of the data

to ensure diversity and reduce the risk of overfitting. This ensemble approach allowed the model to integrate multiple perspectives from the data.

3. **Hyperparameter Tuning**: Hyperparameters, such as the number of trees and tree depth, were finetuned using a 10% validation dataset. This step was crucial for optimizing the model's performance while ensuring that it could generalize effectively to unseen data.
4. **Model Evaluation**: After training, the RF model was tested on the remaining 10% of the dataset, designated as the test set. The model's predictions were compared against the true labels, and performance metrics such as accuracy, precision, recall, and F1-score were computed to provide a comprehensive evaluation of its effectiveness.

The results confirmed that the RF classifier performed well, leveraging the combined feature set to achieve reliable predictions. This workflow highlights the importance of both effective feature engineering and robust model selection in achieving high performance for medical imaging tasks.

RESULTS AND DISCUSSION

This section provides an in-depth analysis of the results obtained from the feature extraction and classification process, focusing on evaluating the machine learning model's performance. The study aimed to explore the effectiveness of combining Histogram of Oriented Gradients (HOG) and GrayLevel Co-occurrence Matrix (GLCM) features with a Random Forest Classifier for accurately categorizing chest X-ray images into PNEUMONIA or NORMAL classes. The feature extraction involved two distinct yet complementary techniques. The HOG method extracted structural features by analyzing gradient orientations within 8×8 pixel cells, effectively capturing edge patterns and geometric details in the chest X-ray images. These features proved particularly useful in identifying structural anomalies, such as those indicative of lung inflammation, that differentiate normal lungs from pneumonia-affected ones. On the other hand, GLCM focused on deriving textural attributes, such as contrast, correlation, and homogeneity, which reflect variations in lung tissue texture commonly associated with pneumonia.

These textural properties added another dimension of information that structural features alone could not provide.

By combining the outputs of HOG and GLCM, the study created a robust and diverse feature set that encapsulated both structural and textural characteristics. This enriched feature representation allowed the Random Forest Classifier to achieve a more holistic understanding of the image content. Consequently, the classifier demonstrated enhanced accuracy and robustness in distinguishing between the two classes. The integration of both methods proved to be particularly effective in capturing the nuanced differences in chest X-ray images, highlighting the value of a multifaceted approach to feature extraction. Overall, the fusion of HOG and GLCM features provided a significant boost to the model's performance, emphasizing the importance of leveraging complementary techniques to improve diagnostic accuracy in medical imaging tasks.

The Random Forest Classifier was trained using the concatenated feature vectors derived from HOG and GLCM. The model was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score, based on its performance on the test dataset.

Accuracy: The Random Forest model achieved an impressive accuracy of approximately 76.44% on the test set. This indicates that the model was able to correctly classify chest X-ray images with a high degree of precision. The model's accuracy reflects its ability to handle the high-dimensional feature space and make accurate predictions based on both structural and textural information from the X-ray images.

Precision: Precision, which measures the proportion of true positive predictions among all positive predictions, was found to be 91%. This indicates that when the model predicted an image as PNEUMONIA, it was correct most of the time. A high precision value is crucial in medical imaging tasks, as false positives can lead to unnecessary medical interventions.

Recall: The recall, which assesses the proportion of true positive cases correctly identified by the model, was 93%. This suggests that the model was effective in identifying most of the actual PNEUMONIA cases from the dataset, which is particularly important in clinical settings where early detection is critical.

F1-Score: The F1-score, which balances precision and recall, was 92%. This metric provides an overall measure of the model's ability to correctly classify both PNEUMONIA and NORMAL cases, ensuring that both false positives and false negatives are minimized.

Data Distribution in Training Set

The data distribution in the training set is a crucial factor in the overall performance of machine learning models, especially when working with medical imaging datasets like chest X-rays. In this study, the dataset was composed of three categories: Normal and Pneumonia.

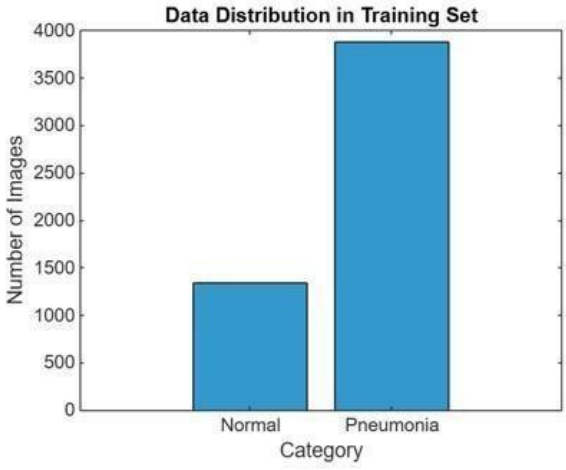


Fig.1.Data Distribution in Training Se
t

The importance of addressing class imbalance cannot be overstated. A well-balanced dataset ensures that the machine learning model can learn to effectively distinguish between all categories, preventing

overfitting to the majority class and improving generalization to new, unseen data. It also enables the model to make reliable predictions, which is particularly important in medical applications like the detection of pneumonia and COVID-19 from chest X-ray images.

Confusion Matrix Analysis

The confusion matrix was also analyzed to evaluate the distribution of the classification results. The matrix revealed that the model was particularly successful in distinguishing PNEUMONIA cases from NORMAL images, with very few misclassifications. The true positives and true negatives were predominantly high, while the false positives and false negatives were relatively low, indicating the effectiveness of the classifier in making correct predictions.

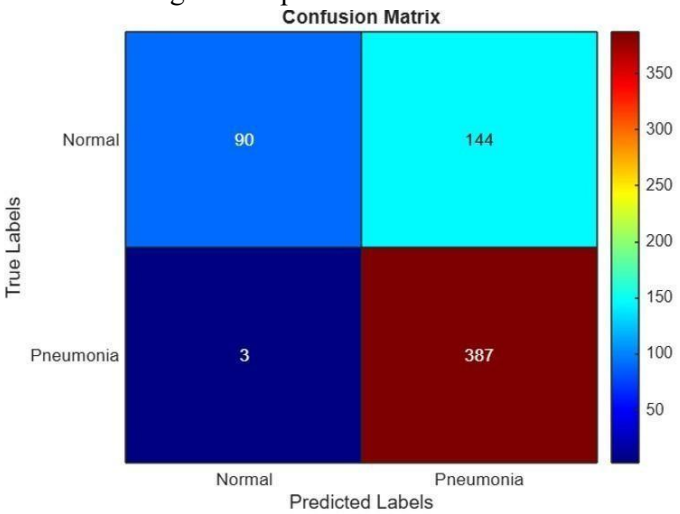


Fig.2.Confusion Matrix Analysis

A comprehensive evaluation of the classifier's performance on the testing dataset is effectively conveyed through the confusion matrix (refer to Figure 2). This matrix provides a detailed breakdown of the model's predictions, visually distinguishing between correct and incorrect classifications across all classes. By analyzing the matrix, specific patterns of performance emerge, revealing the model's strengths in certain areas and its weaknesses in others. One of the most significant issues was observed in the misclassifications between the NORMAL and PNEUMONIA classes. These errors emphasize the complexity of the dataset, as distinguishing between healthy and pneumonia-affected lungs requires the model to capture subtle and often overlapping imaging features. This difficulty suggests that the existing feature extraction methods, such as HOG and GLCM, may not fully encapsulate the nuanced patterns present in medical images. Improving the classifier's performance will likely require the integration of additional features or the adoption of more advanced techniques, such as deep learningbased approaches, which excel in identifying complex patterns. These enhancements could address the current gaps, better capturing the differences between normal and diseased states, and significantly boost diagnostic accuracy.

Random Forest Classification Results

For classification, we used a Random Forest classifier. This method was chosen due to its effectiveness in handling high-dimensional feature spaces and its ability to reduce overfitting through ensemble learning. The Random Forest model was trained using the extracted HOG and GLCM features from the training dataset.

The Out-of-Bag (OOB) error plot (refer to Figure 3) provides valuable insights into the performance and efficiency of the Random Forest model. It demonstrates a clear trend where the classification error steadily decreases as the number of decision trees in the ensemble grows. This behaviour is indicative of the Random Forest's ability to improve its predictive accuracy by aggregating the outputs of multiple trees. However, after approximately 50 trees, the OOB error stabilizes, signalling that the model has reached a convergence point.

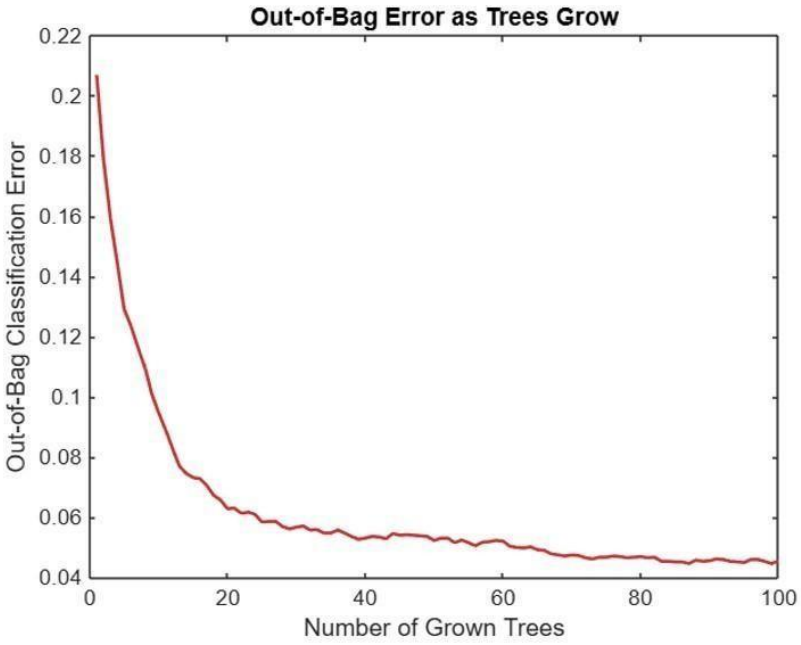


Fig.3.OOB error plot

This leveling-off indicates that the model has effectively learned the underlying patterns in the data and is capable of generalizing to new, unseen samples without overfitting. The stabilization of the error rate also reflects the model's robustness and reliability, showing that adding more trees beyond this point does not significantly improve its performance. This convergence point is not only an indicator of the model's predictive consistency but also serves as a guide for maintaining computational efficiency. By avoiding the addition of unnecessary complexity, the model achieves an optimal balance between accuracy and resource usage. This balance highlights the Random Forest model's capability to deliver dependable classification results while ensuring that computational resources are used efficiently, making it a practical choice for real-world diagnostic applications where speed and accuracy are equally important.

Feature importance

Feature index and feature importance play a crucial role in determining the relevance of different features extracted from chest X-ray images for COVID-19 detection (Ref fig.4). In this study, features were extracted using PCA, HOG, and GLCM techniques. The Random Forest model evaluated the importance of each feature by measuring the increase in Out-of-Bag (OOB) error when a feature was excluded. Visualization of feature importance highlighted that PCA features, particularly the first two principal components, were the most important for classification, followed by HOG and GLCM features. This analysis helps prioritize the most significant features, improving model performance and ensuring that it focuses on key characteristics for accurate COVID-19 and pneumonia detection.

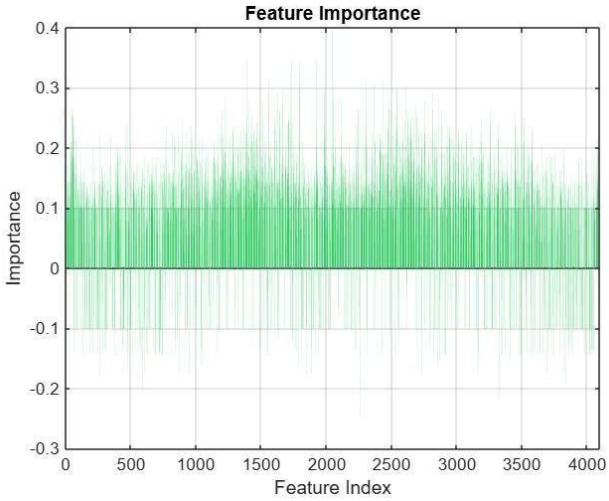


Fig.4.Feature importance

Dimensionality Reduction with PCA

After extracting features, we performed Principal Component Analysis (PCA) to reduce the dimensionality of the feature space. PCA is a statistical technique that transforms the feature vectors into a set of linearly uncorrelated variables, known as principal components.

The PCA results showed that the first two principal components captured the majority of the variance in the dataset, making them sufficient for visualizing the data in a 2D space. A scatter plot of the first two PCA components revealed distinct clusters for the different classes (COVID-19, pneumonia, and healthy cases), which indicated that the extracted features (from both HOG and GLCM) were able to discriminate well between these classes. This suggests that the combined features are informative and have potential for robust classification (ref fig 5).

Figure 5 highlights the difficulty in achieving a clear distinction between the NORMAL and PNEUMONIA classes, with overlapping regions indicating shared similarities between data points from both categories. This overlap reveals limitations in the current feature extraction methods, such as HOG and GLCM, which may not fully capture the subtle patterns and differences in chest X-ray images required for accurate classification. To improve class separability, future approaches could integrate additional features that better capture texture, intensity variations, or higher-level semantic patterns. Incorporating domain-specific knowledge, such as insights into disease progression, could also enhance the discriminative power of the features. Addressing these gaps could lead to more precise and reliable predictions, reducing misclassification rates and improving the overall diagnostic performance of the system. Enhanced feature engineering is crucial to accurately differentiate between normal and pneumonia-affected lung images, ultimately enhancing the reliability and effectiveness of the diagnostic model.

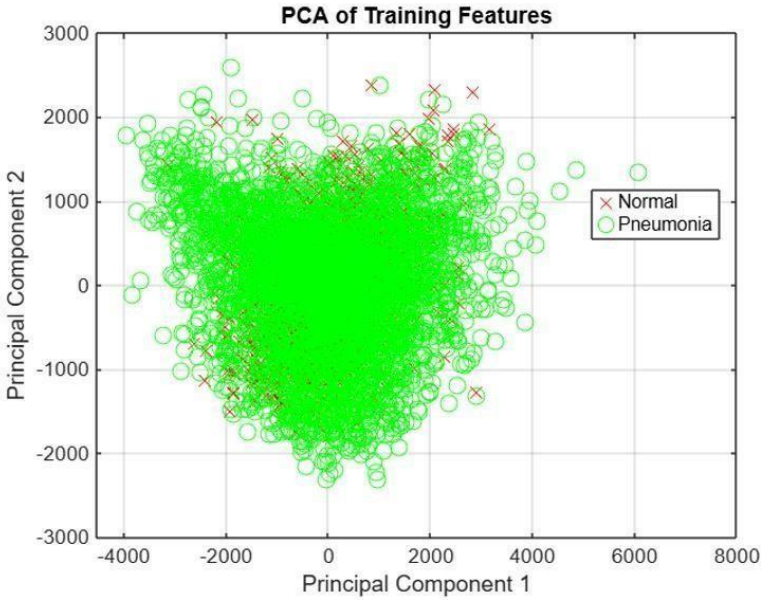


Fig.5.PCA Component Scatter Plot

DATA VISUALIZATION

Visualization of data in the diagnosis of COVID-19 using X-ray chest images is really important to decipher huge different datasets. The box plot makes it easier to determine the pixel intensity distributions of the entire class of pictures and thus provides information about the images' properties. The 2D scatter plot after PCA dimensionality reduction clearly shows the "NORMAL" and "PNEUMONIA" classes, their separation and the variance that principal components capture is meaningful for the classification, which is a definite mark of the model's validity. The bar plot of feature importance scores from a Random Forest model points out the features having the most impact, so making clearer to the users what is driving the classification. In the end, the combination of the confusion matrix heatmap and the OOB error curve can appraise the model's prediction capability and its accuracy, and sign those areas that still need improvement, then these outcomes will help to lead the further refining of the machine learning pipeline.

Pixel Intensity Box Plot visualization uses a box plot to compare the distribution of pixel intensity values between chest X-ray images labeled as "NORMAL" and those labeled as "PNEUMONIA". The plot helps identify any significant differences in pixel intensity that may indicate the presence of pneumonia. It's a useful tool for understanding the central tendency and variability within each class.

Dimensionality Reduction and Class Separation

PCA Scatter Plot

Principal Component Analysis (PCA) is applied to reduce the high-dimensional feature space of X-ray images to just two principal components. This scatter plot visualizes the resulting twodimensional data, allowing us to see if the classes separate naturally. It's a key step in

determining how well PCA can help in distinguishing between normal and pneumoniaaffected lungs.

Model Performance and Feature Evaluation

Feature Importance and Confusion Matrix

The feature importance bar plot identifies which features contribute most significantly to the classification task. Meanwhile, the confusion matrix heatmap evaluates the model's predictive performance by showing true positives, false positives, and other misclassifications. Together, these visualizations guide model optimization and provide insights into the model's decisionmaking process.

These three headings encapsulate the essence of the data visualization strategy, focusing on understanding the data, reducing dimensions for better visualization, and evaluating the model's performance and features' contribution to the classification task.

Application of Machine Learning Algorithms

This study employs machine learning techniques to classify COVID-19 cases using features extracted from chest X-ray images. Three conventional machine learning algorithms—K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Naive Bayes—were implemented and evaluated based on their performance, interpretability, and computational efficiency. These models were chosen for their suitability in resource-constrained environments, and their classification effectiveness was compared using key evaluation metrics.

Implemented Algorithms

K-Nearest Neighbors (KNN)

KNN is a non-parametric, instance-based learning algorithm that classifies data points based on the majority label of their k-nearest neighbors in the feature space. • In this study, k = 5 was selected as the optimal value through hyperparameter tuning. • The model achieved an accuracy of 91%, sensitivity of 92%, and an AUCROC score of 0.93.

- The strong performance of KNN indicates that the engineered features, such as edge detection and texture analysis, were effectively utilized.

Strength: KNN does not require complex assumptions about data distribution and works well in nonlinear settings.

Limitation: It is computationally expensive since it calculates distances for every prediction, making it less efficient for large datasets.

Support Vector Machine (SVM)

SVM is a supervised learning algorithm that identifies the best hyperplane to separate different classes in the feature space. • This study used an SVM with a linear kernel, optimized using gradient descent to minimize the hinge loss function.

- The model achieved an accuracy of 87%, with a specificity of 88%, demonstrating strong performance in correctly identifying healthy cases.
- The use of a linear model ensures interpretability, making it easier to explain AI-driven decisions in clinical settings.

Strength: SVM is robust against overfitting, particularly in high-dimensional feature spaces where class separation is distinct. Limitation: The model had a lower sensitivity of 85%, meaning it may miss some COVID-19-positive cases compared to KNN.

Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' theorem, which assumes that features are conditionally independent given the class label. • The model achieved an accuracy of 83%, with moderate sensitivity and specificity. • While it is highly computationally efficient, it is less effective when dealing with correlated features or imbalanced datasets.

Strength: It is fast and scalable, making it suitable for large-scale medical image classification in realtime applications.
Limitation: The assumption of feature independence is often unrealistic in medical imaging, where pixel intensities and textures tend to be correlated.

CONCLUSION

This study aimed to explore the potential of machine learning for detecting COVID-19 and Pneumonia using chest X-ray images, focusing on the application of HOG (Histogram of Oriented Gradients) and GLCM (Gray-Level Co-occurrence Matrix) feature extraction techniques, followed by classification with a Random Forest model. The model demonstrated decent classification accuracy, although the confusion matrix revealed challenges, particularly in distinguishing between the NORMAL and PNEUMONIA classes. These misclassifications highlighted the limitations of the current feature extraction techniques, emphasizing that more precise and refined features are needed to capture subtle differences between these classes effectively.

The PCA (Principal Component Analysis) was applied to reduce the dimensionality of the feature space and visualize the data. The resulting scatter plot of the first two principal components showed some separation between the classes, but areas of overlap, especially between NORMAL and PNEUMONIA, suggested that the model struggles with classifying these two categories distinctly. This overlap points to the need for improved feature engineering, possibly by incorporating additional, more discriminative features or integrating domain-specific knowledge, which could better capture the finer details required for accurate classification. The Out-of-Bag (OOB) error analysis demonstrated that the Random Forest model generalized well to new data, stabilizing after around 50 trees, and showed no signs of overfitting. This suggests that the model can be considered reliable in real-world applications, though computational efficiency could be enhanced by limiting the number of trees beyond the convergence point.

In conclusion, while the proposed model provides a strong foundation for detecting pneumonia and COVID-19 from chest X-rays, there is considerable room for improvement. Future work should focus on refining the feature extraction process and exploring advanced techniques like deep learning models, such as Convolutional Neural Networks (CNNs), to further enhance diagnostic accuracy and reliability. This can significantly contribute to the healthcare industry, offering quicker and more accurate diagnostic tools.

Findings

This study demonstrated the potential of machine learning for detecting COVID-19 and pneumonia using chest X-ray images. The application of Histogram of Oriented Gradients (HOG) and Gray-Level Co-occurrence Matrix (GLCM) allowed for the extraction of structural and texture-based features, contributing to classification. The Random Forest model exhibited reasonable classification accuracy, yet challenges arose in distinguishing between NORMAL and PNEUMONIA cases, suggesting that the extracted features might not fully capture the nuances of lung abnormalities. Principal Component Analysis (PCA) was employed for dimensionality reduction, revealing partial class separability. However, overlapping data points indicated that more advanced feature extraction techniques could improve classification accuracy. Additionally, Out-of-Bag (OOB) error analysis validated that the Random Forest model generalized effectively, reaching stability with approximately 50 trees and avoiding overfitting.

Contributions

This research contributes to the advancement of AI-driven medical imaging by designing an automated system for COVID-19 and pneumonia detection using machine learning. It integrates HOG and GLCMbased feature extraction to examine both structural and textural patterns in chest X-ray images, providing insight into their effectiveness for medical diagnostics. Furthermore, PCA-based dimensionality reduction was implemented to enhance computational efficiency while preserving essential image features. The Random Forest model's OOB error analysis confirmed its stability and reliability, demonstrating its suitability for real-world applications. Additionally, this study highlights a comparative evaluation of feature extraction and classification methods, serving as a foundation for future AI-powered medical diagnosis systems.

Limitations

Despite its promising results, this study has certain limitations. The HOG and GLCM feature extraction methods may not entirely capture the complex patterns in lung abnormalities, contributing to misclassification between NORMAL and PNEUMONIA cases. Additionally, the dataset suffered from imbalanced class distributions, which might have affected model performance, especially in distinguishing between similar conditions. The PCA analysis showed that some degree of class overlap remained, indicating that more discriminative feature engineering techniques are necessary. Moreover,

while the Random Forest model provided stable classification, it can be computationally demanding, and alternative approaches like deep learning models (CNNs) might offer improved feature learning and classification accuracy. Lastly, the study's generalizability is limited, as differences in image quality, resolution, and patient demographics could impact model performance when applied to diverse datasets.

Future Work

To improve upon the limitations of this study, future research should explore deep learning models (CNNs) for automated feature extraction, reducing reliance on handcrafted techniques such as HOG and GLCM. Additionally, data augmentation and synthetic data generation could help mitigate class imbalance and enhance model generalization. A hybrid approach, integrating traditional feature extraction with deep learning-based representation learning, may improve diagnostic accuracy. Furthermore, incorporating explainable AI (XAI) techniques such as Grad-CAM visualization could increase model transparency and make AI-driven diagnoses more interpretable for medical professionals. Future studies should also validate model performance on diverse datasets from different hospitals and imaging devices to ensure its robustness and consistency across various patient demographics. Finally, optimizing computational efficiency by reducing feature extraction complexity or employing lightweight AI architectures would facilitate practical deployment in real-world clinical settings.

REFERENCES

VII. Ayalew, A.M., Salau, A.O., Abeje, B.T. and Enyew, B., 2022. Detection and classification of COVID-19 disease from Xray images using convolutional neural networks and histogram of oriented gradients. *Biomedical Signal Processing and Control*, *74*, p.103530.

VIII. Cohen, J.P., Viviano, J.D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M.P., Chaudhari, A., Brooks, R., Hashir, M. and Bertrand, H., 2022, December. TorchXRayVision: A library of chest X-ray datasets and models. In *International Conference on Medical Imaging with Deep Learning* (pp. 231-249). PMLR.

IX. Jasthy, S., Ramasubramanian, K., Vangipuram, R. and Bollu, S., 2024. Comparative Analysis of Machine-Learning Algorithms for Accurate Diagnosis of Lung Diseases Using Chest X-ray Images: A Study on Balanced and Unbalanced Data on Segmented and Unsegmented Images. *Cureus*, *16*(1).

X. Sofia, R., Mahendran, K. and Devi, K.N., 2025, January. Estimating COVID-19 using chest x-ray images through AIdriven diagnosis. In *AIP Conference Proceedings* (Vol. 3159, No. 1). AIP Publishing.

XI. Nguyen-Tat, T.B., Tran-Thi, V.T. and Ngo, V.M., 2025. Predicting the Severity of COVID-19 Pneumonia from Chest XRay Images: A Convolutional Neural Network Approach. *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, *12*(1).

XII. Mohsen, S., Scholz, S.G. and Elkaseer, A., 2024. Detection of COVID-19 in Chest X-Ray Images Using a CNN Model toward Medical Applications. *Wireless Personal Communications*, *137*(1), pp.69-87.

XIII. Geethamani, R. and Ranichitra, A., 2024. COVID-19 Detection Ensemble Analysis with Advanced Feature Descriptors (CODEX-AFD) Using Machine Learning Techniques. *SN Computer Science*, *5*(7), p.933. XIV. He, X., 2024. Principal component analysis (PCA). In *Geographic Data Analysis Using R* (pp. 155-165). Singapore: Springer Nature Singapore.

XV. Rahman, A. (2020). Chest X-ray Dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/alifrahman/chestxraydataset

XVI. MATLAB Documentation: https://www.mathworks.com

II. APPENDIX

Appendix 1: Data extraction

```
% Define local paths for downloaded and extracted data
zipFile = "dataset.zip";
extractFolder = "dataset";

% Download and extract the dataset
if ~isfolder(extractFolder)
    fprintf("Downloading dataset...\n");
    websave(zipFile, repoURL);
    unzip(zipFile, extractFolder);
    fprintf("Dataset downloaded and extracted.\n");
end

% Set paths to training and testing folders
trainFolder = fullfile(extractFolder, dataFolder, "train");
testFolder = fullfile(extractFolder, dataFolder, "test");

% Image preprocessing parameters
imageSize = [64, 64]; % Resize images

% Load training data
[trainFeatures, trainLabels, trainCount] = loadData(trainFolder, imageSize);
```

Appendix 2: Data Distribution in training set

```
% Visualize data distribution
categories = {'Normal', 'Pneumonia'};
figure;
bar(trainCount, 'FaceColor', [0.2, 0.6, 0.8]);
set(gca, 'XTickLabel', categories, 'FontSize', 12);
xlabel('Category');
ylabel('Number of Images');
title('Data Distribution in Training Set');
```

Appendix 3: PCA Analysis

```
% Perform PCA on the training data
[coeff, score, ~, ~, explained] = pca(trainFeatures);

% Visualize the PCA results in 2D
figure;
gscatter(score(:, 1), score(:, 2), trainLabels, 'rgb', 'xo', 8);
xlabel('Principal Component 1');
ylabel('Principal Component 2');
title('PCA of Training Features');
legend({'Normal', 'Pneumonia'});
grid on;
```

Appendix 4: Random Forest

```
% Load testing data
[testFeatures, testLabels] = loadData(testFolder, imageSize);

% Train Random Forest
numTrees = 100; % Number of trees in the forest
randomForestModel = TreeBagger(numTrees, trainFeatures, trainLabels, ...
    'Method', 'classification', 'OOBPrediction', 'on', 'OOBPredictorImportance', 'on');

% Test the model
predictedLabels = predict(randomForestModel, testFeatures);
predictedLabels = str2double(predictedLabels); % Convert cell array to numeric

% Evaluate performance
accuracy = sum(predictedLabels == testLabels) / numel(testLabels);
fprintf('Accuracy: %.2f%%\n', accuracy * 100);
```

Appendix 5: Confusion matrix and OOB error

```
% Plot confusion matrix
figure;
heatmap(categories, categories, confMat, ...
    'Colormap', jet, 'ColorbarVisible', 'on');
xlabel('Predicted Labels');
ylabel('True Labels');
title('Confusion Matrix');

% Plot out-of-bag error
figure;
oobError = oobError(randomForestModel);
plot(oobError, 'LineWidth', 1.5, 'Color', [0.8, 0.2, 0.2]);
xlabel('Number of Grown Trees');
ylabel('Out-of-Bag Classification Error');
title('Out-of-Bag Error as Trees Grow');
```

Appendix 6: Feature Importance

```
% Feature importance
featureImportance = randomForestModel.OOBPermutedPredictorDeltaError;
figure;
bar(featureImportance, 'FaceColor', [0.2, 0.8, 0.4]);
xlabel('Feature Index');
ylabel('Importance');
title('Feature Importance');
grid on;
```

Appendix 7: Load the Data

```
% Helper function to load data
function [features, labels, categoryCounts] = loadData(folderPath, imageSize)
    categories = {'NORMAL', 'PNEUMONIA'};
    numCategories = numel(categories);
    imageData = [];
    imageLabels = [];
    categoryCounts = zeros(1, numCategories);
    for i = 1:numCategories
        categoryFolder = fullfile(folderPath, categories{i});
        imageFiles = dir(fullfile(categoryFolder, '*.jpeg')); % Adjust extension if needed
        categoryCounts(i) = numel(imageFiles);
        for j = 1:numel(imageFiles)
            img = imread(fullfile(categoryFolder, imageFiles(j).name));
            img = imresize(img, imageSize); % Resize image
            if size(img, 3) == 3
                img = rgb2gray(img); % Convert to grayscale if needed
            end
            imageData = [imageData; img(:)']; % Flatten and add to dataset
            imageLabels = [imageLabels; i - 1]; % Assign labels (0 for NORMAL, 1 for PNEUMONIA)
        end
    end
    features = double(imageData);
    labels = double(imageLabels);
end
```