

MINOR PROJECT
DISEASE PREDICTION USING SYMPTOMS
Submitted for the partial fulfilment of the requirements for the
award of a Degree
B.TECH IN
DEPARTMENT OF COMPUTER SCIENCE
AND ENGINEERING
UNIVERSITY INSTITUTE OF TECHNOLOGY RGPV
BHOPAL 462033
Session 2021-2025



Submitted By:

Submitted To:

Nikita Kotwal	0101CS211084	Prof. Uday Chorasiya
Nishant Tiwari	0101CS211087	Associate Professor, DoCSE
Shruti Patel	0101CS211115	Dr. Priyanka Dixit
Vedika Walke	0101CS211130	Assistant Professor, DoCSE

**RAJIV GANDHI PROUDYOGIKI VISHWAVIDYALAYA,
BHOPAL**



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
CERTIFICATE

This is to certify **Nikita Kotwal, Nishant Tiwari, Shruti Patel, Vedika Walke** of B.Tech 3th Year,

CSE UIT(RGPV) have completed their Minor Project Synopsis entitled “**Disease Prediction using Symptoms**” during the academic year 2023-2024 under my supervision and guidance.

We approve this project for the submission, for the partial fulfilment of the requirements for the award of a degree in B.Tech. Computer Science and Engineering.

Prof.Uday Chorasiya

Project Guide

Dr. Priyanka Dixit

Project Guide

Dr Manish Kumar Ahirwar
(Head Of Department, DoCSE)

UIT RGPV BHOPAL

DECLARATION BY THE CANDIDATES

We hereby declare that the work which we are presenting in the Report of Major Project entitled — **“Disease Prediction System”** is our own work, submitted for the partial fulfilment of the requirements for the award of a bachelor's degree in Computer Science and Engineering. The work has been carried out at the University Institute of Technology, RGPV, Bhopal, in the session 2023-2024, and an authenticate record of our work which is carried out under the guidance of **Prof. Uday Chorasiya** (Project Guide) & **Dr. Priyanka Dixit** (Project Guide) DoCSE and **Dr. Manish Ahirwar, Head Of Department, Computer Science and Engineering Department, RGPV, Bhopal**. I further declare that, to the best of our knowledge, the matter written in this project is not submitted or used for the award of any other Degrees.

Nikita Kotwal

0101CS211084

Nishant Tiwari

0101CS211087

Shruti Patel

0101CS211115

Vedika Walke

0101CS211130

ACKNOWLEDGEMENT

We would like to offer our heartfelt appreciation to our Project Guides, **Prof. Uday Chourasia** and **Dr. Priyanka Dixit**, for their invaluable advice and assistance. This project would not have been possible without their enthusiasm, hard work, and excellent counsel. Their thorough approach has increased the project's precision and clarity.

We are thankful to **Dr. Manish Ahirwar, Head of the Department of Computer Science & Engineering, University Institute of Technology, RGPV, Bhopal**, for his unwavering support and inspiration for the project, ethics, and morals. The concepts we have acquired from him have always been a source of exponential inspiration for our path in this life.

We are also thankful to all other members and Staff of the Department who were involved in the project directly or indirectly for their valuable co-application.

We are grateful to all our co-workers for inspiring us and creating a nice atmosphere to learn and grow.

Name of the students:

Nikita Kotwal	0101CS211084
Nishant Tiwari	0101CS211087
Shruti Patel	0101CS211115
Vedika Walke	0101CS211130

ABSTRACT

Using Machine learning, our project proposes a **Disease prediction system using Symptoms**. For small problems, the users have to go personally to the hospital for check-up which is more time consuming. Also handling the telephonic calls for appointments is quite hectic. Such a problem can be solved by using disease prediction applications by giving proper guidance regarding healthy living. Over the past decade, the use of the specific disease prediction tools along with the concerning health has been increased due to a variety of diseases and less doctor-patient ratio. Thus, in this system, we are concentrating on providing immediate and accurate disease prediction to the users about the symptoms they enter along with the severity of disease predicted. For prediction of diseases, different machine learning algorithms are used to ensure quick and accurate predictions. In one channel, the symptoms entered will be cross checked with the database. Further, it will be preserved in the database if the symptom is new which its primary work is and the other channel will provide severity of disease predicted. A web/android application is deployed for users for easy portability, configuring and being able to access remotely where doctors cannot reach easily. Therefore, this arrangement helps in easier health management. Keywords: Machine Learning, Decision Tree Algorithm, Random Forest Algorithm, Naive Bayes Algorithm.

Table of Contents

S.NO	TITLE	PAGE NO.
1.	Certificate	2
2.	Declaration	3
3.	Acknowledgement	4
4.	Abstract	5
5.	Table of Contents	6-7
5.	List of figures	8
	Chapter 1 Introduction	
1.1	Overview	9
1.2	Objective of the project	10
1.3	Research Background	11-13
1.4	Technologies used	14
1.5	Working of prediction system	14-15
1.6	Application of prediction system	15
1.7	Motivation	16
1.8	Importance of system	16
	Chapter 2 Literature Survey	
2.1	Methodology	17-19
2.2	Our Approach	20-24
	Chapter 3 Problem Description	
3.1	Problem statement	25
3.2	Our Solution	25
3.3	Future Scope	25-26

3.4	Product Scope	26
3.5	Future Scope	
	Chapter 4 Proposed Work	
4.1	Proposed work	28
4.2	Flowchart of Disease Prediction System using Symptoms	29
4.3	Class Diagram of Disease Prediction System using Symptoms	30
4.4	Sequence Diagram of Disease Prediction System using Symptoms	31
	Chapter 5 Implementation	
5.1	Dataset Of Disease Prediction System	32
5.2	Symptoms Of Disease Prediction System	33-34
5.3	Data Augmentation	34
	Chapter 6 Result and Analysis	
6.1	Result	35-36
6.2	Evaluation Measure dataset Of Disease Prediction System	37
6.3	Result Analysis	38-40
	Chapter 7 Conclusion and Future Work	
7.1	Conclusion	41
7.2	Future Work	41
	References	42

List of Figures

S.NO	Figure	Page No.
1	Working of prediction System	14
2	Flowchart of Disease Prediction System	29
3	Class Diagram of Disease Prediction System	30
4	Sequence Diagram of Disease Prediction System	31
5	Dataset	32
6	Data Preprocessing	33
7	Data Augmentation	34
8	Result and Analysis	35

CHAPTER 1

INTRODUCTION

1.1 Overview

The healthcare industry is changing quickly, and integrating sophisticated technologies has become essential to improving patient outcomes and diagnostic procedures. The Disease

Prediction System utilizing Symptoms is one such invention. It is an advanced tool that makes predictions about possible diseases based on user-reported symptoms. By bridging the gap between initial symptom presentation and precise diagnosis, this approach hopes to give patients and medical professionals access to an invaluable tool for early disease identification. It is impossible to exaggerate the value of early diagnosis in healthcare. Early disease detection frequently results in better patient care, more successful treatments, and noticeably better prognosis. However, because there are so many possible ailments and because symptoms of different conditions frequently overlap, making a diagnosis based only on symptoms can be difficult. The precision and speed of machine learning algorithms can be used to augment the knowledge and experience of healthcare experts, which are crucial in traditional diagnosis procedures.

The foundation of the Disease Prediction System is the idea that a machine learning model may be trained to recognize patterns and connections that may not be immediately obvious by utilizing big datasets of symptoms and related diseases. Using cutting-edge algorithms, this system evaluates symptoms entered and generates a probability-ranked list of possible diseases. By giving individuals early knowledge about their health concerns, this method not only empowers patients but also helps healthcare providers make better decisions. This project has several different goals. First and foremost, the goal is to create a reliable and accurate model that can diagnose illnesses based on a predetermined set of symptoms. Second, the project aims to provide an intuitive user interface that facilitates simple symptom input and transparent results presentation. Lastly, the project aims to support early diagnosis, which is essential for properly controlling and treating illnesses.

Data gathering is the first of several crucial processes in the methodology used for this project. This entails compiling sizable datasets covering a broad spectrum of illnesses and related symptoms from reputable medical databases and literature. After that, the data is preprocessed to guarantee its accuracy and applicability. This includes dealing with missing values, standardizing the data, and correctly classifying the symptoms.

Model development is the main focus of the project after data preprocessing. To create the prediction model, a number of machine learning techniques are used, including Random Forest, Decision Trees, and Naïve Bayes. To understand the complex interactions between symptoms and diseases, the model is trained on the preprocessed information. To make sure the model is reliable and useful, testing and validation are carried out using recognized measures including accuracy, precision, recall, and F1-score. The creation of an intuitive user interface is the project's last part. Because of the interface's easy-to-use design, users may input their symptoms and receive forecasts without encountering any technical challenges. The system's fast and accurate illness forecasts, which can direct additional medical consultation and action, are meant to be a

useful tool for people and healthcare professionals alike.

To sum up, the Disease Prediction System utilizing Symptoms is a noteworthy development in the use of technology in healthcare. This method can facilitate early and precise disease prediction, which can lead to better patient outcomes, improved diagnostic procedures, and more effective healthcare delivery. This report's subsequent sections will go into greater detail on the system's thorough development process, its accomplishments, and its future directions for improvement.

1.2 Objective of the project

- ▶ The main objectives of the Disease Prediction System using EfficientNet are:
- ▶ **Symptom Analysis:** To examine symptoms entered by users and associate them with possible illnesses.
- ▶ **Accurate Prediction:** To create a model with a high degree of accuracy in predicting illnesses.
- ▶ **User-Friendly Interface:** To design an interface that is simple to use and intuitive for patients as well as healthcare professionals.
- ▶ **Early Diagnosis:** To increase the likelihood of effective treatment and management by facilitating early diagnosis.

1.3 Research background

Developments based on ML, which predict disease based on patient symptoms, have received a lot of attention recently due to the challenges associated with accessing healthcare services. With the DBMI dataset [14], which had 133 symptoms and 42 disease types, Gandhi et al. [15] experimented with supervised and unsupervised algorithms. In experiments with the Linear Discriminant Analysis (LDA), random forest, naive Bayes, SVM, KNN, Classification and Regression Trees (CART), and logistic regression algorithms, the accuracy score for logistic regression was the lowest of the group, coming at 80.85%. Agrawal et al. [16] suggested a new ML model in this research that combines a

support vector machine and a genetic algorithm. Additionally, they attempted to reduce the number of features in the dataset, and by using their ML models, they were able to achieve adequate accuracy for all three datasets. They attained the best accuracy of 78.6% for the liver dataset consisting of categorical data. They claim that unstructured medical text data from sources including diagnoses, doctor-patient interactions, medical records, etc. would also be used in future research despite using only structured data in this study. To get over the limitations of ML, Vinitha et al. [17], proposed leveraging big data to predict diseases based on ML.

The idea is to gather information from a hospital that used the Map Reduce (MR) approach and Machine Learning Decision Tree (MLDT) algorithm to analyze data from a forum referred to as structured and unstructured data. The MR algorithm detects the possibility of disease occurrences faster than CNN-UDRP, reaching 94.8% with the standard speed. Kumar, Sharma, and Prakash [18], created a Django-based online application that uses ML algorithms to predict and provide clinical guidance for general disease, heart disease, diabetes, and liver disease. While the results of predictions for common diseases are the names of the diseases, results for predictions for specific diseases, such as heart disease, diabetes, and liver disease, are true or false. In general disease prediction, it is seen that the **highest accuracy score of 90.2%** among the KNN, logistic regression, random forest, and naive Bayes algorithms was **achieved in random forest**. The highest accuracy in heart disease was seen in logistic regression, with 92.3%. While the KNN algorithm gave the highest accuracy with 74% in the liver, it was seen that logistic regression gave the highest accuracy with 78% in diabetes. Mallela, Bhavani, and Ankayarkanni [19], developed a GUI to get the symptoms from the user and they used ML models such as naive Bayes and decision trees. The outputs are the disease, the accuracy of the model, its definition, and the treatment of the particular disease based on the symptoms given by the individual. This paper shows a detailed explanation of how to find the diseases from symptoms; so that the individual can contact the respective doctor of medicine and stay healthy at an early stage. A sample of 4920 patient records with diagnoses for 41 disorders was chosen by Grampurohit and Sagarnal [20] for analysis. 41 diseases made up a dependent variable.

There were 132 independent variables, 95 of which were symptoms closely associated with diseases. The disease prediction system created utilizing ML techniques including decision tree, random forest, and naive Bayes is demonstrated in this research project. Dhabarde et al. [21] use not only structured data but also textbook data, and the dataset used has 230 conditions consisting of the individual's symptoms, age, and gender. In the paper, they conducted experiments with logistic regression, naive Bayes, SVM, random forest, and decision tree algorithms, and the decision tree gave the highest accuracy score, 93.24%.

Alanazi [22], proposes a method for chronic disease prediction using ML algorithms such as Convolutional Neural Network (CNN) and KNN. The proposed system used both structured and unstructured data from real life which were used for dataset preparation.

The performance of the proposed model in the study shows that it is higher than the naive Bayes, decision tree, and logistic regression algorithms and provides 95% accuracy. Uddin et al. [23], conducted a study on different KNN variants (classical one, adaptive, locally adaptive, K-means clustering, fuzzy, reciprocal, ensemble, Hassanat, and generalized mean distance) and their performance comparison for disease prediction. For accuracy measurement, Hassanat KNN shows the highest average accuracy with 83.62%, followed by ensemble approach KNN with 82.34%. For disease prediction with big data in healthcare, Joel and Priya [24], employed extended CNN. The hospital is built using this approach, which offers great accuracy, performance, and convergence speed in the medical industry. The unstructured data is employed with the CNN algorithm, which automatically selects the features, to choose a specific location and then assesses the chronic diseases that contain the structured data which extracted valuable features. The medical data and illness risk model were proposed by the innovative CNN. The suggested approach seeks to forecast the likelihood of liver-focused illness. Therefore, the hospital dataset is concerned with diseases that affect the liver, and it exclusively collects structured data from information on liver diseases.

The proposed approach obtains accuracy by using disease risk modeling. Ibrahim et al. [25] proposed a method for predicting the defervescence day of fever in dengue patients using an artificial neural network. The suggested method primarily depends on clinical symptoms and indicators for detection. Data from 252 patients were collected, of which 4 patients had Dengue Fever (DF) and 248 had Dengue Haemorrhagic Fever (DHF). The neural network toolkit in MATLAB is utilized and the Multi-layer Feed-Forward Neural Network (MFNN) technique is used in this experiment. 90% of the time, MFNN in DF and DHF correctly predicts the day of defervescence of fever. Venkatesh et al. [26], worked on five algorithms,

such as random forest, KNN, naive Bayes, SVM, and decision tree; the highest accuracy score was decision tree, with 95.13%. They also have developed a user interface for patients to input their symptoms and see the disease prediction. Chauhan et al. [27], performed preprocessing on the dataset and then performed experiments on naive Bayes, decision tree, and random forest. When the experiments performed on the non-preprocessed dataset were compared with the results of the preprocessed data, it was seen that the accuracy score of the random forest was the highest in both, increasing to 95.28% in raw data and 97.64% after

processing. Maram, Kumar, and Gampala [28], stored data including 400 symptoms and 147 diseases collected from various repositories in the Hadoop Distributed File System (HDFS).

Among decision trees, random forest, naive Bayes, and a new algorithm proposed in the article, the accuracy of the proposed algorithm showed the best result, with 97.60%. Through the analysis of performance measures, Ferjani [29], identifies patterns among several supervised ML model types for disease diagnosis. The supervised ML algorithms, naive Bayes, decision trees, and KNN, received the greatest attention. According to research, a support vector machine is most effective at spotting Parkinson's illness and kidney ailments. They found that the logistic regression performed well for heart disease prediction. Additionally, CNN and random forest made accurate predictions for common diseases and breast disorders, respectively. For accurate prediction, naive Bayes and KNN algorithms were used in [30] to process the person's life behaviors and check-up data.

The accuracy of heart disease prediction using naive Bayes was shown to be 94.5% greater than KNN. Furthermore, compared to naive Bayes, KNN requires more memory and time. In this work, heart disease was first predicted, and then a risk prediction system using the CNN algorithm was developed to assess the risk of heart disease. The CNN-based Multimodal Disease Prediction (CNN-MDRP) method was developed by Shirsath and Patil [31] to address the limitations of their CNN-based Unimodal Disease Prediction (CNN-UDRP) algorithm, which only analyzes unstructured data. In CNN-MDRP, which focuses on both structured and unstructured data, the accuracy of disease prediction is higher and faster compared to CNN UDRP, with an accuracy score of 94.80%. Nearly 230 diseases were listed by Keniya et al. [32] with over 1000 distinctive symptoms. Various ML algorithms receive as input a person's symptoms, age, and gender. About 230 diseases were predicted using 11 different ML algorithms. The weighted KNN model had a 93.5% accuracy score, which was the highest. For disease prediction, Dahiwade, Patle, and Meshram [33] used KNN and CNN algorithms. The model accepts information from the person's checkups and daily routine as input. With 84.5% accuracy, CNN outperforms the KNN algorithm in general disease prediction. The time and memory requirements for KNN are also higher than for CNN.

1.4 Technologies Used

SOFTWARE USED

- **PYTHON**
- **VSCODE (INTEGRATED DEVELOPMENT ENVIRONMENT)**
- **MACHINE LEARNING**

1.5 Working of prediction System

Using user-inputted symptoms and a systematic method, a symptom machine learning (ML) illness predictor may effectively forecast possible diseases. First, extensive datasets are gathered from multiple sources, such as clinical research, electronic health records (EHRs), and medical databases. To make sure they are ready for study, these datasets—which contain patient demographics, symptoms, diagnoses, and medical histories—are then cleaned, normalized, and processed. To improve the efficacy of the model, features are chosen and categorical data is encoded into numerical representations. The primary function of the system is to choose and train machine learning algorithms on the preprocessed data, such as Random Forests, Decision Trees, Support Vector Machines, Naïve Bayes, and Neural Networks. Metrics like accuracy, precision, recall, and F1-score are used to thoroughly assess and optimize the model. Through an intuitive interface, users communicate with the system by entering their symptoms. After being cleaned up and encoded to conform to the training data format, these inputs are fed into the trained model, which produces a probability-ranked list of possible diseases. The user is presented with the results as well as other data and suggestions. Users can report on the predictability of the results through a feedback mechanism, which is crucial for the model's ongoing improvement. To ensure high accuracy and reliability, this feedback is examined, and the model is updated with new data on a regular basis. A reliable and scalable illness prediction system is ensured by the use of tools and technologies like Python, Scikit-learn, TensorFlow, Pandas throughout the entire process.

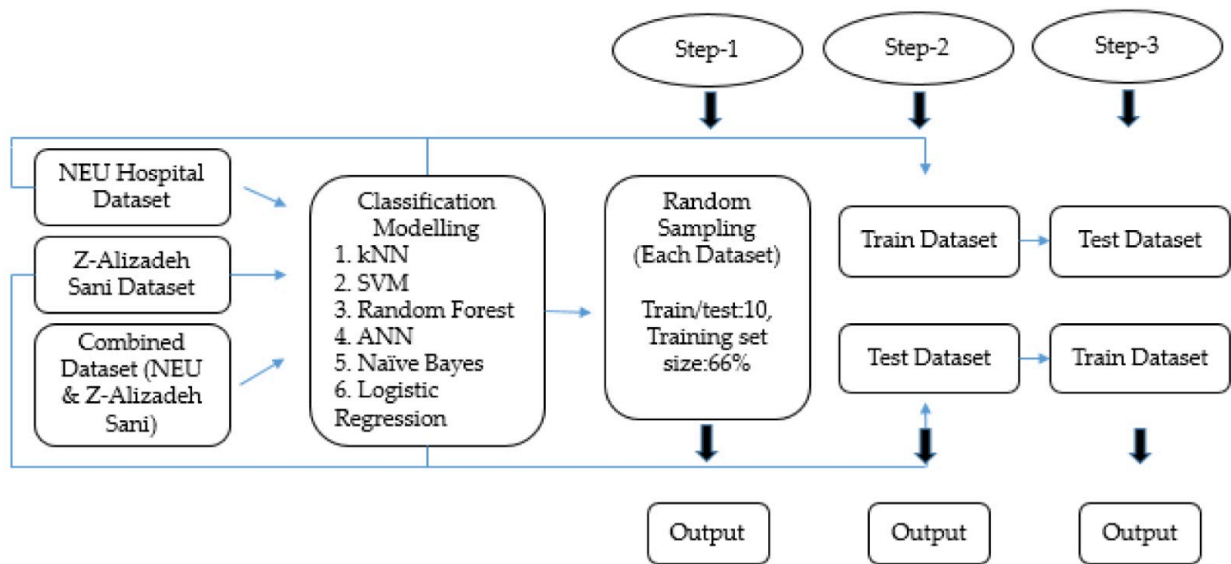


Fig 1.5 Working of prediction System

1.6 Applications of prediction system

The disease prediction system based on the EfficientNet algorithm has several practical applications across healthcare and medical research domains:

1. Early Diagnosis and Prevention

- Early detection of possible illnesses enables timely treatment, which lessens the severity of illnesses and enhances patient outcomes.

2. Clinical Decision Support

- By customizing treatment regimens according to the patient's anticipated conditions, doctors can improve treatment efficacy through prediction-based care.

3. Patient Empowerment and Self-Care

- Through the system, patients can be aware of their health risks and take proactive measures to improve their health by seeing a doctor and seeking guidance.

4. Telemedicine and Remote Healthcare

- By using patient-reported symptoms to make initial diagnosis, the approach improves remote consultations in telemedicine.

1.7 Motivation

The motivation for creating a disease predictor is rooted in improving healthcare outcomes through early diagnosis, especially for conditions like cancer and heart disease. By identifying health issues early, the system enables timely interventions, saving lives and reducing healthcare costs. It also addresses healthcare disparities by providing diagnostic support in remote areas. Leveraging advanced technology, such as machine learning, it optimizes resource allocation and enhances diagnostic accuracy. Additionally, it supports telemedicine, promotes preventative healthcare, and fosters a proactive approach to health management, ultimately aiming to create a more efficient and equitable healthcare system.

Moreover, the economic burden associated with late-stage disease management is significant, making early detection a cost-effective strategy for healthcare systems worldwide. By identifying diseases in their nascent stages, the disease predictor helps minimize the financial strain on healthcare resources and enables more efficient allocation of medical interventions and resources. This is particularly crucial in resource-constrained settings and underserved communities where access to timely healthcare services may be limited.

1.8 Importance of the System:

1. **Early Detection and Prevention:** Enables timely treatment, improves patient outcomes, and increases recovery rates by catching diseases at an early stage when they are more manageable and treatable.
2. **Preventive healthcare:** Promotes general health and well-being by guiding lifestyle modifications and focused screenings to lower the chance of illness onset and transmission.
3. **Enhancing Health Literacy:** Educates patients on their health risks, encouraging proactive management and healthier lifestyle choices.
4. **Public Health Management:** Helps control epidemics and informs policies to effectively address and mitigate health crises.
5. **Research and Development:** Identifies disease patterns and aids in developing new treatments and drugs.

CHAPTER 2

Literature Survey

2.1 Methodology

The Dataset : for getting some dataset and training our model, so far we have made some surveys in the medical field, explored some data on the internet and made an arrow dataset by combining all of that so now we have a dataset. This CSV file contains 5000 rows and 133 columns, 132 columns properties. And last column for the disease class (40 unique disease classes). Some rows of disease with their corresponding symptoms in the dataset.

Data Pre-processing : After collecting that data, as that data is raw data we have to make it suitable for training our machine learning model. By using some python libraries like NumPy, and pandas, we have made that data suitable for machine learning models. Now, our data is ready to use with machine learning algorithms to predict some output. As our problem come under unsupervised machine learning technique, we have used Naive Bayes algorithm.

Model building : After applying these algorithms, we have to select which is most fitted with our dataset and which gives us more accuracy. So, we have used a confusion matrix for that and mapped out the accuracy of each model. And, we have found that all are giving the same 100% accuracy, so we have selected Random Forest Classifier for building our model.

***Naive Bayes Algorithm:** The Naive Bayes algorithm is a probabilistic classifier based on Bayes' Theorem, assuming that the features are independent given the class. It calculates the probability of each class given a set of features and predicts the class with the highest probability.

Key Formula

$$P(C|X) \propto P(C) \cdot \prod_{i=1}^n P(x_i|C)$$

Where:

- $P(C|X)P(C|X)P(C|X)$ is the posterior probability of class CCC given feature vector XXX.
- $P(C)P(C)P(C)$ is the prior probability of class CCC.
- $P(x_i|C)P(x_i|C)P(x_i|C)$ is the likelihood of feature x_i given class CCC.

Steps

1. Training Phase:

- Calculate prior probabilities for each class.
- Calculate likelihoods of each feature given each class.

2. Prediction Phase:

- Compute the posterior probability for each class for a given feature vector.
- Predict the class with the highest posterior probability.

Example Applications

- **Spam Detection:** Classifying emails as spam or not spam.
- **Sentiment Analysis:** Determining if a review is positive or negative.
- **Medical Diagnosis:** Predicting diseases based on symptoms.

***Decision tree classification algorithm:** A decision tree is a supervised learning algorithm used for classification tasks. It splits data into subsets based on feature values, creating a tree-like model of decisions.

1. **Root Node:** Represents the entire dataset.
2. **Decision Nodes:** Points where the data is split based on a feature.
3. **Leaf Nodes:** Terminal nodes that represent the output class.

Steps

1. **Select Best Split:** Choose the feature and value that best split the data using criteria like Gini impurity or information gain.
2. **Split Data:** Divide the dataset into subsets based on the selected feature and value.
3. **Recursively Split:** Repeat the process for each subset until stopping conditions are met (e.g., maximum depth, all instances belong to one class).
4. **Pruning (Optional):** Remove unnecessary branches to prevent overfitting.

Example Applications

- **Spam Detection:** Classifying emails as spam or not spam.
- **Customer Segmentation:** Identifying distinct customer groups.

- **Medical Diagnosis:** Predicting diseases based on symptoms.

* **Random Forest Algorithm:** Random Forest is an ensemble learning algorithm that builds multiple decision trees and combines their predictions to improve accuracy and prevent overfitting.

1. **Ensemble Method:** Combines multiple decision trees.
2. **Bagging:** Each tree is trained on a random subset of data with replacement.
3. **Random Feature Selection:** Random subsets of features are considered for splitting at each node in the trees.

Steps:

1. **Create Subsets:** Generate multiple bootstrap samples from the original dataset.
2. **Train Trees:** Train a decision tree on each sample, using a random subset of features for splits.
3. **Aggregate Predictions:**
 - Classification: Use majority voting to determine the final class.
 - Regression: Average the predictions of all trees.

Example Applications:

- **Spam Detection:** Classifying emails as spam or not spam.
- **Credit Scoring:** Assessing the risk of loan applicants.
- **Disease Prediction:** Predicting diseases based on medical records.

2.2 Our Approach

The main theme of this project is to detect the diseases and to take precautions to avoid or clear that disease.

- **Data Pre-processing** - After collecting that data, as that data is raw data we have to make it suitable for training our machine learning model. By using some python libraries like NumPy, and pandas, we have made that data suitable for machine learning models. Now, our data is ready to use with machine learning algorithms to predict some output.

- **Methodology** -The Dataset : for getting some dataset and training our model, so fore that we have made some surveys in medical field explored some data on internet and made arrow dataset by combining all of that so now we have a dataset

We are using three algorithms to predict the disease:

1. Decision Tree
2. Random Forest
3. Naive Bayes

Here are the definitions and uses for Decision Tree, Random Forest, and Naive Bayes algorithms:

Decision Tree

- Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- The decisions or the test are performed on the basis of features of the given dataset.

- It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
- In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.
- A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

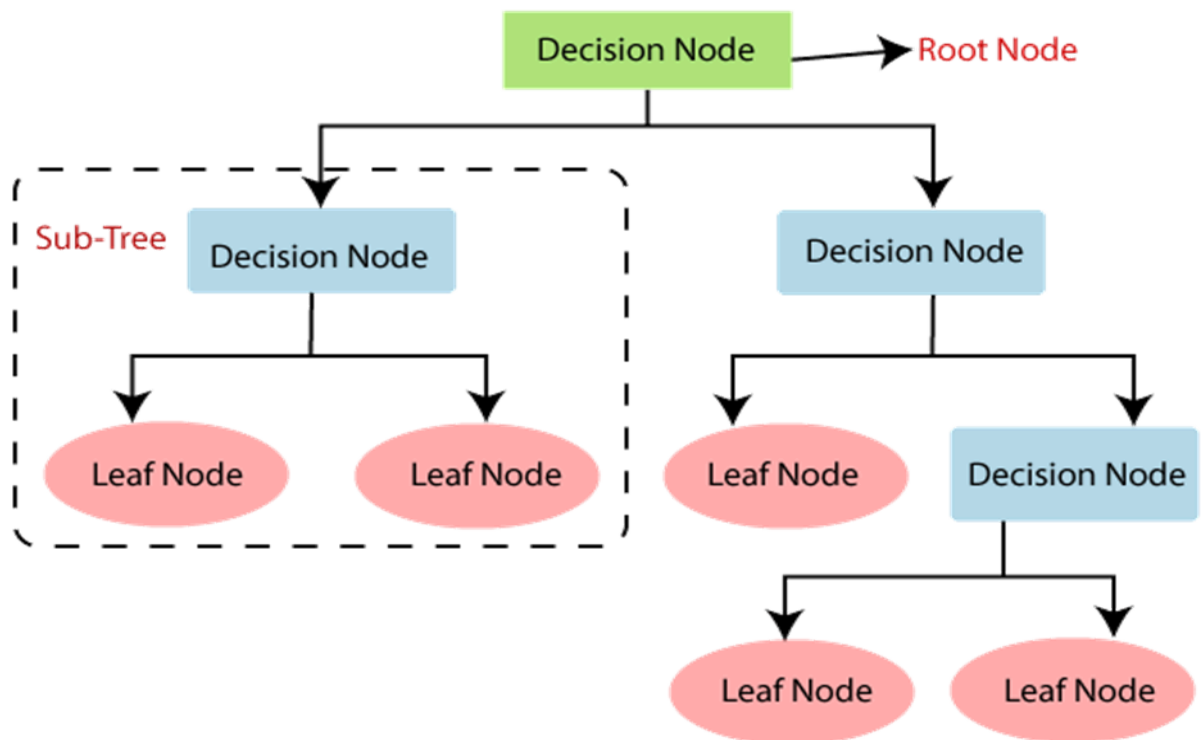


Fig 2.1(a) Decision Tree

Use in Disease Prediction: Decision Trees can be used to predict diseases by learning from past patient data where symptoms (features) and corresponding diagnoses (labels) are known. The tree model can then classify new patients into different disease categories based on their symptoms.

Why use Decision Trees? There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:

- Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
- The logic behind the decision tree can be easily understood because it shows a tree-like structure

Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Why use Random Forest?

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

Use in Disease Prediction: In disease prediction, Random Forest can enhance accuracy and robustness by combining the predictions of multiple decision trees. This method is particularly useful for handling large datasets with many features (symptoms) and can improve the model's ability to generalize to new patient data.

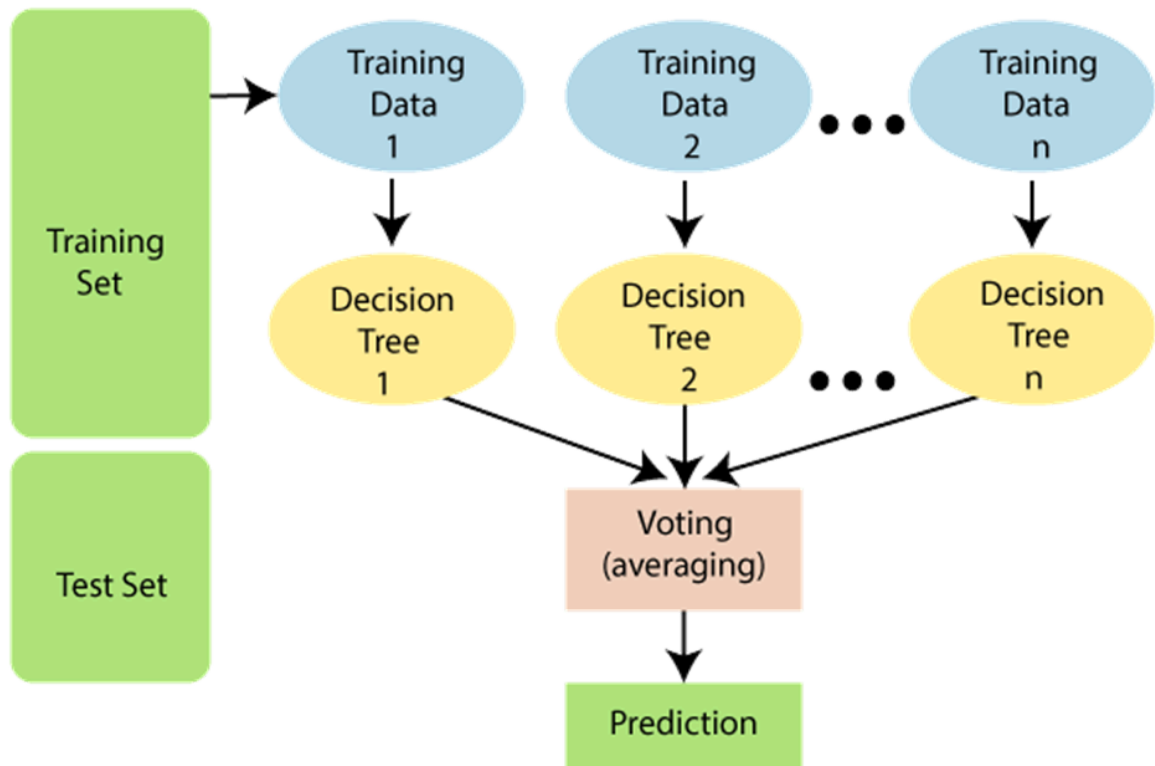


Fig 2.1(b) Random Forest

Naive Bayes

Machine Learning Field Monitoring and its various model decisions there are a wide range of algorithms to predict various diseases with predictive characteristics in tree, neo base and random forests. There Three different models of naïve Bayes, namely Gaussian, multinational and Bernoulli naïve Bayes. Each model has its own accuracy Application and data fitting to assess disease outcome are almost identical in all three models. Since the Gaussian naïve Bayes Relatively easy to understand and much simpler than the other two, this project work was done using it

- Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- It is mainly used in text classification that includes a high-dimensional training dataset.
- Naive Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

- Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.
- Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- It is mainly used in text classification that includes a high-dimensional training dataset.
- Naive Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
- Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.
- Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- It is mainly used in text classification that includes a high-dimensional training dataset.
- Naive Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
- Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

Use in Disease Prediction: Naive Bayes can be used for disease prediction by calculating the likelihood of various diseases based on observed symptoms. It is particularly effective when dealing with high-dimensional data and when the relationships between features are not too complex. The simplicity and speed of Naive Bayes make it suitable for real-time disease prediction system.

CHAPTER 3

Problem Description

3.1 Problem Statement

In healthcare, timely diagnosis is critical for effective treatment and management of diseases. However, many diseases present with overlapping symptoms, making accurate diagnosis challenging for healthcare professionals, especially in resource-constrained settings. To address this challenge, there is a growing interest in developing intelligent systems that can predict diseases based on symptoms reported by patients.

3.2 Our Solution

The Predico: A Symptom-Based Disease Prediction Model Using Machine Learning Approach” does not focus on the prediction of a specific disease; instead, it predicts disease based on the symptoms given by the user. As a result, the user does not need to traverse many models to predict the disease

Even now in some parts of the world there are still some far-flung villages, remote places which lack clinical centers, health facilities. Machines have started to gain popularity and dependency by humans as, without any human mistakes, they could perform duties greater efficaciously and with a steady degree of accuracy. A disease predictor is nothing but a virtual doctor, which can predict the disorder of any affected person without any human errors.

3.3 Future Scope

Real-time Monitoring and Early Detection: Advancements in sensor technology and data analytics can enable real-time monitoring of symptoms and early detection of disease patterns, allowing for timely interventions and improved outcomes.

Integration of Multiple Data Sources: Incorporating data from diverse sources such as wearable devices, electronic health records (EHRs), genetic data, lifestyle factors, environmental factors, and social determinants of health can provide a more holistic view of an individual's health status and improve the accuracy of disease prediction models.

Personalized Medicine: Tailoring disease prediction models to individual patients based on their unique genetic makeup, medical history, lifestyle, and environmental factors can lead to more personalized and effective preventive healthcare strategies.

Collaboration with Healthcare Providers and Researchers: Collaboration between data scientists, healthcare providers, researchers, and policymakers can facilitate the development and

validation of robust disease prediction models, as well as the translation of research findings into clinical practice.

Global Health Impact: Disease prediction models can have a significant impact on global health by helping to identify and prioritize public health interventions, track disease outbreaks, allocate resources efficiently, and improve healthcare access and delivery in resource-limited settings.

Incorporation of Advanced AI Techniques: Leveraging advanced AI techniques such as deep learning, reinforcement learning, and transfer learning can enhance the accuracy and efficiency of disease prediction models by automatically learning complex patterns and relationships from large-scale, high-dimensional data.

3.4 Product Scope

A web application-based **Disease Prediction System**, which predicts diseases that the user may be suffering from, based on the symptoms that the user provides. The User can fill in the Patient details along with the symptoms. The symptoms provided are used by a probability-based algorithm to Display a list of probable Diseases according to their relevance to Symptoms. More efficient and robust Data Mining and Machine Learning algorithms that provide well structure and comparatively larger Datasets that are well known for their accurate prediction can replace the current algorithm. We can also increase records in our disease and symptom database by asking doctors and admin to share with us valuable information. We can also add features such as registering for doctor's appointments from the portal itself.

The system allows the patient to give symptoms and according to those symptoms, the system will predict a disease. Here is a list of symptoms that require medication, for promoting any program, Potential disease involves either accurate or false. This report Explain the nature of some of the diagnoses and related symptoms such as a disease but it may not give complete information about the site Symptoms/diagnosis are not related to the patient or family Record or other factor. The Iliad is an expert disease diagnosis system used to describe a relationship to finding Disease. This system uses the Naïve classification, Decision Tree and Random Forest algorithms one by one to diagnose the disease . Clinical decision support systems are used to identify diagonals the patent was recorded. It has three broad categories.

- **Improve patient safety.**
- **Improve the quality of care.**

· **Improve the efficiency of health care delivery.**

Patient safety in terms of minimizing and correcting errors. The second category describes clinical improvement Documentation and patient satisfaction. Describes the third category: Reduce the price and duplicate list, minimize the negativity of the event.

Use Novell to separate features of all datasets here Classifiers based on bias discrimination function. The hybrid algorithm was used to extract unique features from the throat Biological dataset. Machine learning algorithms are used in the Training set. The main goal is to find a relationship among the features that can be used in decision-making. This is a method of preventing many problems with medical data such as missing Prices, Rare Information, and Temporary Data. Machine learning the algorithm is well suited for this type of data. There are two types of use:

1. To find the relationship between the features.
2. Test prediction for future disorders

Chapter 4

Proposed Work

4.1 PROPOSED WORK

The proposed work aims to develop an effective disease prediction system based on symptoms. This project will begin with comprehensive data collection, sourcing datasets containing symptoms and corresponding diseases. Following data acquisition, preprocessing steps will be implemented to handle missing values, encode categorical variables, and ensure uniformity in data format. Feature engineering techniques will then be applied to extract relevant features conducive to disease prediction, potentially involving feature selection, dimensionality reduction, or feature creation. Subsequently, a variety of machine learning algorithms will be explored for model selection, including Decision Trees, Random Forests, Naive Bayes, Support Vector Machines, and Neural Networks. The chosen model will undergo rigorous training on the preprocessed dataset, with a focus on optimization and preventing overfitting through techniques such as cross-validation. Evaluation metrics such as accuracy, precision, recall, and F1-score will be employed to assess model performance. Upon achieving satisfactory results, the trained model will be deployed through a user-friendly interface, enabling individuals to input their symptoms for disease prediction. Continuous monitoring, feedback integration, and ethical considerations will be paramount throughout the project to ensure the system's reliability, accuracy, and responsible use in healthcare settings.

There are two pages in our system.

- First is the patient info page in which the patient's name, gender and age will be registered.
- Second is the disease prediction page in which patient select its symptoms and our system will predict disease based on those symptoms.

4.2 Flowchart

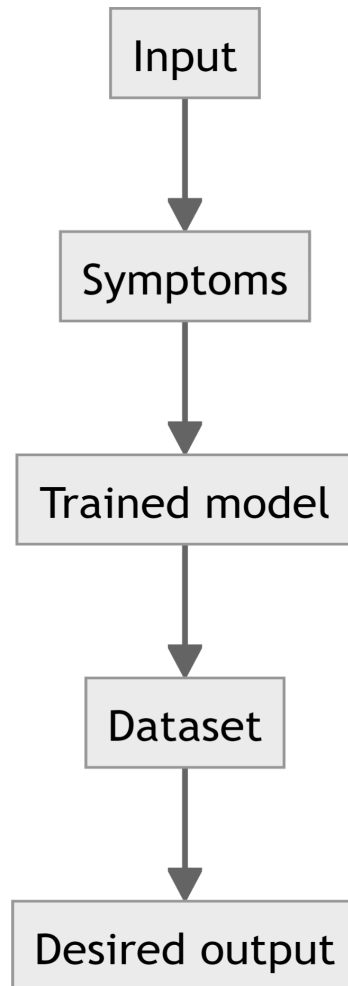


Fig 4.1 Flowchart of Disease Prediction System

4.3 Class Diagram

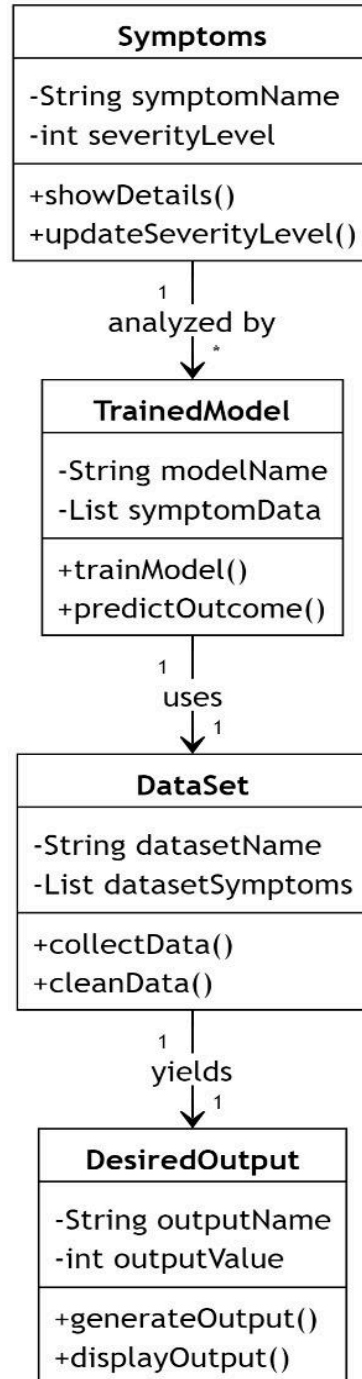


Fig 4.2 Class Diagram of Disease Prediction System

4.4 Sequence Diagram

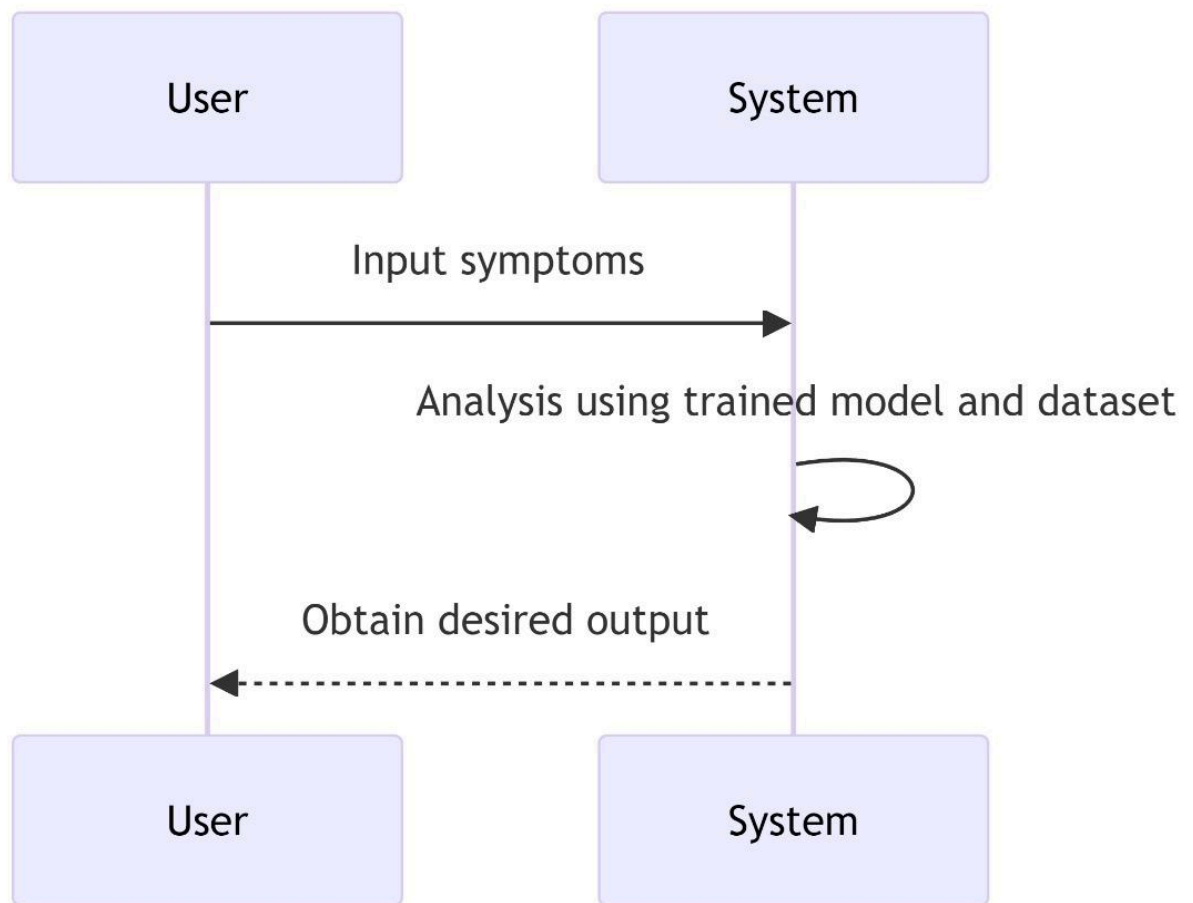


Fig 4.3 Sequence Diagram Of Disease Prediction System

Chapter 5

Implementation

5.1 DATASET OF DISEASE PREDICTION SYSTEM

The dataset was **taken from a study conducted at Columbia University**. It consists of **150 diseases and each disease consists of an average of 8-10 symptoms**. 70% of the dataset used for training was made considering all combinational inputs. The symptoms present for the corresponding disease were marked as 1 and remaining as 0.

It consists of symptoms drop-down options where we have passed a list of symptoms. The user can select any five or upto 10 symptoms and clicking the predict button the disease predicted will be displayed in the text-box.

Prognosis

Fungal infection	hepatitis A	Acne
Allergy	hepatitis B	Urinary tract infection
GERD	hepatitis C	Psoriasis
Chronic cholestasis	hepatitis D	Impetigo
Drug Reaction	hepatitis E	
Peptic Ulcer disease	Alcoholic hepatitis	
AIDS	Tuberculosis	
Diabetes	Common Cold	
Bronchial Asthma	pneumonia	
Hypertension	Dimorphic hemorrhoids(piles)	
Migraine	Heart attack	
Cervical spondylosis	Varicose veins	
Paralysis(brain hemorrhage)	Hypothyroidism	

Jaundice	Hyperthyroidism	
Malaria	Hypoglycemia	
Chicken pox	Osteoarthritis	
Dengue	Arthritis	
Typhoid	(vertigo) Paroxysmal Positional vertigo	

5.2 SYMPTOMS OF DISEASE PREDICTION SYSTEM

1. itching	50. visual disturbances	98. movement_stiffness
2. skin rash	51. receiving blood transfusion	99. spinning_movements
3. nodal skin eruptions	52. receiving unsterile injections	100. loss of balance
4. continuous sneezing	53. coma	101. unsteadiness
5. shivering	54. stomach bleeding	102. weakness of one body side
6. chills	55. distention of abdomen	103. loss of smell
7. joint pain	56. history of alcohol consumption	104. bladder discomfort
8. stomach pain	57. fluid overload	105. foul smell of urine
9. acidity	58. blood in sputum	106. continuous feel of urine
10. ulcers on tongue	59. prominent veins on calf	107. passage of gases
11. muscle wasting	60. palpitations	108. internal itching
12. vomiting	61. painful walking	109. toxic look (typhos)
13. burning micturition	62. pus filled pimples	110. depression
14. spotting urination	63. blackheads	111. irritability
15. fatigue	64. neck_pain	112. muscle pain
16. weight gain	65. scurring	113. altered sensorium
17. anxiety	66. skin peeling	114. red spots over body
18. cold hands and feet	67. silver like dusting	115. belly pain
19. mood swings	68. small dents in nails	116. abnormal menstruation
20. weight loss	69. inflammatory nails	117. dischromic patches
21. restlessness	70. blister	118. watering from eyes
22. lethargy	71. red sore around nose	119. increased appetite
23. patches in throat	72. yellow crust ooze	120. polyuria
24. irregular sugar level	73. swelling of stomach	121. family history
25. cough	74. swelled lymph nodes	122. pain in anal region
26. high fever	75. malaise	123. bloody stool
27. sunken eyes	76. blurred and distorted vision	124. irritation in anus
28. breathlessness		125. mucoid sputum
29. sweating		126. rusty sputum
30. dehydration		127. lack of concentration
31. indigestion		
32. headache		
33. yellowish_skin		

34. dark_urine 35. nausea 36. loss_of_appetite 37. pain_behind_the_eyes 38. back_pain 39. constipation 40. abdominal_pain 41. diarrhoea 42. mild_fever 43. yellow_urine 44. yellowing_of_eyes 45. acute_liver_failure 46. fluid_overload 47. congestion 48. chest_pain 49. pain_during_bowel_movements	77. phlegm 78. throat irritation 79. redness of eyes 80. sinus pressure 81. runny nose 82. dizziness 83. cramps 84. bruising 85. obesity 86. swollen_legs 87. swollen_blood_vessels 88. puffy face and eyes 89. enlarged thyroid 90. brittle nails 91. swollen extremities 92. excessive hunger 93. extra marital contacts 94. drying and tingling lips 95. slurred speech 96. knee pain 97. hip joint pain	128. weakness in limbs 129. fast heart rate 130. pain during bowel movements 131. muscle weakness
---	---	--

5.3 Data Augmentation

By creating additional data points from existing data, a group of techniques known as data augmentation can be used to artificially enhance the amount of data. This includes making minor adjustments to the data or creating new data points using ML models. By creating additional and distinct instances for training datasets, data augmentation helps ML models perform better and more accurately with large datasets. For this reason, it is aimed to obtain more realistic results by adding new data to the dataset used in our previous study [11] and applying data augmentation. Some symptoms play a key role in the prediction of diseases. An example of this is the symptom of loss of smell for Covid disease. For this reason, while creating a new patient history with data augmentation, attention was paid to including these symptoms in each patient's history. In addition, the data augmentation process is aimed to make the data more suitable for real-world cases by completely randomly determining how many times each disease will increase with this process and what symptoms it will contain. As a result of this process, the dataset was increased approximately 15 times, and a new dataset was created with 2006 patient histories.

Chapter 6

Result & Analysis

6.1 Result

TRAINING SET

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	ulcers_on_tongue	muscle_wasting	vomiting	burning_micturition	spotting_urination	fatigue	weight_gain	anxiety	cold_hands_or
2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0
23	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0
24	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0
25	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0
27	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0
29	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0
30	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0
31	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0
32	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
34	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
36	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
37	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
38	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
39	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
40	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
41	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

Dataset of Disease Prediction System

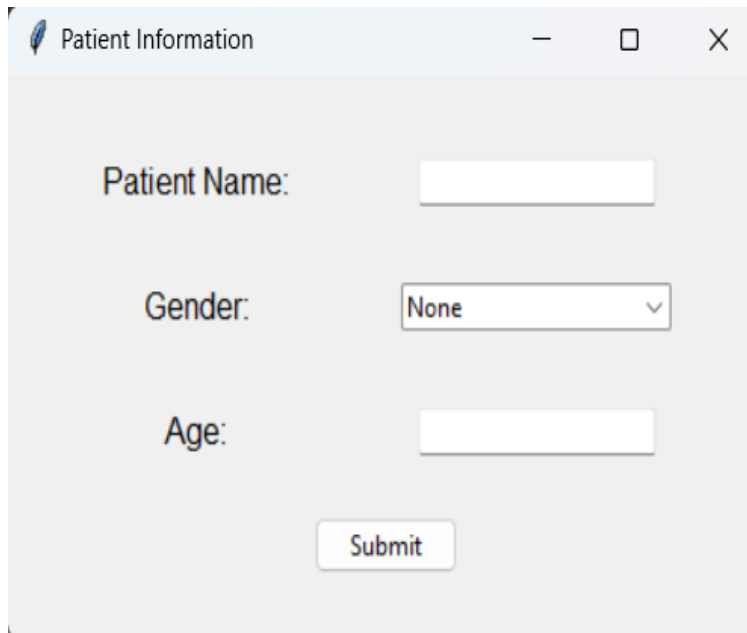
TESTING SET

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
	itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	ulcers_on_tongue	muscle_wasting	vomiting	burning_micturition	spotting_urination	fatigue	weight_gain	anxiety	cold_hands_and_feet	mood_swings
1																			
2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0
5	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
6	1	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
16	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
17	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
18	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
19	0	1	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0
20	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0
21	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
22	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
24	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0
25	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
27	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0
28	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
29	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
35	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0
36	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
39	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
41	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
42	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Dataset of Disease Prediction System

6.2 Evaluation Measures dataset of Disease Prediction System

PATIENT INFORMATION



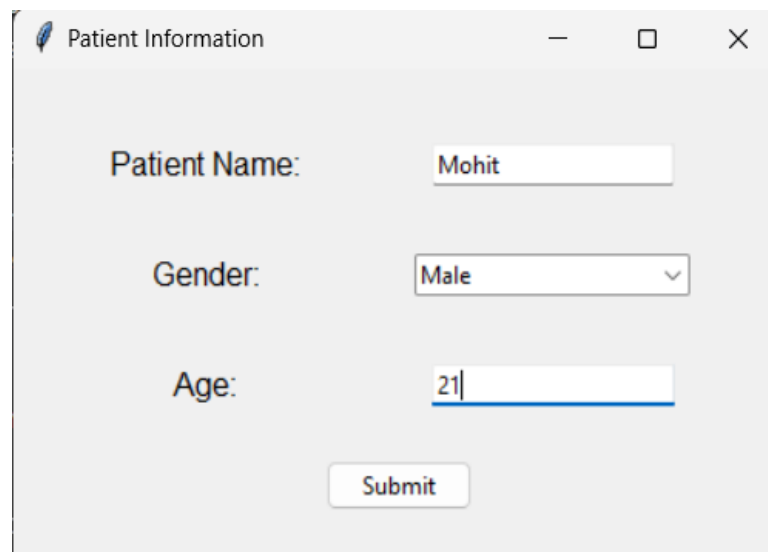
A screenshot of a web form titled "Patient Information". The form has a light gray background and a white border. It contains three input fields: "Patient Name:" with a text box, "Gender:" with a dropdown menu showing "None", and "Age:" with a text box. A "Submit" button is located at the bottom center.

Patient Name:

Gender:

Age:

PATIENT INFORMATION



A screenshot of the same "Patient Information" form, but with data entered. The "Patient Name" field contains "Mohit", the "Gender" dropdown shows "Male", and the "Age" field contains "21". The "Submit" button is still present at the bottom.

Patient Name:

Gender:

Age:

RESULT ANALYSIS

PREDICTED RESULT-1

Disease Predictor using Symptoms

Predico

Symptom 1 abdominal_pain

Symptom 2 back_pain

Symptom 3 belly_pain

Symptom 4 mild_fever

Symptom 5 chest_pain

DecisionTree

Randomforest

NaiveBayes

Clear Entries

Add Symptom

DecisionTree Typhoid

RandomForest Typhoid

NaiveBayes Typhoid

PREDICTED RESULT-1

PREDICTED RESULT-2

Disease Predictor using Symptoms

Predico

Symptom 1: back_pain

Symptom 2: diarrhoea

Symptom 3: loss_of_smell

Symptom 4: cramps

Symptom 5: abdominal_pain

DecisionTree

Randomforest

NaiveBayes

Clear Entries

Add Symptom

DecisionTree: Gastroenteritis

RandomForest: Typhoid

NaiveBayes: Typhoid

PREDICTED RESULT-2

PREDICTED RESULT-3

Disease Predictor using Symptoms

Predico

Symptom 1: back_pain

Symptom 2: chest_pain

Symptom 3: runny_nose

Symptom 4: mild_fever

Symptom 5: muscle_weakness

Buttons: DecisionTree, Randomforest, NaiveBayes, Clear Entries

Add Symptom

DecisionTree: Arthritis

RandomForest: GERD

NaiveBayes: GERD

PREDICTED RESULT-3

Accuracy:

Naive Bayes Algorithms [17] were best with a model accuracy of 94.8%. Following the Naive Bayes model [17] is a weighted KNN model [18] with an accuracy of 93.5%. The research papers using the SVM model [29, 30] where's also very close. However, the suggested model, that is Random forest model, yields the most accurate result, 97% as compared to earlier methods

Precision: Precision is the proportion of correct positive results, and is calculated by

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

Chapter 7

Conclusion and Future Work

7.1 Conclusion

The main aim of this disease prediction system is to predict the disease on the basis of the symptoms. This system takes the symptoms of the user from which he or she suffers as input and generates final output as a prediction of disease. Average prediction accuracy probability of 100% is obtained. Disease Predictor was successfully implemented using the grails framework. This system gives a user-friendly environment and is easy to use. As the system is based on the web application, the user can use this system from anywhere and at any time. In conclusion, for disease risk modeling, the accuracy of risk prediction depends on the diversity feature of the hospital data. This systematic review aims to determine the performance, limitations, and future use of Software in healthcare. Findings may help inform future developers of Disease Predictability Software and promote personalized patient care. The program predicts Patient Diseases. Disease Prediction is done through UserSymbols. In this System Decision tree, Unplanned Forest, the Naïve Bayes Algorithm is used to predict diseases. Forth data format, the system uses the Machine Learning algorithm Process Data on Database Data namely, Random Forest, Decision Tree, Naive Bayes. System accuracy reaches 98.3%. machine learning skills are designed to successfully predict outbreaks.

7.2 Future Scope

- ▶ The future scope of disease prediction using symptoms is quite promising, with several exciting possibilities
- ▶ Precision Medicine
- ▶ Remote Healthcare and Telemedicine
- ▶ Early Detection and Prevention
- ▶ These advancements would contribute to more accurate predictions, wider application, and improved diagnostic capabilities for more Disease.

References

1. Al-Aidaroos, K.M., Bakar, A.A. and Othman, Z.: Medical data classification with Naïve Bayes approach. *Information Technology Journal*. 11(9), 1166 (2012).
2. Asuncion, A. and Newman, D.: UCI machine learning repository downloaded from <https://ergodicity.net/2013/07/>.
3. J Soni, J., Ansari, U., Sharma, D. and Soni, S.: Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, 17(8), 43-48 (2011).
4. Pattekari, S.A. and Parveen, A.: Prediction system for heart disease using Naïve Bayes. *International Journal of Advanced Computer and Mathematical Sciences*. 3(3), 290-294 (2012).
5. Masethe, H.D. and Masethe, M.A.: Prediction of heart disease using classification algorithms. In *Proceedings of the world Congress on Engineering and Computer Science*. 2, 22- 24 International Association of Engineers, Francisco (2014).
- 6 Shouman, M., Turner, T. and Stocker, R.: Using decision trees for diagnosing heart disease patients. In: *Proceedings of the Ninth Australasian Data Mining Conference*, Volume. 121, 23-30. Association of Computing Machinery, Victoria (2011).
7. Shouman, M., Turner, T. and Stocker, R., 2012. Integrating decision tree and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients. In: *Proceedings of the International Conference on Data Science (ICDATA)*, pp. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (World Comp), Monte Carlo (2012).
8. Vembandasamy, K., Sasipriya, R. and Deepa, E., 2015. Heart disease detection using Naive Bayes algorithm. *International Journal of Innovative Science, Engineering & Technology*, 2
- 9, pp.441-444. Er. Harjeet Singh, Assistant Professor and Researcher at LIET Gr. Noida
Mr. Ankit Mehta, Assistant Professor and Researcher at LIET Gr. Noida
<https://www.kaggle.com/neelima98/disease-predictionusingmachine-learning>