MACHINE LEARNING MODEL TO DETECT PNEUMONIA USING CHEST X-RAY

Raju Munavath, Yadaiah Jella, Dr M Venkat Reddy
Assistance Professor, Assistance Professor, Professor
Dept of CSE
Sree Dattha Institute of Engineering and Science

ABSTRACT

Pneumonia, a respiratory infection caused by the inflammation of air sacs due to viruses and bacteria, affects approximately 7% of the global population annually, with 4 million patients facing fatal risks. Early diagnosis is crucial, and typical symptoms include chest pain, shortness of breath, and cough. However, diagnosing pneumonia in children is challenging due to the low sensitivity of tests and weak clinical findings. Chest X-rays have become an important diagnostic tool, but the conventional approach involving manual examination by radiologists is time-consuming, subjective, and can vary in accuracy. To address this, the proposed model leverages machine learning (ML), specifically designed for image analysis, to automatically learn and extract relevant features from chest X-ray images. The dataset consists of annotated chest X-rays collected from diverse patient populations, including both pneumonia-positive and pneumonia-negative cases. This model holds significant implications for the medical field and patient care, as it can rapidly analyze large volumes of chest X-ray images and accurately detect pneumonia patterns with a high level of precision. This will enable healthcare professionals to prioritize urgent cases, expedite diagnosis, and promptly initiate appropriate treatments, leading to improved patient outcomes, reduced hospital stays, and optimized resource allocation within healthcare facilities.

Keywords: Machine Learning Model, Pneumonia Detection, Chest X-Ray, Respiratory Infection, Inflammation, Air Sacs, Viruses, Bacteria, Global Population, Fatal Risks, Early Diagnosis, Symptoms, Chest Pain, Shortness Of Breath, Cough, Children, Low Sensitivity, Diagnostic Tool, Radiologists, Image Analysis, Relevant Features.

1. INTRODUCTION

The number of individuals suffering from pneumonia is approximately more than 450 million a year. It is 7% of the overall population around the globe. Each year more than four million people die from Pneumonia [1]. Pneumonia disease is prevalent among young children below 5 years old [2]. According to the report released by "our World in data" [3], children below five have the highest death rate caused by pneumonia (Fig. 1). In 2017, 808,920 children died due to pneumonia, and this figure is 16 folds more than the deaths caused by cancer a year and ten folds higher than people who died from HIV. According to the report released during World Pneumonia Day, it is estimated that more than 11 million infant children below the age of 5 years are likely to die from pneumonia by the year 2030 [4]. In the early nineteenth century, pneumonia was considered one of the significant causes of death amongst people. In the past, medical doctors relied on several methods such as clinical examination, medical history, and chest X-rays to diagnose patients suffering from pneumonia. Nowadays, Chest-X-rays have become increasingly cheaper due to rapid advancements in technologies such as bio-medical equipment. The Chest X-ray is commonly used in detecting pulmonary diseases like pneumonia. The problem of lack of experts can be addressed through the use of different computer-aided diagnosis techniques. Technological advancements in artificial intelligence (AI) have proven to be helpful in the diagnosis of disease. For instance, techniques like CNN are utilised for classifying Chest-X-rays in order to determine whether pneumonia is present. Some of the exciting research has been done in areas like abnormal-patterns detection [5], biometric recognition [6], trauma seriousness valuation [7, 8], accident prevention at the airport [9], predicting efficiency in information using ANN [10] and diagnoses of bone pathology. However, the higher divergence in the image features impacts the retrieval accuracy. The primary objective is to classify a given CXR image into one of two classes: "Pneumonia Positive," indicating the presence of pneumonia, or "Pneumonia Negative," indicating the absence of pneumonia. To address this problem effectively, a sizable dataset of labelled CXR images is essential. These images should be carefully annotated by medical experts to indicate the presence or absence of pneumonia, and the dataset should encompass various pneumonia types and severity levels. Preprocessing of CXR images is a critical step in preparing the data for ML model training. Images may vary in terms of size, orientation, and quality, necessitating resizing, normalization, and noise reduction to ensure consistent input data. The ML model undergoes training and validation on separate subsets of the dataset. During this phase, hyperparameter tuning and model selection are performed to optimize performance. Evaluation is carried out on a distinct test dataset, measuring metrics like accuracy, precision, recall, and F1-score. Special attention is given to false positives and false negatives, as they have different clinical implications.

2.LITERATURE SURVEY

2.1 Introduction

Pneumonia detection using machine learning models has revolutionized the medical field by offering efficient and accurate diagnostic capabilities. Pneumonia, an acute respiratory infection affecting the lungs, poses significant health risks, especially in regions with limited access to healthcare professionals. Pneumonia detection using ML has emerged as a informative technology in the field of medical safety. These systems leverage artificial intelligence to enhance early detection, reduce false assumptions, and provide valuable analytics for risk management. AI-based systems can detect diseases at an earlier stage than traditional detectors. AI algorithms can distinguish between real one and fake more effectively. Early diagnosis of pneumonia is critical for effective treatment and reducing mortality rates associated with the disease. Machine learning-assisted prediction models based on non-invasive measures like biomarkers and physical features have shown promising results in predicting pneumonia accurately In conclusion, the application of machine learning algorithms in pneumonia detection represents a significant advancement in medical diagnostics. These innovative approaches not only enhance diagnostic accuracy but also streamline the detection process, making it more accessible even in resource-constrained settings. The ongoing research focus on improving these models further underscores their potential to revolutionize pneumonia diagnosis and improve patient care outcomes.

2.2 Surveys

Ren et. al [1] studied two cases (a)investigated the performance disparities between geriatric and younger patients when using chest X-ray images to detect pneumonia, and (b)developed and tested a multimodal model called CheXMed that incorporated clinical notes together with image data to improve pneumonia detection performance for older people. Accuracy, precision, recall, and F1-score were used for model performance evaluation. CheXMed outperformed baseline models on all evaluation metrics. The accuracy, precision, recall, and F1-score were 0.746, 0.746, 0.740, 0.743 for CheXMed, 0.645, 0.680, 0.535, 0.599 for CheXNet, 0.623, 0.655, 0.521, 0.580 for DenseNet121, and 0.610, 0.617, 0.543, 0.577 for ResNet18.

Linghua et. al [2] proposed an anchor-free object detection framework and RSNA dataset based on pneumonia detection. First, a data enhancement scheme was used to preprocess the chest X-ray images; second, an anchor-free object detection framework was used for pneumonia detection, which contained a feature pyramid, two-branch detection head, and focal loss. The average precision of 51.5 obtained by Intersection over Union (IoU) calculation showed that the pneumonia detection results obtained in this study could surpass the existing classical object detection framework, providing an idea for future research and exploration.

Nalluri et. al [3] proposed AHGOA (Archimedes-assisted Henry Gas Optimization Algorithm) model selected the best characteristics from the retrieved features. It was the best option for avoiding the dimensionality curse. The selected EC + AHGOA method obtained good accuracy (~0.95) for tuning percentage 70 in pneumonia diagnosis from Chest X-ray images than some other previous techniques, including EC + AOA (~0.92), EC + HGSO (~0.93), EC + HGS (~0.88), EC + PRO (~0.90), and EC + BES (~0.89).

Mann's et. al [4] proposed model, image preprocessing was performed using Student's t distribution, a compact probability density function (cPDF), for better sampling and segregation between the healthy and infected part of lungs, to improve the predictions. Further, a hybrid deep convolutional neural network model was built to extract image features by fine-tuning the pretrained models, viz. Resnet-50, EfficientNet, VGG-16, MobileNetV2 and DenseNet to achieve better results of diagnosis. The proposed hybrid model was analyzed using Grad-CAM visualization, which produced a course localization map, highlighting the infected region in the image used for prediction. The proposed hybrid model was evaluated based on governing parameters, viz. precision, recall, F1-score and accuracy. The results showed that the proposed model achieved a precision of 97.47%, recall of 98.09%, F1-score of 97.77%, and overall accuracy of 97.69% as compared to other existing models.

Udbhav et. al [5] proposed a project used ResNet, which performed well in image-recognition-related tasks and was an important part of a deep convolutional neural network. The project focused on proper classification, causes detection, and helped people avoid pneumonia infection by the use of (ANN) artificial neural networks and (CNN) convolutional neural network systems and their working methodologies. By the use of these techniques, many deaths could be stopped, helping people to live freely.

Lowie et. al [6] proposed the project, the data set contained 5,800 images, which was considerably less compared to deep-learning standards. In many situations, the count could even be hundreds or thousands, which was difficult to process on local machines. The model used in the project was quite heavy and would consume a lot of time to run on a normal CPU. To solve the issue, the project used Google Collaborator, which helped very much in learning the deep-learning model. Google Collab was a component of the Google Research project that supported machine-learning research and education and was hosted on a Google Cloud, which was available to everyone at no cost. By use of Collab, deep-learning needs could easily be fulfilled. For the image-based dataset, the project used ResNet, which performed well in image-recognition-related tasks and was an important part of a deep convolutional neural network. The project focused on proper classification, causes detection, and helped people avoid pneumonia infection by the use of (ANN) artificial neural networks and (CNN) convolutional neural network systems and their working methodologies. By the use of these techniques, many deaths could be stopped, helping people to live freely.

Omar et. al [7] studied, E. coli and K. pneumonia were isolated from infected urine samples, and the existence of the CTX-M resistant genes in ESBL-producing bacterial isolates of E. coli and K. pneumonia was determined. The antibacterial effect of Neem plant (A. indica) ethanolic extract against these bacteria was also determined. For phenotype detection of ESBLs, the hybrid disc method was used, and 12 available commercial antibiotics were used in the antibiotic susceptibility test. The tested bacterial isolates showed high resistance to these antibiotics. The PCR results confirmed that the ESBL-producing isolates E. coli and K. pneumonia had the CTX-M gene and the CTX-M-1 gene. The antibacterial activity of ethanolic extract of Neem plant against these bacteria was evaluated by the agar well diffusion method at different concentrations (50, 100, 200, and 300mg/ml). The inhibition zone diameters were (2, 4, 5, and 6mm) against E.coli (ESBL+) and (0, 2, 5, and 6mm) against K. pneumonia (ESBL+), respectively, while E.coli (ESBL-) was (7, 6, 4, and 3 mm) respectively, and with K. pneumonia (ESBL-) it gives (7, 6, 3, and 2 mm) respectively. The MIC for E.coli (ESBL+) was 125mg/ml, while the MIC for E.coli (ESBL-) was 62.5 mg/ml. The MIC for K. pneumonia (ESBL-) was at a concentration of 31.25mg/ml and for K. pneumonia (ESBL+) was at 62.5mg/ml. The ethanolic extract of Neem plant at different concentrations had antibacterial activity against (ESBL+) and (ESBL-) bacterial isolates of E. coli and K. pneumonia.

Chengxiang Zhang et. al [8] proposed a DBM-ViT model for deep learning. The model utilized CXR/CT lung images for effective health detection of normal, COVID-19, and other types of pneumonia. The model employed depthwise convolutions with different expansion rates to efficiently capture global information from CXR/CT lung images. Then, the lung feature maps with combined sequences were fed into the ViT module to capture local information. Multi-scale features combined with global and local information ensured maximum feature learning. The results showed that the detection accuracy of the DBM-ViT model in the CXR/CT image dataset reached 97.25%/98.36%. This method could effectively capture global and local information in lung images with high detection accuracy and could be used for rapid auxiliary diagnosis of pneumonia types.

3. PROPOSED METHODOLOGY

3.1 Overview

A respiratory infection called pneumonia has the potential to be fatal if it is not identified and treated quickly. Through the analysis of medical images, such as chest X-rays, machine learning (ML) algorithms can help detect pneumonia by finding patterns that are suggestive of the illness. The main goal of this project is to detect pneumonia using the Random Forest (RF) algorithm.

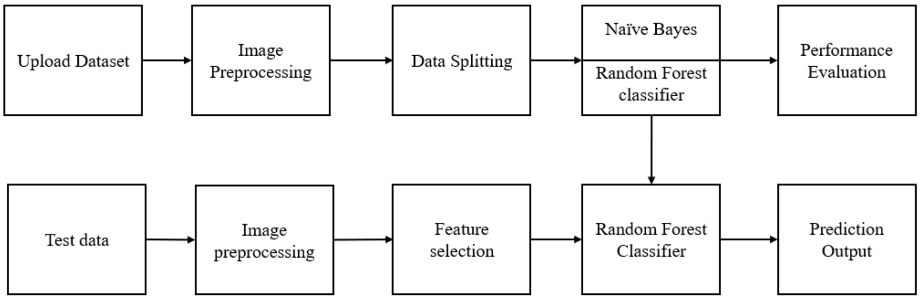


Figure: Proposed Block diagram of Pneumonia detection

The RF model will be trained and evaluated using a dataset of chest X-ray pictures. Chest X-Ray Images (CXR) is a frequently used dataset that includes many chest X-ray images with the labels "Normal" or "Pneumonia."

Firstly, we gathered relevant medical data, including patient histories, physical examination results, and diagnostic tests, such as chest X-rays and blood tests. This data served as the foundation for training our machine learning model.

Next, we pre-processed the collected data by cleaning and organizing it to ensure its quality and compatibility with the machine learning algorithm. This step involved handling missing values, normalizing features, and encoding categorical variables, among other preprocessing techniques. We then proceeded to train our Random Forest model using the pre-processed data. This algorithm is well-suited for classification tasks and has shown promising results in medical diagnosis. During the training phase, the model learned to distinguish between pneumonia and non-pneumonia cases based on the provided features. After training, we evaluated the performance of our model using various metrics, such as accuracy, precision, recall, and F1-score. This evaluation helped us assess the effectiveness of our system in accurately detecting pneumonia cases. To further enhance the accuracy and generalization capability of our model, we performed hyperparameter tuning, optimizing the parameters of the Random Forest algorithm. By finding the best combination of parameters, we aimed to improve the model's performance and reduce the risk of overfitting. Once satisfied with the model's performance, we deployed it in a user-friendly interface or integrated it into a larger healthcare system. This allowed healthcare providers to input patient data and receive real-time predictions on the likelihood of pneumonia. Throughout the project, we prioritized the interpretability and explainability of our model. This involved analyzing the significance of the features used in the prediction process, allowing healthcare professionals to understand the factors contributing to the pneumonia diagnosis.

3.2 Random Forest Classifier

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

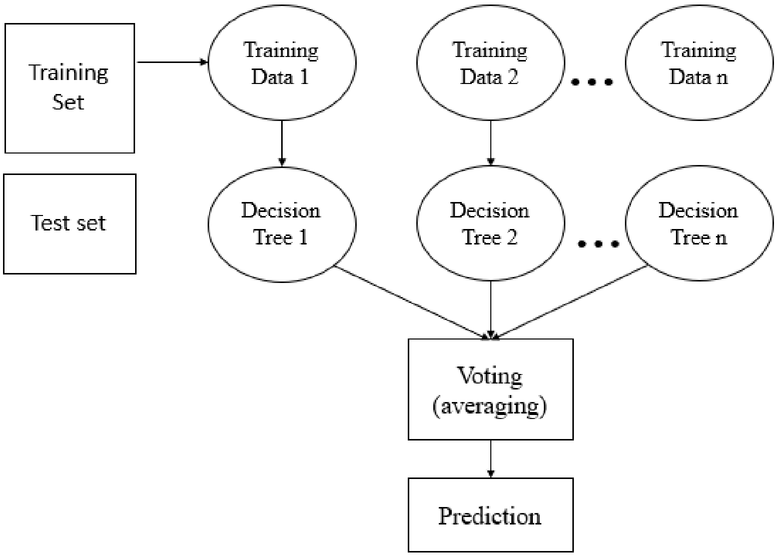


Figure: Random Forest algorithm

Random Forest algorithm

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

Important Features of Random Forest

- **Diversity**- Not all attributes/variables/features are considered while making an individual tree, each tree is different.
- **Immune to the curse of dimensionality**- Since each tree does not consider all the features, the feature space is reduced.
- **Parallelization**-Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.
- **Train-Test split**- In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.
- **Stability**- Stability arises because the result is based on majority voting/ averaging.

3.2.1 Assumptions for Random Forest

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Below are some points that explain why we should use the Random Forest algorithm

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

3.2.2 Types of Ensembles

Before understanding the working of the random forest, we must look into the ensemble technique. Ensemble simply means combining multiple models. Thus, a collection of models is used to make predictions rather than an individual model. Ensemble uses two types of methods:

Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest. Bagging, also known as Bootstrap Aggregation is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.

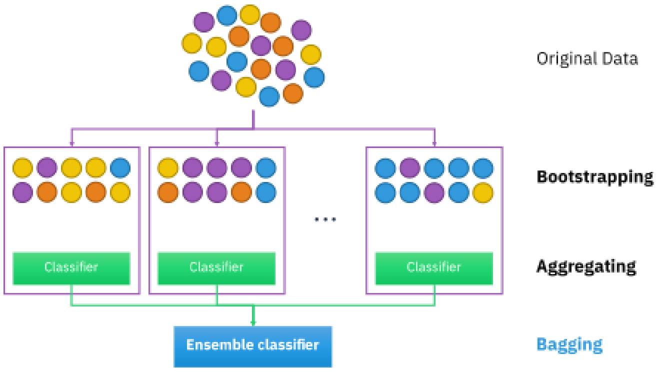


Figure. RF Classifier analysis.

Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

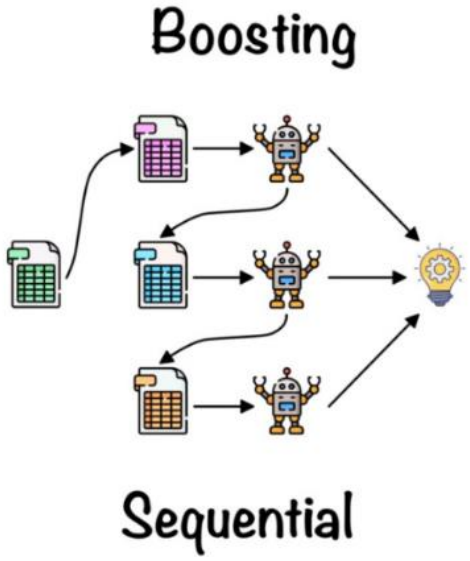


Figure. Boosting RF Classifier.

Advantages of Random Forest

- It can be used in classification and regression problems.
- It solves the problem of overfitting as output is based on majority voting or averaging.
- It performs well even if the data contains null/missing values.
- Each decision tree created is independent of the other thus it shows the property of parallelization.
- It is highly stable as the average answers given by a large number of trees are taken.
- It maintains diversity as all the attributes are not considered while making each decision tree though it is not true in all cases.
- It is immune to the curse of dimensionality. Since each tree does not consider all the attributes, feature space is reduced.

Applications of Random Forest: There are mainly four sectors where Random Forest mostly used:

- Banking: Banking sector mostly uses this algorithm for the identification of loan risk.
- Medicine: With the help of this algorithm, disease trends and risks of the disease scan be identified.
- Land Use: We can identify the areas of similar land use by this algorithm.
- Marketing: Marketing trends can be identified using this algorithm.

4.RESULTS

Figure 1 shows a selection of images from the dataset that are classified as belonging to the "pneumonia" class. These images likely exhibit characteristics associated with pneumonia in chest X-ray images.

Figure 2 displays sample images from the dataset categorized as "normal." These images are likely examples of chest X-ray images with no signs of pneumonia or abnormalities.

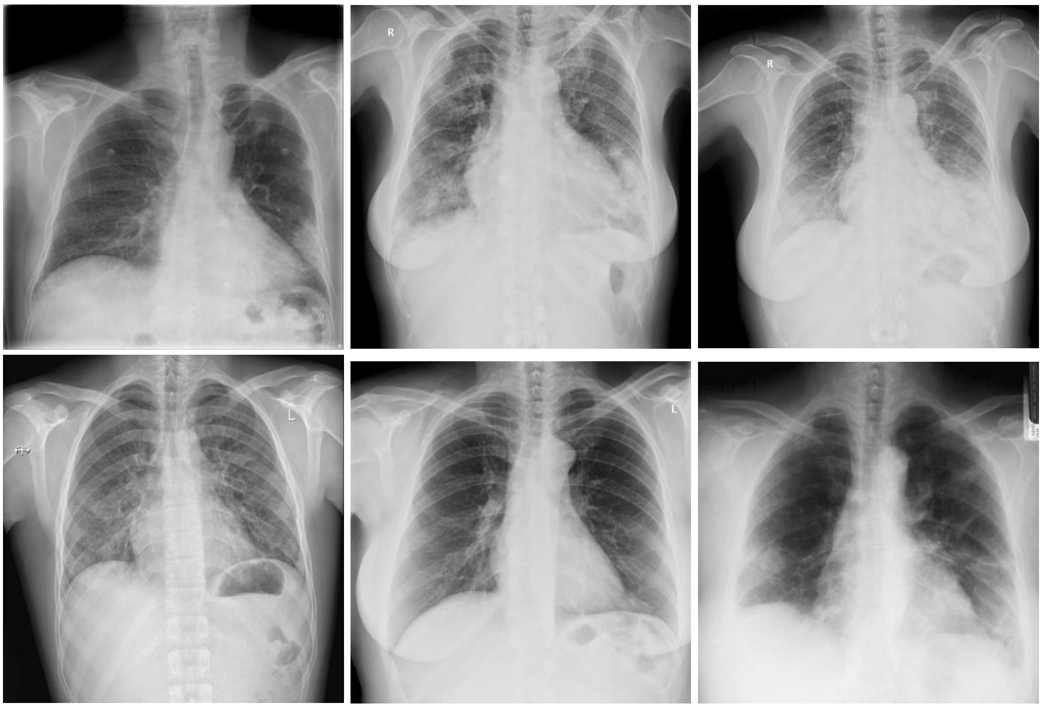


Figure 1: Sample images of dataset with pneumonia class.

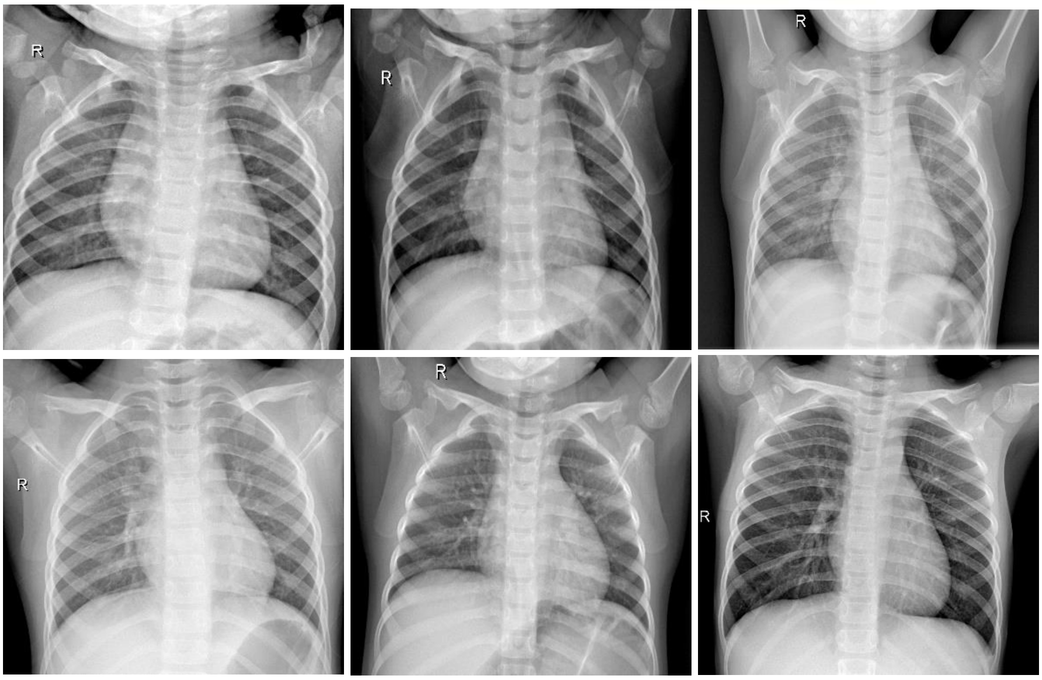
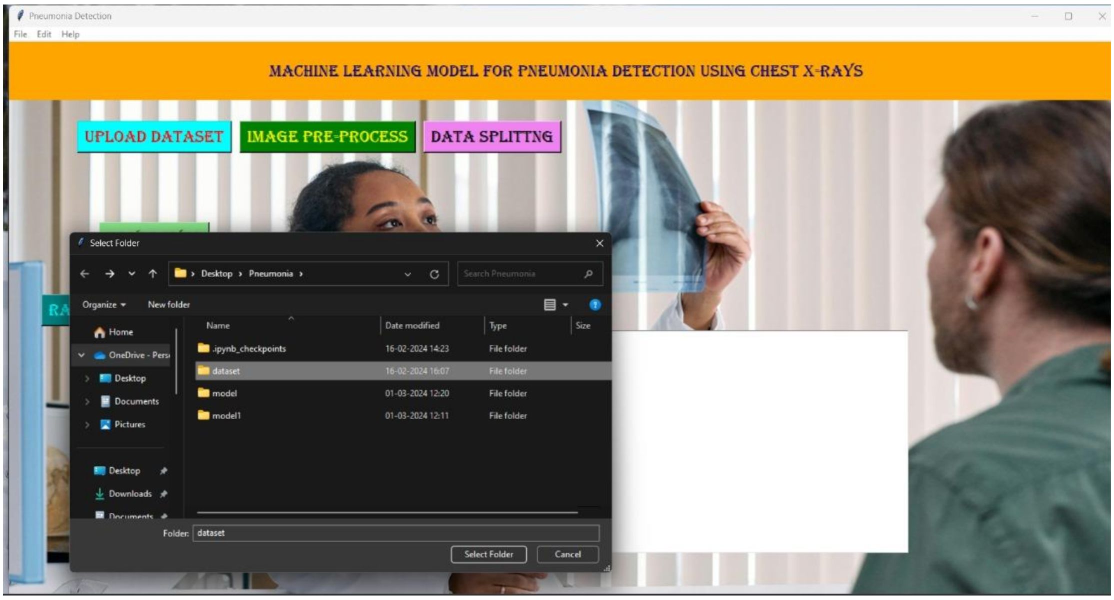


Figure 2: Sample images from dataset with normal class.



Name	Date modified	Type	Size
.ipynb_checkpoints	16-02-2024 14:23	File folder	
dataset	16-02-2024 16:07	File folder	
model	01-03-2024 12:20	File folder	
model1	01-03-2024 12:11	File folder	

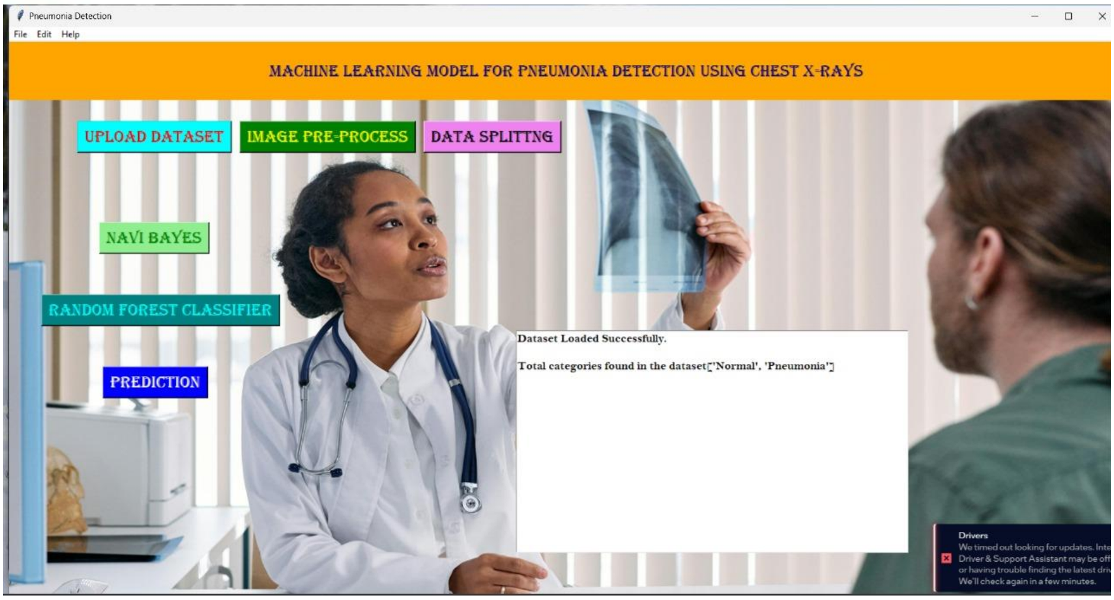
Folder: dataset

Figure 3: UI Shows uploading Dataset

Figure 3 depicts a user interface (UI) where a dataset is being uploaded for processing. It's the initial step in the data processing pipeline.

Figure 4 UI displays the UI after the uploaded data, presumably chest X-ray images, has undergone preprocessing steps such as resizing, normalization, or enhancement.

Figure 5 demonstrates the UI after the dataset has been split into training, validation, and testing subsets, a crucial step in training and evaluating ML models.



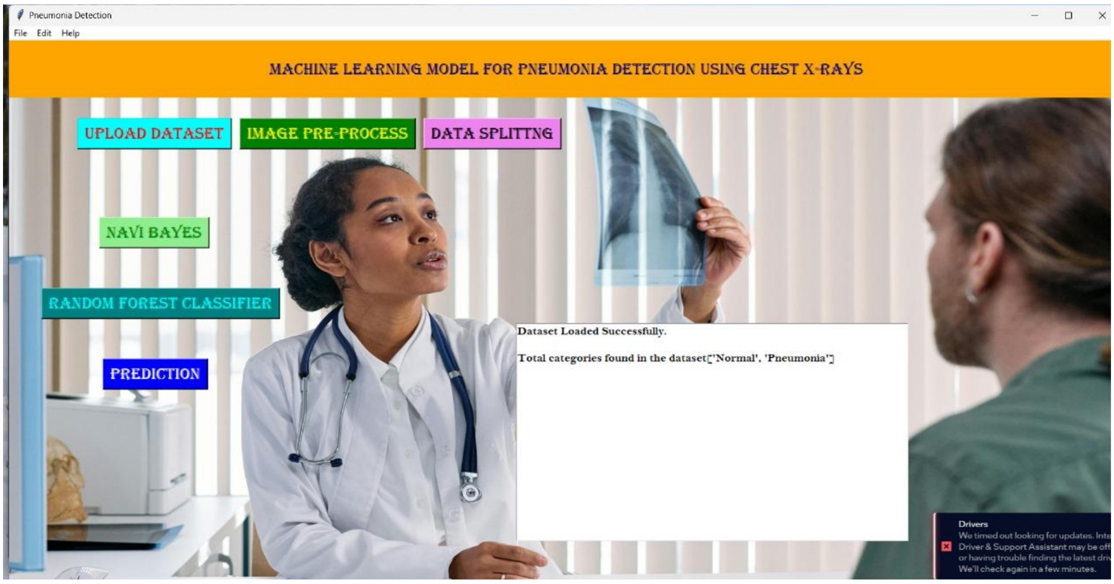
```
Dataset Loaded Successfully.

Total categories found in the dataset['Normal', 'Pneumonia']
```

```
Drivers
We timed out looking for updates. Inte
Driver & Support Assistant may be off
or having trouble finding the latest dri
We'll check again in a few minutes.
```

Figure 4:UI shows the Data after Image Pre-Processing

Figure 6 presents the accuracy metrics and confusion matrix associated with a Random Forest classifier. It shows how well the classifier performed in terms of correctly predicting classes and identifying misclassifications.

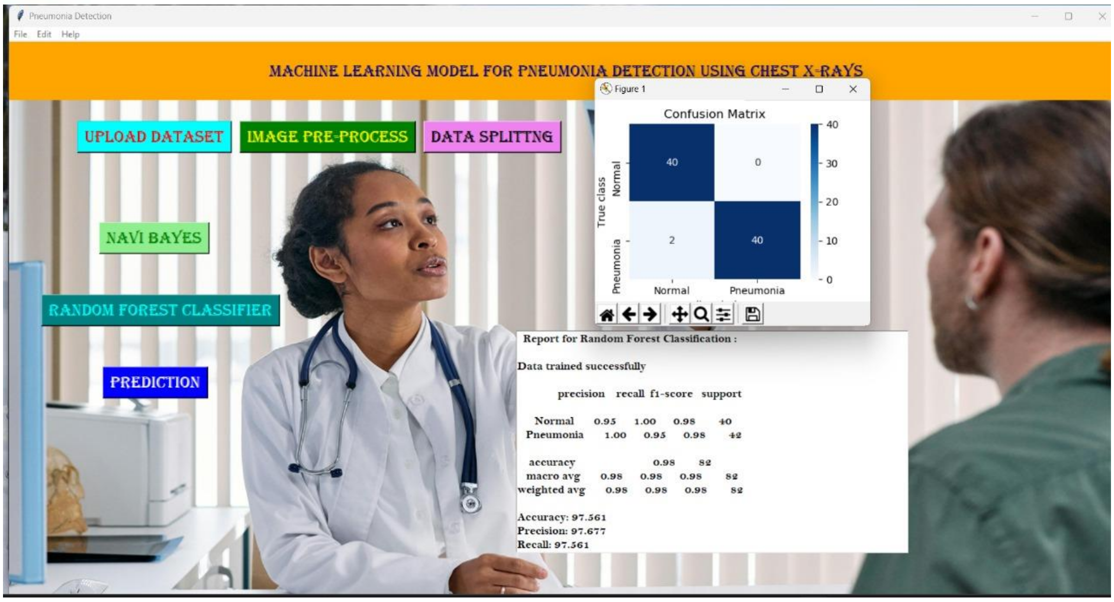


```
Dataset Loaded Successfully.

Total categories found in the dataset['Normal', 'Pneumonia']
```

```
Drivers
We timed out looking for updates. Inte
Driver & Support Assistant may be off
or having trouble finding the latest dri
We'll check again in a few minutes.
```

Figure 5:UI shows the data after data Splitting



```
Report for Random Forest Classification :

Data trained successfully

          precision   recall  f1-score   support

    Normal     0.95      1.00      0.98        40
 Pneumonia     1.00      0.95      0.98        42

  accuracy                         0.98        82
 macro avg     0.98      0.98      0.98        82
weighted avg    0.98      0.98      0.98        82

Accuracy: 97.561
Precision: 97.677
Recall: 97.561
```

Figure 6: Random Forest Classifier Accuracy and confusion matrix

Figure 7 illustrates the predicted outputs generated by the Random Forest classifier. It could show a comparison between the actual labels and the predicted labels for a subset of the data.

Table 2: Provides an overview of the performance metrics (accuracy, precision, recall, and F1-score) for the proposed ML models, namely Random Forest and Naïve Bayes. It indicates that the Random Forest model outperforms the Naïve Bayes model in terms of accuracy.

Table 3 is a detailed breakdown of the performance metrics (precision, recall, and F1-score) for each class (Normal and Pneumonia) predicted by the Random Forest and Naïve Bayes classifiers. It shows consistency in performance metrics across classes for each model.

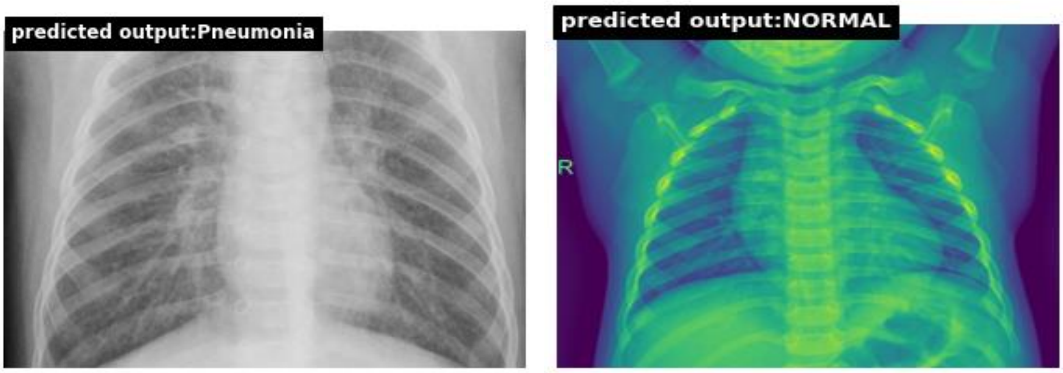


Figure 7: Predicted output using RandomForest Classifier

Table 2: Overall performance comparison of proposed ML models.

Model name	Accuracy	Precision	Recall	F1-score

Random Forest	0.97	0.97	0.97	0.97
Naïve Bayes	0.78	0.74	0.81	0..75

Table 3: Class-wise performance comparison of proposed ML models.

Model name	Random Forest		Naïve Bayes classifier	
	Normal	Pneumonia	Normal	Pneumonia
Precision	0.98	0.98	0.78	0.78
Recall	0.98	0.98	0.78	0.78
F1-score	0.98	0.98	0.78	0.78

5. CONCLUSION

In conclusion, the proposed machine learning model for automated pneumonia detection from chest X-rays offers a promising solution to the challenges faced in diagnosing this critical respiratory infection, particularly in children. By leveraging advanced image analysis techniques, this model can swiftly and accurately identify pneumonia patterns, facilitating early diagnosis and prompt initiation of treatment.

The implications of such a model are significant for the medical field and patient care. Rapid analysis of chest X-ray images enables healthcare professionals to prioritize urgent cases, leading to improved patient outcomes by reducing the time to diagnosis and treatment initiation. Moreover, the model can aid in optimizing resource allocation within healthcare facilities by streamlining the diagnostic process and potentially reducing hospital stays.

REFERENCES

1. Ren, Hao, Fengshi Jing, Zhurong Chen, Shan He, Jiandong Zhou, Le Liu, Ran Jing et al. "CheXMed: A multimodal learning algorithm for pneumonia detection in the elderly." *Information Sciences* 654 (2024): 119854.
2. Wu, Linghua, Jing Zhang, Yilin Wang, Rong Ding, Yueqin Cao, Guiqin Liu, Changsheng Liufu et al. "Pneumonia detection based on RSNA dataset and anchor-free deep learning detector." *Scientific Reports* 14, no. 1 (2024): 1-8.
3. Nalluri, Sravani, and R. Sasikala. "Pneumonia screening on chest X-rays with optimized ensemble model." *Expert Systems with Applications* 242 (2024): 122705.
4. Mann, Palvinder Singh, Shailesh D. Panchal, Satvir Singh, Guramritpal Singh Saggu, and Keshav Gupta. "A hybrid deep convolutional neural network model for improved diagnosis of pneumonia." *Neural Computing and Applications* 36, no. 4 (2024): 1791-1804.
5. Udbhav, Milind, Robin Kumar Attri, Meenu Vijarania, Swati Gupta, and Khushboo Tripathi. "Pneumonia Detection Using Chest X-Ray with the Help of Deep Learning." In *Concepts of Artificial Intelligence and its Application in Modern Healthcare Systems*, pp. 177-191. CRC Press, 2024.

6. Lowie, Thomas, J. Vandewalle, Giles Hanley-Cook, Bart Pardon, and Jade Bokma. "Circadian variations and day-to-day variability of clinical signs used for the early diagnosis of pneumonia within and between calves." *Research in Veterinary Science* 166 (2024): 105082.

7. Omar, Mariam Moneim, Ebtesam Naeim Hosseny, Eman Mohammed Handak, Faisal AL-Sarraj, and Ahmed Mahmoud El-Hejin. "Detection of (CTX-M) Resistance Genes in Extended Spectrum Beta-lactamases Bacterial Isolates of Escherichia coli and Klebsiella pneumonia and the Antibacterial Effect of Ethanolic Extract of Neem plant (Azadirachta indica) Against these Bacteria." *Egyptian Journal of Chemistry* 67, no. 3 (2024): 337-345.

8. Hao, Yong, Chengxiang Zhang, and Xiyan Li. "DBM-ViT: A multiscale features fusion algorithm for health status detection in CXR/CT lungs images." *Biomedical Signal Processing and Control* 87 (2024): 105365.