“MULTIPLE LUNG DISEASE DETECTION FROM CHEST XRAY IMAGES USING MACHINE LEARNING FOR BALANCED DATASET”

-Basani Sneha Latha Reddy

-Tanoj Kumar Anapana

ABSTRACT

According to the latest WHO data published in 2020 Lung Disease Deaths in India reached 879,732 or 10.38% of total deaths. It is important to accurately diagnose the lung diseases to take necessary actions and lower the probability of deaths due to them. Chest X-Ray (CXR) is the best choice to diagnose many lung diseases due to its low cost and easy availability. In this project, various algorithms are compared to identify the algorithms that give better results.

Models generated from six different Machine Learning algorithms are studied: Logistic Regression, Naive Bayes, k-Nearest Neighbors, Decision Tree, Random Forest and Support Vector Machine. The models are built on both unsegmented and segmented chest X-Ray images. Three classes are used in this project: Normal, Pneumonia and Other. All the three classes have equal number of images, i.e., the dataset is balanced. The performance is compared for both before and after attribute reduction.

Two dimensionality reduction methods are used: Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). The best results without attribute reduction are observed for Random Forest with a 97.17% accuracy. When the attributes were reduced using PCA, the best accuracy of 90.39% is observed using Logistic Regression. With the dimensions being reduced using LDA, an accuracy of 84.78% with a model built using Support Vector Machine is observed.

Keywords: Lung Diseases, X-Ray, Logistic Regression, Naive Bayes, k-Nearest Neighbors, Decision Tree, Random Forest and Support Vector Machine, PCA, LDA

1. INTRODUCTION

Lung disease is any problem in the lungs that prevents the lungs from working properly. These are some of the most common medical conditions in the world. Tens of millions of people have lung disease in the U.S. alone. Smoking, infections, and genes cause most lung diseases. A few types of lung diseases are:

- Affecting the airways:
  The trachea branches into tubes called bronchi and then much smaller tubes called bronchioles. Some of the diseases that affect the airways are asthma, chronic obstructive pulmonary disease, chronic bronchitis, etc.
- Affecting the alveoli:
  The bronchioles end in clusters of air sacs called alveoli. Some of these diseases include pneumonia, tuberculosis, pulmonary edema, pneumoconiosis, etc.
- Affecting the interstitium:
  The thin and delicate lining between the alveoli is called interstitium. Various diseases affect the interstitium like pneumonia, pulmonary edema and interstitial lung disease.
- Affecting the pleura:
  The thin lining that surrounds the lung is called pleura. Pleural effusion, pneumothorax and mesothelioma are some diseases that affect the pleura.

Pneumonia is an acute respiratory infection that makes it hard for the person to breathe and limits their oxygen intake. It is the largest infectious cause of death in children worldwide. Pneumonia killed 740,180 children under the age of 5 in 2019. It is observed although pneumonia affects people everywhere, the deaths are highest in Southern Asia and sub-Saharan Africa.

The goal of this project is comparing the various algorithms used and identifying the best one. For this, the Chest X-Ray (CXR) images are being used. The models are built for segmented as well as unsegmented images.

1.1 . Problem Definition

It is very important to correctly diagnose any lung diseases since they are major symptoms for many other deadly diseases. If the detection of the disease is done using a Machine

Learning model, it could reduce the burden on doctors and also provide accurate results. It

is important to know which Machine Learning algorithm is giving better results. The project “Multiple Lung Disease Detection from Chest X-Ray Images using Machine Learning for Balanced Dataset” uses various Machine Learning algorithm and identifies which gives better results.

1.2 . Existing System

There are many projects which studied the accuracy of various algorithms when applied on a chest X-Ray dataset. But all these projects used imbalanced dataset and very few number of images for each class. Due to the dataset being imbalanced, the accuracy calculated is biased towards the class with maximum images.

1.3 . Proposed System

In this project, a dataset with 9000 images is used and these images are used to build various

Machine Learning models using algorithms like Logistic Regression, Naïve Bayes, kNearest Neighbors, Decision Tree, Random Forest and Support Vector Machine. Three different classes are used, i.e., Normal, Pneumonia and Other. The models are built for segmented and unsegmented images. The models are built for before and after attribute reduction. Two dimensionality reduction techniques are used: PCA and LDA.

1.4 . Requirements Specification

1.4.1 Software Requirements

The software requirements for the successful working of the project are:

Operating System: Windows 10 or 11

Technologies Used: Python, OpenCV

Tools: Jupyter Notebook, TensorFlow, Scikit-learn

1.4.2 Hardware Requirements

The hardware requirements for the successful working of the project are:

RAM: 8GB or more

Processors: Intel/AMD Ryzen processor, NVIDIA GTX/RTX GPU

Input Devices: Keyboard, Mouse

3. DESIGN METHODOLOGY OF LUNG DISEASE DETECTION

3.1. Block Diagram of Lung Disease Detection

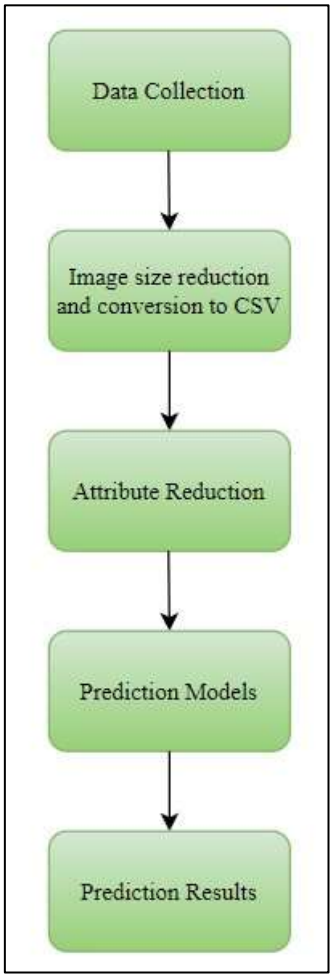

Figure 3.1: Block Diagram for Lung Disease Detection

Firstly, the image dataset is read from the storage. Each class has 3000 images and there are 3 different classes. Since the images could be of different sizes, they are resized to 25x25. Then, these images are converted to CSV file format, i.e., the pixel values are recorded. There are 625 attributes for each image. Then, the labels for each row are added, i.e., they are Normal, Pneumonia or Other.

In the next step, the records are divided into training and testing datasets. The attributes are reduced using PCA and LDA. Next, the training dataset is used to train the model. After training the model, the model is evaluated using test dataset. After the evaluation of the model, the performance metrics (Accuracy, Precision, Recall and F1-Score) are calculated, which indicate how well the model can predict.

The ordinary X-Ray images are segmented using a CNN model. Then the segmented images are resized and converted to CSV format with 625 attributes. The same process is followed for the segmented data as was for unsegmented data.

3.2. Dataset

The dataset used in this project is a combination of various open-source datasets from Kaggle. The normal lung images are from the COVID-19 Radiography Database. Pneumonia images are from the Pneumonia X-Ray Images dataset. And the other lung disease images are from the NIH Chest X-rays database. Total 9000 images are used in this project of which 7200 are used for training and 1800 are used for testing.

All the 9000 ordinary images are segmented using a model built on CNN using U-Net Architecture.

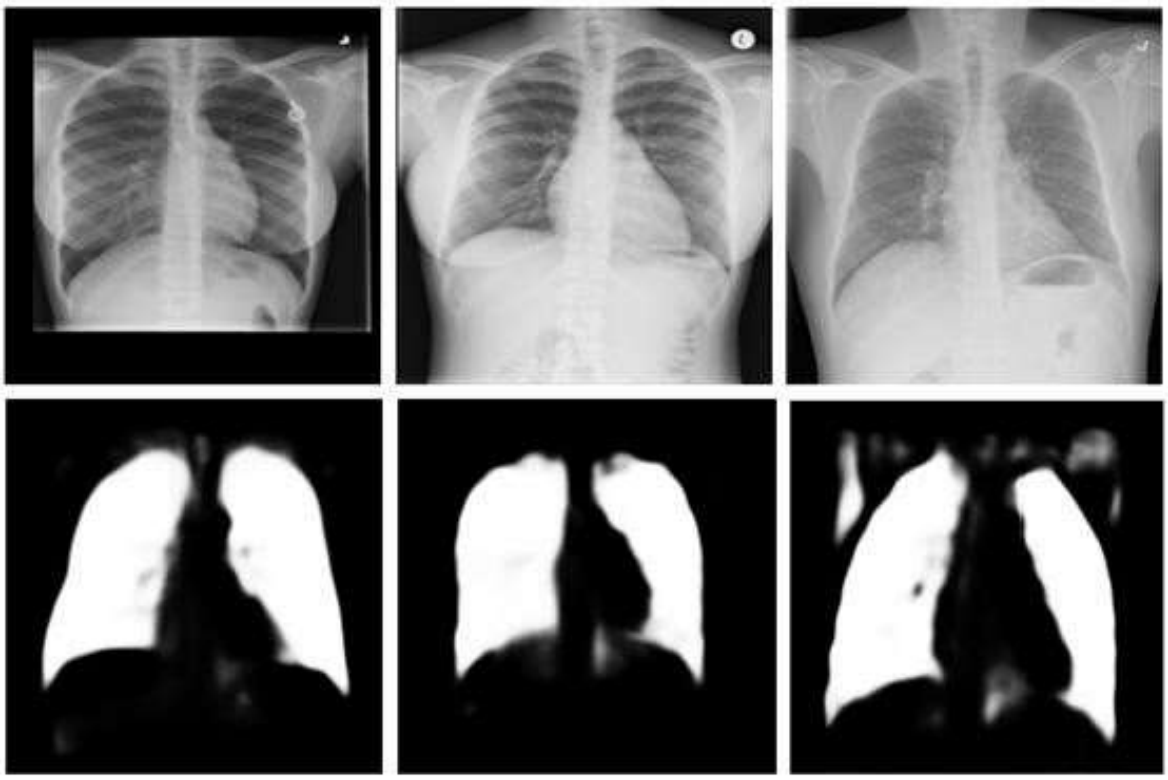

Figure 3.2: Normal – Unsegmented and Segmented Images

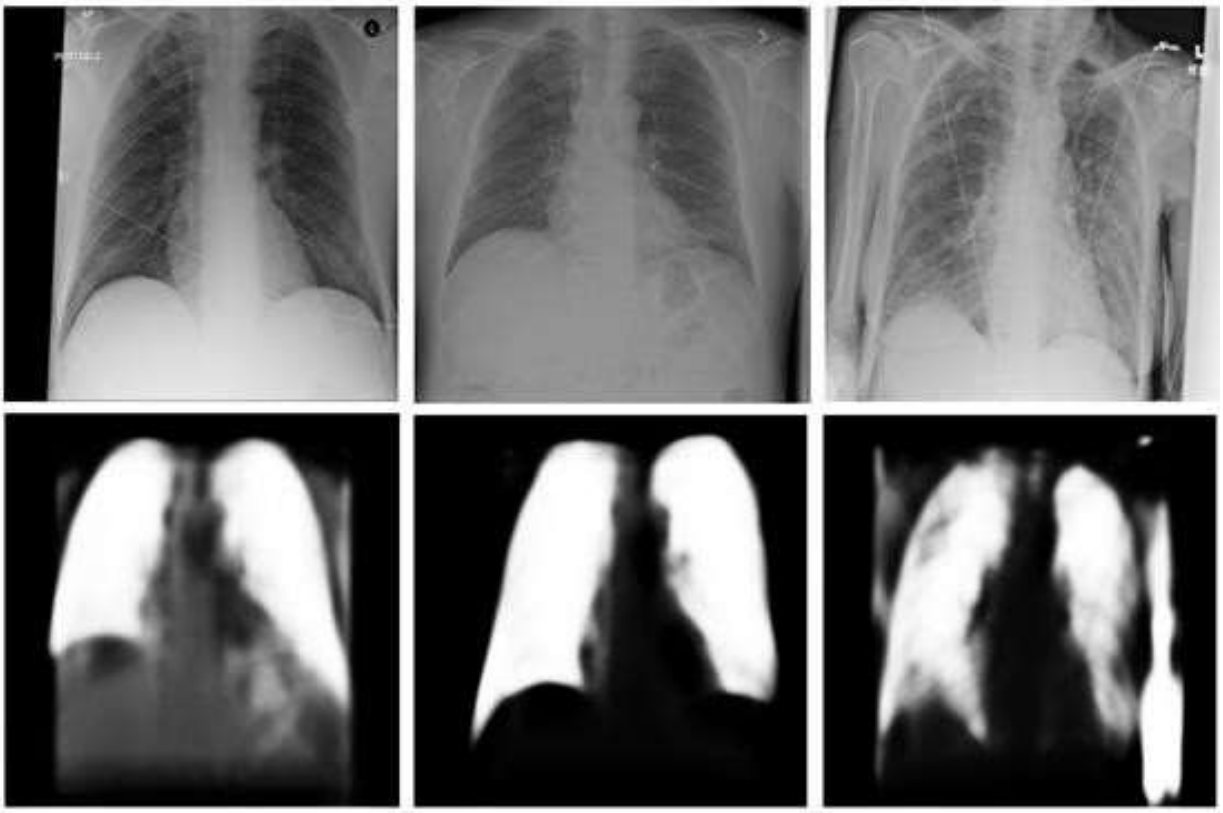

Figure 3.3: Other – Unsegmented and Segmented Images

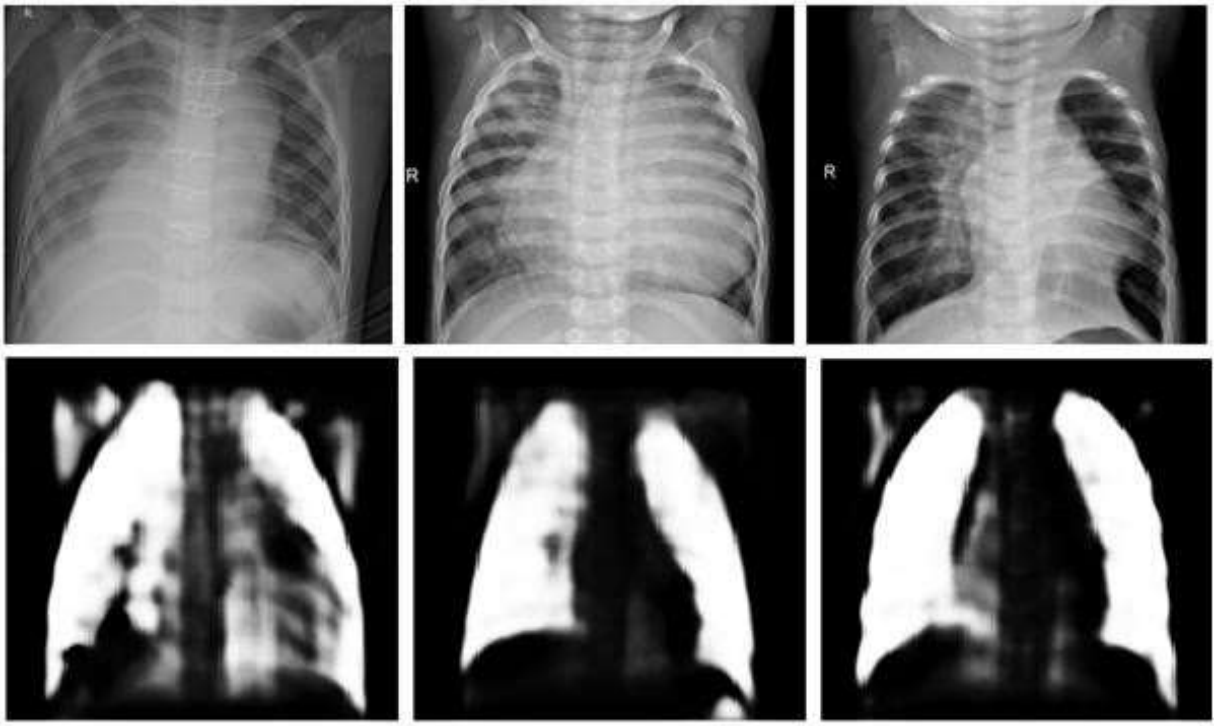

Figure 3.4: Pneumonia – Unsegmented and Segmented Images

3.4. Dimensionality Reduction

3.4.1. Principal Component Analysis

Principal Component Analysis (PCA) is a technique for reducing the dimensionality of a dataset. It is used when there are a large number of features. In this project, there are 625 attributes for each record. The number of features are reduced using PCA. The n_components parameter is given 0.90, i.e., 90% of the information in the old features is preserved. The number of new attributes extracted from the old attributes are 36 for unsegmented images and 120 for segmented images.

3.4.2. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a method in statistics that is used to find a linear combination of features such that they separate the different classes. LDA takes into account the labels of the classes while separating the data. It maximizes the distance between means of different classes and minimizes the variance within the individual classes. LDA can be used as a linear classifier or for dimensionality reduction. Here, it is used for dimensionality reduction. The number of components parameter of the function should be less than or equal to the minimum of the number of features and one less than number of classes. Here, n_components is selected as 2.

3.5. Training

The various libraries imported are Scikit-learn, OpenCV, NumPy, Matplotlib and TensorFlow.

The total number of samples in the data are 9000 with each class having 3000 samples. After the images are converted to pixel values, the dataset is split into 80% training and 20% test samples. Therefore 7200 samples are used for training and 1800 for testing. Then the data is standardized.

The results are observed for reduced as well as unreduced attributes. When the dimensionality of pre-processed data was reduced using PCA, it gave a total of 36 components for unsegmented data and 120 for segmented data. Using LDA on the pre-processed data for dimensionality reduction, 2 components were extracted.

The model is trained and built using various Machine Learning algorithms. The different algorithms used are Logistic Regression, Naïve Bayes, k-Nearest Neighbors, Decision Tree, Random Forest and Support Vector Machine.

3.6. Machine Learning Techniques

3.6.1. Logistic Regression

When Logistic Regression is applied on unreduced attributes, the accuracy observed for unsegmented images is 78.17% and for segmented images it is 89.39%. When PCA is used for dimensionality reduction, the accuracy is 78.17% for unsegmented images and 90.39% for segmented images. Using LDA, an accuracy of 78.33% and 84.61% is observed for unsegmented and segmented images, respectively.

3.6.2. Naïve Bayes

When Naïve Bayes is applied on unreduced attributes, the accuracy observed for unsegmented images is 65.33% and for segmented images it is 67.89%. When PCA is used for dimensionality reduction, the accuracy is 72.33% for unsegmented images and 53.22% for segmented images. Using LDA, an accuracy of 78.61% and 84.56% is observed for unsegmented and segmented images, respectively.

3.6.3. k-Nearest Neighbors

The number of neighbors, i.e., the value of k is given as 3, since 3 gave better results than 5. The accuracy observed for unsegmented images is 81.17% and for segmented images it is 79.56% when attributes are not reduced. Using PCA, an accuracy of 80.83% and 79.5% is observed for unsegmented and segmented images, respectively. When LDA is used for dimensionality reduction, the accuracy is 76.11% for unsegmented images and 81.72% for segmented images.

3.6.4. Decision Tree

The model built has an average of 30 depth for the trees. The accuracy observed for unsegmented images is 71% and for segmented images it is 91.67% when attributes 625. Using PCA, an accuracy of 71.78% and 69.22% is observed for unsegmented and segmented images, respectively. When LDA is used for dimensionality reduction, the accuracy is 73.17% for unsegmented images and 79.17% for segmented images.

3.6.5. Random Forest

When Random Forest is applied on unreduced attributes, the accuracy observed for unsegmented images is 83.22% and for segmented images it is 97.17%. When PCA is used for dimensionality reduction, the accuracy is 80.5% for unsegmented images and 80.78% for segmented images. Using LDA, an accuracy of 76.83% and 82.67% is observed for unsegmented and segmented images, respectively.

3.6.6. Support Vector Machine

The accuracy observed for unsegmented images is 83.78% and for segmented images it is 89.06% when attributes 625. Using PCA, an accuracy of 83.72% and 88.33% is observed for unsegmented and segmented images, respectively. When LDA is used for dimensionality reduction, the accuracy is 78.55% for unsegmented images and 84.78% for segmented images.

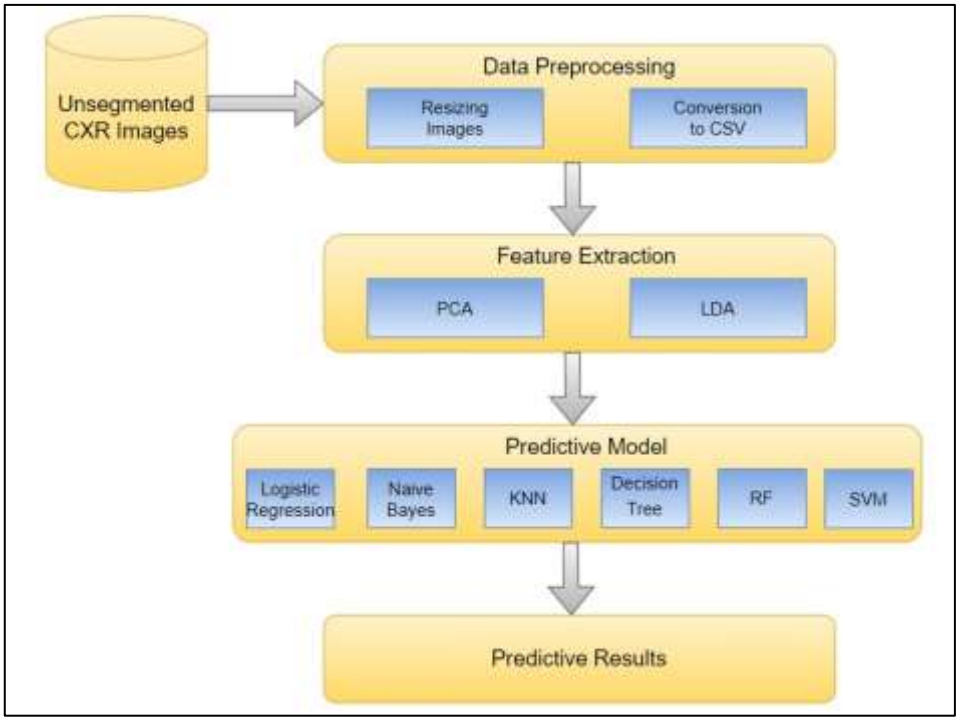

Figure 3.5: System Architecture – Unsegmented Images

First, the stored lung Chest X-Ray images are pre-processed, i.e., they are resized and converted to CSV format. Then, they are split into training and testing data. Then the features are extracted using PCA and LDA. Later, predictive models are built using different Machine Learning algorithms. The samples in test data are used to evaluate the models and their performance measures are compared.

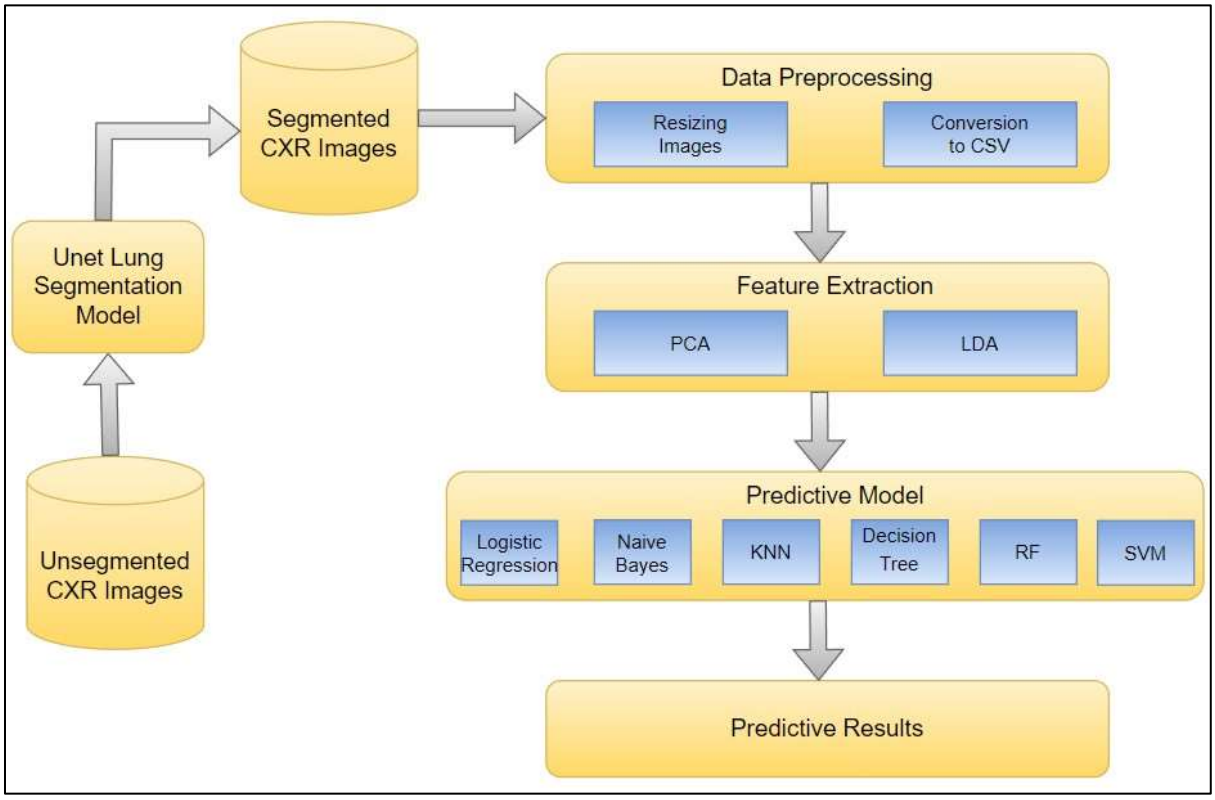

Figure 3.6: System Architecture – Segmented Images

First, the stored lung Chest X-Ray images are segmented so that lung masks can be obtained. These segmented images are then pre-processed, i.e., they are resized and converted to CSV format. Then, they are split into training and testing data. Then the features are extracted using PCA and LDA. Later, predictive models are built using different Machine Learning algorithms. The samples in test data are used to evaluate the models and their performance measures are compared.

4. IMPLEMENTATION AND RESULTS

4.1. Implementation

The implementation of Detection of Lung Diseases involves the execution of the code in Jupyter notebook.

First, the dataset is uploaded. Then, pre-processing of the images takes place in which they are resized and converted to CSV format. Then, dimensionality reduction techniques are applied. Later, the training of the model takes place followed by testing and recording the performance metrics of each algorithm.

4.2 Results

Algorithm	Attribute Selector	Original Attributes	Attributes After Reduction	Performance Metrics Before Attribute Reduction	Performance Metrics After Attribute Reduction
Logistic Regression	PCA	625	36	Accuracy: 78.17% Precision: 0.7814 Recall: 0.7817 F1 Score: 0.7805	Accuracy: 78.17% Precision: 0.7809 Recall: 0.7817 F1 Score: 0.7776
	LDA	625	2		Accuracy: 78.33% Precision: 0.7822 Recall: 0.7833 F1 Score: 0.7822
Naïve Bayes	PCA	625	36	Accuracy: 65.33% Precision: 0.6298 Recall: 0.6533 F1 Score: 0.6197	Accuracy: 72.33% Precision: 0.7188 Recall: 0.7233 F1 Score: 0.7195
	LDA	625	2		**Accuracy: 78.61% Precision: 0.7866 Recall: 0.7861 F1 Score: 0.7839**
K Nearest Neighbors	PCA	625	36	Accuracy: 81.17% Precision: 0.8095 Recall: 0.8117 F1 Score: 0.8097	Accuracy: 80.83% Precision: 0.8060 Recall: 0.8083 F1 Score: 0.8067

	LDA	625	2		Accuracy: 76.11% Precision: 0.7595 Recall: 0.7611 F1 Score: 0.7600
Decision Tree	PCA	625	36	Accuracy: 71.00% Precision: 0.7104 Recall: 0.71 F1 Score: 0.7099	Accuracy: 71.78% Precision: 0.7171 Recall: 0.7178 F1 Score: 0.7174
	LDA	625	2		Accuracy: 73.17% Precision: 0.7303 Recall: 0.7317 F1 Score: 0.7310
Random Forest	PCA	625	36	Accuracy: 83.22% Precision: 0.8317 Recall: 0.8322 F1 Score: 0.8297	Accuracy: 80.5% Precision: 0.8073 Recall: 0.805 F1 Score: 0.8009
	LDA	625	2		Accuracy: 76.83% Precision: 0.7675 Recall: 0.7683 F1 Score: 0.7675
Support Vector Machine	PCA	625	36	**Accuracy: 83.78% Precision: 0.8391 Recall: 0.8378 F1 Score: 0.8369**	**Accuracy: 83.72% Precision: 0.8388 Recall: 0.8372 F1 Score: 0.8357**
	LDA	625	2		Accuracy: 78.55% Precision: 0.7861 Recall: 0.7855 F1 Score: 0.7838

Table 4.1: Results – Unsegmented Lung Images

The best results for unsegmented images is given by the Support Vector Machine with almost 83% accuracy.

Algorithm	Attribute Selector	Original Attributes	Attributes After Reduction	Performance Metrics Before Attribute Reduction	Performance Metrics After Attribute Reduction
Logistic Regression	PCA	625	120	Accuracy: 89.39% Precision: 0.8942 Recall: 0.8939 F1 Score: 0.8940	**Accuracy: 90.39% Precision: 0.9045 Recall: 0.9039 F1 Score: 0.9041**

	LDA	625	2		Accuracy: 84.61% Precision: 0.8467 Recall: 0.8461 F1 Score: 0.8464
Naïve Bayes	PCA	625	120	Accuracy: 67.89% Precision: 0.7125 Recall: 0.6789 F1 Score: 0.6785	Accuracy: 53.22% Precision: 0.6045 Recall: 0.5322 F1 Score: 0.5042
	LDA	625	2		Accuracy: 84.56% Precision: 0.8461 Recall: 0.8456 F1 Score: 0.8456
K Nearest Neighbors	PCA	625	120	Accuracy: 79.56% Precision: 0.7983 Recall: 0.7956 F1 Score: 0.7963	Accuracy: 79.5% Precision: 0.7975 Recall: 0.795 F1 Score: 0.7959
	LDA	625	2		Accuracy: 81.72% Precision: 0.8194 Recall: 0.8172 F1 Score: 0.8181
Decision Tree	PCA	625	120	Accuracy: 91.67% Precision: 0.9167 Recall: 0.9167 F1 Score: 0.9167	Accuracy: 69.22% Precision: 0.6952 Recall: 0.6922 F1 Score: 0.6935
	LDA	625	2		Accuracy: 79.17% Precision: 0.7938 Recall: 0.7917 F1 Score: 0.7924
Random Forest	PCA	625	120	**Accuracy: 97.17%** **Precision: 0.9717** **Recall: 0.9717** **F1 Score: 0.9716**	Accuracy: 80.78% Precision: 0.8066 Recall: 0.8078 F1 Score: 0.8060
	LDA	625	2		Accuracy: 82.67% Precision: 0.8283 Recall: 0.8267 F1 Score: 0.8273
Support Vector Machine	PCA	625	120	Accuracy: 89.06% Precision: 0.8902 Recall: 0.8906 F1 Score: 0.8904	Accuracy: 88.33% Precision: 0.8830 Recall: 0.8833 F1 Score: 0.8832

	LDA	625	2		**Accuracy: 84.78% Precision: 0.8490 Recall: 0.8478 F1 Score: 0.8483**

Table 4.2: Results – Segmented Lung Images

The best results for segmented images is given by the Random Forest with almost 97% accuracy.

5. CONCLUSION AND FUTURE SCOPE

5.1. Conclusion

The best results without attribute reduction are observed for Random Forest with a 97.17% accuracy on segmented images and 83.78% on unsegmented images for Support Vector Machine. When the attributes were reduced using PCA, the best accuracy of 90.39% is observed using Logistic Regression on segmented images and 83.72% using Support Vector Machine on unsegmented images. With the dimensions being reduced using LDA, an accuracy of 84.78% with a model built using Support Vector Machine is observed on segmented images and 78.61% using Naïve Bayes on unsegmented images. It can be observed that segmented images give better results than unsegmented images.

5.2. Future Scope

In the future, many other lung diseases can be considered for detection. The number of images used could also be increased. Many other Machine Learning algorithms can also be used and compared to know which gives best results.