Bachelor of Science in Computer Science & Engineering

Developing a Machine Learning Based Assistance System for Cough Type Disease Diagnosis and Treatment

by

Upol Chowdhury

ID: 1604117

Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)

Chattogram-4349, Bangladesh.

August, 2022

Developing a Machine Learning Based Assistance System for Cough Type Disease Diagnosis and Treatment

Submitted in partial fulfilment of the requirements for
Degree of Bachelor of Science
in Computer Science & Engineering

by

Upol Chowdhury
ID: 1604117

Supervised by

Dr. Mahfuzuhoq Chowdhury
Associate Professor
Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)
Chattogram-4349, Bangladesh.

The thesis titled '**Developing a Machine Learning Based Assistance System for Cough Type Disease Diagnosis and Treatment**' submitted by ID: 1604117, Session 2019-2020 has been accepted as satisfactory in fulfilment of the requirement for the degree of Bachelor of Science in Computer Science & Engineering to be awarded by the Chittagong University of Engineering & Technology (CUET).

Board of Examiners

Chairman

Dr. Mahfuzuhoq Chowdhury
Associate Professor
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Member (Ex-Officio)

Prof. Dr. Md. Mokammel Haque
Professor & Head
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Member (External)

Prof. Dr. Kaushik Deb
Professor
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Declaration of Originality

This is to certify that I am the sole author of this thesis and that neither any part of this thesis nor the whole of the thesis has been submitted for a degree to any other institution.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. I am also aware that if any infringement of anyone's copyright is found, whether intentional or otherwise, I may be subject to legal and disciplinary action determined by Dept. of CSE, CUET.

I hereby assign every rights in the copyright of this thesis work to Dept. of CSE, CUET, who shall be the owner of the copyright of this work and any reproduction or use in any form or by any means whatsoever is prohibited without the consent of Dept. of CSE, CUET.

Signature of the candidate

Date:

Acknowledgements

I would like to acknowledge and show tremendous appreciation to all individuals without whom, it would be impossible to finish this study. First and foremost, I offer my heartfelt gratitude to my supervisor Dr. Mahfuzulhoq Chowdhury, Associate Professor, Department of Computer Science and Engineering. I am indebted to him for his motivation, proper guidance, constructive criticism and immense support towards the progress of this study. All these things have helped me to grow as a machine learning practitioner.

I also want to express my gratitude to every person who made this study possible, including Dr. M. A. Sattar, Professor & Head (Department of Medicine), Chittagong Medical College, Hospital; Dr. Abhijit das, Dr. Shakil Iftekhar, Dr. Salman abdullah, Dr. Al Muhaimen Azam, Dr. Farhin Etu, Dr. Diptanil Das, Dr. Joy Deb, Dr. Aninda Kushal Das, Dr. Nahida Sultana and many others who helped create the dataset by gathering patient symptoms and providing expert support.

Last but not the least, I want to show my special gratitude to late Dr. Sandipan Das, Cardiologist & Junior consultant (Department of Cardiology), Chittagong Medical College Hospital for his support in the initial phase of this research study. This study is dedicated to him.

Abstract

The effectiveness of machine learning to diagnose patients' illnesses has been the subject of extensive research over the past ten years. Machine learning has been found to considerably lower the probability of inaccurate diagnoses when incorporated into modern diagnostic procedures. This study proposes a method for diagnosing the three most similarly symptomized cough-type chronic diseases: COPD, Bronchial Asthma and Pneumonia, with the goal of having an influence on the healthcare system. The symptoms of cough-type chronic disease patients, admitted to the hospital were collected from eight medical colleges spread out over Bangladesh in order to construct the classifying model.

The study also proposes a set of 28 attributes for appropriately classifying cough-type chronic diseases. The obtained data is examined, and the main research findings are explained. Several machine learning methods are tested using the dataset. The research has found and suggests Gradient Tree Boosting to be the most effective, with a classification accuracy of 91%, despite the fact that previous studies have identified Support Vector Machine and Random Forest to be the most efficient models in these kind of classification tasks.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence. 8, 38

API Application Programming Interface. 27

APK Android Application Package. 27

CNN Convolutional Neural Network. 10

COPD Chronic Obstructive Pulmonary Disease. 1, 2, 4, 6, 8, 9, 18, 19, 40

CSS Cascading Style Sheet. 14

DL Deep Learning. 10

DNN Deep Neural Network. 10, 40

ED Emergency Department. 9

HTML HyperText Markup Language. 14, 27

KNN K-Nearest Neighbour. 8, 10, 18, 40

ML Machine Learning. 8, 10, 40

RL Reinforcement Learning. 50

SVM Support Vector Machine. 8, 10, 11, 18, 40

WSGI Web Server Gateway Interface. 15

Chapter 1

Introduction

1.1 Introduction

The fact that cough may have such a dramatic and negative impact on a patient's quality of life makes it one of the most prevalent symptoms for which people visit primary care physicians. Three categories of coughing exist: acute (lasting less than three weeks), subacute (lasting between three and eight weeks), and chronic (lasting more than eight weeks) [1]. Occupational chronic cough may be viewed as one of the most preventable forms of the disease among the wide range of subtypes of chronic cough, either defined by their clinical or pathogenetic reasons. The next most common cause of chronic cough is obstructive airway illnesses, like chronic obstructive pulmonary disease or asthma [2, 3]. Pneumonia is a disease mostly causes because of acute cough but it can lead to COPD in elder citizens, it also highly related to bronchial asthma [4]. So name of pneumonia is taken side by side with chronic diseases related to cough as they have strong interconnection [3, 4].

More than 95 percent of the 5.6 million children under the age of 5 who die each year around the world die of pneumonia, which is the leading cause of death among children. Pneumonia is responsible for 15,000 of the 119,000 deaths of children under the age of five that occurred in Bangladesh in 2015. Nevertheless, more than 88 percent of pneumonia deaths in Bangladesh occur in people aged 70 and older. Actually, pneumonia is the fifth most common reason for death for seniors [5]. 3.23 million people died from chronic obstructive pulmonary disease in 2019, making it the third most common killer worldwide. Nearly 90 percent of COPD fatalities in people under the age of 70 take place in low- and middle-income nations [6]. Asthma-related mortality in Bangladesh reached 8,893 in

2020, or 1.24 percent of all fatalities, according to the most recent WHO data. Bangladesh is ranked 74 in the world due to its 7.70 per 100,000 population age-adjusted Death Rate [7].

In terms of the creation of models, machine learning is a specific approach to data analysis that automates model generation. It is important to stress that in machine learning, we are not instructing the machines where to search. Instead, the machines learn to use certain techniques to uncover hidden insights from data. Machine learning is an iterative process that enables the computer to modify its approaches and results as it is presented with fresh scenarios and data [8]. The evolution of the healthcare system over the past few years has been one of the many uses of machine learning technology that has shown promising results [9, 10]. A hidden pattern in the complicated medical data that is having an effect on clinical diagnosis systems can be found by machine learning.

In this case, we want to take use of advancements in machine learning and widespread smartphone use. Our goal is to create an application powered by a classification and prediction based machine learning solution to identify chronic diseases like (COPD, Asthma, Pneumonia) from data gathered from people labeled by doctors. Through which we will provide initial treatment suggestions and primary prediction about the respiratory situation to the user. Based on an assessment of the sensitivity, specificity and accuracy of the created classification model, we will provide the user with information.

Patients fill out a questionnaire that asks about their symptoms. According on the symptoms, doctors have assigned the corresponding respiratory illnesses. These preprocessed, marked data were used to train several classification models. We have developed a special model via which we are able to predict cough-based respiratory disease that exhibits chronic behavior after evaluating the accuracy, specificity, and sensitivity of all utilized classification methodologies. We obtained good accuracy with our classification model when compared to the specified label. Then, based on the user symptoms that we collected, we designed an application where the forecast findings are displayed. Through our app, we will also offer doctor verification of our forecast and provide contact information for doctors.

1.2 Work Overview

Data is needed to feed the machine learning model in machine learning systems. Traditional machine learning models are sensitive to the choice of features and the preprocessing of the data. Therefore, choosing features before collecting data is a crucial step. Patients' symptoms are noted in accordance with the feature. For the purpose of computation, the acquired string data is mapped to its corresponding numerical data. To understand the structure of the numerical data, preprocessing and visualization are used.

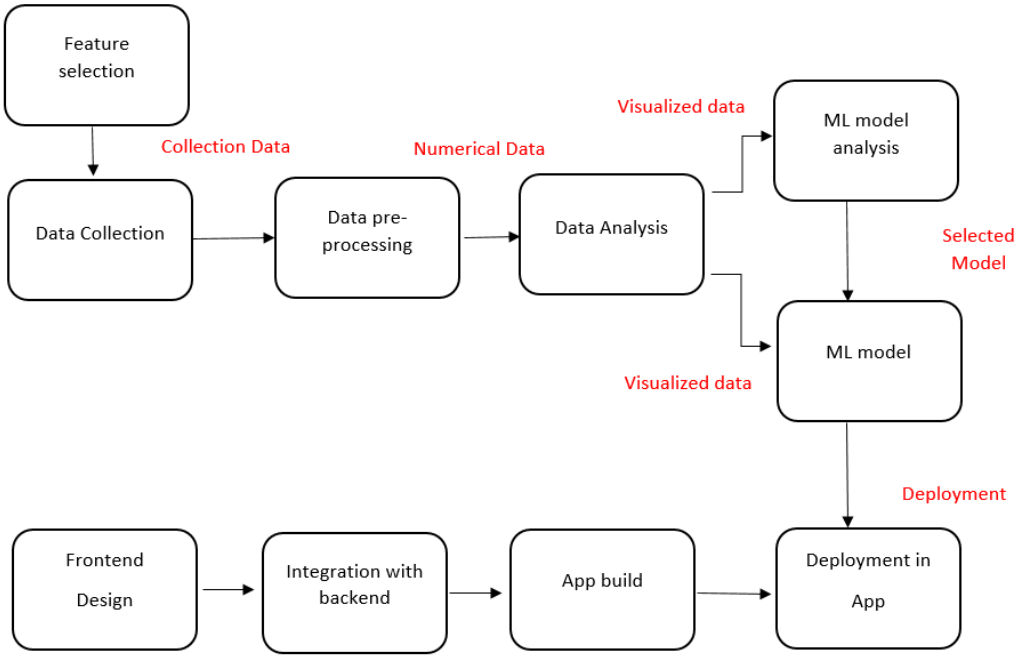


Figure 1.1: System Overview

The information is then fed into a variety of machine learning models for analysis and selection of the model with the highest accuracy for the system's improved performance. Frontend and backend web technologies are used to create an app. With a focus on excellent usability, the chosen machine learning model is introduced to the application. From figure 1.1 we can see the overview of it.

1.3 Challenges

The process of detecting chronic disease cases based on coughing has encountered a number of challenges.

First, it takes careful planning and feature selection to distinguish between the three diseases with the closest clusters of symptoms. Accurate learning outcomes collection is a crucial phase of mission definition [11]. The gold-standard labels used for supervised prediction activities are frequently generated from outcomes.

Secondly, all of them have strong relation with Covid-19. Those with severe COVID-19 results had a higher rate of respiratory illnesses than patients with non-severe outcomes, according to findings. Without prior Covid-19 history, it is exceedingly challenging to identify patients with cough-based chronic disorders (COPD, Asthma, Pneumonia). Since they all have similar symptoms [12, 13]. Patients with strong clinical symptoms of the aforementioned disorders but who are unrelated to Covid-19 make difficult participants for data collection.

Thirdly, the outcomes are based on the data that the machine learning model has encountered itself. However, information gathered from hospitalized patients is necessary for delicate concerns like diagnosis. Once more, data must be gathered in accordance with the selected feature [14].

Finally, incorrectly classified data can significantly harm a machine learning model's performance. Models aim to establish a connection between the features chosen and the final output. Predictions may become useless due to information leaks.

1.4 Applications

Machine learning technology is widely used in the healthcare system in a number of disciplines due to its ability to identify relationships in statistical data that is not structured. Future developments are probably going to have a bigger effect on the clinical system using this technology. As more data is produced and collected, the technological impediment that is currently present in the system

may be significantly diminished. The following method can be used to quantify how the work was applied:

- A diagnostic tool that provides data or suggests diagnosis to assist doctors in the moment.
- In locations without diagnostic facilities, rapid diagnostic results can be attained.
- People from different areas can get primary information about their diseases and also can connect with respective doctors and elementary treatment information is followed by.
- The treatment procedure is a significant factor of the healthcare system. The suggested system can be used to construct the treatment procedure.

1.5 Motivation

Modern clinical systems have been significantly impacted by machine learning and artificial intelligence. Healthcare systems are becoming more and more dependent on machine learning technology [15]. Experts have identified three categories into which AI applications in the healthcare sector can be categorized: patient-oriented, clinician-oriented, and operational-oriented. The following list of reasons motivates our work :

1. Cases that have gone undiagnosed or have been misdiagnosed in rural areas
2. Low clinical care productivity and quality
3. Improper or delayed medical care
4. Expensive and inefficient solution

1.6 Contribution of the thesis

The work has been completed to accomplish a particular set of objectives. The major goal of this effort is to use the capabilities of machine learning and mobile application technologies to improve and have an impact on the healthcare system.

Additionally, make a contribution to the machine learning community for the evolution of the medical system. Following is an overview of the contributions:

1. Created a dataset of respiratory patient symptoms gathered from patients hospitalized across Bangladesh.
2. Setting out a list of features (28 Features) to distinguish three diseases with similar symptoms (COPD, Asthma, Pneumonia).
3. Analysis of data gathered from hospitals.
4. Comparing the performance of various machine learning models with the data that has been gathered and suggesting the best model for the task.
5. Model's deployment via a mobile application.
6. Offer primary health care via mobile app.

1.7 Thesis Organization

This is how the rest of the report is organized:

The researchers' exploration on using machine learning for disease diagnosis in the healthcare industry is covered in the next chapter. There is a list of the experimental methods, data gathering procedure, faults, and algorithms utilized in the literature for the findings.

The strategy employed to achieve the work's objective is thoroughly detailed and discussed in chapter 3. The collecting of fundamental attributes is where the approach starts.

We explained how our strategies produced the outcomes in chapter 4. At the start of the chapter, the dataset is described. The outcomes of data collecting and the key findings drawn from the data are also presented.

In chapter 5, we came to a conclusion and gave a brief overview of the results of our work and possible future works.

1.8 Conclusion

Detailed information about the work has been offered in this chapter. In the chapter, the reader is introduced to connected ideas. Additionally, a quick discussion is made about the difficulty faced and the use of this effort. In the section dedicated to motivation, the significance of the work is discussed, along with its contribution. The previous research in this field of study is examined in the following chapter.

Chapter 2

Literature Review

2.1 Introduction

An overview of the work and its significance are given in the previous chapter. Discussion also includes the work's contribution and applications. In order to increase the effectiveness of the current system, a lot of research has been done during the last ten years on the application of ML and AI to the healthcare industry, particularly for patient disease diagnosis. This chapter evaluates the researchers' relevant research and provides an overview of the structure. Furthermore, the main conclusions of the studies and the techniques used to draw the conclusions are also highlighted. To comprehend the situation now, the results of the academic articles are examined.

2.2 Related Literature Review

A number of diseases threaten human well-being by adversely affecting life expectancy and general prosperity. Chronic diseases, such as chronic obstructive pulmonary disease (COPD), lung cancer, Pneumonia, and Asthma, are among them and are regarded as significant causes of death in both developed and developing countries. Experts agree that the earlier a disease is identified and treated, the better the patient's prospects of recovery. In this literature, Vora [16] has tried to classify the severity of COPD, as they used the SVM and KNN methods. Utilizing a test dataset from Prajna Health Care Medical Hospital, these approaches were evaluated. It has been documented in the literature that researchers were able to attain a test dataset accuracy of 96.97% for SVM and 92.30% for KNN. However, the work has a serious flaw. Patients' symptoms were

only collected from one hospital, hence the dataset only includes COPD symptoms from a narrow area, which resulted in a lack of diversity. And also they have used pathological test reports as test data. No study has been reported regarding clinical features of the patients as those are not valued as test data.

The clinical decision support systems used in healthcare are examined in this research by Dimitri [17], specifically in relation to the prevention, diagnosis, and treatment of respiratory disorders as asthma and chronic obstructive pulmonary disease. The observational pulmonology analysis with sample size=133 makes an effort to identify the key elements that contribute to the diagnosis of these illnesses. The results of machine learning demonstrate that in the case of chronic obstructive pulmonary disease, Random Forest classifier beats other methods with 97.7% precision, where smoking, forced expiratory, age, and forced vital capacity being the most important characteristics for diagnosis. With the Random Forest classifier yet again, the best precision for asthma is 80.3%. Because of the inadequate sample size, there was a dearth of diversity. Additionally, because samples came from particular clinic, a variety of features could not be discovered. There was a major risk of overfitting in the case of COPD because the depth of the tree was so high, and they haven't mentioned how they dealt with this.

For the purpose of predicting two clinical outcomes (critical care and hospitalization) in ED patients with asthma or COPD exacerbation, Tadahiro Goto [18] compared the effectiveness of various machine learning algorithms. 3202 patients have actively participated in this work. A prediction of who would require critical care or hospitalization was made. Gradient Tree boosting outperformed all four machine learning models with an accuracy rate of 80% for critical care outcomes, and Random Forest outperformed them all with an accuracy rate of 83% for hospitalization outcomes. They didn't measure a lot of clinical characteristics. Additionally, there is no validation performed before to training when labeling data instances.

The study [19] made use of a dataset of 4500 individuals (1500 with bronchitis and 3000 with pneumonia) that included details on the demographics, symptoms, and outcomes of laboratory tests. Manual feature selection was done, concentrating

on clinical signs that are simple to measure in public places. Logistic regression, Decision trees, and SVM were put to the test and contrasted. Through a prolonged process of training, validating, and testing, models were created. They concentrated on six patient symptoms that were clinically significant and simple to interpret as the best predictors of pneumonia. The decision tree they used as their final model had an accuracy rate of 83 percent.

To save patients' lives and make doctors' jobs easier, the early and correct diagnosis of lung disorders has become essential. Using DL and ML approaches,DNN, and KNN methodologies, this research by Yahyaoui [20] focuses on the prediction of specific lung diseases such as Pneumonia and Asthma. Utilizing a confidential data set from the pulmonary diseases division of the Diyarbakir hospital in Istanbul, these methods are assessed. Each of the 212 samples is defined by 36 input characteristics. With a detection accuracy of 90% for the KNN approach and 84.35% for the DNN method, the findings obtained demonstrated the potency of these methods to identify pulmonary illnesses.

Lung X-ray pictures that can be utilized to diagnose pneumonia were employed in this investigation done by Ergen [21]. In order to complete this specific assignment, the was used as a feature extractor. AlexNet, VGG-16, and VGG-19 were three of the available CNN models that were used. Then, for each deep model, the lowest duplication maximum relevancy approach was used to minimize the number of deep features from 1000 to 100. As a result, we were able to extract 100 deep features from each deep model, and we combined these features to create a useful feature set that had 300 deep features in total. The decision tree, k-nearest neighbors, linear discriminant analysis, linear regression, and support vector machine learning models were used in this step of the experiment to process this feature set. Finally, while all models produced positive results, linear discriminant analysis produced the best outcomes with a 99.41% accuracy rate.

There hasn't been a thorough analysis of the possible use of data on patient disease trajectory for early exacerbation prediction in adult asthma patients. This study of Joseph Finkelstein [22] set out to determine whether telemonitoring data might be used to create machine learning algorithms that may foretell asthma

exacerbations before they happened. The study's data set included 7001 records from daily self-monitoring reports provided by adult asthma patients during home telemonitoring. Predictive modeling involved the creation of stratified training datasets, the choice of predictive features, and the evaluation of the classifiers that were produced. An asthma exacerbation could be predicted using a Naive Bayesian classifier, an Adaptive Bayesian network, and SVM with sensitivity of 0.80, 1.00 and 0.84 and specificity of 0.77, 1.00 and 0.80 and accuracy of 0.77, 1.00 and 0.80. Scope of improvement is left as feature selection process was not optimized enough.

2.3 Machine Learning Algorithms

2.3.1 Support Vector Machine

Support vector machines are supervised machine learning methods that can be applied to both regression and classification applications [23]. They mostly deal with classification issues. Each instance of obtained data is plotted by a support vector method as a point on an n-dimensional space or graph, where n is the total number of features in the data. A particular coordinate on the graph represents the value of each data point. To divide the data instances, the SVM creates an n-1 dimensional hyperplane. The kernel function, a method used in the algorithm, changes data so that an ideal boundary can be shown on it to divide the instances.

2.3.2 Decision Tree

One of the most widely used supervised machine learning techniques for solving a classification or regression problem is the decision tree. This algorithm seeks to create a model that forecasts the value of a target variable, and the decision tree resolves the issue by utilizing a tree representation, where the leaf node corresponds to a class label and characteristics are expressed on the internal node of the tree. A tree is created by subdividing the source set, the tree's root node, into subsets, the tree's successor children [24]. A set of classification feature-based

splitting laws serve as the base for the splitting. Decision trees are one of the most popular machine learning algorithms due to their clarity and simplicity.

2.3.3 Naive Bayes

Naive Bayes is a precise approach for creating classifiers, models that assign class labels to problem cases represented as vectors of feature values and choose the class labels from a finite set. The Bayes probability theorem, a popular tool for solving statistical issues and classification problems, serves as the foundation for the method [25]. Since all Naive Bayes classifiers assume that the value of one function is essentially independent of every other attribute, provided the class variable, there is no unique technique for training such classifiers. Instead, there is a family of algorithms based on the same premise.

2.3.4 K-Nearest Neighbors

K-Nearest Neighbors is one of the most elementary yet important classification algorithms in machine learning [26]. Applications it finds in the supervised learning domain include pattern recognition, data processing, and intrusion detection. Due to its non-parametric nature, which implies that it makes no assumptions about how data will be delivered, it is frequently utilized in real life scenarios. Prior data are also known as training data is provided, which splits coordinates into categories depending on an attribute. It initially selects n centroids, where n is the total number of projected class labels. Nearer to one centroid data points are regarded as belonging to the same class. Each data point belongs to a specific assigned cluster.

2.3.5 Logistic Regression

Another statistical technique that machine learning has adopted is logistic regression [27]. One of the most popular Machine Learning algorithms used in Supervised Learning is logistic regression. A categorical dependent variable can be estimated using this technique using a number of independent factors. Logistic regression is known as a robust machine learning algorithm because it can seek

out new data using both continuous and discrete datasets and can also has probability. The only real difference between linear regression and logistic regression is how they are applied. While logistic regression is used to address classification challenges, linear regression is used to address regression-related challenges.

2.3.6 Random Forest

The supervised learning approach is used by the well-known machine learning algorithm Random Forest [28]. It can be applied to classification and regression issues in machine learning. It is based on ensemble learning, a technique for combining several classifiers to tackle challenging problems and improve model accuracy. The technique creates several decision trees for the prediction task, hence the name "forest" in English. The random forest uses the predictions from each decision tree and anticipates the ultimate performance based on the predictions that received the most votes, as opposed to relying just on one decision tree. The forest is more accurate and the overfitting problem is eliminated when there are the more trees in it.

2.3.7 Gradient Tree Boosting

In Gradient Boosting, the word "Gradient" refers to the existence of two or more derivatives of the same function. Gradient Boosting is a functional gradient algorithm that iteratively selects a function that points in the direction of the negative gradient, or a weak hypothesis, in order to reduce a loss function. Gradient boosting has been used in a variety of technological domains over the years. The method may appear complex at first, but it often only uses one fixed configuration for classification and one for regression. Of course, this configuration can be changed to meet specific needs.

2.4 Web Technology

2.4.1 Frontend Technology

- **HTML**

The fundamental markup language for documents meant to be used in a web browser is HTML, or HyperText Markup Language. Online browsers transform HTML files into multimedia web pages after receiving them from a web server or locally stored files. HTML first offered hints for the document's display and defined the semantic structure of a web page. The elements that make up HTML pages are known as HTML elements. Using HTML structures, images and other artifacts, such as interactive forms, can be added to the produced page. HTML enables the creation of well-organized texts by designating structural semantics for text elements including headers, paragraphs, links, quotations, and other objects.

- **CSS**

  The appearance of a text written in a markup language like HTML can be specified using the style sheet language CSS. CSS is an essential part of the World Wide Web, along with HTML and JavaScript. With the help of a style sheet called CSS, you may separate presentation from text, including fonts, colors, and layout. By specifying the pertinent CSS in a separate.css file, this separation facilitates improved content accessibility, more flexibility and control in the specification of presentation characteristics, and the ability for multiple web pages to share formatting. This separation also reduces complexity and repetition in the structural content and enables the.css file to be cached to improve page load speed between the pages that have the same format and similar file.

- **Javascript**

  A simple, interpreted programming language is JavaScript. The development of network-centric applications is its intended use. In addition to complementing Java, it also works with it. Because it is integrated with HTML, JavaScript is relatively simple to use. It is free and platform-independent. Because it is the most widely used programming language in the world, Javascript is a fantastic option for programmers. Learning Javascript makes it easier to create excellent front-end and back-end applications utilizing

a variety of Javascript-based frameworks, including jQuery, Node.JS, and others.

2.4.2 Backend Technology

- **Django**

  Django is a popular Python web application framework that sticks to the "batteries-included" concept. Batteries-included is based on the idea that common functionality for creating web applications should be included with the framework rather than available as separate libraries. Simply put, the extensibility and batteries-included philosophies represent two distinct approaches to framework development. Inherently, neither philosophy is superior to the other. Django handles a lot of the hassle associated with web development, allowing one to concentrate on developing app without having to create the wheel. It is open source and free, has a strong community, excellent documentation, and a variety of free and paid support options.

  1. **WSGI**: For creating Python online applications, the Web Server Gateway Interface (WSGI) has emerged as the industry standard. A specification for a common interface between the web server and web applications is the Web Service Gateway Interface (WSGI).

  2. **Werkzeug**: It is a WSGI toolkit that manages requests, responses, and other regular operations. As a result, a web application may be constructed on top of it. One of the pillars of the Django system is Werkzeug.

  3. **Jinja2**: Among Python templating engines, Jinja2 is the most popular. For the purpose of producing dynamic web pages, a web templating system combines a template with a specific data source.

2.4.3 Matrices used

- **Accuracy**: The accuracy is calculated by dividing the total number of predictions in the dataset by the number of accurate forecasts (ACC). The

accuracy levels range from 0.0 to 1.0, with 1.0 being the most accurate.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision**: Precision is obtained by dividing the total number of accurate positive forecasts by the total number of positive predictions (PREC). Another term for it is positive predictive value (PPV). The best and worst precision values are 1.0 and 0.0, respectively.

  $$Precision = \frac{TP}{TP + FP}$$

- **Recall**: Sensitivity is calculated by dividing the total number of correct positive predictions by the total number of positives (SN). It is often referred to as the true positive rate (TPR) or the recall rate (REC) (TPR). The sensitivity ranges from 0.0 to 1.0, with 1.0 being the most.

  $$Recall = \frac{TP}{TP + FN}$$

- **F1-score**: The F-score is a harmonic mean of recall and precision..

  $$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

2.4.4 Visualization

Many insights that data alone cannot offer are made possible by data visualization. Python offers some of the most interesting tools for data visualization. While the simpler plot types are found in many libraries, some are exclusive to a few. We have utilized the Matplotlib and Seaborn visualization tools.

2.5 Conclusion

The research that has been done about applying machine learning in the healthcare system to diagnose disease is described in this chapter. The algorithms utilized for the results as given in the literature are identified, together with the experimental procedures, data gathering strategy, and flaws. Also offered are several potential machine learning algorithms. The conventional machine learning algorithms that are employed in this context are briefly outlined after that.

The methods which are employed in this work is briefly detailed in the following chapter.

2.5.1 Implementation Challenges

The preparation of datasets, gathering student data, using machine learning algorithms, visualizing the gathered data, and evaluating the performance of various algorithms are a few implementation problems.

Chapter 3

Methodology

3.1 Introduction

Researchers have been studying various aspects of artificial intelligence over the past ten years in an effort to successfully and effectively integrate it into the healthcare system and improve the standard of clinical practice. These studies produced highly beneficial findings and had an effect on the existing therapeutic system. Machine learning and artificial intelligence are being used in the realm of medical science to diagnose sickness in patients.

The results of the earlier studies are shown in the previous chapter. The research methodology is examined, and its shortcomings are noted. This chapter provides a brief explanation of the approach used in this work to obtain the results.

3.2 Overview of Framework

The procedure of feature selection is necessary for the implemented methodology to diagnose patients with COPD, Bronchial Asthma, and Pneumonia. Data collection is finished after carefully choosing features from medical books and also with the help of experts. The above-mentioned illness' symptoms were directly collected from patients admitted to hospitals around Bangladesh. Data preprocessing came after the data collection phase. This stage involves converting textual data or raw data into numerical data for computation.

Then, machine learning algorithms like Gradient Boosting, Decision Tree, Random Forest, KNN, SVM, and Logistic Regression have been implemented. Since different algorithms perform better in various situations, multiple algorithms are used. Depending on the algorithms' specialization, some operate better with less

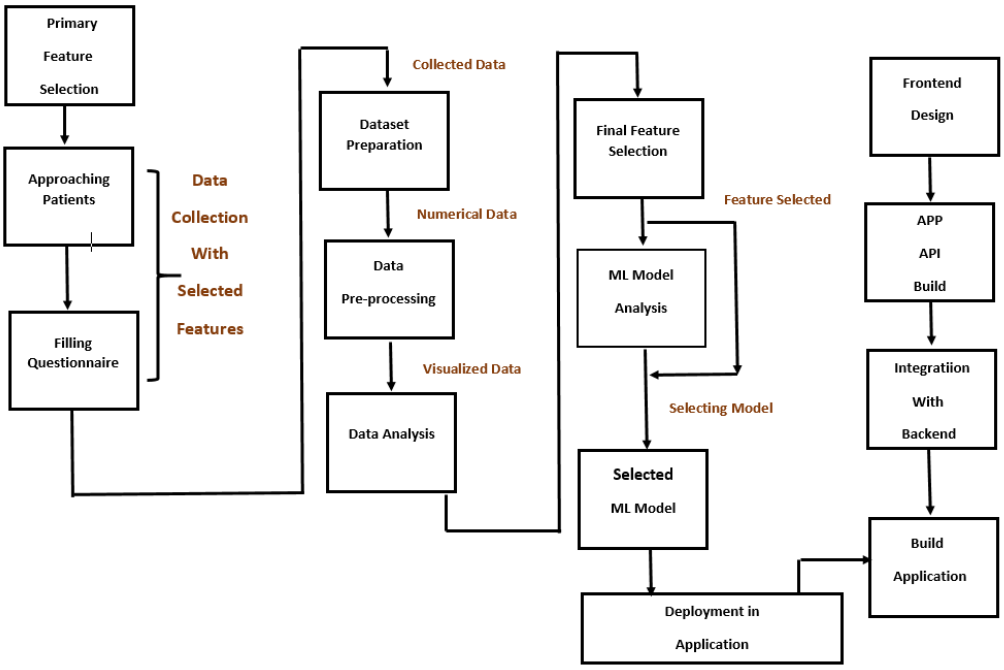


Figure 3.1: Methodology Overview

data while others are effective with larger datasets. We want our module to provide us with the highest level of accuracy when determining whether a patient has the aforementioned diseases or not.

3.3 Detailed explanation of implementation

The implementation approach has been briefly stated in this part in the sequential order shown in the figure3.1.

3.3.1 Primary Feature Selection

The procedure of choosing the major features is the most significant aspect of the entire work. The main characteristics in this case are those of patients with cough-type chronic diseases [29]. Many features that must be considered to diagnose these kinds of patients have been designed and proposed by researchers in the past. However, the medical expert advised that a significant number of features with a cautious selection procedure is needed in order to distinguish between the three diseases with the closest cluster of symptoms, such as COPD, Asthma, and

Pneumonia. 28 characteristics are chosen for our effort to classify the diseases. The attributes are detailed below:

- Age
- Gender
- Condition of cough
- Sore throat
- Wheezing
- Chills
- Abdominal Pain
- Vomiting
- Fever
- Headache
- Nausea
- Tiredness
- malaise
- Body ache
- Anorexia
- Shortness of breath
- Convulsion
- Weight loss
- Face condition
- Fauces condition
- Chest pain
- Shivering
- Sweating
- Shoulder pain
- Herpes Labialis found
- Periodic asthmatic suffering
- Nocturnal episode of Dyspnea

3.3.2 Collection of Data

Collection of data is e single most important stage in tackling every machine learning challenge. However, it is a huge obstacle for plenty of academics and computer scientists. It usually takes a long time to process data, which mainly comprises data collecting, data labeling, and upgrading existing data or models. Some constrained are also followed while collecting data. Which are described below.

- To make the data as straightforward and anonymous as possible, no personally identifiable information (such as Name, Contact Number, or Address) is registered.

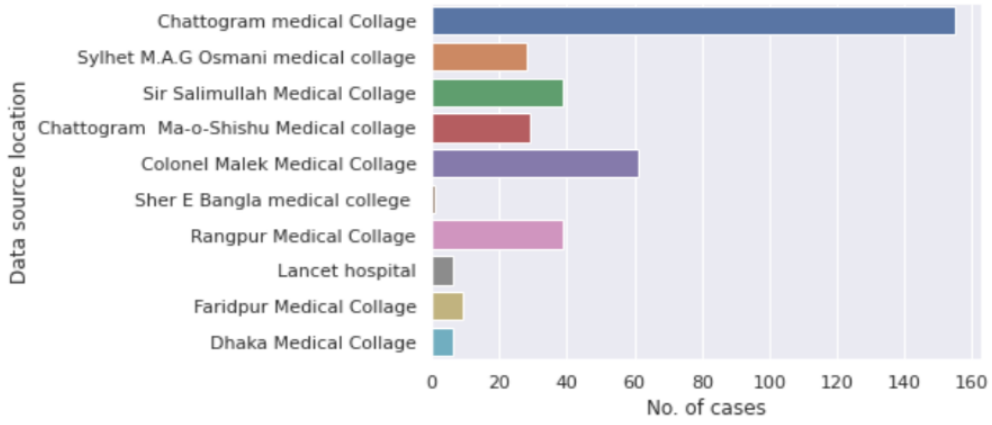


Figure 3.2: Number of cases from respective medical collages

- It took 5 to 10 minutes to complete each data collection procedure.
- Through a questionnaire, patients were asked about the clinical symptoms and signs of the cough-type chronic disease they were experiencing.
- A detailed record of the information is made, and instances are given the appropriate labels.

3.2 shows the number of patients from respective medical collages involved in data collection process.

3.3.3 Data Preprocessing

Any machine learning approach that involves data preprocessing translates or encodes the data to make it easier for the computer to process. In other words, the algorithm can now comprehend the qualities of the input quickly. Raw data cannot be understood by machines. Prediction cannot be made by computing string or text data, either. Again, for algorithms to classify, data must be plotted in a graph or plane. The obtained raw data must therefore be transformed into its corresponding numerical data so that machine learning techniques can be used to produce the desired outcome.

As seen in the 3.3, 3.4, 3.5, each raw data point that was gathered was converted into a numerical value for computation. Data cleaning is a significant part of the conventional data preprocessing stages. There's a good chance that a lot of the data is useless or incomplete. This part is managed by performing data cleaning.

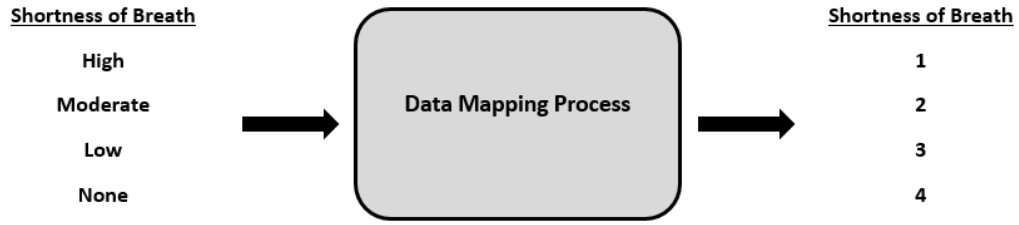


Figure 3.3: Data mapping example(Shortness of Breath)

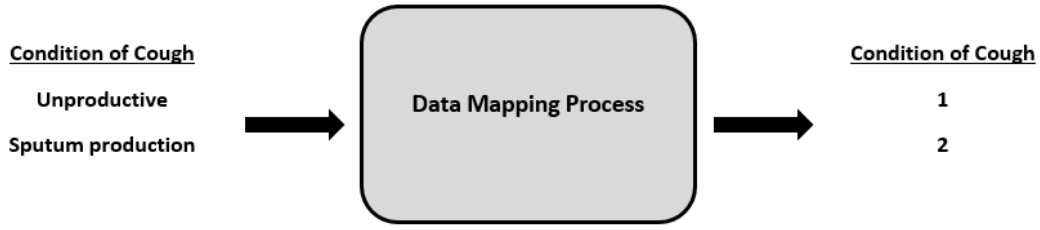


Figure 3.4: Data mapping example(Condition of Cough)

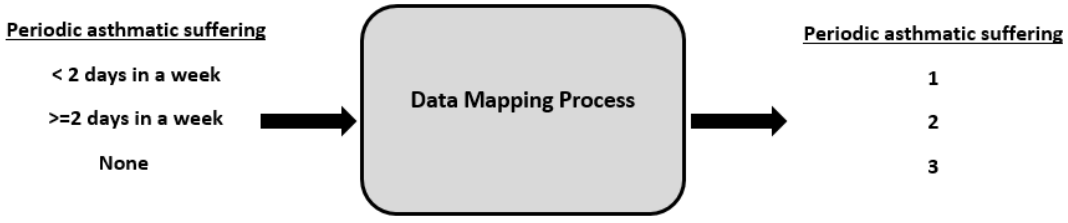


Figure 3.5: Data mapping example(Periodic Asthmatic suffering)

Dealing with imperfect data, noisy data, and other issues is necessary. However, in order to ensure the quality of our dataset, we make sure there are no missing data. Data cleaning is therefore not required in this case. Normalizing the data won't have a significant impact on machine learning performance because the transformed value ranges are minimal. No normalization is done as a result. Additionally, there is no dimensionality reduction.

3.3.4 Data analysis

Data processing requires data analysis as a critical step. The data engineers choose the final features for the machine learning algorithms that will be used to anticipate the outcome with the assistance of data analysis. This process makes the relationship between the features themselves clear. Additionally, data analysis aids the data engineer in locating redundant features or factors that can be eliminated. The outcome of the analysis helps to comprehend the narrative that the data is attempting to convey and establishes relevance.

Trial and error methods have long been used in data analysis, however these methods become impractical when dealing with large and heterogeneous data sets. Big data was criticized for being overhyped precisely for this reason. The complexity of developing new, accurate predictive models is strongly correlated with the amount of data that is available.

The following chapter, Chapter 4, describes the outcomes of the data analysis performed utilizing the gathered dataset. In that section, the data analysis result is briefly discussed. However, the most important discoveries were that each element that was considered had an effect on a certain aspect of the ultimate result. As a result, before feeding data to the machine learning model, no features are eliminated or altered.

3.3.5 Final Feature Selection

High-dimensional data analysis is a major obstacle for engineers and scientists working in the disciplines of data mining and Machine Learning (ML). By removing superfluous and unnecessary data, feature selection provides a straightforward

yet efficient solution to this problem. The accuracy of learning is increased, computation time is decreased, and a better comprehension of the learning model or data is made possible by removing irrelevant data. The majority of the time, not all the parameters in the dataset are relevant when creating a real-world ML model. The addition of redundant variables reduces the model's capacity for generalization and may also reduce a classifier's overall precision. A model also becomes more complex if additional variables are added to it. That's why feature selection is introduced.

The process of limiting the amount of input variables when building a predictive model is called feature selection. Reduced processing costs and, in some situations, increased model accuracy can both be achieved by reducing the number of input variables. Utilizing statistics, the correlation between each input variable and the goal variable is assessed, and the input variables with the strongest correlation are chosen. These methods can be quick and effective because the selection of statistical measurements is dependent on the data form of the input and output variables. In order to reduce training time, avoid the dimensionality curse, and simplify machine learning models, features are selected.

We created a correlation matrix to identify the features that are equally important and irrelevant in order to examine the correlation of major features. By removing elements that are not crucial for prediction, model performance can be improved. However, we didn't find any such elements in our work, thus nothing was removed or altered.

3.3.6 Evaluation and Selection of ML Model

Using advanced machine learning libraries like scikit-learn and Keras, it is easy to fit a range of machine learning models to a predictive modeling dataset. As a result, selecting a model from a range of alternatives for a given problem is the complexity of applied machine learning. Models of various types and models of the same type with various model hyperparameters can be compared using a process called machine learning model selection. A model's generalization error will be found through model evaluation. Unsurprisingly, a successful machine

learning model performs well both on data that was not observed during training and on data that was. As a result, before putting a model into development, we should have confidence that the results won't be compromised by the presence of fresh data.

With the preprocessed data that was gathered for this work, several machine learning models were trained for the model selection procedure. The following is a list of trained and assessed machine learning algorithms:

- Support Vector Machine
- Naive Bayes
- Decision Tree
- Random Forest
- Logistic Regression
- K-Nearest Neighbors
- Gradient Tree Boosting
- AdaBoost

The Random Forest model is chosen and deployed through Web-view technology after evaluating the performance of the machine learning algorithms.

3.3.7 App Development

App development is another segment of our work which will work as an interface of our output, HTML, CSS, Java-script are used for frontend and for backend Django and PostgreSQL have been used. We have designed our workflow according to the figure shown in 3.6

3.3.7.1 Frontend Development

In front-end web development, also known as client-side development, HTML, CSS, and JavaScript are created and written for a website or Web application so that a user can view and interact with it directly. The challenge with front-end

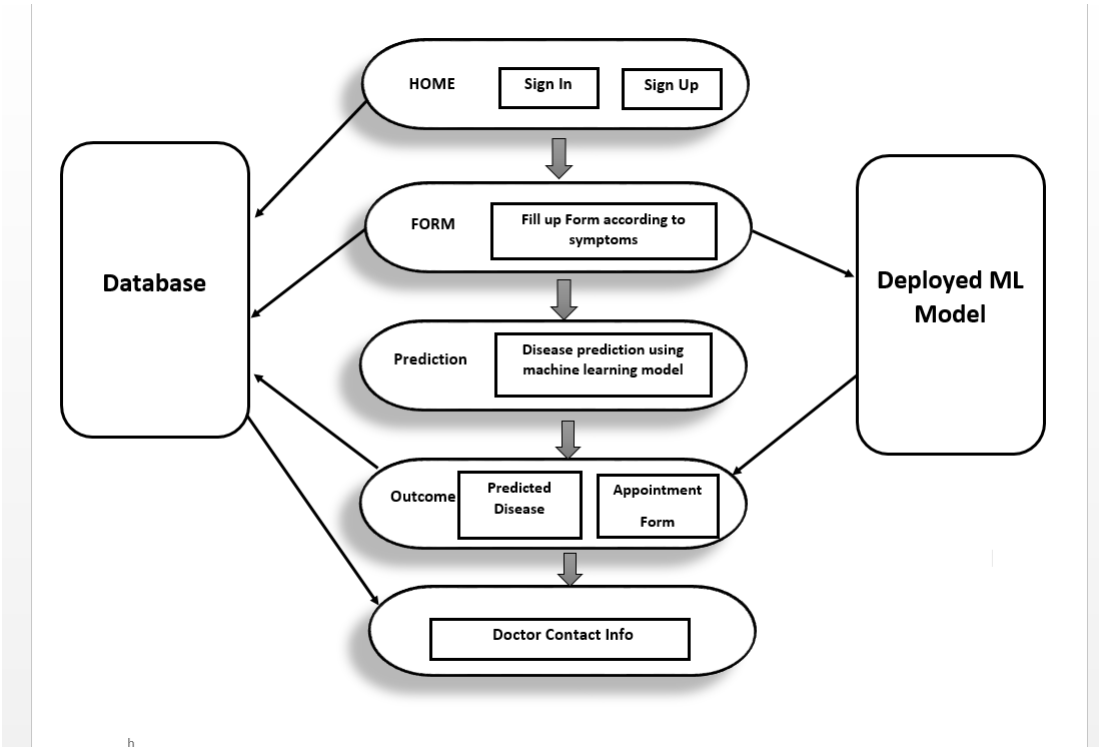


Figure 3.6: Block Diagram of mobile application

development is that the techniques and approaches used to create a website's front end are always changing, requiring the developer to maintain constant awareness of the direction in which the field is headed. Making sure that content is presented in an understandable and easy-to-read manner for website visitors is the aim of website design.

This is made even worse by the fact that consumers already use a variety of gadgets with various screen sizes and resolutions, which forced the web designer to take these things into account while designing the site. The application's frontend has been developed using such technology. It contains multiple pages. Several pages' interface like user sign-up, user sign-in, Form to collect symptoms, prediction result, doctor appointment have been developed using HTML. CSS and Java-script have been used to enhance user interface view.

3.3.7.2 Backend Development

Backend development refers to the server side of technology, which is mostly focused on how the web works. The main duties will be managing the site's

functionality and making adjustments and changes. With this type of web construction, a server, an app, and a database are often included. The code that transmits database knowledge to the viewer is written by backend engineers. A backend developer is in charge of things like databases and servers that are hidden from view. Backend developers may have positions that are recognized as programmers or web developers.

PostgreSQL is an open-source relational database management system that works with objects as a relational component. It is also known as Postgres because it employs Structured Query Language (SQL) to retrieve the data in the database's tables. Some of this database's standout characteristics include its extreme robustness and dependability, ease of recovery, and low cost and human labor requirements for maintenance. The PostgreSQL Global Development Group, a team of PostgreSQL developers, created and maintains it.

The backend of the application is created using Python programming and Django microservices. A pickle file is created after training the machine learning model that was selected after the model evaluation procedure. In the backend codebase, the model is then loaded. Use of Postgresql for DBMS was prefered. There are multiple APIs are developed in the backend. One API is used to fetch user information to the backend. When the submit button is pressed, the information is first received using an HTML form at the client-side of the App and then fetched via an API to the backend. As soon as the button is pressed, the app accesses the predict API and passes the patient data it has received through the trained model to produce a prediction. The frontend of the app is then rendered with the outcome. The appointment form is also developed using HTML and and an API is developed to fetch information from that form to the backend. When button is pressed the app accesses Doctors' information from backend and display that in the userend of the application. Another API is used to do this.

For hosting, Heroku free hosting facilities has been used. Web into app technology is used to convert the app into an APK file. Which is then can be operated like an android application.

3.3.8 Model Deployment

Only once the target audience for which a machine learning model was created has frequent access to the ideas it generates will that model begin to be useful. A machine learning model that has met certain criteria is deployed, and then users or other programs can access its predictions. Model discovery, model verification, and other routine machine learning tasks like feature engineering are not the same as deployment. Model implementation is among the most challenging aspects of reaping the rewards of machine learning. The Django microservice and REST API are used in our effort to deploy the machine learning model.

3.4 Conclusion

This chapter explains and thoroughly discusses the approach used to accomplish the work's objective. The selection of the primary feature starts the process. After carefully choosing the feature, information is collected from the hospitalized patient in accordance with the feature. For the purpose of calculation, the gathered data is preprocessed and transferred to a numerical value. The final feature selection is then made after analyzing the numerical data. No feature is eliminated or altered in this step. The data and the chosen features are then fed into a number of machine learning models, which are then evaluated to see which model performs the best. The model is delivered to the web app, which is then tested for usability. Results of the implementation are covered in the following chapter.

Chapter 4

Results and Discussions

4.1 Introduction

The end result is a part that summarizes the main conclusions of a study, whereas the discussion evaluates those outcomes and examines their significance for readers. The Outcome and Discussion section of the paper describes the findings of the exploration and statistical analysis. It provides an overview and predictions based on the data gathered. This section helps other researchers decide what is feasible and what ought to be anticipated when a particular technique is applied.

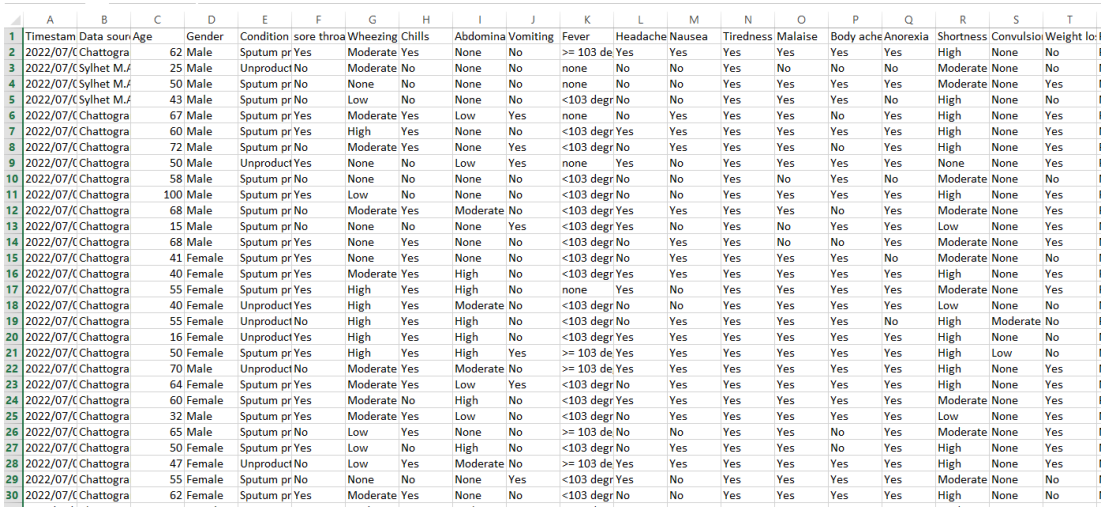
The approach used to extract the desired result is covered in great detail in the preceding chapter. The procedure for gathering data, choosing features, and analyzing the gathered data are all described. Additionally, a description of machine learning algorithm assessment is provided.

The outcomes of the method that was used are illustrated in this chapter. As well as presenting the information produced from the approach, the main findings are also discussed. Relevant figures are used to display the data. As a complement to the calculated result, contextual analyses are provided.

4.2 Dataset Description

The dataset includes the symptoms of 373 hospital patients who were either diagnosed with COPD, asthma, or pneumonia after being admitted. Patients who shared related symptoms but did not have those specific diseases are also mentioned. According to feature design, each instance consists of 28 variables that have an effect on the prediction outcome and effectively categorize three closely

symptomized cough-type chronic diseases. Figure 4.1 shows a visual description of the dataset.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Timestam	Data sour	Age	Gender	Condition	sore throa	Wheezing	Chills	Abdomina	Vomiting	Fever	Headache	Nausea	Tiredness	Malaise	Body ache	Anorexia	Shortness	Convulsio	Weight lo
2	2022/07/(	Chattogra	62	Male	Sputum pr	Yes	Moderate	Yes	None	No	>= 103 de	Yes	Yes	Yes	Yes	Yes	Yes	High	None	No
3	2022/07/(	Sylhet M.A	25	Male	Unproduct	No	Moderate	No	None	No	none	No	No	Yes	No	No	No	Moderate	None	No
4	2022/07/(	Sylhet M.A	50	Male	Sputum pr	No	None	No	None	No	none	No	No	Yes	Yes	Yes	Yes	Moderate	None	Yes
5	2022/07/(	Sylhet M.A	43	Male	Sputum pr	No	Low	No	None	No	<103 degr	No	No	Yes	Yes	Yes	No	High	None	No
6	2022/07/(	Chattogra	67	Male	Sputum pr	Yes	Moderate	Yes	Low	Yes	none	No	Yes	Yes	Yes	No	Yes	High	None	Yes
7	2022/07/(	Chattogra	60	Male	Sputum pr	Yes	High	Yes	None	No	<103 degr	Yes	Yes	Yes	Yes	Yes	Yes	High	None	Yes
8	2022/07/(	Chattogra	72	Male	Sputum pr	No	Moderate	Yes	None	Yes	<103 degr	No	Yes	Yes	Yes	No	Yes	High	None	Yes
9	2022/07/(	Chattogra	50	Male	Unproduct	Yes	None	No	Low	Yes	none	Yes	No	Yes	Yes	Yes	Yes	None	None	Yes
10	2022/07/(	Chattogra	58	Male	Sputum pr	No	None	No	None	No	<103 degr	No	No	Yes	No	Yes	No	Moderate	None	No
11	2022/07/(	Chattogra	100	Male	Sputum pr	Yes	Low	No	None	No	<103 degr	No	No	Yes	Yes	Yes	Yes	High	None	Yes
12	2022/07/(	Chattogra	68	Male	Sputum pr	No	Moderate	Yes	Moderate	No	<103 degr	Yes	Yes	Yes	Yes	No	Yes	Moderate	None	Yes
13	2022/07/(	Chattogra	15	Male	Sputum pr	No	None	No	None	Yes	<103 degr	Yes	No	Yes	No	Yes	Yes	Low	None	Yes
14	2022/07/(	Chattogra	68	Male	Sputum pr	Yes	None	Yes	None	No	<103 degr	No	Yes	Yes	No	No	Yes	Moderate	None	Yes
15	2022/07/(	Chattogra	41	Female	Sputum pr	Yes	None	Yes	None	No	<103 degr	No	Yes	Yes	Yes	Yes	No	Moderate	None	No
16	2022/07/(	Chattogra	40	Female	Sputum pr	Yes	Moderate	Yes	High	No	<103 degr	Yes	Yes	Yes	Yes	Yes	Yes	High	None	Yes
17	2022/07/(	Chattogra	55	Female	Sputum pr	Yes	High	Yes	High	No	none	Yes	No	Yes	Yes	Yes	Yes	Moderate	None	Yes
18	2022/07/(	Chattogra	40	Female	Unproduct	Yes	High	Yes	Moderate	No	<103 degr	No	No	Yes	Yes	Yes	Yes	Low	None	No
19	2022/07/(	Chattogra	55	Female	Unproduct	No	High	Yes	High	No	<103 degr	No	Yes	Yes	Yes	Yes	No	High	Moderate	No
20	2022/07/(	Chattogra	16	Female	Unproduct	Yes	High	Yes	High	No	<103 degr	Yes	Yes	Yes	Yes	Yes	Yes	High	None	No
21	2022/07/(	Chattogra	50	Female	Sputum pr	Yes	High	Yes	High	Yes	>= 103 de	Yes	Yes	Yes	Yes	Yes	Yes	High	Low	No
22	2022/07/(	Chattogra	70	Male	Unproduct	No	Moderate	Yes	Moderate	No	>= 103 de	Yes	Yes	Yes	Yes	Yes	Yes	High	None	Yes
23	2022/07/(	Chattogra	64	Female	Sputum pr	Yes	Moderate	Yes	Low	Yes	<103 degr	No	Yes	Yes	Yes	Yes	Yes	High	None	Yes
24	2022/07/(	Chattogra	60	Female	Sputum pr	Yes	Moderate	No	High	No	<103 degr	Yes	Yes	Yes	Yes	Yes	Yes	Moderate	None	Yes
25	2022/07/(	Chattogra	32	Male	Sputum pr	Yes	Moderate	Yes	Low	No	<103 degr	No	Yes	Yes	Yes	Yes	Yes	Low	None	Yes
26	2022/07/(	Chattogra	65	Male	Sputum pr	No	Low	Yes	None	No	>= 103 de	No	No	Yes	Yes	No	Yes	Moderate	None	Yes
27	2022/07/(	Chattogra	50	Female	Sputum pr	Yes	Low	No	High	No	<103 degr	No	Yes	Yes	Yes	No	Yes	High	None	Yes
28	2022/07/(	Chattogra	47	Female	Unproduct	No	Low	Yes	Moderate	No	>= 103 de	Yes	Yes	Yes	Yes	Yes	Yes	High	None	Yes
29	2022/07/(	Chattogra	55	Female	Sputum pr	No	None	No	None	Yes	<103 degr	No	No	Yes	Yes	Yes	Yes	Moderate	None	No
30	2022/07/(	Chattogra	62	Female	Sputum pr	Yes	Moderate	Yes	None	No	<103 degr	No	No	Yes	Yes	Yes	Yes	High	None	No

Figure 4.1: Dataset

Figure 4.2 shows how the 373 data records are distributed throughout the classes. Minor unbalance in the dataset may not be noticeable in this instance. Bronchial Asthma disease is less frequently recorded in the dataset than in other datasets, as can be observed.

Dataset was gathered and recorded from various medical colleges in Bangladesh. The dataset includes a substantial amount of variance from this standpoint because it was gathered over a wide geographic region.

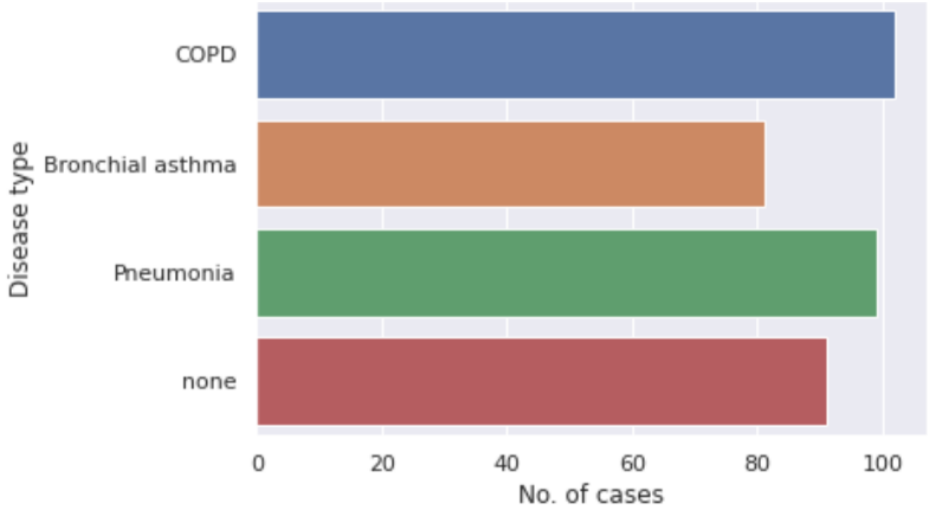


Figure 4.2: No of cases for respective diseases

Patients of all ages provided their symptoms, which were gathered. The patient

count is plotted versus the age range of the patients included in the dataset in Figure 4.3. The dataset contains the most symptoms of patients over 50 who have been admitted that have been recorded.

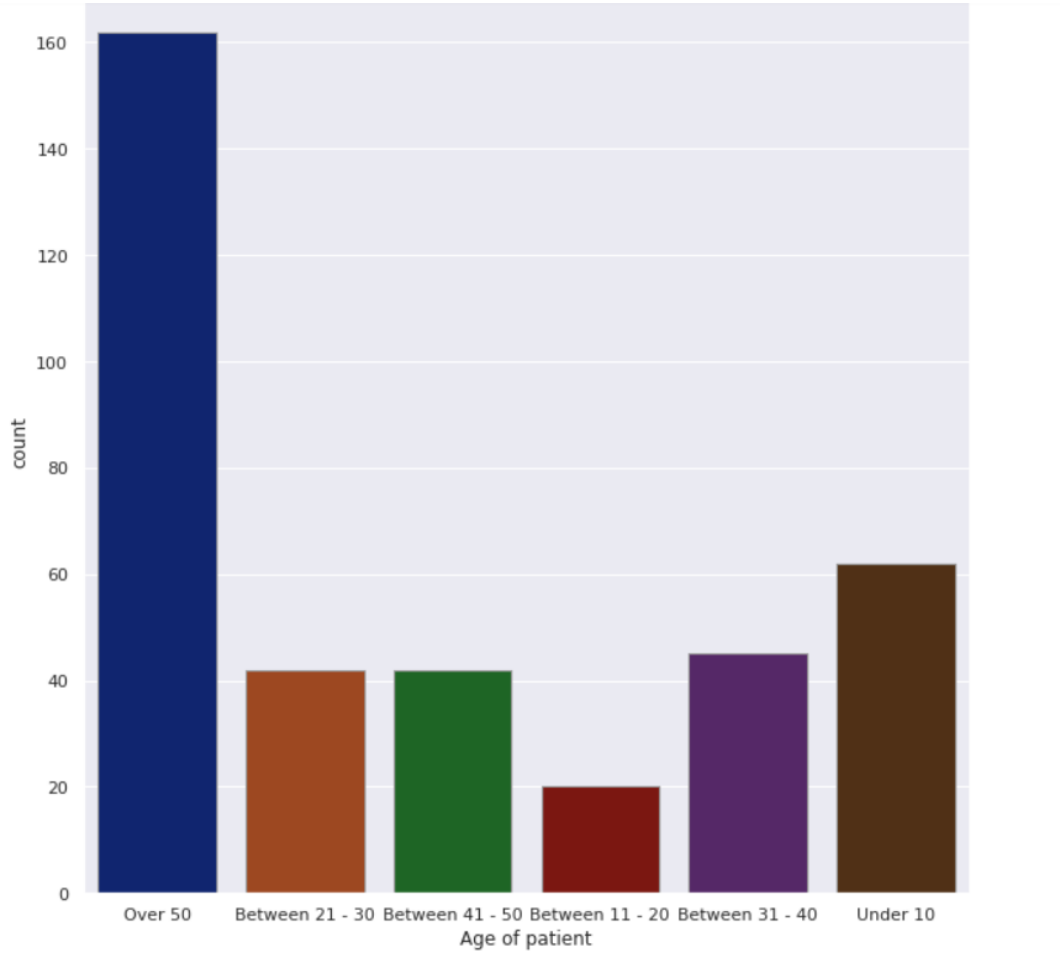


Figure 4.3: No. of Records(Patient Age)

In order to obtain greater understanding, we have categorized the data by gender in figure 4.4 and displayed the average age of the patients who had been diagnosed with COPD, Bronchial Asthma, Pneumonia, or none of those diseases. As we have known age as one of the relevant factors while diagnosing particular diseases [30] .

One of the features examined in figure 4.5 is implemented in the machine learning model. The cough pie chart in subfigure 4.5a shows that 59.1% of patients had a wet cough, which is also referred as as sputum production or blood strained (clinically known as hemoptysis), while 40.9% of patients had an unproductive cough, also known as a dry cough, as a clinical symptom of the diseases. The

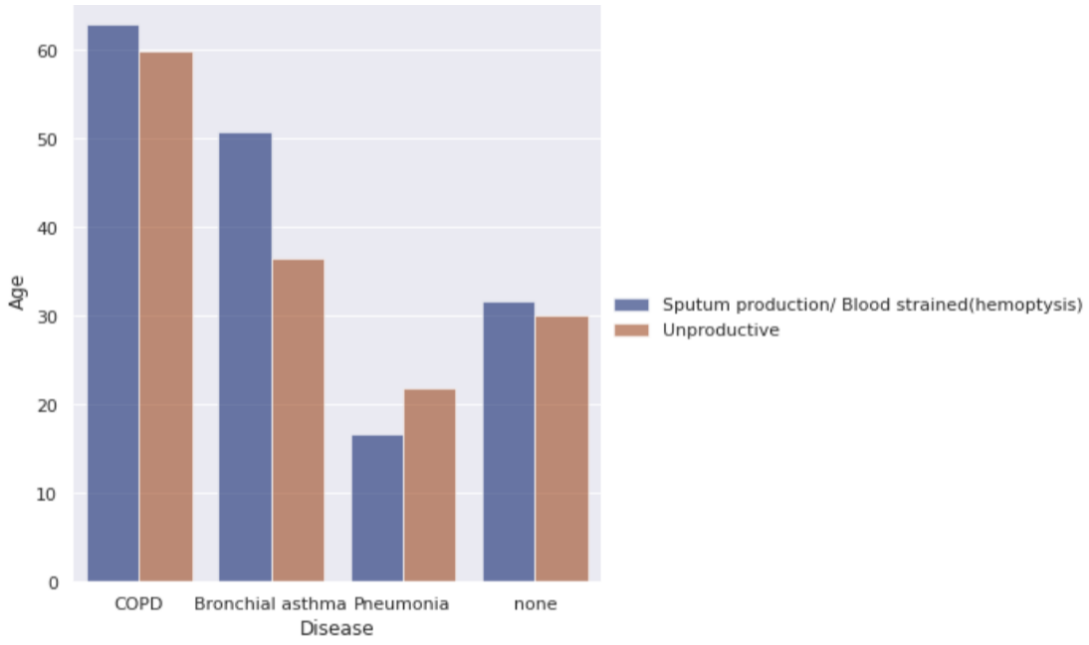


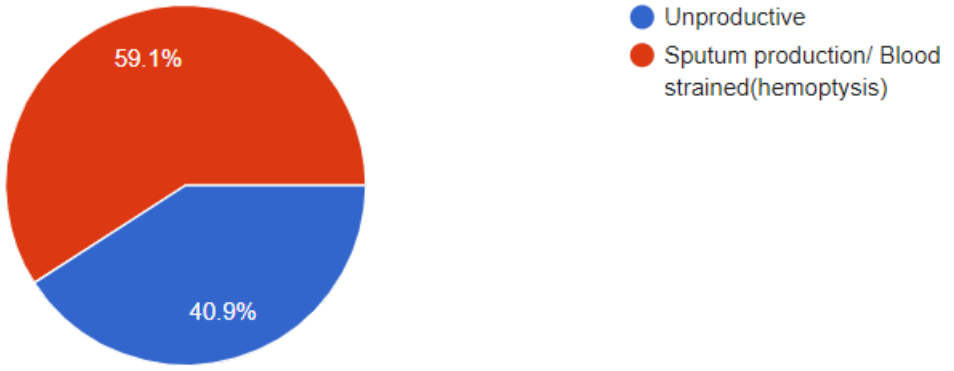
Figure 4.4: Age vs Symptoms

number of patients who experienced both types of cough, however, are categorized by the predicted diseases in subfigure 4.5b.

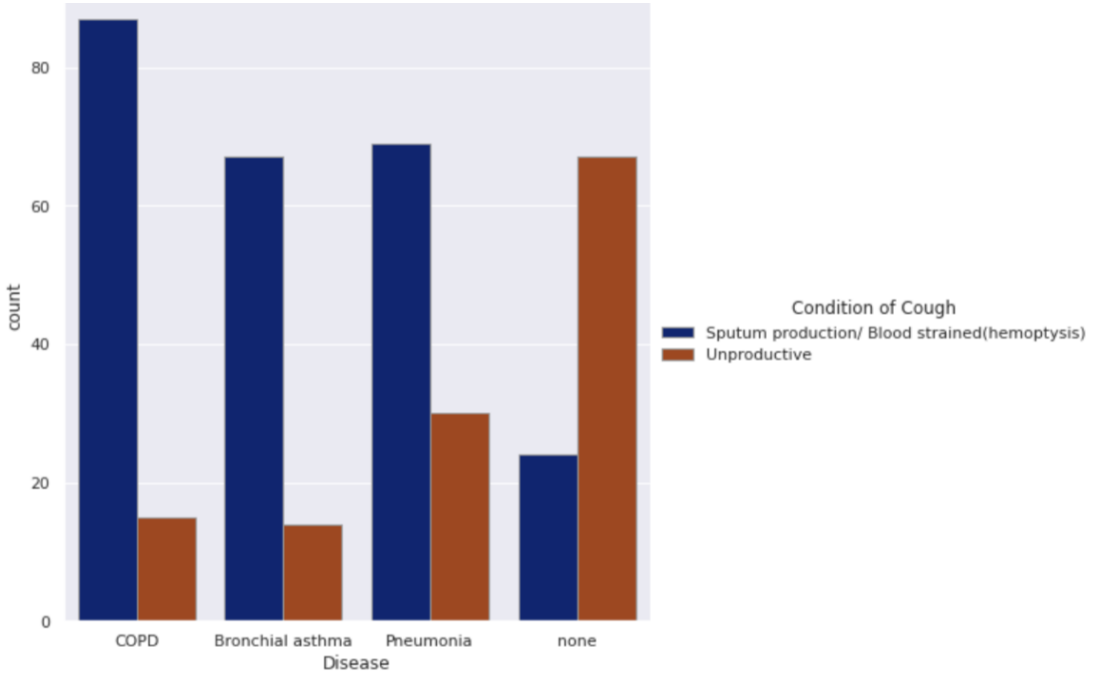
Here, another interesting phenomena we have noticed, if we discuss subfigure 4.5b, we can see cases which are leveled as none of the diseases are mostly related to unproductive or dry regarding cough condition. Majority patients of predicted diseases had experienced wet cough which is also known as sputum production or blood strained (hemoptysis). That means cough had been a severe feature among the participants of this study.

One of the traits examined in figure 4.6 is one that was included in the machine learning model. The abdominal pie chart in subfigure 4.6a shows that, overall, 62.9% of patients did not have abdominal pain as a disease symptom. Subfigure 4.6b on the contrary, displays the number of patients—grouped by the disease recognized reported having or not having abdominal pain. Apart from cases of COPD majority number of patients have face or experienced no abdominal pain.

Chest pain is one of the common clinical feature for respiratory and chronic diseases. In the figure 4.7 we can observe 39.3% patients among 373 have had no chest pain, where 27.8% among all the individuals had experienced severe chest pain which is labeled as high.

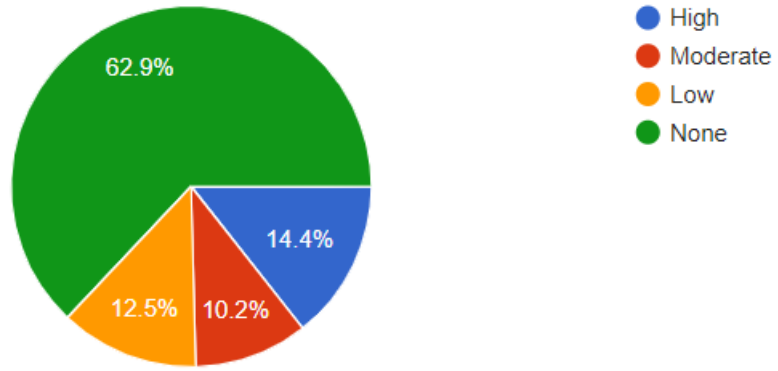


(a) Condition of Cough Pie Chart

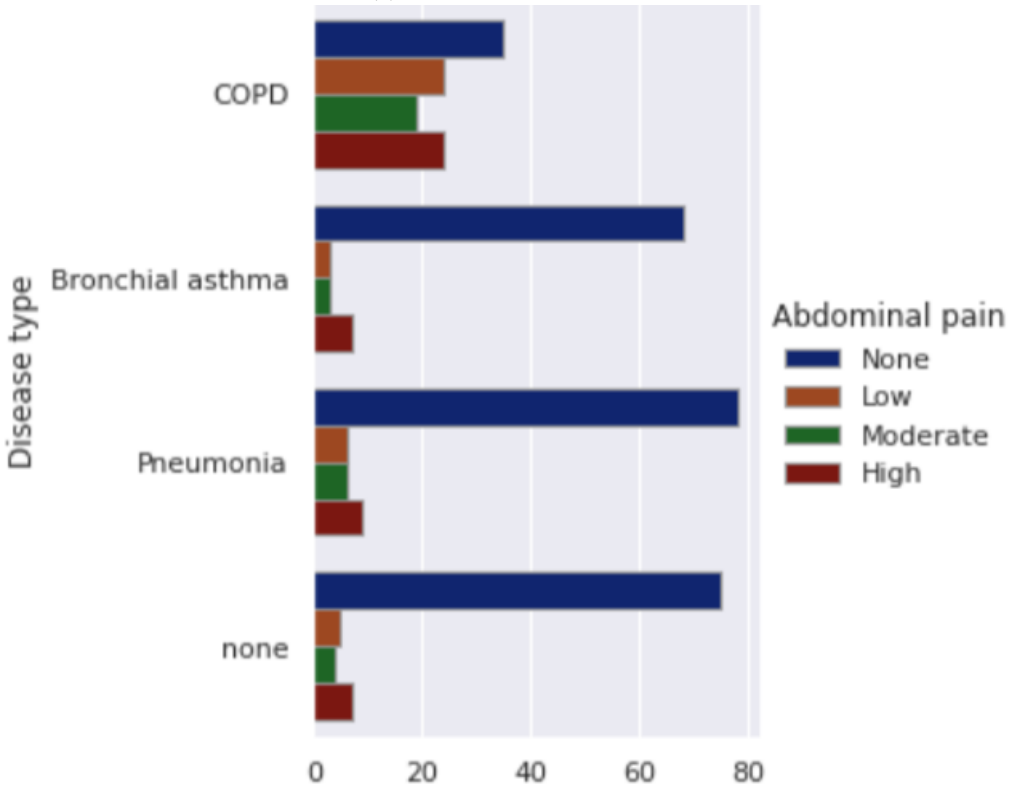


(b) Condition of cough grouped by disease

Figure 4.5: Condition of Cough Exploration

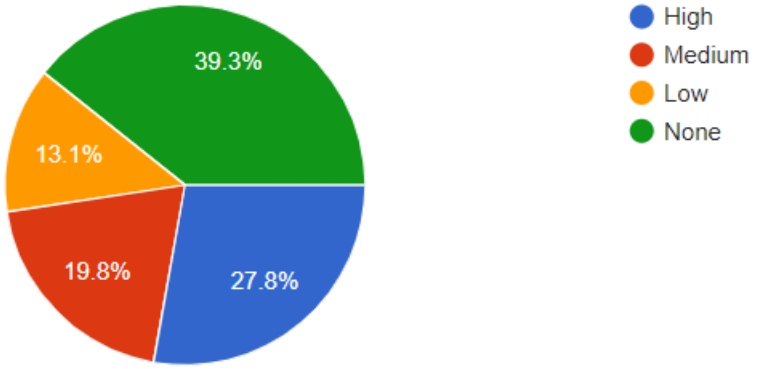


(a) Abdominal Pain Pie Chart

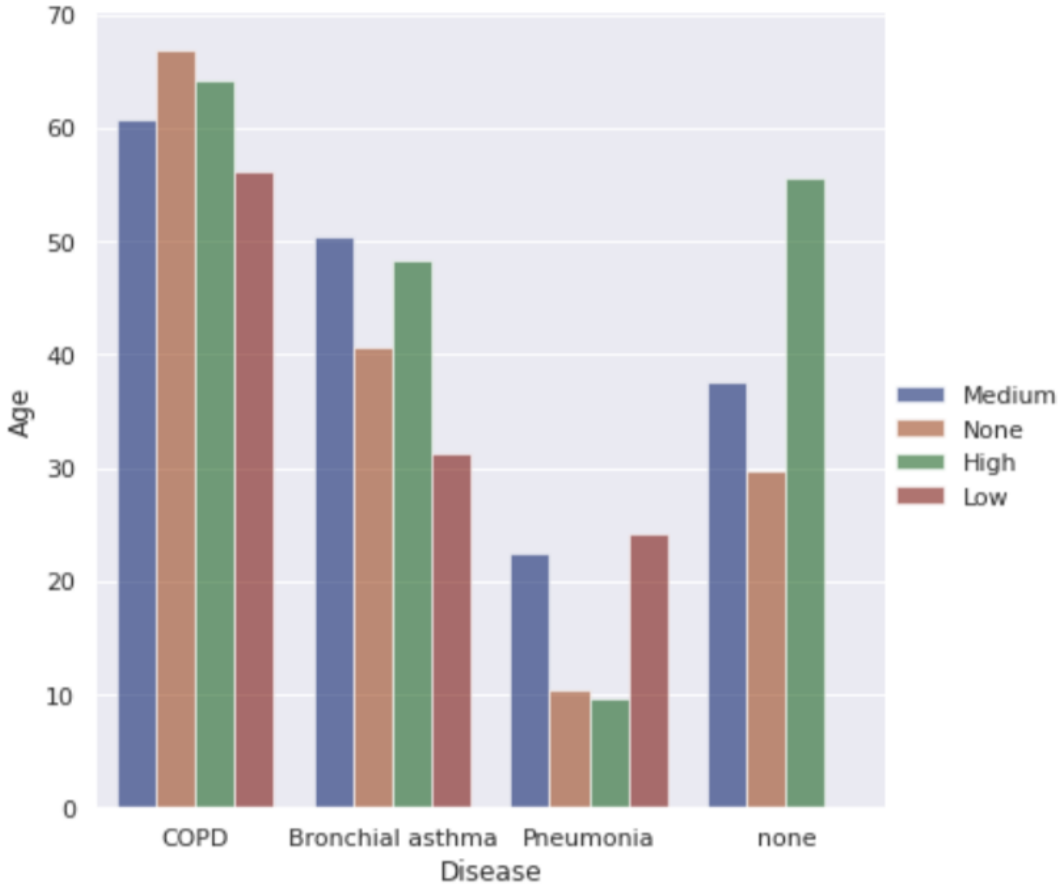


(b) Abdominal Pain grouped by disease

Figure 4.6: Abdominal Pain Exploration



(a) Chest Pain Pie Chart



(b) Abdominal Pain grouped by disease

Figure 4.7: Abdominal Pain Exploration

From the subfigure 4.7b, it is visible that, apart from cases of Pneumonia the chest pain ranges from high to medium mostly in all other cases considering the none ones also. We are assuming those none cases where chest pain is high are mostly diagnosed with heart diseases.

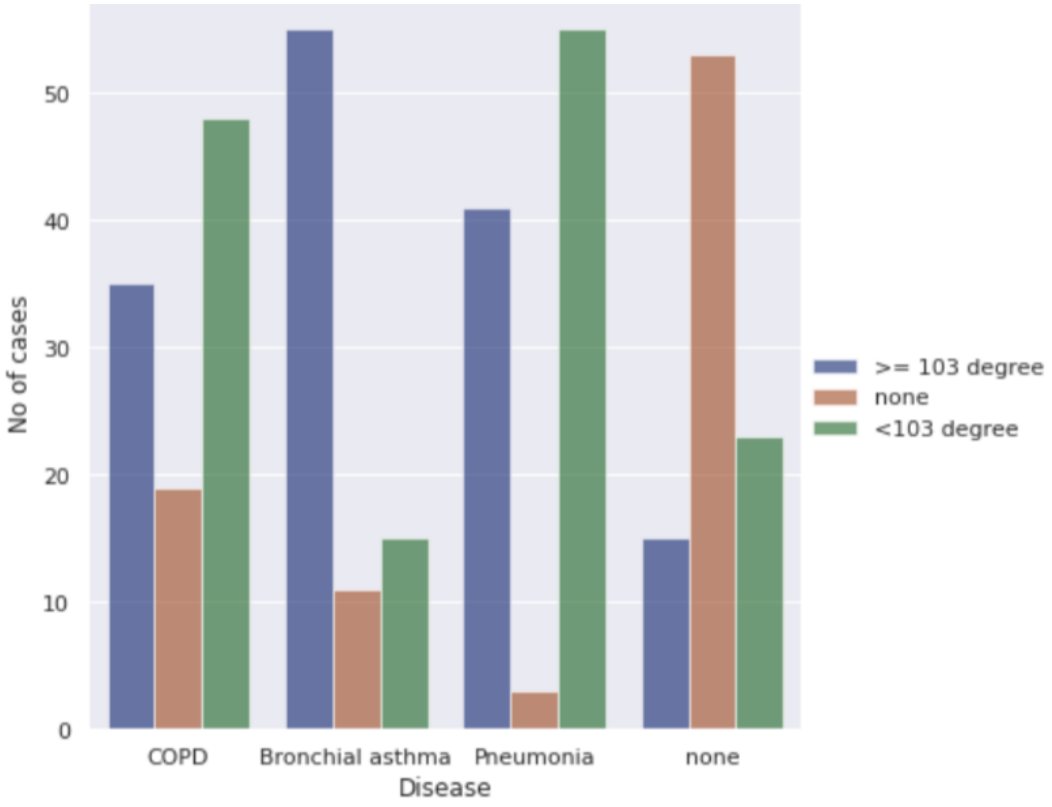


Figure 4.8: Fever grouped by disease

As it shows in figure 4.8 Bronchial Asthma patients had experienced fever more than 103 degree in most cases. For pneumonia experienced individual, less than 103 degree is found in most cases.

In subfigure 4.9 and it is clearly visible that periodic asthmatic suffering and nocturnal episode of dyspnea is very common among the patients diagnosed with Bronchial Asthma.

The correlation matrix of the features is shown in figure 4.10. No features are eliminated or changed for the machine learning model because the correlation matrix clearly shows that each feature contributes in some way to the final output.

The top important variables that have a stronger connection to the target variable can also be determined from the correlation matrix.

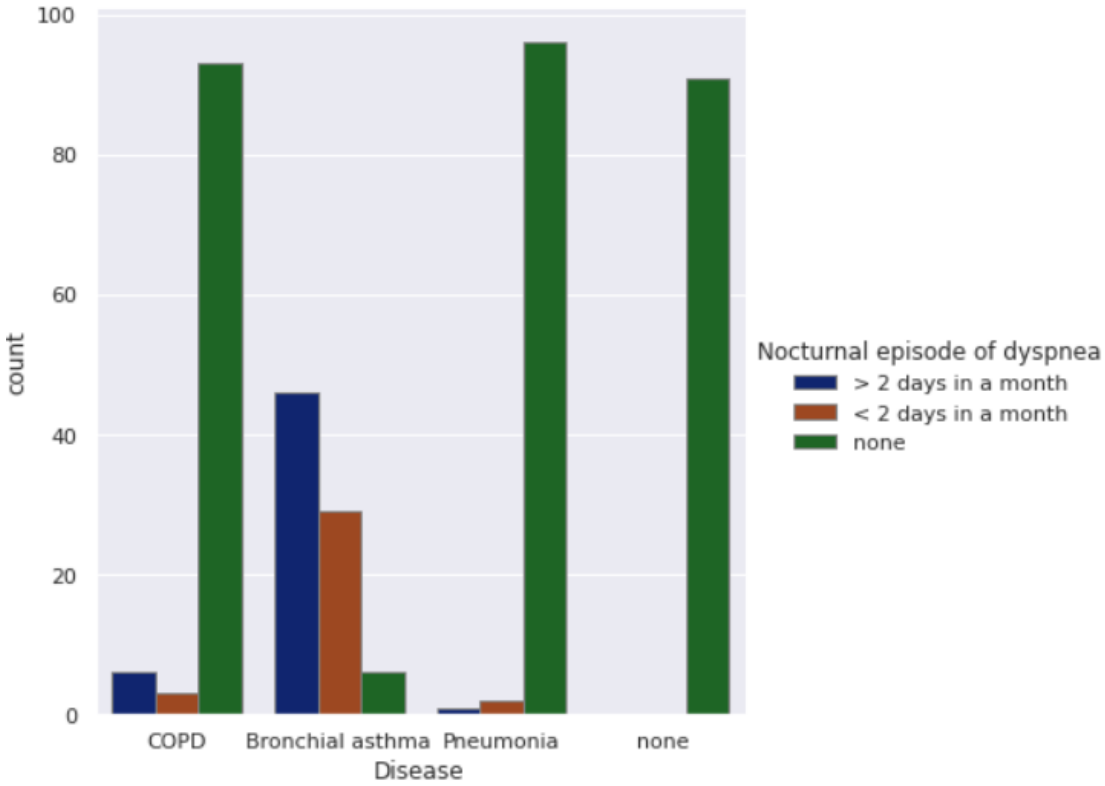


Figure 4.9: Periodic asthmatic episode

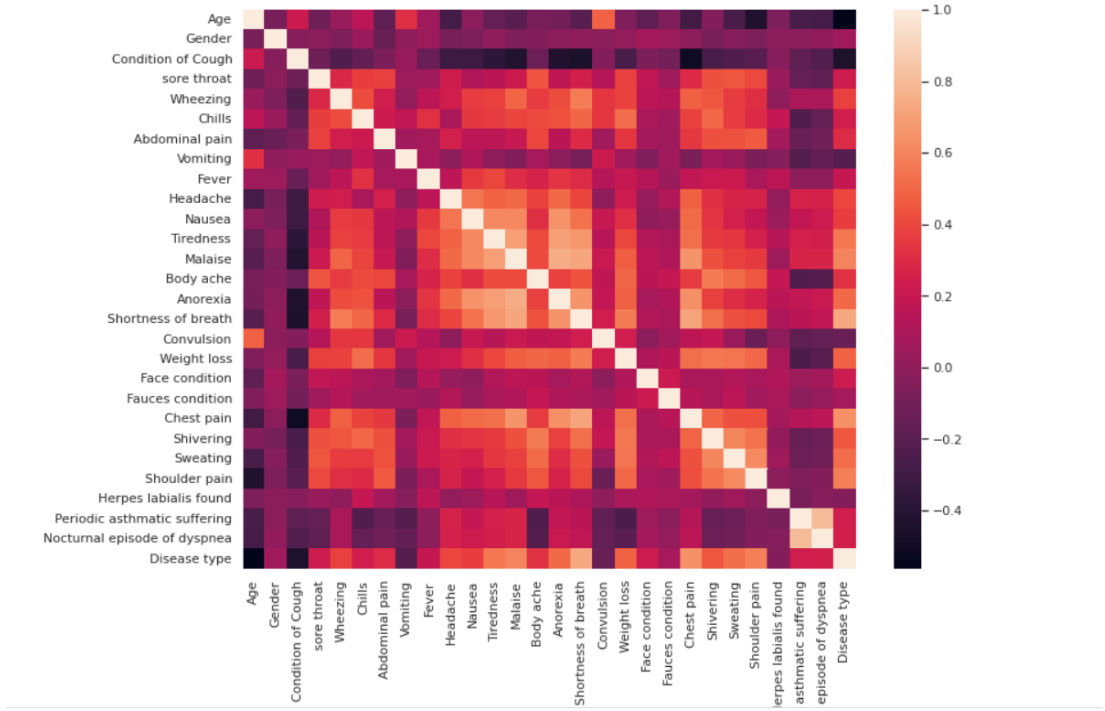


Figure 4.10: Fever grouped by disease

Feature	Relation Strength Factor
Shortness of Breath	0.81
Chest pain	0.79
Condition of cough	0.76
Shoulder pain	0.73
Malaise	0.70

Table 4.1: Key Factors

The table above 4.1 is an analysis of the important aspects and their relationship to the output.

4.3 Impact Analysis

The future of machine learning is rapidly approaching. Machine learning is a broad area of study that, in recent years, has begun to have an impact on every industry one by one. It is not only relevant to the healthcare system. The healthcare industry's future however, is data-driven, and machine learning is essential for allowing artificial intelligence in the health sector. The following are some ways to measure the work's impact:

1. An application for diagnostic support systems that provides real-time assistance to doctors by accessing data and recommending diagnosis.
2. The effectiveness of the present disease detection systems can be upgraded with Machine Learning and AI.
3. Rapid diagnostic assessments can be gathered in areas where hospital facilities are not readily available.

4.4 Evaluation of Performance

Performance evaluation is a crucial step in the machine learning process. However, it is a challenging task. Therefore, it needs to be done carefully while implementing machine learning. Any research must include an evaluation of a machine learning algorithm. A machine learning algorithm may yield good results when evaluated using an accuracy score metric, but it may yield substandard results while evaluated using other metrics, such as logarithmic loss or any other

metric. Classification accuracy is frequently used to evaluate a learning model's performance, although this is insufficient to evaluate a model completely.

In our research, we used the eight most well-liked machine learning classifiers and trained them to test the performance of machine learning models against our gathered and preprocessed dataset. We have computed the model accuracy, precision, recall, F1 score, sensitivity, specificity, error rate, and MAPE of each trained model for the evaluation process and to identify the best-performing model. We have divided the dataset into 75% and 25% for training and testing sequentially in order to evaluate performance in a high-quality manner. Every model was tested with 25% of the data samples after it had been trained with 75% of the shuffled data samples.

For evaluating models and finding optimum result we have used K-Fold Cross Validation. Where value of K is 10. It is found in previous studies, that 10-fold cross validation is better for limited size dataset [31]. As in our case we have also dealt with relatively smaller dataset, we have used 10-fold cross validation.

Classifier	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Specificity**	**MAPE**
SVM	83.77%	0.82	0.82	0.84	0.83	0.12
Naïve Bayes	78%	0.79	0.77	0.79	0.93	0.19
K-Neighbors	83.1%	0.82	0.79	0.79	0.93	0.16
Decision Tree	85.5%	0.84	0.83	0.84	0.95	0.23
Random Forest	89%	0.87	0.88	0.87	0.96	0.07
Logistic Regression	83%	0.81	0.82	0.81	0.94	0.16
Gradient Tree	91%	0.90	0.90	0.90	0.98	0.09
AdaBoost	83.5%	0.81	0.83	0.82	0.94	0.15

Table 4.2: Performance Evaluation of ML Algorithms

Classifier	**Accuracy**	**Standard Deviation**
SVM	91.77%	6.78%
Naïve Bayes	87.84%	6.61%
K-Neighbors	88.17%	7.83%
Decision Tree	90.71%	6.02%
Random Forest	91.21%	7.22%
Logistic Regression	92.14%	6.55%
AdaBoost	82.17%	9.25%
Gradient Tree	92.27%	7.55%

Table 4.3: Performance Evaluation after 10-fold cross validation

From analyzing 4.2 , it is seen that, Gradient tree outperformed other 7 algorithms with 91% accuracy. It also stands alone with better precision, recall and F1 score. From table 4.3 we can see, after 10-Fold cross validation four algorithms stand out

among 8 ML algorithms which are Random Forest, SVM, Logistic Regression and Gradient tree with accuracy of 91.21%, 91.77%, 92.14% and 92.27% respectively. In this case also, Gradient Tree performs slightly better then others. Performance evaluation is also displayed in figure 4.12.

In previous works, Vora [16] classified severity of COPD only, suggested SVM and KNN for better result with very good accuracy, though the dataset had narrow diversity and also no clinical feature had been taken care of. Dimitri [17] had observed 97% precision for COPD but 80.3% for Asthma cases with Random Forest classifier where sample size was too small consists of 133 samples only. Goto [18] had observed 3202 patients of Asthma and COPD exacerbation and also compared performance of various ML models' performance for classifying diseases for both critical care and hospitalization outcomes. For critical care and hospitalization outcome, he proposed Random Forest Classifier with accuracy of 80% and 83% respectively. The work of Yahyaoui [20] had focused on predicting Pneumonia and Asthma exacerbation. They received detection accuracy of 90% with KNN approach and 84.35% with DNN approach. In our research, we believe there is diversity in our dataset, and we have received good accuracy and precision with Gradient Tree.

4.5 Model Justification

In this section, we will consider our used ML models to visualize their accuracy and error rate. It is measured with the help of confusion matrix. Confusion matrix of the test cases is plotted for error analysis purpose. Here, class 0, class 1, class 2, class 3 represents COPD, Bronchial Asthma, Pneumonia and none cases respectively. As Gradient Tree has outnumbered other models in our case, figure 4.11 is the confusion matrix through which we will evaluate it's performance. Among 94 test cases it has successfully predicted 24 cases of COPD, 22 cases of Bronchial Asthma and 19 cases of Pneumonia. It has also identified none cases of 21. 4 cases of Pneumonia was predicted as COPD and one case of Bronchial Asthma and COPD had been predicted wrong. The number of

inaccurate predictions is comparatively quite small given how closely diseases are symptomized.

With Gradient Tree, We have achieved 91% accuracy following that error rate is only 9%. We have achieved 90% for precision, Recall, F1-score respectively. Gradient Tree boosting is proved to be good for small size dataset and though it has been prone to overfitting but we have handled parameters with sincerity to avoid that. We have gotten good accuracy. In this case learning rate is kept lower for keeping the model robust. Also the number of estimator is 100 in our case with learning rate 0.25. After cross validation the accuracy has increased to 92.27%

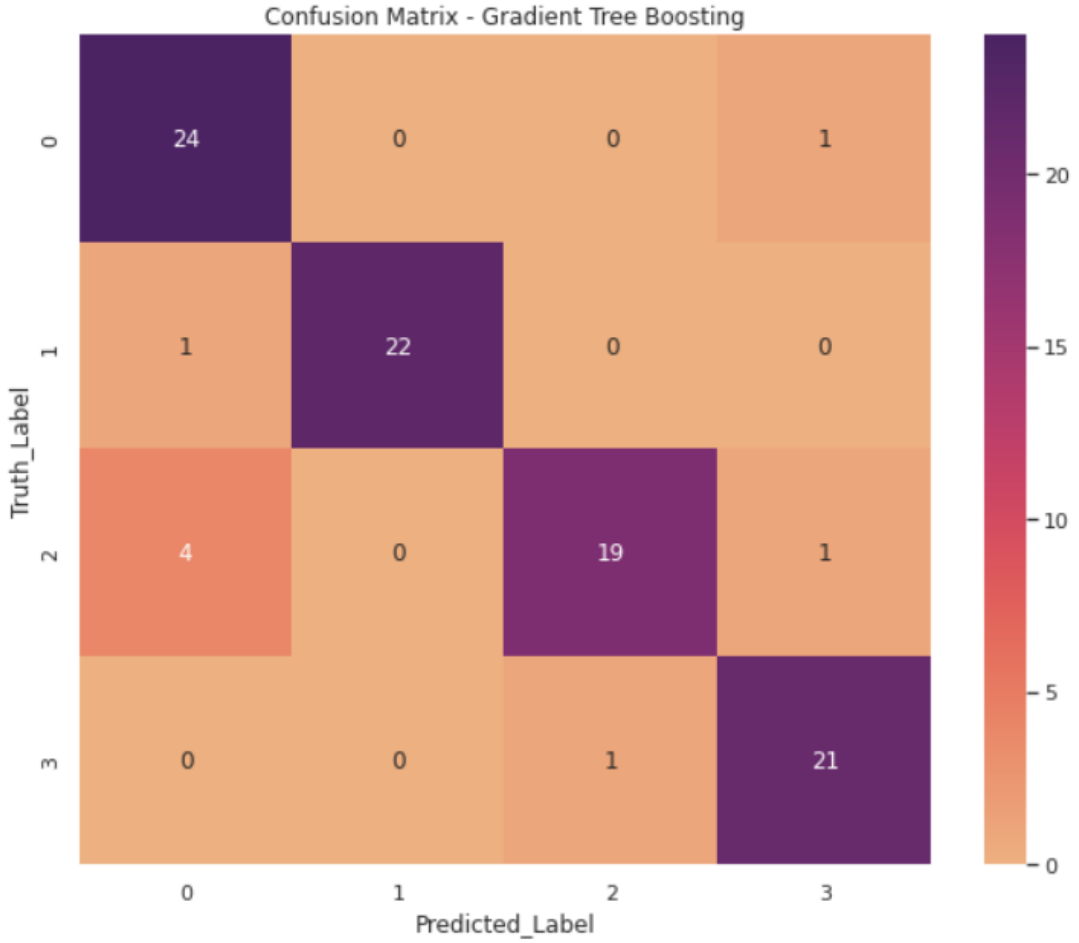


Figure 4.11: Confusion Matrix (Gradient Tree Boosting)

We have also examined Principle Component Analysis, where the 27 features are compressed to 11 key features. After splitting the dataset by 75% train set and 25% test set, we have applied pca to all eight model. And the achieved result is discussed in 4.4

Classifier	Accuracy	Precision	Recall	F1 Score
SVM	82.5%	0.84	0.82	0.84
Naïve Bayes	79%	0.80	0.79	0.79
K-Neighbors	83%	0.83	0.83	0.82
Decision Tree	79%	0.79	0.78	0.78
Random Forest	84%	0.84	0.83	0.83
Logistic Regression	84%	0.84	0.84	0.84
Gradient Tree	83%	0.83	0.82	0.82
AdaBoost	79%	0.78	0.77	0.77

Table 4.4: Performance Evaluation (after applying pca)

We can see from here that the result is fallen. Simply finding a low-dimension set of variables that summarize data is the basic premise of PCA. PCA just considers the variation of each feature since it is logical to expect that features with large variance are more likely to have a good split between classes. PCA does not take class information into consideration. The algorithm actually creates a new set of attributes by combining the existing ones. However, there are situations when features with high variation have little effect on predicting the result, which is happened in our case.

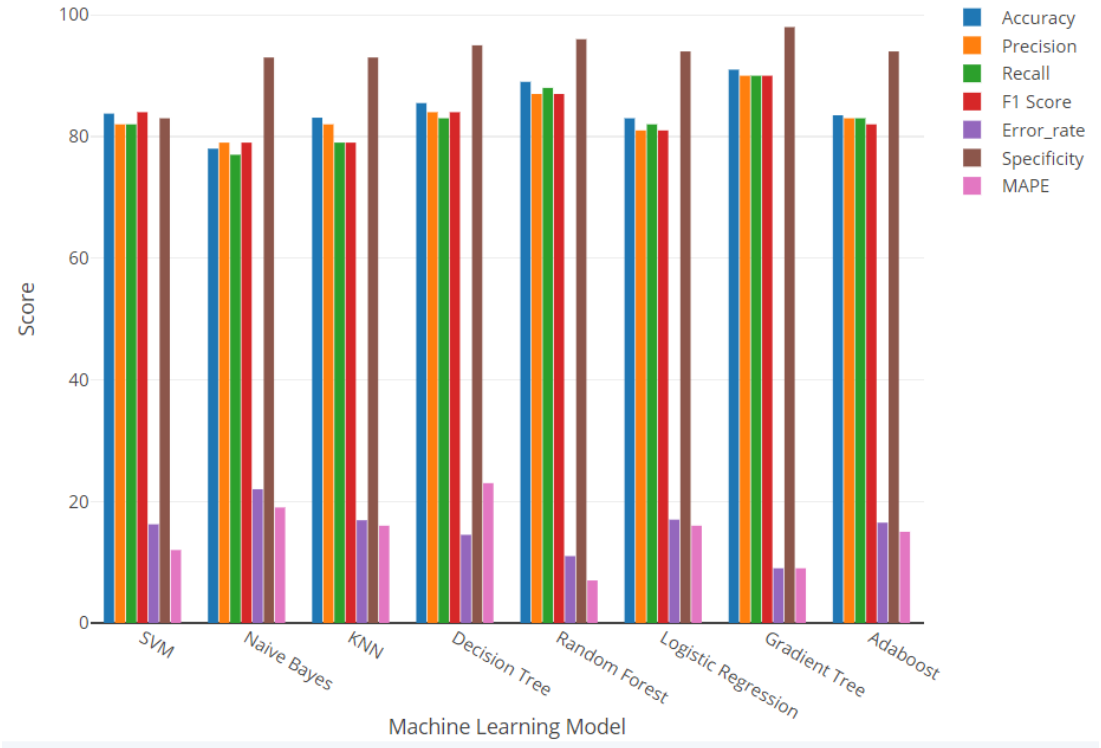


Figure 4.12: Performance evaluation

4.6 Health Care Application

In this section the interface and feature of our mobile application where we have deployed our model with the help of Django micro-service will be shown and displayed.

4.6.1 User Interface

Our user interface is consists of home page and login sign-up functionality. Users profile can be created in this interface. who will further proceed to share their symptoms with us. The figure 4.13 displays the rest.

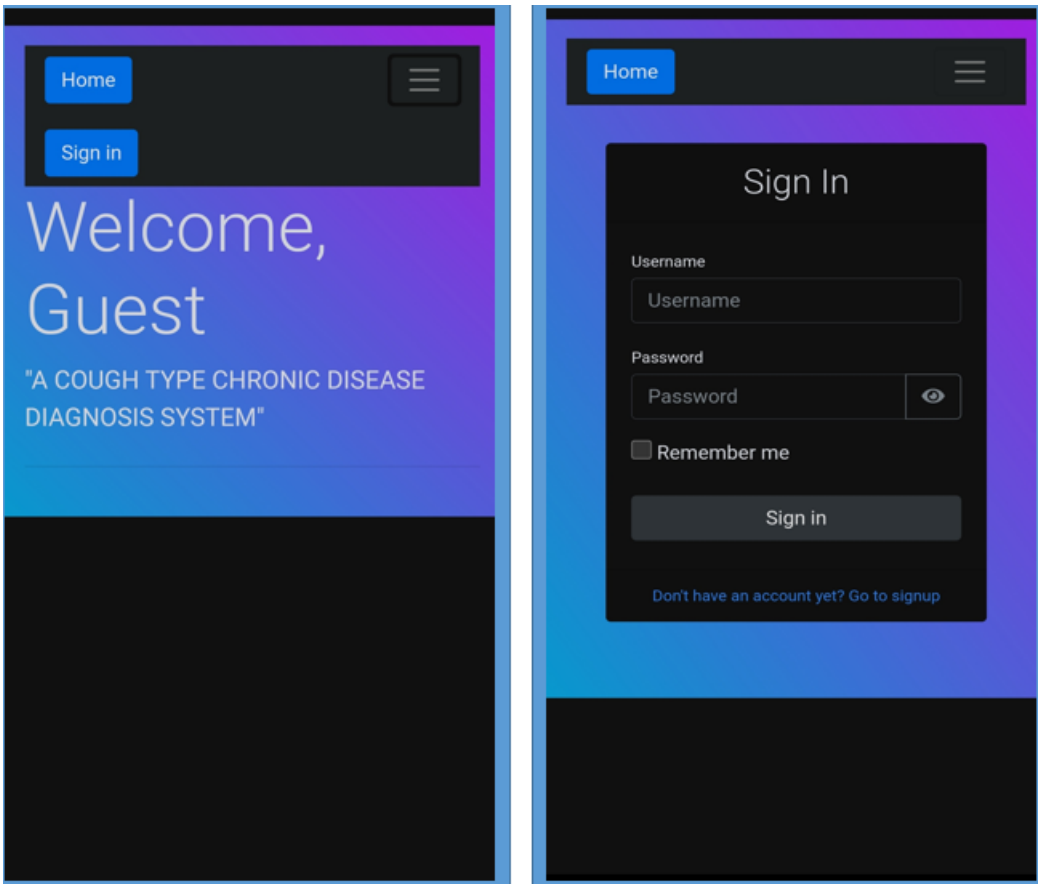
4.6.2 Prediction Interface

In this section, where figure 4.14 shows that, the logged in users will able to share their clinical and visual symptoms with our system by filling up the form. And their given information will be taken to the backend where the model is deployed through API. Then the model's result will be shown through the application via another API.

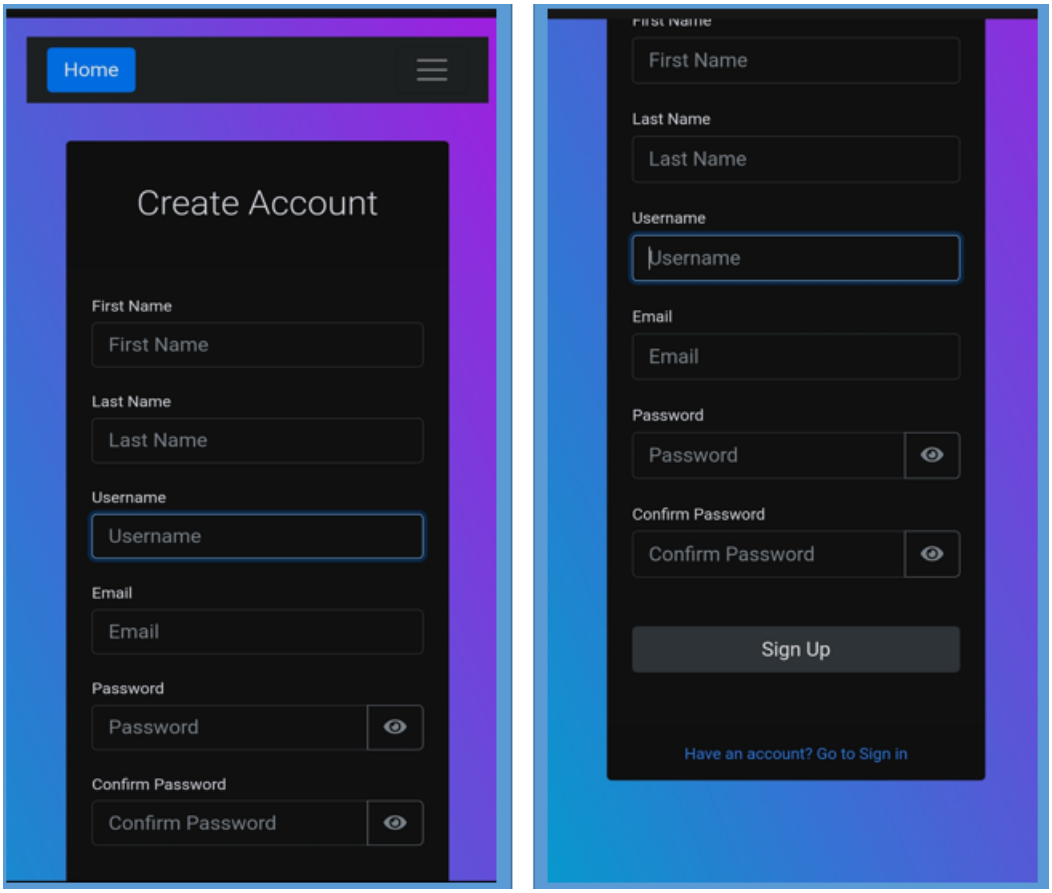
Then the application will show weather they have been affected with such diseases or not. Then it will move to the next part and where patient can take appointment of nearby doctor or can get doctors' phone number through app to directly contact with them. The appointment part us also integrated with the backend and using API doctors' information is shown through the application interface. So that it will be simple solution for rural people to connect with the doctor.

4.6.3 Doctor Appointment

In subfigure 4.14b Doctor appointment system is shown where nearby doctors' chamber address and contact information will be shared according to the appointment form of the patient through who which they can take medical help from the doctors in both online and offline method. Figure 4.15 shows the doctors' information which is available in our system.

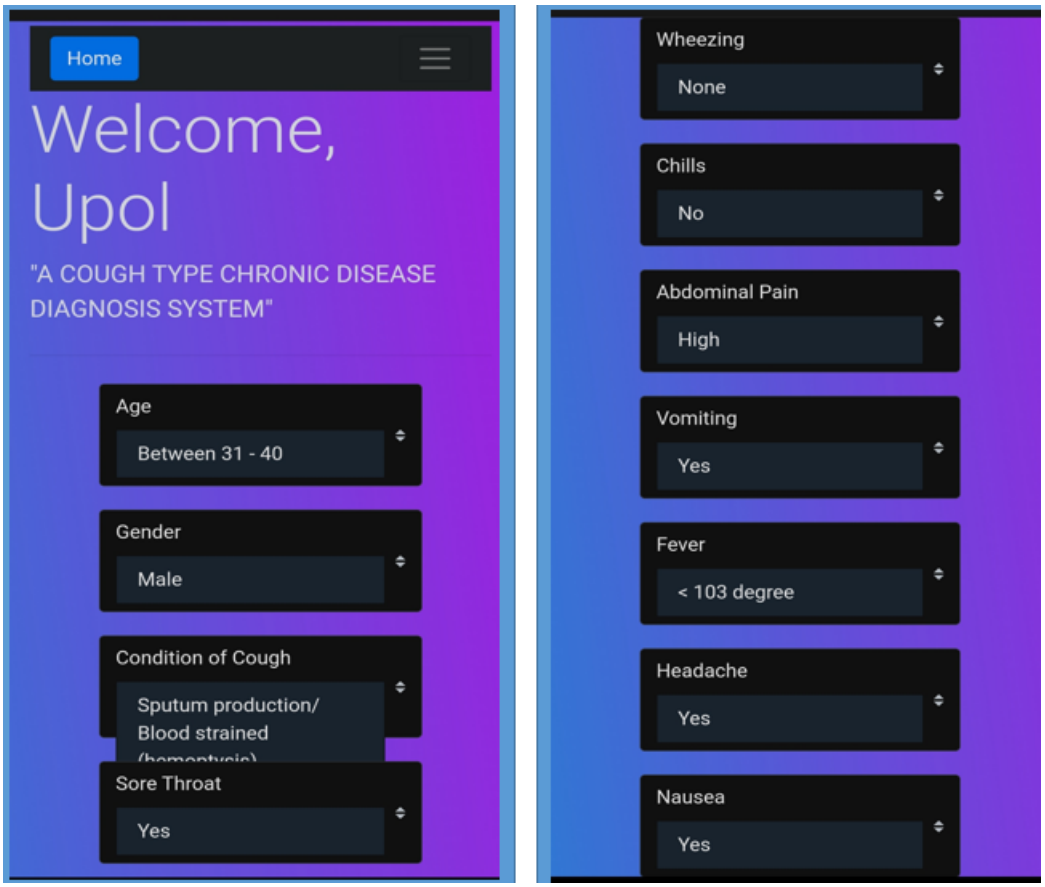


(a)

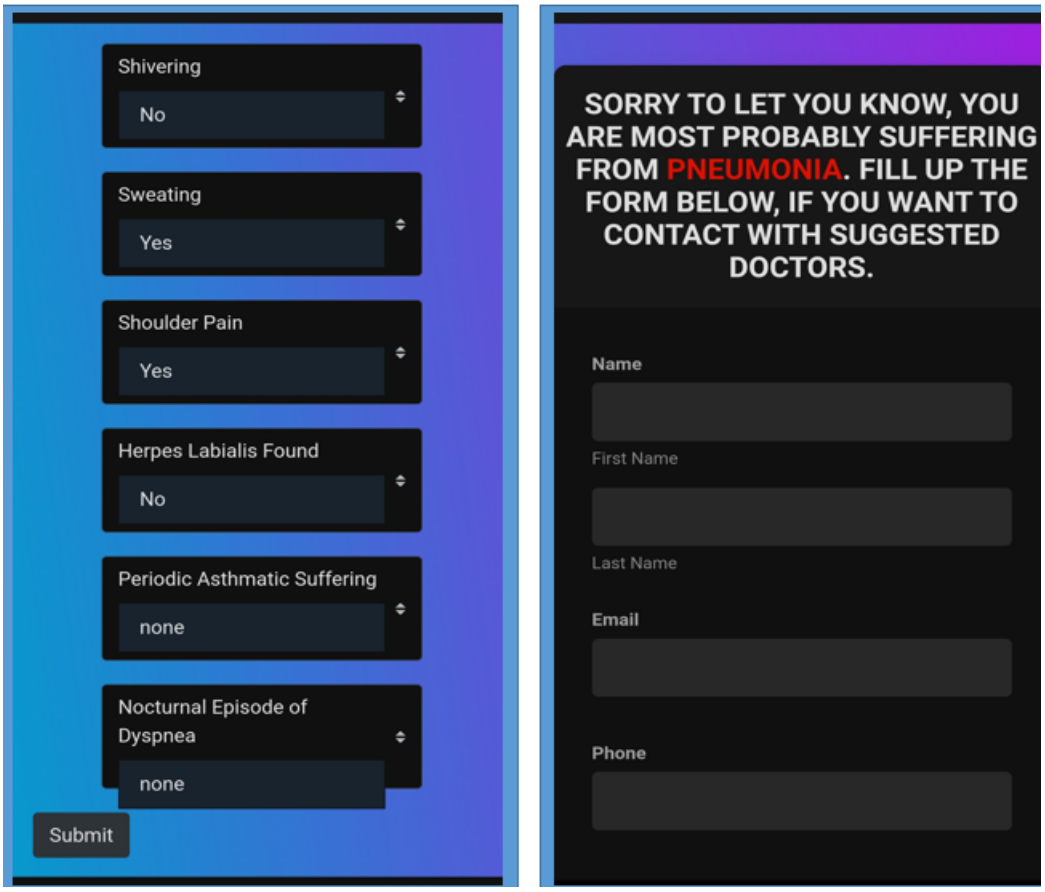


(b)

Figure 4.13: User Interface



(a)



(b)

Figure 4.14: User Interface

Through the application doctors will be able to connect with the patients. The treatment will be easier.

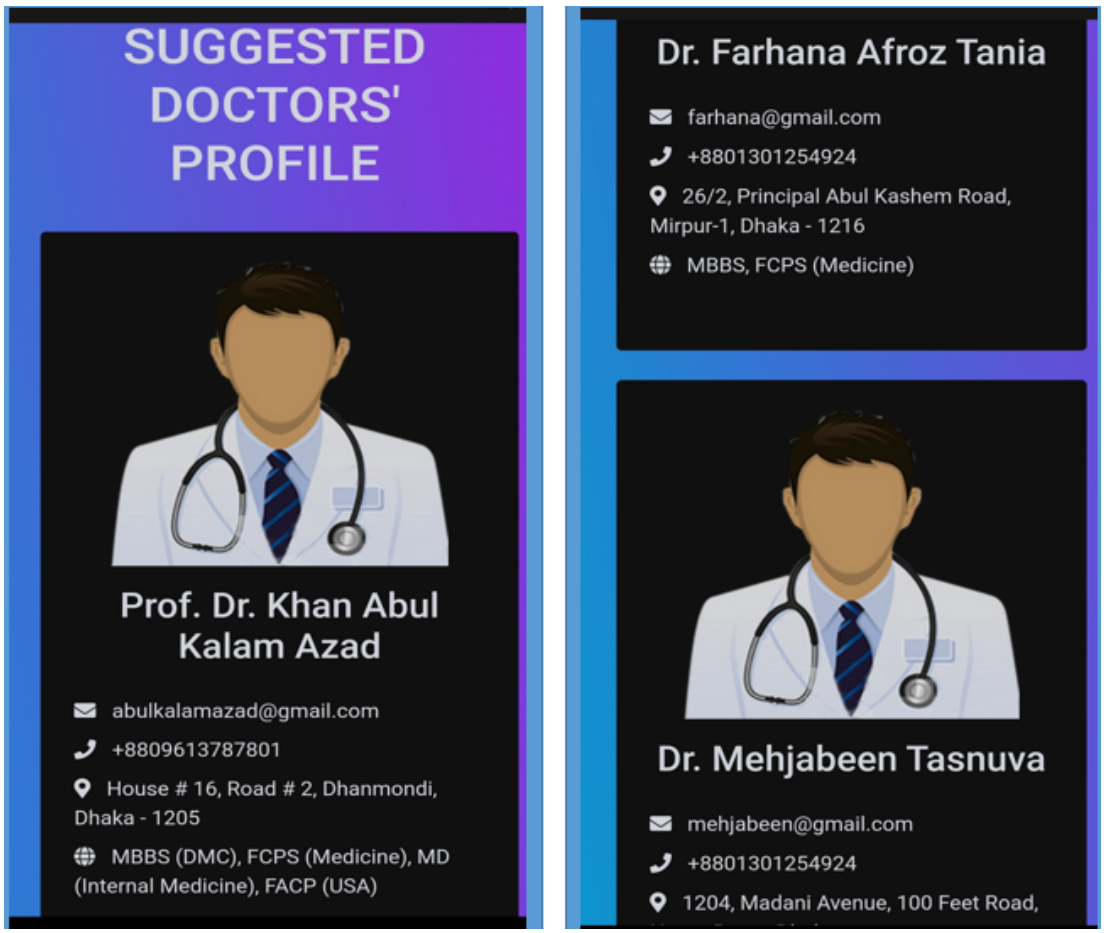


Figure 4.15: Doctor information

4.6.4 User Review

Our system is reviewed by 70 users, our login system has gotten good reviews from 55 person and 47, 18, 4, 1 persons have rated our prediction system by giving good, medium, no comment rating respectively.

4.7 Conclusion

In theory, a model's anticipated performance can provide insight into how well it works with unseen data. Making predictions based on potential outcomes is the main concern we strive to solve. It's crucial to take the context into account when choosing a metric because each machine learning algorithm uses a different dataset to try to solve a challenge with a different purpose.

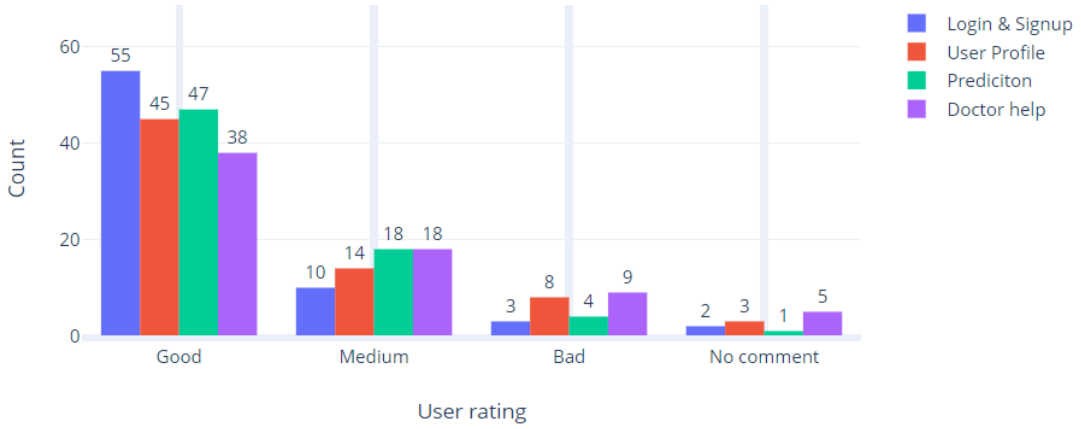


Figure 4.16: User Review

We have described the outcome that was obtained utilizing our methodology inside this chapter. The dataset is explained at the beginning of the chapter. Additionally, the results of the data analysis and significant findings are presented. The results of the model evaluation and our research's findings are then shown. In addition, a comparative assessment of our results with those of earlier researchers is provided. Also shown is the interface figure that is intended to improve usability. Our effort comes to a close in the following chapter with a conclusion.

Chapter 5

Conclusion

5.1 Conclusion

In order to improve patient outcomes, machine learning techniques are used in the healthcare industry. These methods have many advantages, but they also come with numerous disadvantages. The three main areas where machine learning is applied are medical imaging, natural language interpretation of medical records, and common knowledge. The diagnosis, classification, and prediction are major topics in many of these disciplines. Although there is currently a vast system of medical devices that produce data, there is also no infrastructure that allows for the effective use of the data. The multiplicity of formats used for health information could make data processing more difficult and contribute to more noise.

As an instance, consider the Undiagnosed Diseases, which receives funding from the federal government and collaborates with elite medical schools like Harvard and Stanford [32] to identify and treat rare disorders. After examining a vast quantity of data from the US and the UK, developers at Google Health used deep learning to build a model that can identify breast cancer. Over 28,000 women in both nations provided information for this study. Surprisingly, compared to board-certified radiologists, the system was capable of learning and identify breast cancer with 5.7% fewer false positives and 9.4% fewer false negatives [33]. The majority of the breast tumors diagnosed by the artificial intelligence model in the Stanford University study were intrusive, and based on the results, it has huge potential to reduce misdiagnosis and medical errors.

By developing an approach to identify cough-type chronic diseases, we attempted

to have an impact on the current healthcare system. The work is completed with the intention of enhancing the current clinical system.

A broad outline of the study was provided in the first chapter. In this chapter, related subjects were introduced to the reader. The problem encountered and how this task was implemented are both briefly discussed. The significance and contribution of the work are highlighted under the section on inspiration.

The study that the researchers conducted into using machine learning for diagnosis of diseases in the healthcare industry is covered in the following chapter. The experimental methods, data gathering procedure, faults, and algorithms used in the research for the results are mentioned. Also various machine learning algorithms which has been proposed by the researchers are addressed. Later, a brief discussion will be made of conventional machine learning methods that are applied in this case.

The strategy employed to achieve the work's objective is thoroughly stated and discussed in chapter 3. The collection of fundamental attributes is where the approach starts. According to the functionality selected after careful consideration, data are collected from patients who are admitted to the hospital. The acquired data is preprocessed and converted to a numeric values for computing. The final component set is then determined by analyzing the numerical data that has been accumulated. Nothing is altered or removed throughout this process. The attributes and data are then loaded into various machine learning models, that are then tested to see which model performs the best. The idea is implemented into a mobile application and tested for usability.

We described in detail in chapter 4, that how our methodology produced the results we saw. At the start of the chapter, the dataset is discussed. Additionally, the findings of the data gathering and key findings from the data have been analyzed. The results of the model evaluation are then presented, together with our research's conclusions. We also provide a quick description of how our study compares to those of earlier scholars. In order to increase usefulness, an interface representation is also shown.

5.2 Future Work

This study has a lot of potential to be upgraded in future. They are outlined below:

1. **Dataset Quantity:** There are only 373 instances in the dataset. If there had been additional data available for the procedure, the analysis's findings would have been more precise.
2. **Dataset Variance:** Not only should the number of data instances be increased, but data should also have been gathered from a variety of sources to enhance diversity.
3. **Applying RL:** Implementation of Reinforcement Learning agent instead of conventional machine learning model can be done and discussed.

References

[1] R. S. Irwin and J. M. Madison, 'The diagnosis and treatment of cough,' *New England Journal of Medicine*, vol. 343, no. 23, pp. 1715–1721, 2000 (cit. on p. 1).

[2] D. A. Groneberg, D. Nowak, A. Wussow and A. Fischer, 'Chronic cough due to occupational factors,' *Journal of Occupational Medicine and Toxicology*, vol. 1, no. 1, pp. 1–10, 2006 (cit. on p. 1).

[3] R. Cavallazzi and J. Ramirez, 'Community-acquired pneumonia in chronic obstructive pulmonary disease,' *Current Opinion in Infectious Diseases*, vol. 33, no. 2, pp. 173–181, 2020 (cit. on p. 1).

[4] S. R. Zaidi and J. D. Blakey, 'Why are people with asthma susceptible to pneumonia? a review of factors related to upper airway bacteria,' *Respirology*, vol. 24, no. 5, pp. 423–430, 2019 (cit. on p. 1).

[5] F. Ferdous *et al.*, 'Pneumonia mortality and healthcare utilization in young children in rural bangladesh: A prospective verbal autopsy study,' *Tropical Medicine and Health*, vol. 46, no. 1, p. 17, May 2018, ISSN: 1349-4147. DOI: 10.1186/s41182-018-0099-4. [Online]. Available: https://doi.org/10.1186/s41182-018-0099-4 (cit. on p. 1).

[6] W. H. Organization, *6 most common diseases during monsoon: Dengue, malaria, typhoid, cholera and more- symptoms and prevention*, (accessed 26 July 2022). [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd) (cit. on p. 1).

[7] WORLD HEALTH RANKINGS, *BANGLADESH: INFLUENZA AND PNEUMONIA*, https://www.worldlifeexpectancy.com/bangladesh-influenza-pneumonia?fbclid=IwAR3uPn76nWmS9aQwxtOkveq3uxRBd2E9xySRHg8PDjrB45hA1v1Jt7zCCp8, (accessed 09 July 2021) (cit. on p. 2).

[8] R. Bhardwaj, A. R. Nambiar and D. Dutta, 'A study of machine learning in healthcare,' in *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, IEEE, vol. 2, 2017, pp. 236–241 (cit. on p. 2).

[9] J. Wiens and E. S. Shenoy, 'Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology,' *Clinical Infectious Diseases*, vol. 66, no. 1, pp. 149–153, 2018 (cit. on p. 2).

[10] A. Callahan and N. H. Shah, 'Machine learning in healthcare,' in *Key Advances in Clinical Informatics*, Elsevier, 2017, pp. 279–291 (cit. on p. 2).

[11] G. Caramori *et al.*, 'Molecular mechanisms of respiratory virus-induced asthma and copd exacerbations and pneumonia,' *Current medicinal chemistry*, vol. 13, no. 19, pp. 2267–2290, 2006 (cit. on p. 4).

[12] D. C. Sanchez-Ramirez and D. Mackey, 'Underlying respiratory diseases, specifically copd, and smoking are associated with severe covid-19 outcomes: A systematic review and meta-analysis,' *Respiratory medicine*, vol. 171, p. 106 096, 2020 (cit. on p. 4).

[13] C. M. A. d. O. Lima, *Information about the new coronavirus disease (covid-19)*, 2020 (cit. on p. 4).

[14] T. M. Mitchell, *Machine learning*. MacGraw-Hill, 1997 (cit. on p. 4).

[15] M. Atul Kaushal, M. Ken Abrams, M. David Sklar and M. Bill Fera, *The future of artificial intelligence in health care*, Dec. 2019. [Online]. Available: `https://www.modernhealthcare.com/technology/future-artificial-intelligence-health-care` (cit. on p. 5).

[16] S. Vora and C. Shah, 'Copd classification using machine learning algorithms,' *Int. Res. J. Eng. Technol*, vol. 6, pp. 608–611, 2019 (cit. on pp. 8, 40).

[17] D. Spathis and P. Vlamos, 'Diagnosing asthma and chronic obstructive pulmonary disease with machine learning,' *Health informatics journal*, vol. 25, no. 3, pp. 811–827, 2019 (cit. on pp. 9, 40).

[18] T. Goto, C. A. Camargo Jr, M. K. Faridi, B. J. Yun and K. Hasegawa, 'Machine learning approaches for predicting disposition of asthma and copd exacerbations in the ed,' *The American journal of emergency medicine*, vol. 36, no. 9, pp. 1650–1654, 2018 (cit. on pp. 9, 40).

[19] K. Stokes *et al.*, 'A machine learning model for supporting symptom-based referral and diagnosis of bronchitis and pneumonia in limited resource settings,' *Biocybernetics and Biomedical Engineering*, vol. 41, no. 4, pp. 1288–1302, 2021 (cit. on p. 9).

[20] A. Yahyaoui and N. Yumuşak, 'Deep and machine learning towards pneumonia and asthma detection,' in *2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, 2021, pp. 494–497. DOI: `10.1109/3ICT53449.2021.9581963` (cit. on pp. 10, 40).

[21] M. Toğaçar, B. Ergen, Z. Cömert and F. Özyurt, 'A deep feature learning model for pneumonia detection applying a combination of mrmr feature selection and machine learning models,' *Irbm*, vol. 41, no. 4, pp. 212–222, 2020 (cit. on p. 10).

[22] J. Finkelstein and I. C. Jeong, 'Machine learning approaches to personalize early prediction of asthma exacerbations,' *Annals of the New York Academy of Sciences*, vol. 1387, no. 1, pp. 153–165, 2017 (cit. on p. 10).

[23] C. Cortes and V. Vapnik, 'Support-vector networks,' *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995 (cit. on p. 11).

[24] X. Wu *et al.*, 'Top 10 algorithms in data mining,' *Knowledge and information systems*, vol. 14, no. 1, pp. 1–37, 2008 (cit. on p. 11).

[25] A. McCallum, 'Graphical models, lecture2: Bayesian network represention,' *PDF). Retrieved*, vol. 22, 2019 (cit. on p. 12).

[26] E. Fix, *Discriminatory analysis: nonparametric discrimination, consistency properties*. USAF school of Aviation Medicine, 1985, vol. 1 (cit. on p. 12).

[27] J. Tolles and W. J. Meurer, 'Logistic regression: Relating patient characteristics to outcomes,' *Jama*, vol. 316, no. 5, pp. 533–534, 2016 (cit. on p. 12).

[28] T. K. Ho, 'Random decision forests,' in *Proceedings of 3rd international conference on document analysis and recognition*, IEEE, vol. 1, 1995, pp. 278–282 (cit. on p. 13).

[29] A. Torres, F. Blasi, N. Dartois and M. Akova, 'Which individuals are at increased risk of pneumococcal disease and why? impact of copd, asthma, smoking, diabetes, and/or chronic heart disease on community-acquired pneumonia and invasive pneumococcal disease,' *Thorax*, vol. 70, no. 10, pp. 984–989, 2015 (cit. on p. 19).

[30] R. K. Shukla, S. Kant, B. Mittal and S. Bhattacharya, 'Comparative study of gst polymorphism in relation to age in copd and lung cancer,' *Tuberk Toraks*, vol. 61, no. 4, pp. 275–282, 2013 (cit. on p. 31).

[31] H. L. Vu, K. T. W. Ng, A. Richter and C. An, 'Analysis of input set characteristics and variances on k-fold cross validation for a recurrent neural network model on waste disposal rate estimation,' *Journal of Environmental Management*, vol. 311, p. 114 869, 2022 (cit. on p. 39).

[32] P. Agyemang-Gyau, 'Artificial intelligence in healthcare and the implications for providers,' *On-Line Journal of Nursing Informatics*, vol. 25, no. 2, 2021 (cit. on p. 48).

[33] J. Abbasi, 'Artificial intelligence improves breast cancer screening in study,' *JAMA*, vol. 323, no. 6, pp. 499–499, 2020 (cit. on p. 48).