ABSTRACT

The proposed model uses data mining for deriving useful information and relationship among data items i.e. patient’s data, thereby predicting the different lung diseases and determining the severity of the lung diseases. Nowadays data mining is used in the medical sector for saving patients’ lives. Data mining has two primary goals. One is prediction which involves finding unknown and future values of some variables or fields of interest using some fields in the dataset whose value is known. The other is description which refers to finding pattern in the dataset that can be interpreted by humans.

Our project aims at developing a robust model which will perform data mining and analytics on the historic dataset of patients having lung related problems by considering their medical test parameters and general parameters. The model is developed by using several classification algorithms like Logistic regression, Binary regression, Decision Tree, SVM, Naive Bayes, Artificial neural network.

S.No. Title of Chapter Page No.

CHAPTER 1

INTRODUCTION

1.1 Motivation

Our project aims at early detection of lung diseases.

The tests those are currently conducted require a lot of co-operation from the patients. If the patients are not able to provide correct input, the test results will not calculated accurately and patients receives medication which is not appropriate for their problems. Currently for the detection of lung diseases some medical tests are performed. One of major tests is pulmonary function tests (PFTs).

This test is used to measure lung volume, capacity, rates of flow, and gas exchange. There are 2 types of lung disorders:

- Obstructive: This occurs when air has trouble flowing out of the lungs.
- Restrictive: This occurs when the lung tissue and/or chest muscles can’t expand enough.

Through our model, we will perform data mining and analytics on real time dataset containing historical test data and additional general data of thousands of patients. The model will make robust predictions of various types of lung diseases at an early stage, so that patient will be able to get correct medication at correct time.

1.1 Problem Definition

Problem Statement: To develop model using different data classification techniques to predict and analyze different lung diseases.

Goals:

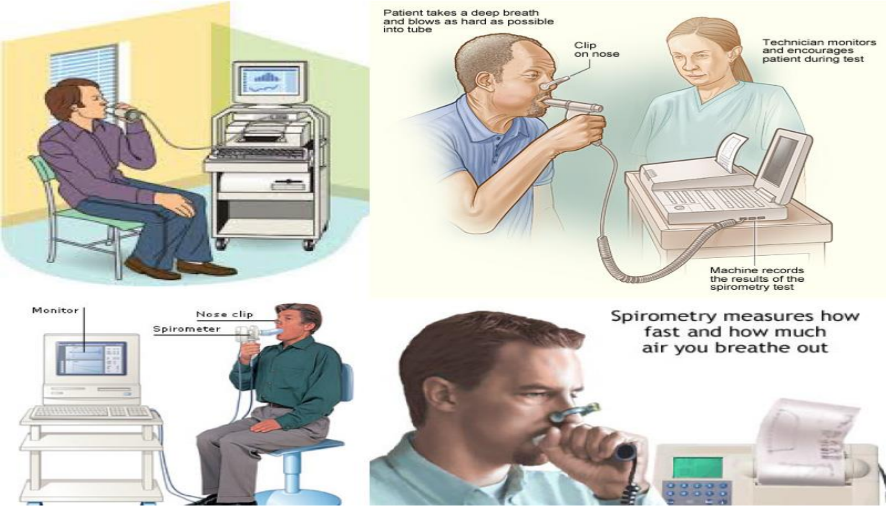
- This application is proposed and developed in detecting particular lung disease accurately using data analytics techniques.
- A huge amount of Medical data is collected and stored every year. Through our model we are aiming to perform data mining and analytics on that data and derive useful information which would be used further for generating reports.
- To provide the system that does not require any external hardware thereby making the system efficient.
- Some medical tests are not feasible for patients below 5 years or patients having heart problems. For such patients our model will consider their general parameters along with the test data to give accurate prediction of their problem for further treatment.

CHAPTER 2

LITERATURE SURVEY:

Pulmonary Function Test:

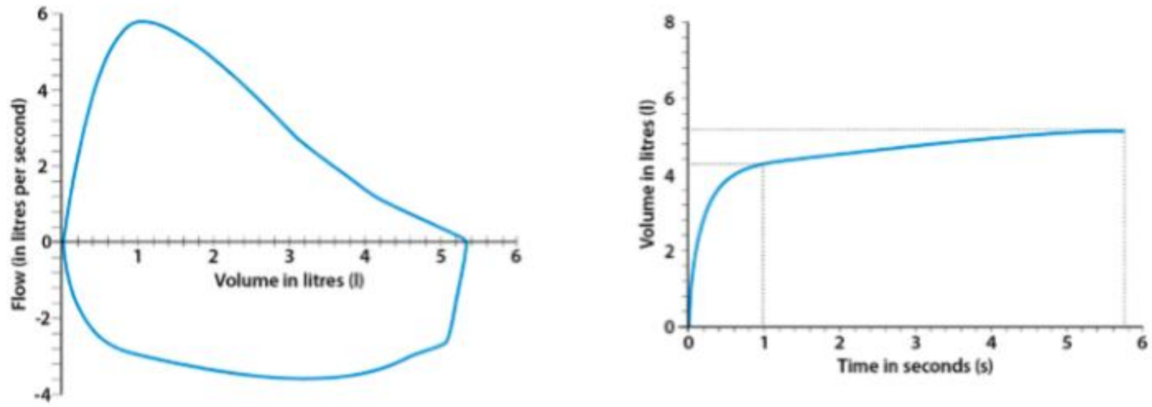
Pulmonary function tests (PFTS) are an important tool in investigating and monitoring of patients with respiratory pathology. They provide information regarding large and small airways, pulmonary parenchyma and other disorders. Pulmonary Function Test consist of various test such as Spirometry etc.


Patient takes a deep breath and blows as hard as possible into tube
Clip on nose
Technician monitors and encourages patient during test
Machine records the results of the spirometry test
Monitor
Nose clip
Spirometer
Spirometry measures how fast and how much air you breathe out

Spirometry:

Spirometry is a simple test which is used to diagnose and monitor certain lung conditions by measuring amount of air you can breathe out in one forced breath. It is carried out using a device called a spirometer, a small machine attached by a cable to a mouthpiece. The spirometry test results are as follows:

An example of a (normal) spirometry result for a hospital spirometer



	Min	Ref	Max	Best	%Ref	SR
FEV1 [L]	3.76	4.31	4.99	4.31	100	0.0
FVC [L]	4.71	5.35	5.81	5.35	100	0.0
VC [L]	4.82	5.47	5.92	5.47	100	0.0
FEV1/VC [%]	68.1	78.8	-	78.8	100	0.0
		Normal range		Your best effort		

Parameters range for different lung conditions

Restrictive stage COPD	FEV1/FVC >= 70% and FEV1 < 80%
Spirometry within normal limits	FEV1/FVC >= 95% and FVC > 80%
Early small airway obstruction	FEF = 25-75 and PEFR < 70%
Mixed blockage	(FEV1/FVC)%pred < 95% and (FVC % pred) < 80%
Moderate Restriction	(FEV1/FVC)%pred < 95% and (FVC % pred) < 64%
Mild Restriction	(FEV1/FVC)%pred < 95% and (FVC % pred) < 80%
Mild Obstruction	(FEV1/FVC)%pred < 95% and (FVC % pred) > 80%
Severe Restriction	(FEV1/FVC)%pred < 95% and (FVC % pred) < 44%

COPD Stages:

Mild	Stage 1	FEV1 % >= 80%
Moderate	Stage 2	50 < FEV1 < 80
Severe	Stage 3	30 < FEV1 < 50

Asthma Stages:

Mild	FEV1 >= 80%	Normal (FEV1/FVC)
Moderate	60 < FEV < 80	Reduced 5% (FEV1/FVC)
Severe	FEV < 60	Reduced 5% (FEV1/FVC)

Normal Readings:

Age group	FEV1/FVC
8-9 years	85%
20-39 years	80%
40-59 years	75%
60-80 years	70%

In paper “Survey on Asthma Prediction Using Classification Technique” early prediction and proper diagnosis of the asthma is described. The approach used in this paper is based on different data mining classification techniques such as Decision tree, Naïve Bayes and ANN. One dependency augmented Naïve Bayes classifier (ODANB) and naïve creedal classifier 2 (NCC2) were used for data pre-processing which gave more efficient decisions. The Dataset consist of certain parameters such as Age, Gender, Air pollution, Alcoholic use, allergy, Generic risk, Smoking, Chest pain and disease symptoms. The data classifiers were applied forming various models and these models were compared to the efficient one among them. While comparing the data classifiers used, they found naïve Bayes classifier with maximum accuracy.

In paper “Prediction model for exacerbations in different COPD patients using Data mining techniques” the predicators of COPD exacerbations depending on COPD population were discovered. The proposed system in this paper is able to overcome the drawback of previous models. The system described in paper is able to distinguish high risk patients based on previous exacerbations results along with specified exacerbation rate by degree of air flow obstruction. Binary regression is one of the data classification technique used for prediction of COPD (Chronic Obstructive Pulmonary Disease). The Exacerbations observed in patients was defined as significant symptom to predict the severity of COPD separately for total and severe exacerbations. The annual Exacerbation rate was predicted using negative binomial regression with total risk time as offset variable and exacerbation rate as outcome. FEV1 predicted, previous exacerbations and disease quality life are important predictors regardless of the COPD severity. Other predictors such as age, low body mass index, cardiovascular disease and emphysema were found useful in secondary care patients.

In paper “Evaluation of Bayesian classifiers in asthma exacerbation prediction after medication discontinuation” evaluates the performance of Bayesian network classifiers used for predicting asthma exacerbation. The main advantage of using Bayesian classifier is that relationships between predicators can be shown graphically. Hence compared to other classifier BNC is more helpful for multifactorial disease such as asthma. In the paper it was observed that semi-naïve network classifier was able to predict high risk of asthma exacerbations in future.

CHAPTER 3

SOFTWARE REQUIREMENT SPECIFICATION

3.1 Introduction

3.1.1 Project Scope:

We are building a model to predict various lung diseases by performing data analytics on a real time dataset containing medical tests results.

Spirometry Test:

It is a simple test used to help diagnose and monitor certain lung conditions by measuring amount of air breathed out in one forced breath**.** Spirometry is the most commonly used for measuring of lung function. It is used to measure the amount (volume) and speed (flow) of air that can be inhaled and exhaled. The most common measurements used are:

- Forced expiratory volume in one second (FEV1). This is the amount of air blown out within one second.
- Forced vital capacity (FVC). The total amount of air that can be blown out in one breath.
- FEV1 divided by FVC (FEV1/FVC). It is the ratio of the total amount of air exhaled out in one breath to the amount of air exhaled out in one second.
- Diseases that could be predicted by the model:
    - Asthma
    - Chronic Obstructive Pulmonary Disease.
    - Pneumonia.
    - Pulmonary Tumor

Dataset Description:

The model will consider general parameters of patients as well besides spirometry test parameters.

Parameters:

- Age.
- Habits: 1. Smoking/non-smoking.
- BMI (weight, height).
- FEV1
- FVC
- FEV1/FVC ratio.
- Lung Age.
- Gender
- Disease
- Severity of Disease

Spirometry Test Parameters:

- Forced vital capacity (FVC): Amount of air forcefully exhaled after quick inhalation.
- Forced expiratory volume (FEV): Amount of air exhaled during the first, second, and third seconds of the FVC test.
- Forced expiratory flow (FEF): Average flow rate during the middle half of the FVC test.
- Peak expiratory flow rate (PEFR): Fastest rate to force air out of lungs.

3.1.2 User Classes and Characteristics

There are mainly two user classes for our project.

1. Project Users
    - The user will enter the general as well as tests data through the user interface.
    - The user will be able to view the report generated by the model and take appropriate measures to accurately detect and diagnose the disease.

2. Model
    - The model will classify the dataset using multiple classification techniques and give the prediction based the best technique.
    - The model will predict the disease depending on the input provided by the user.
    - The model will also predict the severity of the disease and analyze its results.

3.1.3 Assumptions

Assumption

- The prediction made by our model considering all general as well as test parameters would be very precise.
- The dataset used for training the model would be sufficient to build a robust model for deploying it in real time.
- Different classifiers are used for prediction of lung diseases and thereby analyze its performance.

3.2 Functional Requirements:

3.2.1 Input Parameters:

- User will provide input to the system. The input consist of general data as well medical test data of patients.

3.2.2 Data Classification techniques performed on data**:**

- Our model will classify lung diseases like Asthma, COPD etc. It will consider the classification techniques like Multiple Logistic Regression, Naïve-Bayes and Random Forest, SVM depending on various performance measures of the classification technique it will choose best technique and make prediction using the same.

3.2.3 Report generation**:**

- Our system will generate report as an outcome which will contain prediction of lung disease and detailed analysis for the same.

3.2.4 Data analysis and visualization**:**

- The analysis will help us to know about the distribution of classes among the parameters in the dataset. Various data visualization techniques are being used for different parameters.
- Analysis performed to determine best suited algorithm for each disease. Determine severity of patient for disease by using this classifier.

3.3 External Interface Requirements:

3.3.1 User Interface

- The user will enter the input parameters through UI.
- The UI will display the fields regarding the test parameters of various tests and general parameters.
- The UI will display the results generated by the model running in the backend.

3.3.2 Hardware Interfaces

Sr. No.	Parameter	Minimum Requirement	Justification
1	CPU Speed	2.1 GHz Minimum	2.1 GHz required
2	RAM	4 GB Minimum	4 GB Required

3.3.3 Software Interfaces:

Platform:

1. Operating System: Windows 10

2. IDE: Anaconda/Flask

3. Programming Languages: Python

3.4 Non- Functional Requirements:

3.4.1 Performance Requirements –

- The disease prediction should be precise and accurate.
- The data used for training the model should be variable and large.
- Execution of operations must be fast with minimum response time.

3.4.2 Software quality attributes –

- Performance: The performance of our system is the most important attribute of our application. For the system to perform well, the disease prediction should be fast and precise. Thus, the response time of the application should be as less as possible.
- Accuracy: Accuracy of the model should be as high as possible because the prediction results are going to act as assistance for doctors for diagnosis of patients in the correct direction. The algorithms should be able to handle all the possible test cases accurately thereby, providing with the desired output.
- Efficiency: Efficiency of the system is directly interlinked to the performance and accuracy. Better performance and more accuracy together make the system efficient. Loss of any one of the quality will directly affect the efficiency of the system.
- Reliability: It is the most important aspect in quality assurance. It can be defined as the probability of failure free operation of the application in the specified environment for specified time. This application is reliable as far as the basic software and the hardware requirements are fulfilled.
- Usability: Usability is a convenience and practical use of an application. This application is user friendly and it uses GUI which is simple.
- Availability: Application can be available on any of the desktop/laptops.
- Installability: The application can be installed on any desktop or laptop which fulfills all the hardware requirements.

3.5 System Requirements:

3.5.1 Software Requirements:

1. Operating System: Windows 10

2. IDE: Anaconda

3. Programming Languages: Python

3.5.3 Hardware Requirements:

Sr. No.	Parameter	Minimum Requirement	Justification
1	CPU Speed	2.1 GHz Minimum	2.1 GHz required
2	RAM	4 GB Minimum	4 GB Required

3.6 Analysis Model SDLC to be applied:

For our model we are planning to use Agile SDLC model as the model aims to achieve high accuracy of prediction of lung diseases.

Through Agile Methodology we will be able to improve the performance by modifying model's parameters and algorithms.

Agile Method:

The Agile Software Development model was proposed in the mid-1990s. The main advantage of agile model is to adapt to change request quickly. Project is completed rapidly due to the model's agility. Agility is achieved by keeping the necessary activities in project thereby eliminating the ones which could cause time and effort wastage.

Agile Model consist of group of development processes which share basic characteristic with subtle differences among themselves.

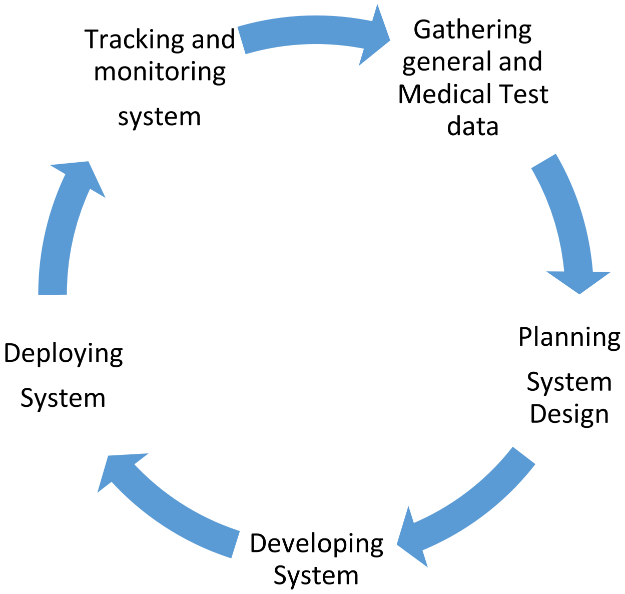
In this model, the requirements are decomposed into several small portions which can be incrementally developed. As the model adopts iterative environment each part is developed over an iteration. The iterations are small and manageable. One iterations can be planned, developed and deployed to customers at a time.

Agile model is the model derived from incremental and iterative process models. Steps involved in agile SDLC models are:

- Requirement gathering
- Requirement Analysis
- Design
- Developing/Coding
- Deployment
- Tracking and monitoring

Principles of Agile model:

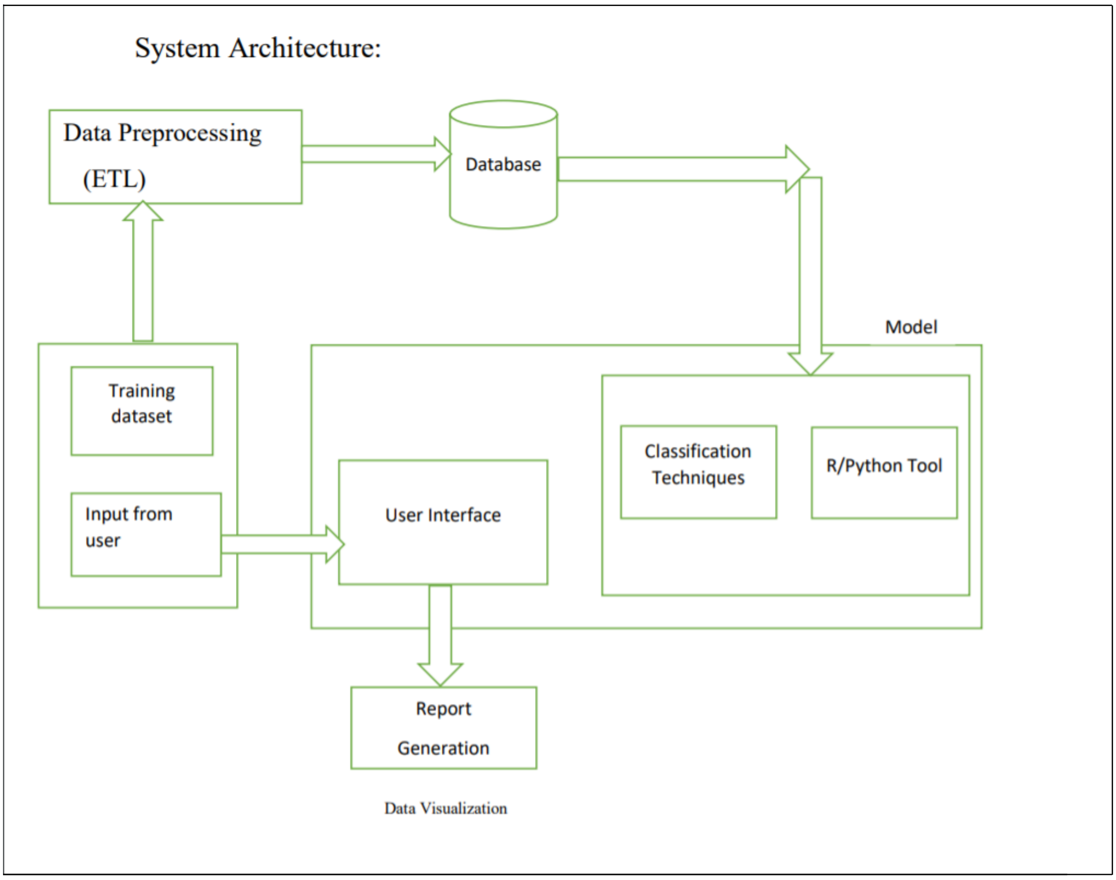
- Agile project requires a customer representative on the team for establishing contacts with customer and for better understanding of customer's requirements. The stakeholders and customer representative review the progress and re-evaluate the requirements at the end of each iteration.
- Agile model mainly relies on working deployed software than the documentation.
- Incremental versions of software are delivered to customer representative frequently
- The change requests can be incorporated efficiently depending on the customer's requirements.
- The team members efficiency and communication are the important factors while forming a team.
- To have collaborative work environment the size of development team should be kept small (5-9 people).
- Pair programming is used for deployment in Agile development process. In pair programming, two programmers work together where they can switch roles after particular intervals.


Gathering general and Medical Test data
Planning
System Design
Developing System
Deploying
System
Tracking and monitoring
system

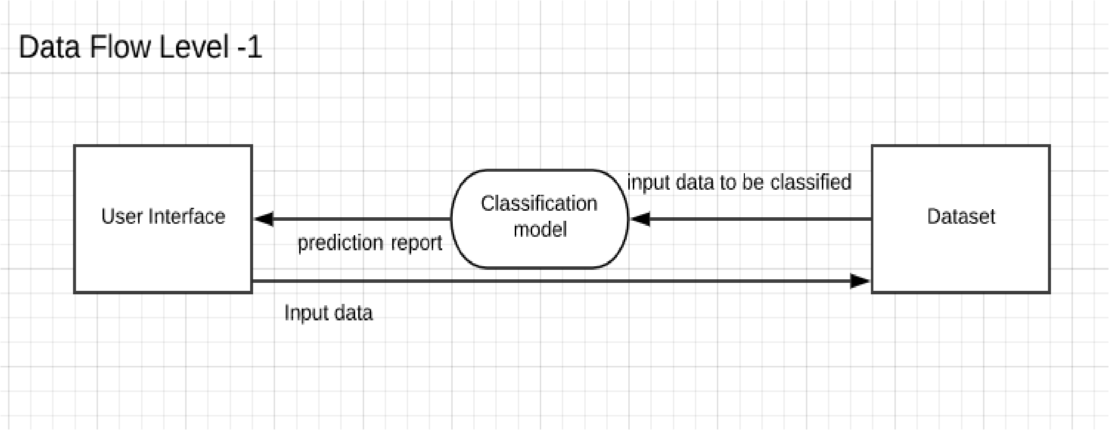
CHAPTER 4

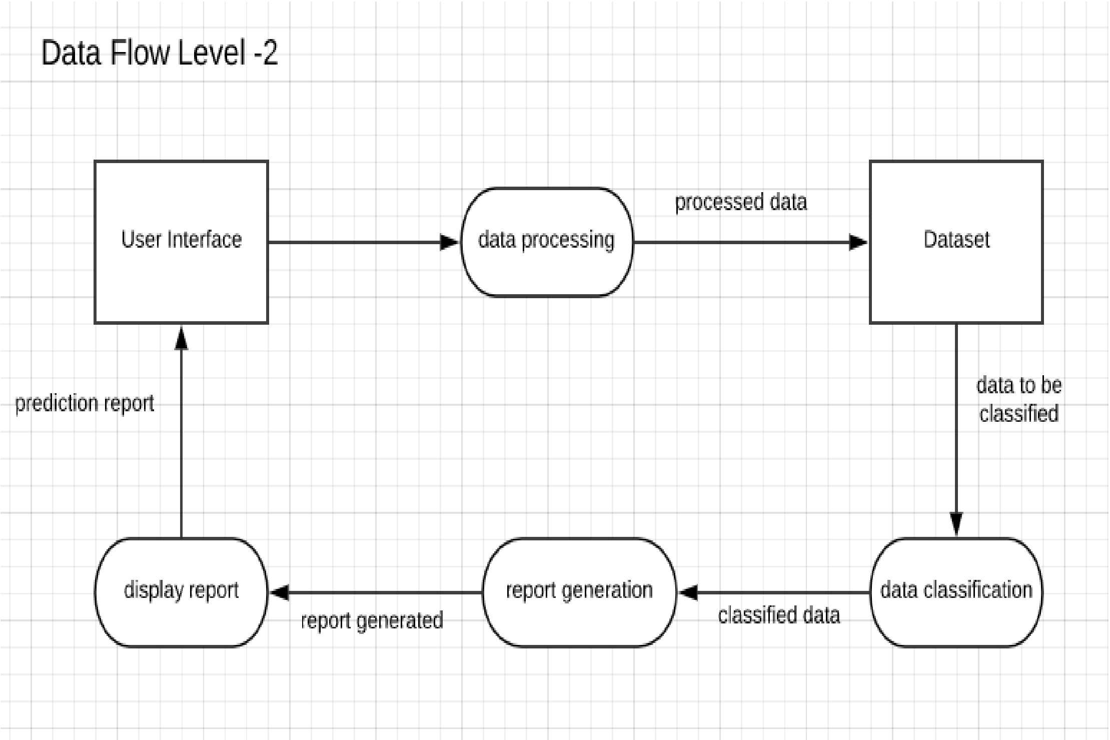
SYSTEM DESIGN

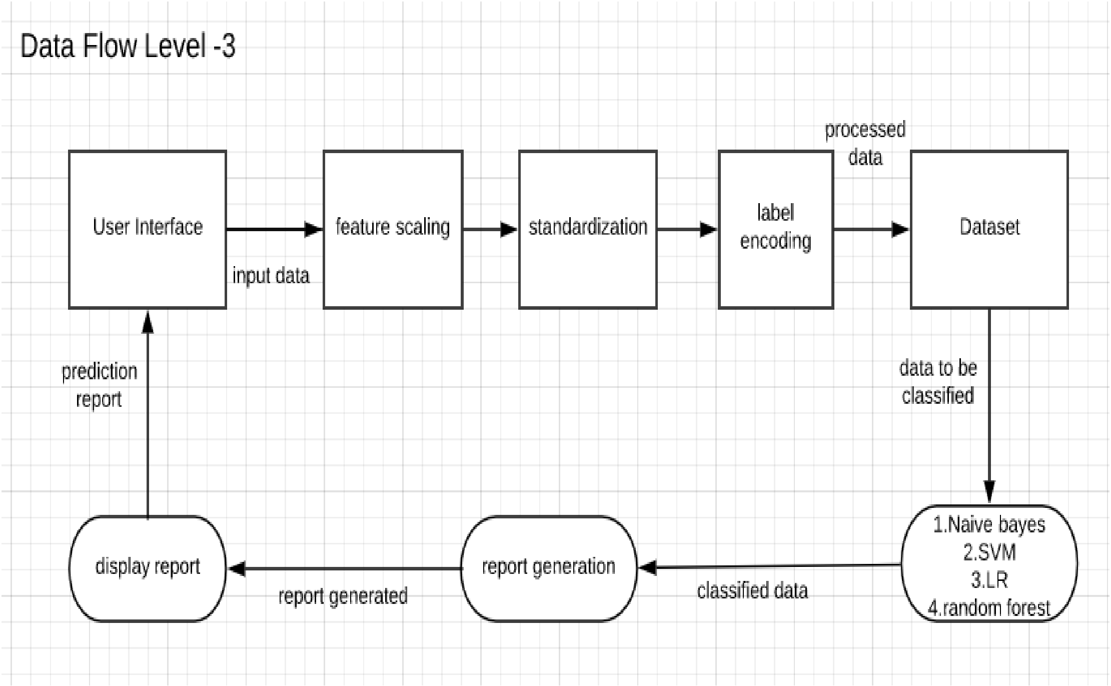
4.1 System architecture:



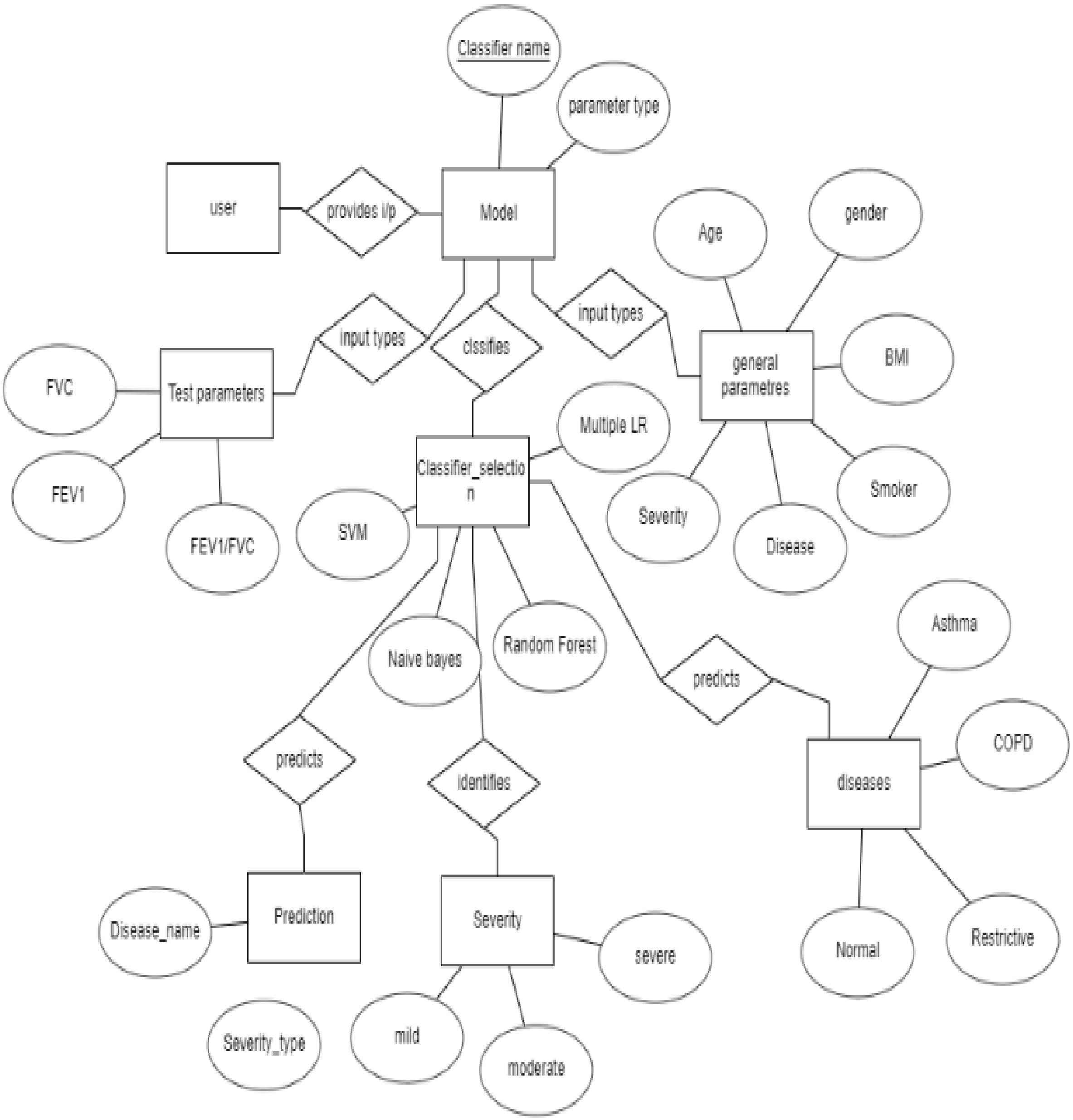
4.2 Data Flow Diagram





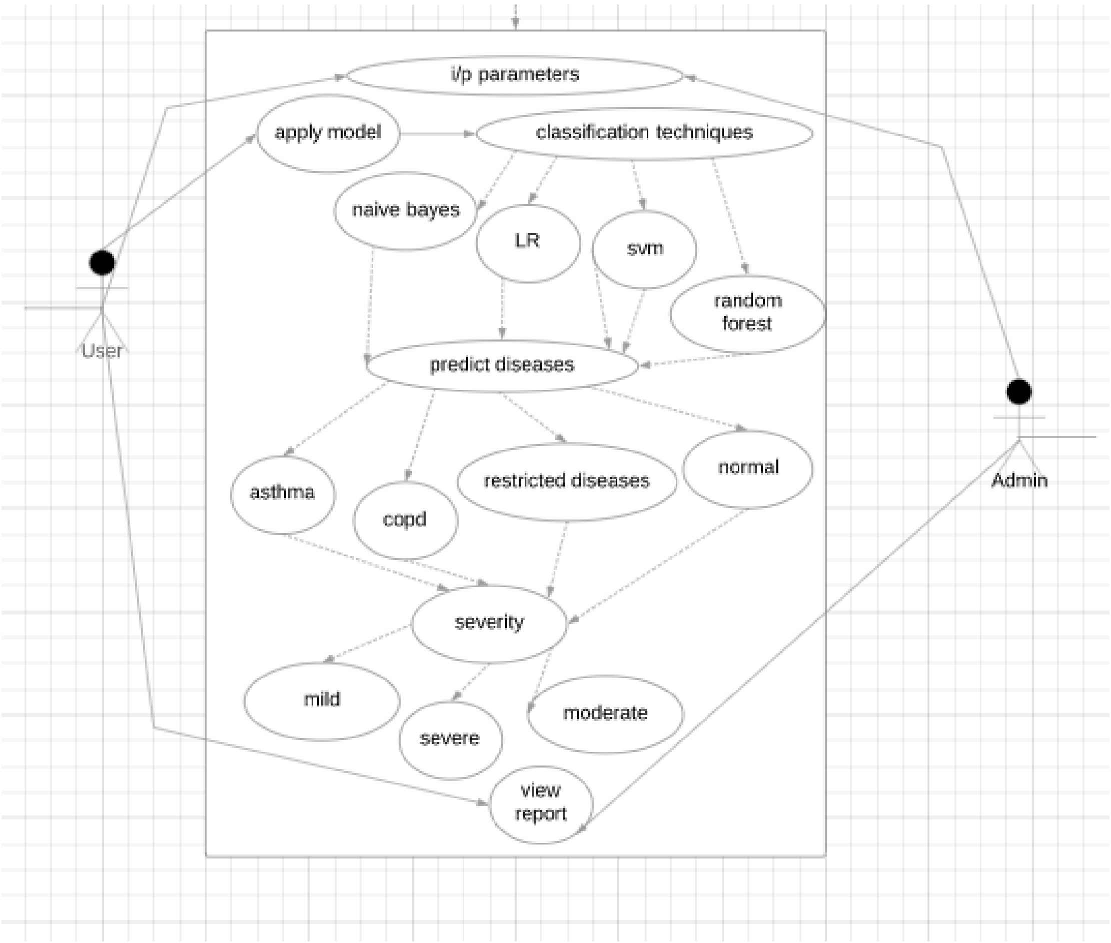

Data Flow Level -3
User Interface
input data
feature scaling
standardization
label encoding
processed data
Dataset
data to be classified
1.Naive bayes
2.SVM
3.LR
4.random forest
classified data
report generation
report generated
display report
prediction report

4.3 ER Diagram :

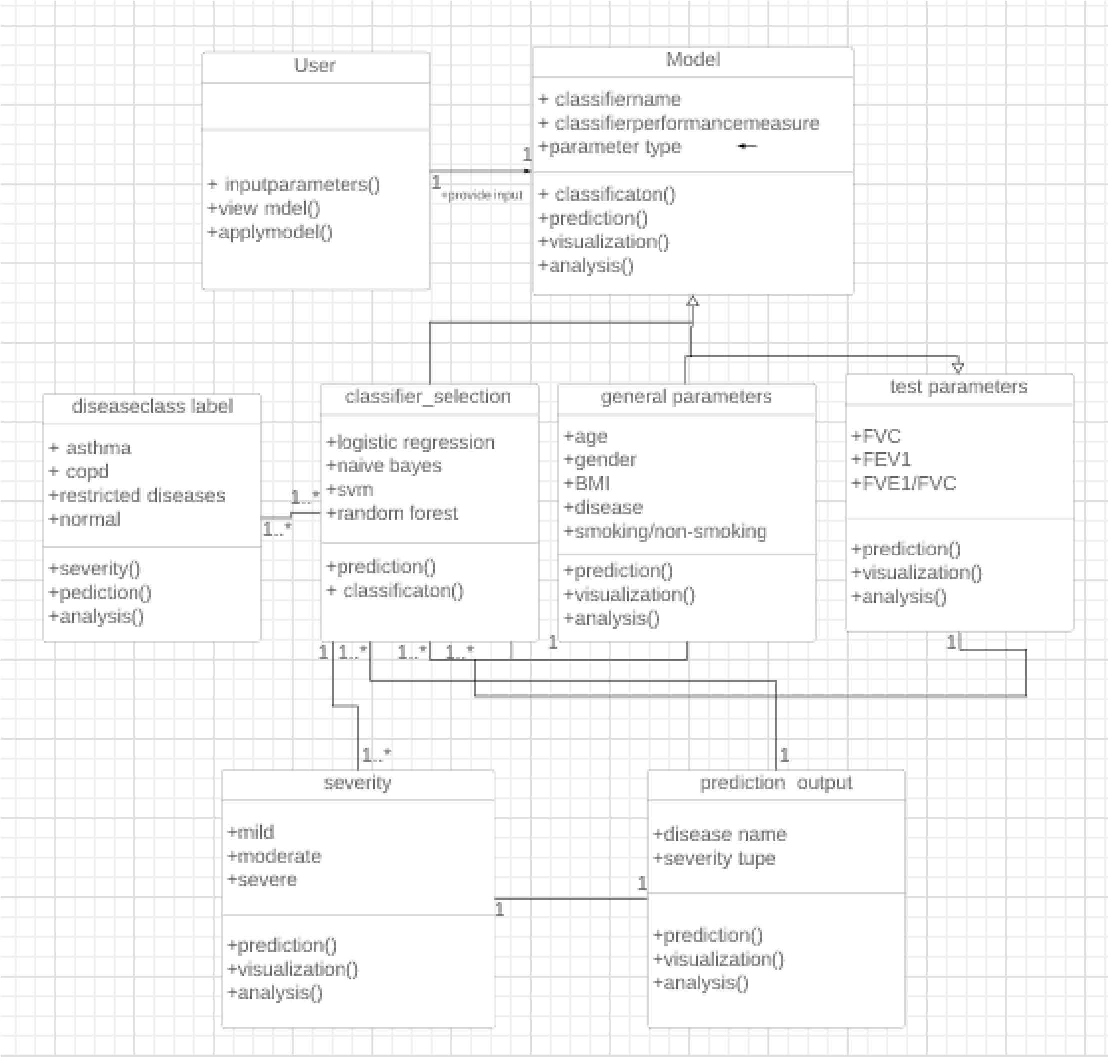


4.4 UML Diagram:

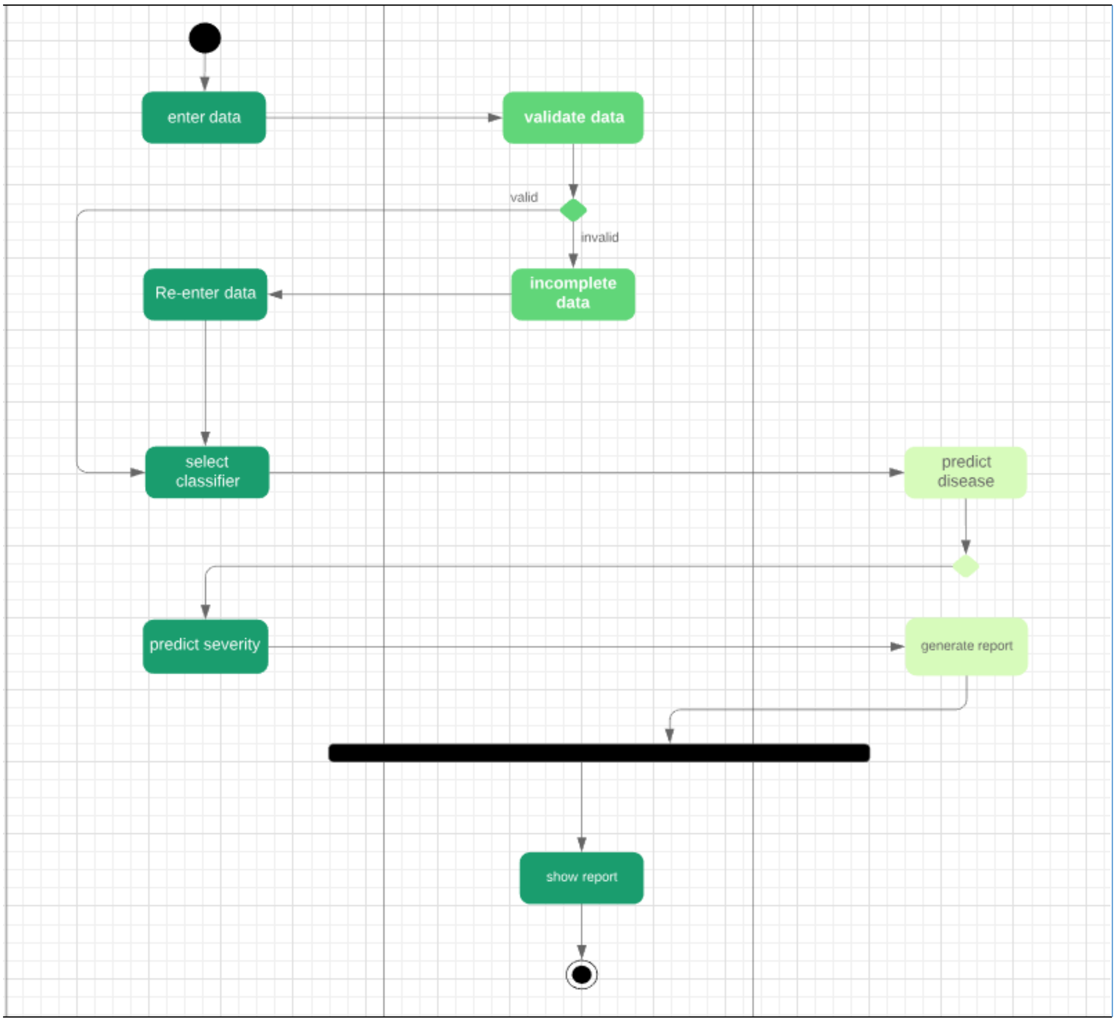
4.4.1 Use- case diagram :



4.4.2 Class Diagram



4.4.3 Activity Diagram



CHAPTER 5

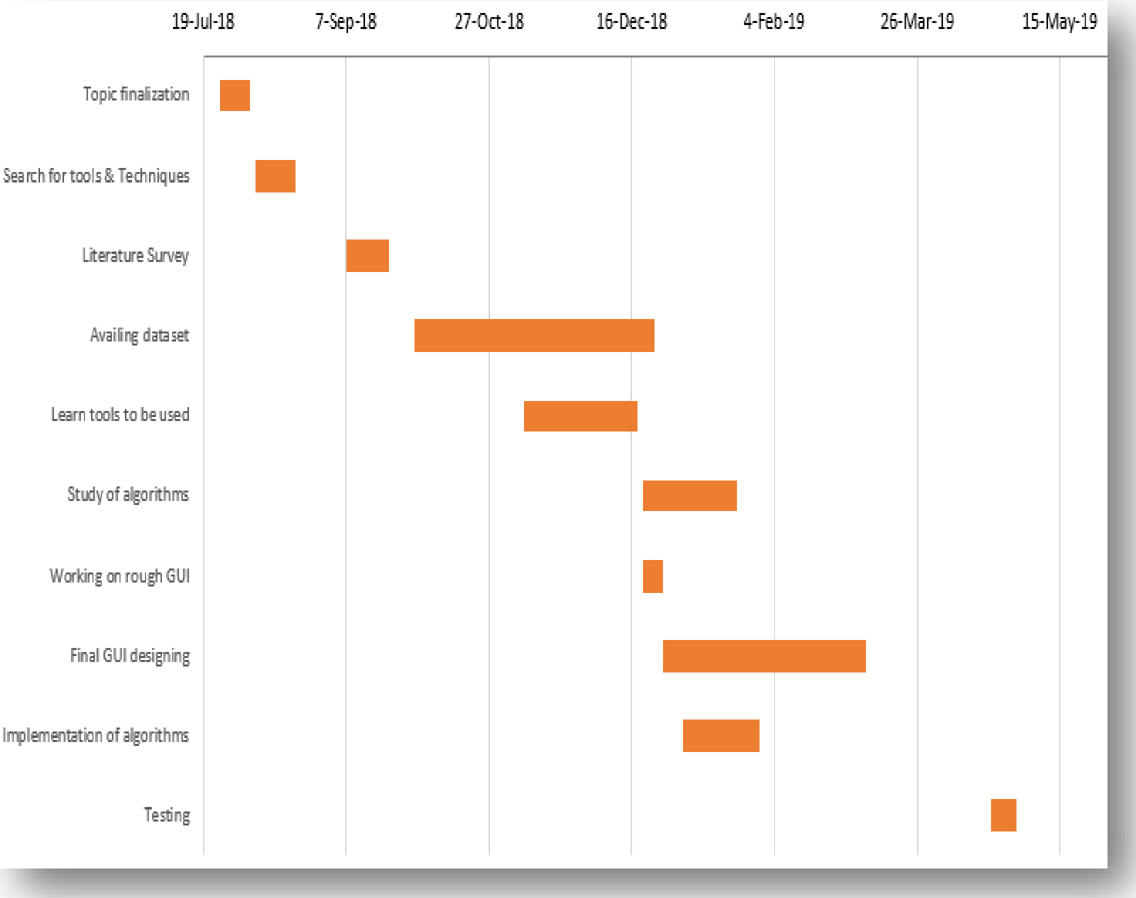
5.1 Project schedule

5.1.1 Project task set

Sr. No.	Task/Activity	Month/Period	Expected deliverables	Work done/Target achieved
1	Study of existing system in detail	2 weeks (July)	Detail knowledge of system	Studied in detail
2	Research of PFT domain	3 weeks (July – august)	Study of PFT domain through research papers and discussion with Pulmonologists.	All the necessary details regarding PFT to be used in our project was gathered.
3	Project topic finalization	1 week (august)	Topic selection	“Prediction of Lung Diseases using Data Analytic Techniques” was chosen as the final topic.
4	Finding drawbacks and limitations of existing system.	2 week (august)	Limitations should be properly identified	All merits, demerits and features of existing system and future Scope,etc studied.
5	Searching different tools and techniques.	2 week (august)	Explore about related technologies.	Anaconda and Flask were finalized.
6	Installation of Anaconda and Flask.	1 week (September)	Anaconda and Flask should be installed.	Successfully installed.
7	Exploring Anaconda/ Flask and different UI	1 week (September)	Explore about the tools and how they work.	We learned how required tools would

	software.			be applied to our project.
8	Surveying papers to compare different algorithms.	2 weeks (September)	Finding different approaches	Papers surveyed
9	Availing of Dataset	October-December	Finding Dataset	Dataset made available.
10	Working on rough GUI	December (last week)	Designing of rough GUI.	Rough GUI idea is ready.
11	Implementation of classification algorithm	January	Implementation and execution of algorithm.	Algorithms are successfully implemented.
12	Final GUI designing	December-March	User friendly GUI should be ready.	GUI ready which client can easily understand.
13	Apply different testing methods on final project	April (last week)	Test report will be ready.	Testing done successfully.

5.1.2 Timeline Chart



5.2 Team Structure

We plan to divide the workload equally at the technical side. The basic Structure of workload of team as follows:

Mamta Ingle : Developing backend of the system and availing dataset

Ashwini Khedkar : Developing backend of the system and availing dataset

Supriya Kshirsagar : Designing of user interface of the system and availing dataset

CHAPTER SIX

IMPLEMENTATION

6.1 Overview of Project Modules

6.1.1 Input from User:

- User will provide input to the system. The input consists of general data as well medical test data of patients.

6.1.2 Data Classification techniques performed on data**:**

- Our model will classify lung diseases like Asthma, COPD etc. It will consider the classification techniques like Multiple Logistic Regression, Naïve-Bayes and Random Forest, SVM. Depending on various performance measures of the classification technique it will choose best technique and make prediction for severity of disease.

6.1.3 Report generation**:**

- Our system will generate report as an outcome which will contain prediction of lung disease, its severity , accuracy and confusion matrix of each classifier and detailed analysis on real time dataset.

6.1.4 Data analysis and visualization**:**

- The analysis will help us to know about the distribution of classes among the parameters in the dataset. Various data visualization techniques are being used for different parameters.
- Analysis performed to determine best suited algorithm for each disease. Determine severity of patient for disease by using this classifier.

6.2 Tools and Technologies Used

6.2.1 Python

Python is a widely used high-level programming language for general-purpose programming. Python features a dynamic type system and automatic memory management and supports multiple programming paradigms, including object-oriented, imperative, functional programming, and procedural styles. It has a large and comprehensive standard library .

6.2.2 NumPy

NumPy is a library for the Python programming language, adding support for large, multidimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays .

6.2.3 Pandas

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It offers wide range of features including Data-frame object for data manipulation with integrated indexing, tools for reading and writing data between in-memory data structures and different file formats, data alignment and integrated handling of missing data and more .

6.2.4 Matplotlib

Matplotlib is a multi-platform data visualization library built on NumPy arrays . It has a wide variety of plots such as scatter plot, line plots ets. Plots helps to understand trends, patterns, and to develop correlations typically instruments for reasoning about quantitative information.

6.2.5 scikit_learn

Sci-kit learn is open-source simple and efficient tool for data mining and analytics . It is built on numpy,sci-py and matplotlib libraries of python. It is used widely in machine learning for various techniques such as classification,regression,clustering ,dimensionality reduction,model selection and data pre-processing.

6.2.6 Flask

Flask is python micro-web framework as it does not require any libraries, or particular tools .Flask is easy to implement due to its little boilerplate code for running simple application. Flask framework is most widely used web framework due to functionalities it provides such as integrating applications with Front-End frameworks.

6.2.7 HTML/CSS

Hypertext Markup Language (HTML) is the standard markup language for creating web pages and web applications. Web browsers receive HTML documents from a web server or from local storage and render them into multimedia web pages. HTML describes the structure of a web page semantically and originally included cues for the appearance of the document. Cascading Style Sheets (CSS) is a style sheet language used for describing the presentation of a document written in a markup language. It is most often used to set the visual style of web pages and user interfaces written in HTML.

6.2.8 JavaScript

JavaScript (JS) is a high-level, dynamic, weakly typed, object-based, multi-paradigm, and interpreted programming language. Alongside HTML and CSS, JavaScript is one of the three core technologies of World Wide Web content production. It is used to make web pages interactive [10].

6.3 Algorithms:

6.3.1 Data-preprocessing:

Various data-preprocessing techniques are applied to the real time dataset consisting of patients general and medical test parameters.

1. Feature scaling

   When dataset contains of attributes with varying scales, we rescale the attributes to same scale which helps in attaining optimized algorithm. Rescaling is done using scikit-learn library.

2. Standardization

   Standardization is a useful technique to transform attributes with a Gaussian distribution and differing means and standard deviations to a standard Gaussian distribution with a mean of 0 and a standard deviation of 1. We can standardize data using scikit-learn with the standardscaler class.

3. Label encoding

Label encoding deals with dataset containing multiple lables in one or moe columns. With label encoding we convert labels into numeric data so it can be converted further into machine readable form.Label encoding is an important pre-processing method for structured dataset.

6.3.2 Classifiers:

6.3.2.1 Multi-nominal Logistic regression:

Multi-nominal Logistic Regression is one of the most simple and commonly used Machine Learning algorithms. It describes and estimates the relationship between one dependent binary variable and independent variables having three or more nominal values. Logistic Regression predicts the probability of occurrence of a binary event utilizing a logistic function or sigmoid function.

Algorithm steps:

1. Load the data from CSV file , perform data preprocessing techniques and split it into training and test datasets.

2. Summarize the properties/features in the training dataset so that we can calculate probabilities and make predictions.

3. Calculate by probability of occurrence of an event using logic function or sigmoid function and summarizing of the dataset to generate a single prediction.

Hypothesis is calculated as follows:

$$\mathrm{h(x)} = \frac{1}{1+e-\Phi(tx)}$$

4. Generate predictions given a test dataset using prediction() method and a summarized training dataset.

5. Evaluate the accuracy of predictions made for a test dataset as the percentage correct out of all predictions made.

6.3.2.2 Naïve Bayes

The Naive Bayes algorithm is an supervised learning method that uses the probabilities of each attribute belonging to each class to make a prediction.

It simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. By multiplying the conditional probabilities together for each attribute for a given class value, we have a probability of a data instance belonging to that class.

To make a prediction we can calculate probabilities of the instance belonging to each class and select the class value with the highest probability.

Let **(x_1, x_2, …, x_n)**be a feature vector and **y** be the class label corresponding to this feature vector.

Applying bayes Theorem,

$$P(Y|x1, x2 \ldots.xn) = \frac{\mathrm{P(y) * P(x1, x2 \ldots xn|Y)}}{\mathrm{P(x1, x2 \ldots xn)}}$$

Algorithm :

1. Load the data from CSV file , perform data preprocessing techniques and split it into training and test datasets.

2. Summarize the properties/features in the training dataset so that we can calculate probabilities and make predictions.

3. Calculate class probabilities by summarising of the dataset to generate a single prediction.

4. Generate predictions given a test dataset using prediction() method and a summarized training dataset.

5. Evaluate the accuracy of predictions made for a test dataset as the percentage correct out of all predictions made.

6.3.2.3 Support Vector Machine (SVM)

Support Vector Machine"(SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the k classes.

Algorithm steps:

1. Load the data from CSV file , perform data preprocessing techniques and split it into training and test datasets.

2. For a training set (x_1,y_1) ... (x_n,y_n) with labels y_i in [1..k], it finds the solution of the following optimization problem during training.

$$\min 1/2 \Sigma_{i=1..k} w_i*w_i + C/n \Sigma_{i = 1..n} \xi_i$$

$$\text{s.t. for all } y \text{ in } [1..k]\text{: } [x_1 \bullet w_{yi}] >= [x_1 \bullet w_y] + 100*\Delta(y_1,y) - \xi_1$$

$$\ldots$$

$$\text{for all } y \text{ in } [1..k]\text{: } [x_n \bullet w_{yn}] >= [x_n \bullet w_y] + 100*\Delta(y_n,y) - \xi_n$$

C is the usual regularization parameter that trades off margin size and training error. $\Delta(y_n,y)$ is the loss function that returns 0 if y_n equals y, and 1 otherwise.

3. Calculate the prediction by using linear function in case of linear SVM.

4. Generate predictions given a test dataset using prediction() method and a summarized training dataset.

5. Evaluate the accuracy of predictions made for a test dataset as the percentage correct out of all predictions made.

6.3.2.4 Random Forest

Random forest is an extension of bagged decision trees. Samples of the training dataset are taken with replacement, but the trees are constructed in a way that reduces the correlation between individual classifiers. Specifically, rather than greedily choosing the best split point in the construction of the tree, only a random subset of features are considered for each split. Random Forest model for classification can be constructed using the RandomForestClassifier class.

Algorithm :

1. Load the data from CSV file , perform data preprocessing techniques and split it into training and test datasets.

2. Summarize the properties/features in the training dataset so that we can calculate probabilities and make predictions.

3. Calculate prediction by using randomclasssifier class in sci-kit learn.

4. Generate predictions given a test dataset using prediction() method and a summarized training dataset.

5. Evaluate the accuracy of predictions made for a test dataset as the percentage correct out of all predictions made.

Chapter 7

Software Testing

7.1 Type of Testing

Unit Testing:

Unit Testing is a software testing technique by means of which individual units of software i.e. group of computer program modules, usage procedures and operating procedures are tested to determine whether they are suitable for use or not. It is a testing method using which every independent modules are tested to determine if there are any issue by the developer himself. It is correlated with functional correctness of the independent modules. Unit testing is typically performed by the developer.

Integration Testing:

Integration testing is the process of testing the interface between two software units or module. It's focus on determining the correctness of the interface. The purpose of the integration testing is to expose faults in the interaction between integrated units. Once all the modules have been unit tested, integration testing is performed.

System Testing:

It is a level of software testing where a complete and integrated software is tested. The purpose of this test is to evaluate the system's compliance with the specified requirements.

7.2 TEST CASES AND TEST RESULTS

Module	Test ID	Test Case Name	Test Case Description	Steps	Expected output	Test Result (Pass or Fail)
User Input	TC_01	Validate Input parameters	To Verify all the medical parameters are in numeric format.	Enter the Input Parameters.	Error message , ”Please enter values in numeric format.” displayed	Pass
Logistic Regression Classifier	TC_02	Feature Selection	To select features from dataset to achieve maximum accuracy for prediction.	Trial and error strategy used for feature Selection	Prediction of disease and accuracy.	Pass
Naïve Bayes Classifier	TC_03	Feature Selection	To select features from dataset to achieve maximum accuracy for prediction.	Trial and error strategy used for feature Selection	Prediction of disease and accuracy.	Pass
SVM Classifier	TC_04	Feature Selection	To select features from dataset to achieve maximum accuracy for prediction.	Trial and error strategy used for feature Selection	Prediction of disease and accuracy.	Pass
Random Forest Classifier	TC_05	Feature Selection	To select features from dataset to achieve maximum accuracy for prediction.	Trial and error strategy used for feature Selection	Prediction of disease and accuracy.	Pass

Severity	TC_06	Feature Selection	To select features for particular disease to get accurate severity.	Trial and error strategy used for feature Selection	Prediction of Severity (Mild, Moderate, Severe)	Pass
Severity	TC_07	Classifier selection	Classifier giving best performance was selected for predicting severity	The accuracies of all the classifiers were compared and the most accurate classifier was used for predicting severity.	Prediction of Severity (Mild, Moderate, Severe)	Pass
Visualization	TC_08	Visualization of Analysis of dataset.	Different visualization techniques were used and the best was selected.	The visualization technique which gave a clear understanding of analysis were used.	Pie Chart showing the summary of analysis.	Pass
Classification	TC_09	Integration of all Classifiers.	All classifiers were integrated in one code and prediction performed and their performances were compared.	All classifiers were separately trained on the dataset and their predictions were recorded. Their accuracies were compared.	The prediction of disease given by the best classifier is displayed.	Pass

System Testing	TC_10	To test system for all functionalities.	To test the system by giving different user inputs and check the accuracy of predictions and report generated.	Give user input. Check the disease predicted. Check the severity predicted. View the analysis. View the report generated.	Display Lung disease, its severity, analysis of dataset based on age, gender, smoking habits	Pass

CHAPTER 9

CONCLUSION:

- Data mining classification algorithms are used for classifying different lung diseases based on the accuracy of the algorithm.
- Early prediction based upon the symptoms and medical tests helps doctor to take necessary measures at appropriate time.

- Use of machine learning enabled us to convert existing system into an expert system that is capable of recognizing lung diseases.
- Due to incapability of many patients unable to provide correct input during medical tests, the detection of disease becomes difficult. Our Model will provide an assistance to doctors to have better detection.

FUTURE WORK:

- On Similar approach, we will build model using Impulse Oscillometry test results
- We will compare this model with the one which uses spriometry test results and give more accurate prediction using much more robust model which will be a combination of both the models.

Appendix A:

NP hard or NP- complete problem:

NP-Complete is a complexity class which represents the set of all problems x in NP for which it is possible to reduce any other NP problem y to x in polynomial time. It deals with problems that may not be possible to solve in polynomial time but are made solvable by applying certain constraints. In this project, user/doctor is able to interact with the interface and predict lung disease as early as possible using data mining techniques. We are using several data classification techniques to achieve maximum accuracy. This application detect lung disease using patients general as well medical test data as general parameters to the system and make this problem as NP-complete.

APPENDIX B

Research papers:

- Martine Hoogendoorn,[1] Talitha L Feenstra,[2,3] Melinde Boland,[1] Andrew H Briggs,[4] Sixten Borg,[5] Sven-Arne Jansson,[6]Nancy A Risebrough,[7] Julia F Slejko,[8] and Maureen PMH Rutten-van Mölken,"Prediction models for exacerbations in different COPD patient populations: comparing results of five large data sources" *Int*

J Chron Obstruct Pulmon Dis. 2017;12:3183-3194. Published 2017 Nov 1. doi:10.2147/COPD.S142378

- Ionnis I.Spyroglou , Gunter Spock, Alexnadros G.Rigas and E.N. Paraskakis, “Evaluation of Bayesian classifiers in asthma exacerbation prediction after medication discontinuation.” *BMC research notes*vol. 11,1 522. 31 Jul. 2018, doi:10.1186/s13104-018-3621-1

- Mrs. J. Cathrin Princy 1 ,Mrs. K. Sivaranjani , “Survey on Asthma Prediction Using Classification Technique”, IJCSMC, Vol. 5, Issue. 7, July 2016, pg.515 – 518

.

Web-Links:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5677310/

https://www.hopkinsmedicine.org/healthlibrary/test_procedures/pulmonary/pulmonary_function_tests_92,p07759

https://www.nhlbi.nih.gov/health-topics/pulmonary-function-tests