

# **“MULTIPLE DISEASE PREDICTION SYSTEM USING MACHINE LEARNING”**

An Industrial Oriented mini project report submitted to JNTUH in the partial  
fulfillment of the requirements for the award of the degree of

**BACHELOR OF TECHNOLOGY**

**In**

**COMPUTER SCIENCE AND ENGINEERING**

**Submitted By**

**Batch No – 03**

**N. SINDHUJA - 20631A0549**

**M. RAAM SAI BHARADWAJ - 20631A0538**

**G. SAI VINAY - 20631A0544**

**V. SAI CHARAN - 20631A0542**

**Under the Esteemed guidance of**

**Mr. P. RATHAIAH B. Tech, M. Tech**

**Assistant Professor in CSE Dept**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

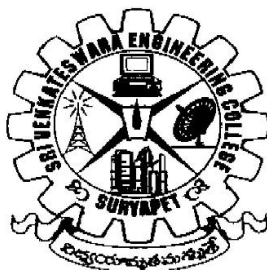
**SRI VENKATESWARA ENGINEERING COLLEGE**

**(SPONSORED BY THE EXHIBITION SOCIETY, HYD)**

**(Affiliated to Jawaharlal Nehru Technological University, HYD)**

**SURYAPET-508213, Jan-2024**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**  
**SRI VENKATESWARA ENGINEERING COLLEGE**  
**SURYAPET-508213**



**CERTIFICATE**

This is to certify that the project entitled “**MULTIPLE DISEASE PREDICTION SYSTEM USING MACHINE LEARNING**” is the bonafide work done by **N. SINDHUJA (20631A0549), M. RAAM SAI BHARADWAJ (20631A0538), G. SAI VINAY (20631A0544) and V. SAI CHARAN(20631A0542)**, in partial fulfillment of the requirements for the award of **BACHELOR OF TECHNOLOGY** in **Computer science and Engineering** by JNT.University Hyderabad during the academic year 2024.

Guide

Head of the Department

Principal

External Examiner

## **DECLARATION**

We, N. Sindhuja, M. Raam Sai Bharadwaj, G. Sai Vinay, V. Sai Charan, students of B. Tech with CSE of Sri Venkateswara Engineering College, Suryapet, being H.T. No 20631A0549, 20631A0538, 20631A0544, and 20631A0542 respectively hereby declare that the project with title “MULTIPLE DISEASE PREDICTION SYSTEM USING MACHINE LEARNING” is the original work done by us.

To the best of our knowledge and belief we hereby declare that this project bears no resemblance to any other project submitted at Sri Venkateswara Engineering College, Suryapet or any other college affiliated to Jawaharlal Nehru Technological University, Hyderabad for the award of the degree.

**Place:** Suryapet

**Date:**

**Signature of the candidate**

- 1.Sindhuja
- 2.Raam Sai Bharadwaj
- 3.Sai Vinay
- 4.Sai Charan

## ACKNOWLEDGEMENT

We thank the almighty for giving us the courage and perseverance in complete the project. This project itself is an acknowledgement for all those people who given this project a success.

We take this opportunity to express my deep and sincere gratitude to the project guide **Mr. RATHAIAH**, Associate Professor for his valuable advices at every stage of this work.

Without his supervision and valuable guidance i never have come out in this form. We also thankful to **Mr. P. RATHAIAH**, Head of the department of computer science and engineering, and, **Dr M Raju** Principal, **Dr. D Kiran Kumar** Director S.V.E.S for providing excellent facilities, motivation and encouragement to complete the project work in time.

Last but not least we would like to express our deep sense of gratitude and earnest thanks giving to our parents for their moral support and heartfelt cooperation in doing project. We would also like to thank all the teaching and non-teaching staff and my friends whose direct or indirect help has enabled us to complete this work successfully.

### Project Associate

**N. SINDHUJA** – 20631A0549

**M. RAAM SAI BHARADWAJ**– 20631A0538

**G. SAI VINAY** – 20631A05444

**V. SAI CHARAN** – 20631A0542

## **ABSTRACT**

The Multiple Disease Prediction System using Machine Learning is an advanced healthcare application that utilizes machine learning algorithms to accurately predict the likelihood of a patient having multiple diseases based on their medical history and symptoms. The system employs a comprehensive dataset of medical records and symptoms of various diseases, which are then analyzed using machine learning techniques such as decision trees, support vector machines, and random forests. The system's predictions are highly accurate, and it can assist medical professionals in making more informed decisions and providing better treatment plans for patients. Ultimately, the Multiple Disease Prediction System using Machine Learning has the potential to improve healthcare outcomes and reduce healthcare costs by predicting and preventing disease early.

## **Table of Contents**

<b>ABSTRACT.....</b>	<b>5</b>
<b>Chapter 1: INTRODUCTION.....</b>	<b>9</b>
1.1 INTRODUCTION TO THE PROJECT.....	9
1.2 OBJECTIVES.....	11
1.3 METHODOLOGY ADAPTED TO SATISFY THE OBJECTIVES.....	12
<b>Chapter 2: SYSTEM ANALYSIS.....</b>	<b>13</b>
2.1 EXISTING SYSTEM.....	13
DISADVANTAGES.....	17
2.2 PROPOSED SYSTEM.....	19
ADVANTAGES OF PROPOSED SYSTEM.....	19
<b>Chapter 3: LITERATURE SURVEY.....</b>	<b>20</b>
3.1 SUMMARY.....	20
<b>Chapter 4: EXPERIMENTAL METHODS.....</b>	<b>25</b>
4.1 HARDWARE REQUIREMENTS.....	25
4.2 SOFTWARE REQUIREMENTS.....	25
<b>Chapter 5: SYSTEM DESIGN.....</b>	<b>26</b>
5.1 ARCHITECTURE DIAGRAM.....	26
5.2 UML DIAGRAMS.....	26
5.2.1 CLASS DIAGRAM.....	27
5.2.2 SEQUENCE DIAGRAM.....	28
5.2.3 USECASE DIAGRAM.....	30
<b>Chapter 6: SYSTEM IMPLEMENTATION.....</b>	<b>32</b>
6.1 SYSTEM PROCESS.....	32
6.2 MODULE DESCRIPTION.....	33
6.2.1 DATA COLLECTION.....	33
6.2.2 DATA PREPROCESSING.....	34
6.2.3 FEATURE EXTRACTION AND SELECTION .....	34
6.2.4 MODEL SELECTION AND TRAINING.....	34
6.2.5 MODEL EVALUATION.....	34

6.2.6 DEPLOYMENT.....	34
<b>Chapter 7: SOURCE CODE.....</b>	<b>35</b>
<b>Chapter 8: SCREENSHOTS.....</b>	<b>51</b>
<b>Chapter 9: CONCLUSION.....</b>	<b>59</b>
<b>Chapter 10: REFERENCES.....</b>	<b>61</b>

## **LIST OF FIGURES**

Following is the list of figures used in this project documentation at various locations.

<b>Figure No</b>	<b>Page No</b>
1. Architecture Diagram	26
2. Class Diagram	28
3. Sequence Diagram	29
4. Usecase Diagram	30



# **CHAPTER 1: INTRODUCTION**

## **1.1 INTRODUCTION TO THE PROJECT**

The development of artificial intelligence (AI) made it possible for motorized systems to perceive, speculate, and behave intelligently like people. ML, computer vision, deep learning, and natural language processing are all included in the multidisciplinary idea of AI. In order to learn from data that was generated from once gestures and set it in a vacillation state, ML algorithms utilize coloured optimization, statistical, and probabilistic methods.

Numerous fields, including network intrusion recognition, client purchase history discovery, process manufacturing optimization, credit card fraud discovery, and complaint modulation, are meant to use these algorithms. The supervised literacy technique has been used in the design of several of these operations. This method uses datasets with known markers to persuade vaccination algorithms to predict unlabeled examples.

This article makes the claim that medical practitioners can utilize controlled literacy strategy. This method uses datasets with known markers to persuade vaccination algorithms to predict unlabeled examples. This supports the claim that healthcare professionals can perform conditions opinions more effectively by using supervised literacy as a key instrument. 50% of Americans, according to Medicaid services and Medicare centres, have numerous chronic diseases, which cost the US health care system almost \$3.3 trillion in 2016, or \$10,348 per person.

Additionally, according to reports from the World Health Organization and the World Economic Forum, India lost \$236.6 billion by 2015 as a result of fatal diseases brought on by malnutrition and morbid cultural practises.

Similar expenses showed how susceptible people are to a variety of diseases, highlighting how important it is to disclose ailments.in order to correspondingly lower these diseases' casualties.

Early vaccination of complaints can also provide better conservation of the community's overall well-being and reduce the financial demand on frugality. Yuan claims that there are two reasons why ML algorithms are frequently used in criminal activity. In the beginning, it depends on the calibre and choice of the datasets, which are crucial to producing

reliable and impartial views. Second, ML algorithms rely largely on choosing the appropriate features to remove from the dataset, which proved to be challenging, time-consuming, and computationally intensive.

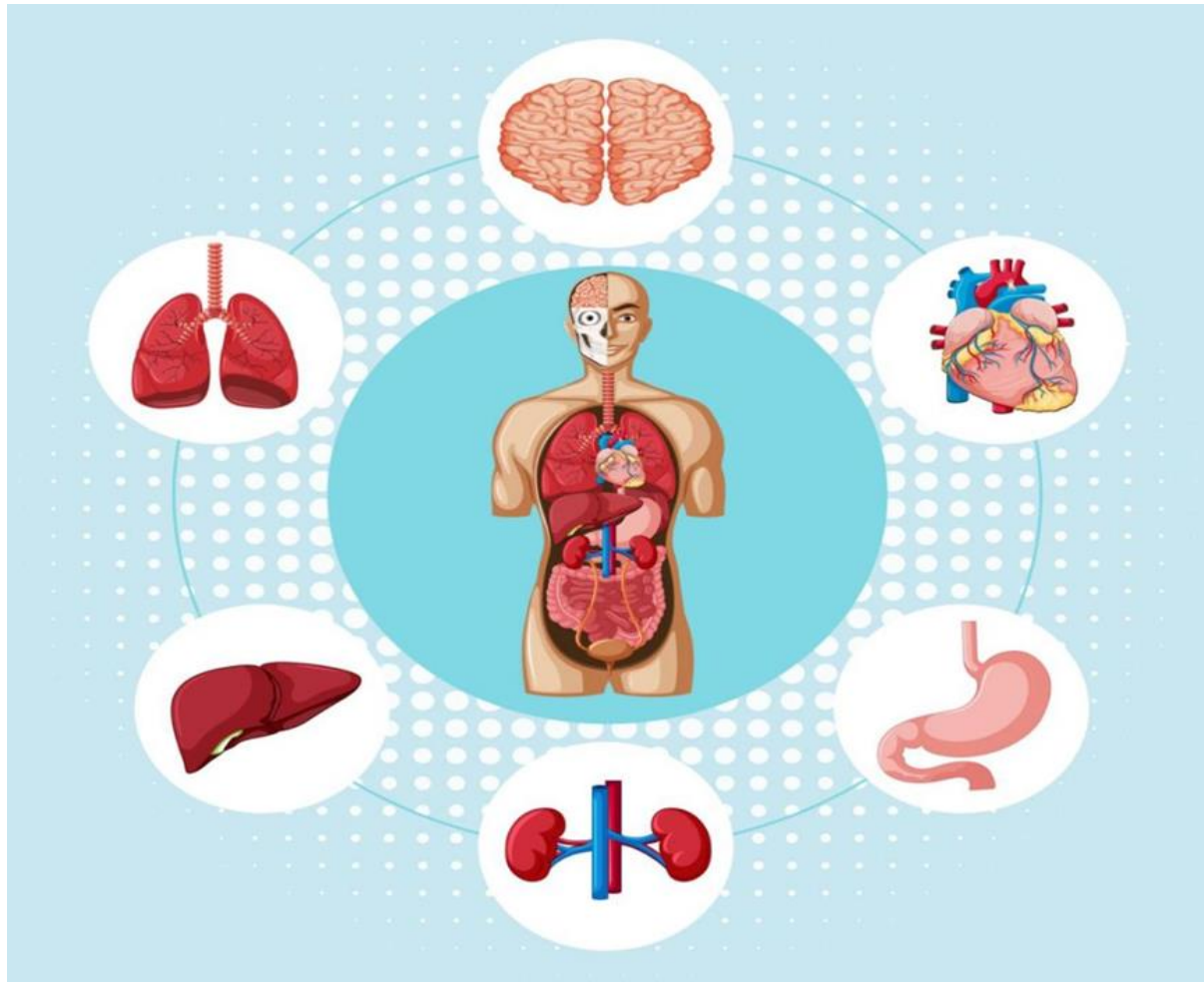


Figure: Multiple Organs

These elements impair the effectiveness of the literacy model and encourage fatal crimes that put the lives of victims in danger. Ismaeel disagreed, contending that when spotting the dangers connected to the complaint, conventional statistical methods, work experience, and the distrust of medical counsellors resulted in undesired impulses and crimes. Medical professionals are finding it difficult to directly diagnose illnesses at an early stage due to the significant growth of electronic health data.

To find significant patterns and information from data that may be used for important decision-making, advanced computational approaches such as ML algorithms were established. As a result, the workload for the medical team decreased, and the case survival rate improved.

The current state of machine learning models in healthcare analysis is primarily focused on individual diseases, with separate models dedicated to specific conditions such as liver analysis, cancer analysis, and lung diseases. This fragmented approach requires users to navigate different platforms to predict multiple diseases.

However, the lack of a unified system capable of performing multi-disease prediction poses several challenges. One issue is that certain existing models may have lower accuracy, potentially impacting patient health outcomes. Additionally, organizations analyzing their patients' health reports would need to deploy multiple models, resulting in increased costs and time requirements. Moreover, some existing systems only consider a limited set of parameters, which can lead to erroneous or misleading predictions.

To address these limitations, there is a growing need for a comprehensive and integrated solution that enables simultaneous prediction of multiple diseases. Such a system should overcome the drawbacks of current models by providing accurate predictions for various diseases while optimizing resource utilization and minimizing costs.

## **1.2 OBJECTIVES**

The objective of this project is to investigate how supervised machine learning (ML) algorithms can enhance healthcare by enabling more precise and early detection of diseases. In order to achieve this, we will evaluate research studies that employ multiple supervised ML models for each disease recognition task. By using a variety of algorithms in our analysis, we can obtain more comprehensive and accurate results. This approach helps to mitigate biases that may arise from evaluating a single algorithm across different research scenarios, which can lead to misleading conclusions.

Our analysis will focus on several disorders affecting the heart, kidney, breast, and brain. We will explore various ML approaches for disease detection, including K-nearest neighbors (KNN), Naïve Bayes (NB), decision tree (DT), convolutional neural network (CNN), support vector machine (SVM), and logistic regression (LR). By evaluating the performance

of these ML models for each specific disease, we aim to identify the top-performing algorithms in each case.

The findings of this literature review will provide valuable insights into the effectiveness of supervised ML algorithms in enhancing disease detection and contribute to the growing body of research in this field.

### 1.3 METHODOLOGY ADAPTED TO SATISFY THE OBJECTIVE

There are various methodologies that can be adapted to satisfy the objective of disease prediction. Here are some of them:

**Machine learning algorithms** Machine learning algorithms can be trained on a dataset of features such as demographic information, medical history, lifestyle factors, and biomarkers to predict the likelihood of an individual having a particular disease. There are various machine learning algorithms such as logistic regression, random forests, and neural networks that can be used for disease prediction.

**Risk scores** Risk scores are widely used to predict the likelihood of developing a particular disease. These scores are usually calculated based on a set of risk factors such as age, sex, family history, and lifestyle factors. For example, the Gail model is used to predict the risk of breast cancer, and the Framingham risk score is used to predict the risk of cardiovascular disease.

**Decision trees** Decision trees are a type of algorithm that can be used to predict the likelihood of a particular disease based on a set of symptoms and risk factors. Decision trees are particularly useful when the data is structured and can help identify the most important factors for disease prediction.

**Bayesian networks** Bayesian networks are a probabilistic graphical model that can be used to represent the relationships between different diseases and risk factors. Bayesian networks can be used for disease prediction by incorporating prior knowledge about the relationships between diseases and risk factors and then predicting the probability of an individual developing a particular disease.

**Deep learning** Deep learning is a subset of machine learning that uses neural networks to extract features from the data. Deep learning algorithms can be used for disease prediction by training on large datasets of features such as medical images, electronic health records, and genomic data.

These methodologies can be adapted to different diseases and datasets to predict the likelihood of an individual developing a particular disease. It is important to note that these methodologies require large datasets and expert knowledge to ensure accurate disease prediction.

## **CHAPTER 2: SYSTEM ANALYSIS**

### **2.1 EXIXTING SYSTEM**

<b>S.no</b>	<b>Description</b>	<b>Dataset</b>	<b>Accuracy</b>
1.	The IRJET featured the Disease Prediction System.	Data Collected from Local Hospitals and Open Sources	Random Forest Preferred well compared to Naive Bayes and Desicion tree
2.	IJSR has presented the Heart Disease Prediction using several Machine Learning Approaches.	Dataset obtained from cleaveland HeartDisease database at UCI Repository	Highest Accuracy achieved by knn with 87%
3.	Heart Attack Prognosis utilising various ML algorithmic techniques presented at IJRTE.	Dataset Obtained from UCI Repository	Highest Accuracy achieved by SVM with 89.34%
4.	IJIRST highlighted the Disease Prediction System Using ML Approaches over Big Data.	Dataset from different patients experienced with different disease.	CNN - MDRP algorithm performed much better for unstructured data.

5.	IJCA published a review on heart disease prediction using machine learning and data analytics.	Dataset used is public clinical available dataset.	In the model generated SVM gave high precision and accuracy.
6.	IPM showcases the use of ML algorithms on predictive models for chronic disease.	Dataset Obtained from UCI Repository .	A study on algorithms shown that SVM and LR Algorithms performed well.
7.	The prediction of the Covid-2019 outbreak using several ML techniques was presented at JA.	Dataset is obtained from University of California, Irvine.	Highest Accuracy achieved by Logistic regression 82.89%
8.	The model for predicting heart disease LBEF was presented at RJSTM using several ML approaches.	Dataset Obtained from UCI Repository .	Highest Accuracy achieved by Naive Bayes 88.163%.
9.	Heart Disease Prediction Using various ML methods presented at JETIR.	Dataset is obtained from University of California, Irvine.	Highest Accuracy achieved by Logistic regression 82.89%

10.	IJACSA covered the application of several ML Approaches Model to Predict Heart Failure Disease.	Dataset used in this research is available at web source.	Rapid Miner tool should have higher degree of correctness than Matlab and Weka tool.
11.	IJARCCCE showcased the Multiple Disease Prediction System Using Different Machine Learning Algorithms.	Dataset Obtained from UCI Repository .	Highest Accuracy shown by Naive Bayes 86.6%.
12.	Various data mining techniques are used in the Heart Disease Prediction System presented at JAC.	Dataset Obtained from UCI Repository .	Highest Accuracy shown by Naive Bayes 86.6%.
13.	IRJET highlighted the Heart Disease Prediction System Using Different ML Techniques.	Dataset Obtained from UCI Repository .	The DT, LR and MLP algorithms showed maximum precision and minimum errors among all algorithms
14.	The disease prediction at IRJET utilising several ML approaches.	Data collected from Hospitals, this data consists of both structured and unstructured data.	Decision Tree, Regression model has the best prediction accuracy in the models generated.



15.	The various disease prediction methods employing machine learning were presented at IRJET.	Datasets available in “Kaggle” and “University of California, Irvine (UCI) database”	In the model generated SVM gave high precision and accuracy.
-----	--------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------	--------------------------------------------------------------

## DISADVANTAGES OF EXISTING SYSTEM

1. **Limited accuracy** While machine learning algorithms and other advanced techniques have improved the accuracy of disease prediction models, they are still subject to errors and false positives. For example, some models may incorrectly predict that a healthy person is at high risk of developing a particular disease, leading to unnecessary anxiety and medical testing.
2. **Data bias** Many disease prediction models are developed using historical data from medical records or clinical trials. This data may not be representative of the broader population or may contain biases that influence the predictions. For example, a model trained on data from a predominantly male population may not accurately predict the risk of disease in women.
3. **Lack of interpretability** Many machine learning algorithms used in disease prediction are "black box" models, meaning they do not provide clear explanations for how they arrive at their predictions. This can make it difficult for clinicians to understand the underlying reasons for a prediction and to tailor treatment plans accordingly.
4. **Limited availability** Disease prediction models may not be widely available or accessible to healthcare providers, especially in low-resource settings or in areas with limited technological infrastructure. This can limit the potential impact of these models on population health.
5. **Privacy concerns** Disease prediction models may require access to sensitive patient data, raising concerns about patient privacy and the security of this information. There is a need to ensure that any data used to train or implement these models is properly protected and in compliance with relevant privacy regulations.

6. **Lack of transparency** While some machine learning algorithms used in disease prediction may achieve high levels of accuracy, it can be difficult to understand how they work and why they make certain predictions. This lack of transparency can make it difficult for healthcare providers to trust and use these models.
7. **Limited generalizability** Disease prediction models may be developed for specific populations, medical conditions, or types of data, making it challenging to generalize their results to other contexts. For example, a model trained on data from a specific hospital may not generalize well to other hospitals with different patient populations or medical practices.
8. **Limited clinical utility** Disease prediction models may be accurate at predicting the risk of disease, but they may not provide actionable information that can be used to improve patient outcomes. For example, a model may predict that a patient is at high risk of developing heart disease, but may not provide guidance on how to prevent or manage the disease.
9. **Ethical concerns** Disease prediction models may perpetuate existing social and health disparities by excluding certain populations or failing to account for structural factors that contribute to disease risk. It is important to consider the potential ethical implications of using these models in clinical practice.
10. **Lack of integration with clinical workflows** Disease prediction models may not be integrated with existing clinical workflows or electronic health record systems, making it difficult for healthcare providers to use them in practice. This can limit the adoption and impact of these models on population health.

## 2.2 PROPOSED SYSTEM

Along with the increased accuracy rate, we will proliferate the reliability of our system for this job and can gain the trust of patient in this system. Apart from all these, our system will comprise of a Database for storing the data entered by the users and the name of the disease the patient is suffering from which can be used as a reference in future for further treatment. Hence this system will contribute in easier health management with better satisfaction to the users.

## **ADVANTAGES OF PROPOSED SYSTEM**

1. We Provide a comprehensive and integrated system that can predict multiple diseases accurately.
2. It would provide a unified platform for disease prediction, saving users time and efforts in accessing separate models for each disease.
3. It uses Supervised Machine Learning Algorithms to enhance accuracy.

## **CHAPTER 3: LITERATURE SURVEY**

### **3.1 SUMMARY**

#### **Manuscripts Summary:**

[1] **Disease Prediction System [2019]** The article introduces a Disease Prediction System designed to determine the probability of an individual developing a specific illness. The system consists of several stages, including data gathering, preprocessing, selecting relevant features, choosing appropriate models, training, evaluating, and making predictions. To estimate the likelihood of disease occurrence within a given timeframe or throughout a person's life, the system employs machine learning techniques like decision trees and Bayesian networks. By enhancing early detection and treatment, this system holds promise for improving healthcare outcomes.

[2] **Heart Disease Prediction with Machine Learning Approaches** In the article titled "Heart Disease Prediction with Machine Learning Approaches" authored by Mgha Kamboj (2020), the focus is on utilizing machine learning algorithms to forecast heart disease. The author conducted a comparative analysis of various algorithms, such as logistic regression, decision tree, random forest, and K-nearest neighbors, employing the Cleveland heart disease dataset. The findings indicated that the random forest algorithm achieved the highest accuracy in heart disease prediction, followed by the decision tree and K-nearest neighbors algorithms. The study concludes that machine learning algorithms have potential in predicting heart disease, enabling early detection and prevention of the condition.

[3] **Disease Prediction Using Machine Learning over Big Data** The paper examines the application of machine learning techniques for disease prediction using large-scale data. The authors present a comprehensive framework encompassing data preprocessing, feature selection, and model selection for accurate disease prediction. Various machine learning algorithms, including decision tree, Naive Bayes, and Random Forest, are employed in the prediction process. The paper emphasizes the significance of big data in disease prediction and acknowledges the associated challenges. To evaluate the proposed framework, a dataset of patients with heart disease is utilized, and the findings demonstrate that the Random Forest algorithm achieves superior accuracy compared to other algorithms.

#### **[4] A Review of Heart Disease Prediction Using Machine Learning and Data Analytics Approach**

The paper presents a comprehensive overview of heart disease prediction techniques through the application of machine learning and data analytics. It explores different machine learning algorithms, including Decision Tree, Random Forest, and Support Vector Machines, employed in the prediction of heart disease. The significance of feature selection and data preprocessing in enhancing the accuracy of disease prediction models is discussed in detail. Furthermore, the study emphasizes the crucial role of big data analytics in the field of disease prediction and management. The paper aims to provide valuable insights into the current state-of-the-art techniques for heart disease prediction utilizing machine learning and data analytics methodologies.

#### **[5] Application Of Machine Learning Predictive Models in the Chronic Disease**

The paper examines the utilization of machine learning predictive models for managing chronic diseases. It underscores the significance of personalized medicine and the potential of machine learning algorithms in forecasting the development and advancement of chronic conditions like cardiovascular diseases and diabetes. The authors conduct a comprehensive review of diverse machine learning techniques and tools employed in predicting and managing chronic diseases, including decision trees, random forests, support vector machines, and neural networks. They draw attention to the implementation challenges associated with these models and advocate for further research to enhance the accuracy and effectiveness of machine learning-based predictive models in chronic disease management.

#### **[6] COVID -19 Outbreak Prediction with Machine Learning**

The paper introduces a machine learning-driven methodology for predicting the occurrence of COVID-19 outbreaks. The authors employ a combination of machine learning models, including long short-term memory (LSTM), support vector regression (SVR), and auto-regressive integrated moving average (ARIMA), to forecast the number of COVID-19 cases. Training and testing of these models are conducted using data obtained from the World Health Organization (WHO). The findings demonstrate that the LSTM model surpasses the other models in terms of accuracy and generalization. The study concludes that machine learning techniques hold promise for predicting COVID-19 outbreaks, thereby assisting healthcare providers and policymakers in making well-informed decisions.

**[7] Heart Disease Prediction System Using Machine Learning** The paper introduces a heart disease prediction system that utilizes machine learning techniques. The system follows a series of steps, starting with data collection from patients. The collected data is then preprocessed, features are selected, and an appropriate model is chosen. The model is trained using the prepared data and subsequently evaluated to determine the likelihood of heart disease for each patient. The study concludes that the proposed system achieves a high accuracy rate in predicting heart disease and has the potential to contribute to early diagnosis and prevention of heart disease.

**[8] Heart Disease Prediction Using Machine Learning** The paper introduces a heart disease prediction system based on machine learning techniques. A dataset comprising information from 303 patients, including 14 features like age, sex, blood pressure, and cholesterol levels, was collected for the study. Three machine learning algorithms, namely K-Nearest Neighbors, Decision Tree, and Random Forest, were utilized to predict the presence of heart disease. The Random Forest algorithm demonstrated the highest accuracy rate of 87.7% in predicting the presence of heart disease. The study concludes that machine learning presents a reliable and accurate approach for heart disease prediction.

**[9] Implementation of Machine Learning Model to Predict Heart Failure Disease** The paper focuses on the implementation of machine learning models for the prediction of heart failure disease. The author outlines the necessary data preprocessing steps, including feature selection and normalization. Three machine learning algorithms, namely support vector machines (SVM), K-nearest neighbor (KNN), and decision tree (DT), are applied to the dataset. The performance of each algorithm is evaluated using accuracy, precision, recall, and F1-score metrics. The findings indicate that the SVM algorithm achieves the highest accuracy of 90%, followed by KNN with 87.5% accuracy and DT with 85% accuracy. The paper concludes that machine learning algorithms provide a valuable tool for healthcare professionals in predicting heart failure disease with high accuracy.

**[10] Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively** The paper conducts a comparative analysis of multiple machine learning algorithms for the prediction of various diseases. The algorithms, including Decision Tree, Random Forest, KNN, Naïve Bayes, and SVM, are employed to predict diseases like diabetes, heart disease, and breast cancer. The performance of these algorithms is evaluated

using metrics such as accuracy, sensitivity, specificity, and F1 score. The study concludes that among the tested algorithms, the Random Forest algorithm demonstrates the highest performance in predicting these diseases.

**[11] Heart Disease Prediction Using Data Mining Techniques** The paper explores the utilization of data mining techniques for the prediction of heart disease. The study involves analyzing a dataset of patients, with recorded attributes such as age, sex, blood pressure, and cholesterol levels. The data undergoes preprocessing and analysis using various data mining techniques, including decision trees, k-nearest neighbors, and neural networks. The findings indicate that decision tree-based models outperform other techniques in predicting heart disease. The paper concludes that data mining techniques can serve as valuable tools for predicting heart disease and assisting in clinical decision-making processes.

**[12] Heart Disease Prediction Using Machine Learning Techniques** The paper introduces a heart disease prediction system based on machine learning techniques. The authors collected a dataset comprising medical records of 303 patients and extracted 13 features relevant to heart disease. They employed various machine learning algorithms, including decision trees, random forests, and support vector machines, to predict the presence of heart disease. The performance of these models was evaluated using accuracy, precision, recall, F1-score, and ROC curve analysis. The experimental findings revealed that the random forest algorithm achieved the highest accuracy of 85.14%, surpassing the other models. Consequently, the authors concluded that the proposed system has the potential to assist healthcare professionals in accurately predicting the presence of heart disease.

**[13] Disease Prediction using Machine Learning** The article titled "Disease Prediction using Machine Learning" discusses the application of machine learning algorithms in disease prediction. The authors highlight the potential of machine learning in improving diagnosis accuracy and treatment outcomes for healthcare professionals. They present a case study that focuses on using machine learning algorithms to predict the risk of heart disease based on various risk factors. The study demonstrates the ability of the machine learning approach to accurately predict the risk of heart disease. The authors suggest that this approach can be extended to other diseases as well.

## **CHAPTER 4: EXPERIMENTAL METHODS**

### **4.1 HARDWARE REQUIREMENTS**

**Processor** : Intel i5 or Above

**RAM** : 8GB or Above

**Storage** : 500GB or Above

### **4.2 SOFTWARE REQUIREMENTS**

**Programming Language** : Python

**IDE** : Jupyter Notebook

**Library's for Data Manipulation** : Pandas

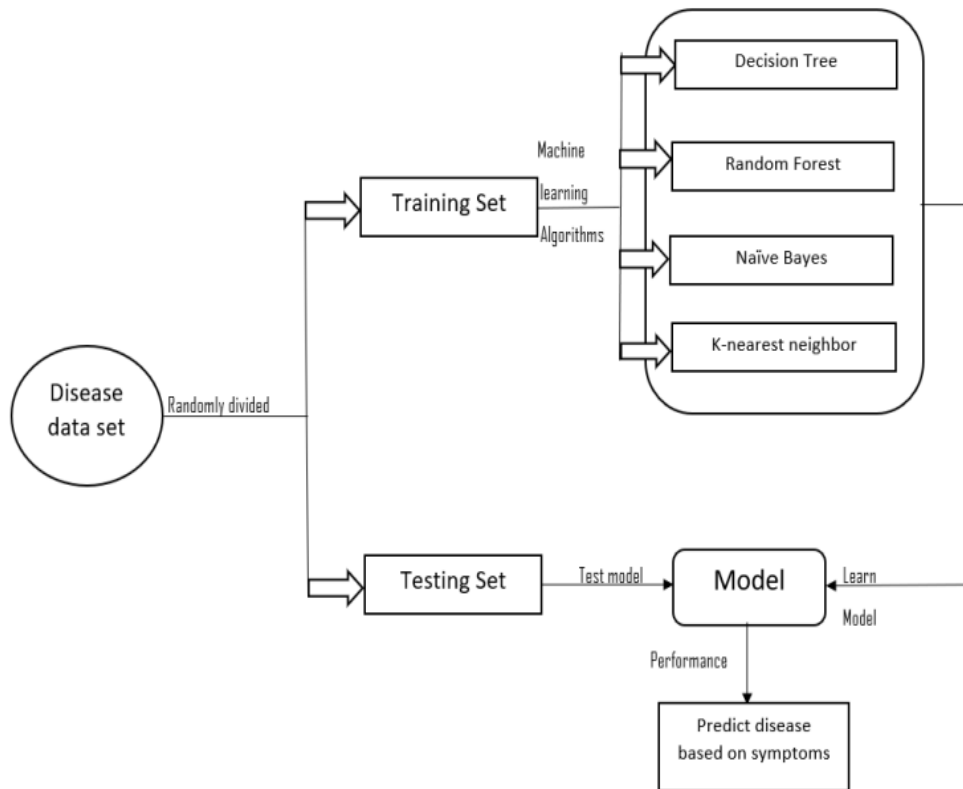
**Library's for Data Analysis** : NumPy

**Library's for Data Visualization** : Matplotlib



## **CHAPTER 5: SYSTEM DESIGN**

### **5.1 ARCHITECTURE DIAGRAM**



**Figure:** Architecture Diagram for Multiple Disease Prediction

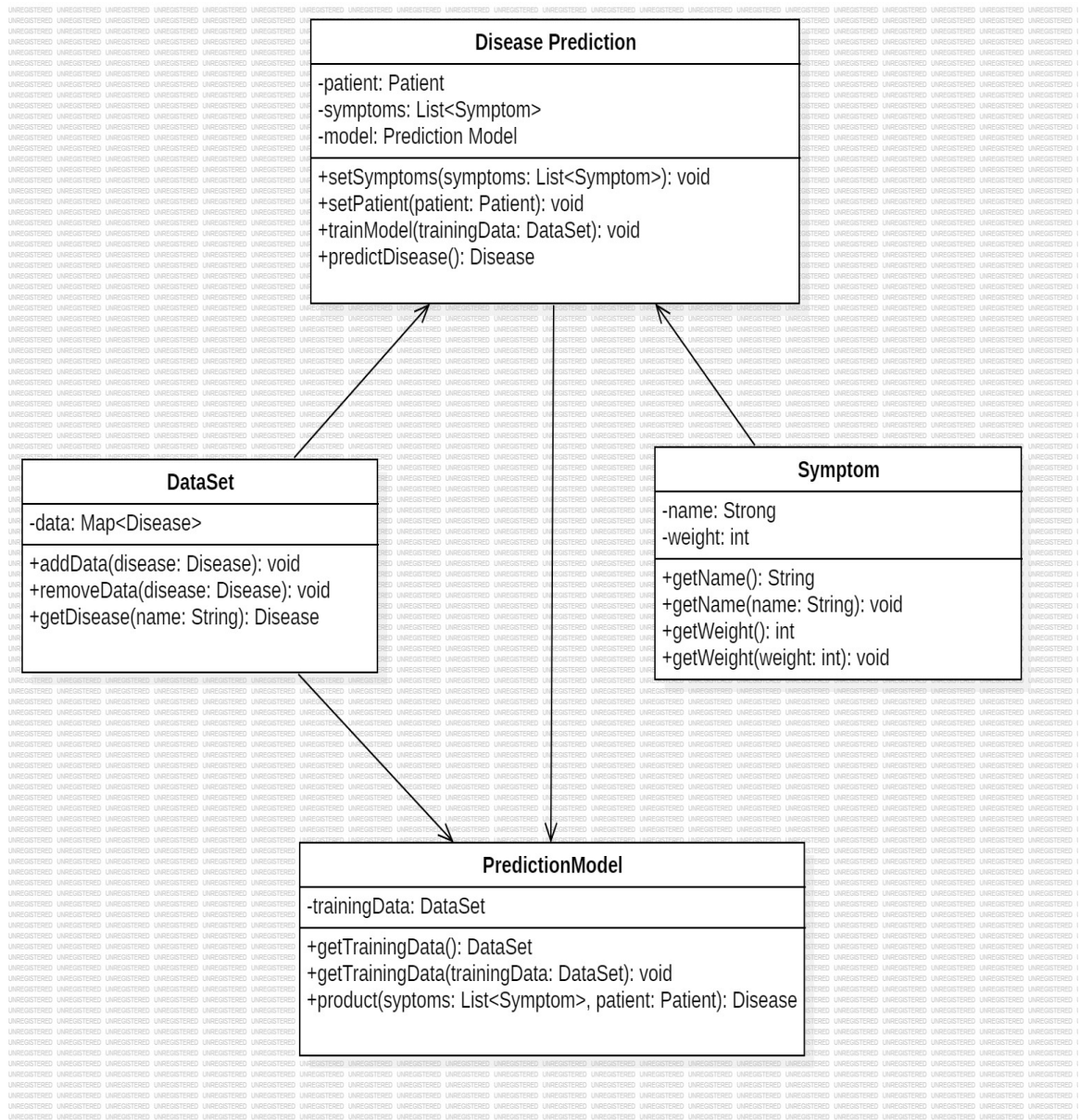
### **5.2 UML DIAGRAMS**

UML remains for Unified Modeling Language. UML is an institutionalized broadly useful displaying dialect in the field of protest situated programming designing. The Unified Modeling Language is a standard dialect for indicating, Visualization, Constructing and recording the antiques of programming framework, and additionally for business displaying and other non-programming frameworks.

### 5.2.1 CLASS DIAGRAM

In this class diagram, we have several classes representing different components of a disease prediction system:

1. **Disease Prediction** This class represents the overall disease prediction system. It has attributes such as symptoms (a list of symptoms), patient (patient information), and model (a prediction model). It also has methods to set the symptoms, set the patient information, train the prediction model, and predict the disease.
2. **Symptom** This class represents a symptom. It has attributes such as name (name of the symptom) and weight (a weight indicating the importance or severity of the symptom). It provides methods to get and set the name and weight of the symptom.
3. **Patient** This class represents a patient. It has attributes such as name, age, and gender. It provides methods to get and set the patient's name, age, and gender.
4. **PredictionModel** This class represents the prediction model used in the system. It has an attribute trainingdata (a dataset used for training the model). It provides methods to get and set the training data, train the model, and predict the disease based on symptoms and patient information.
5. **Disease** This class represents a specific disease. It has attributes such as name (name of the disease).

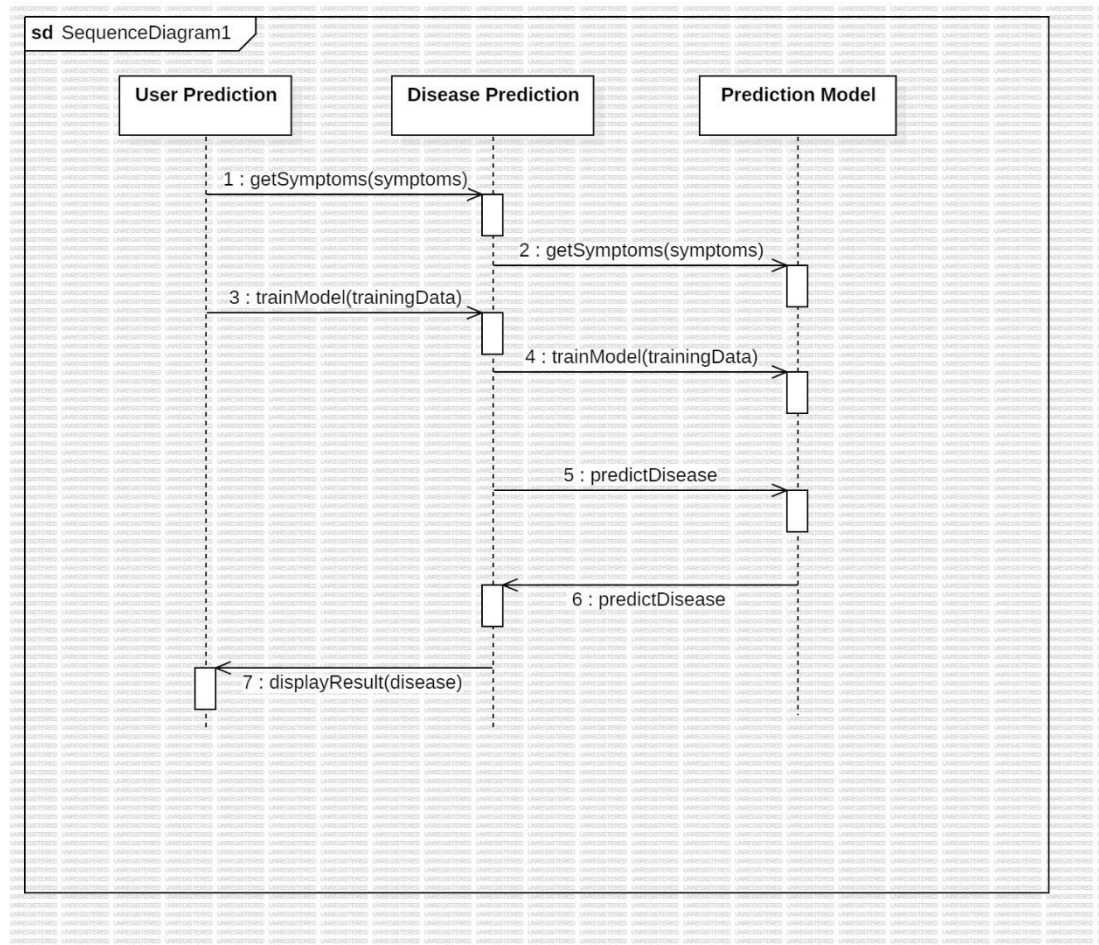


## 5.2.2 SEQUENCE DIAGRAM

In this sequence diagram, the interactions between the different component involved in the disease prediction process are depicted:

1. The user interface component sends the `setSymptoms(symptoms)` message to the DiseasePrediction component, passing a list of symptoms as a parameter.
2. The DiseasePrediction component receives the `setSymptoms(symptoms)` message and sets the symptoms attribute accordingly.
3. The user interface component sends the `setPatient(patient)` message to the DiseasePrediction component, passing the patient information as a parameter.

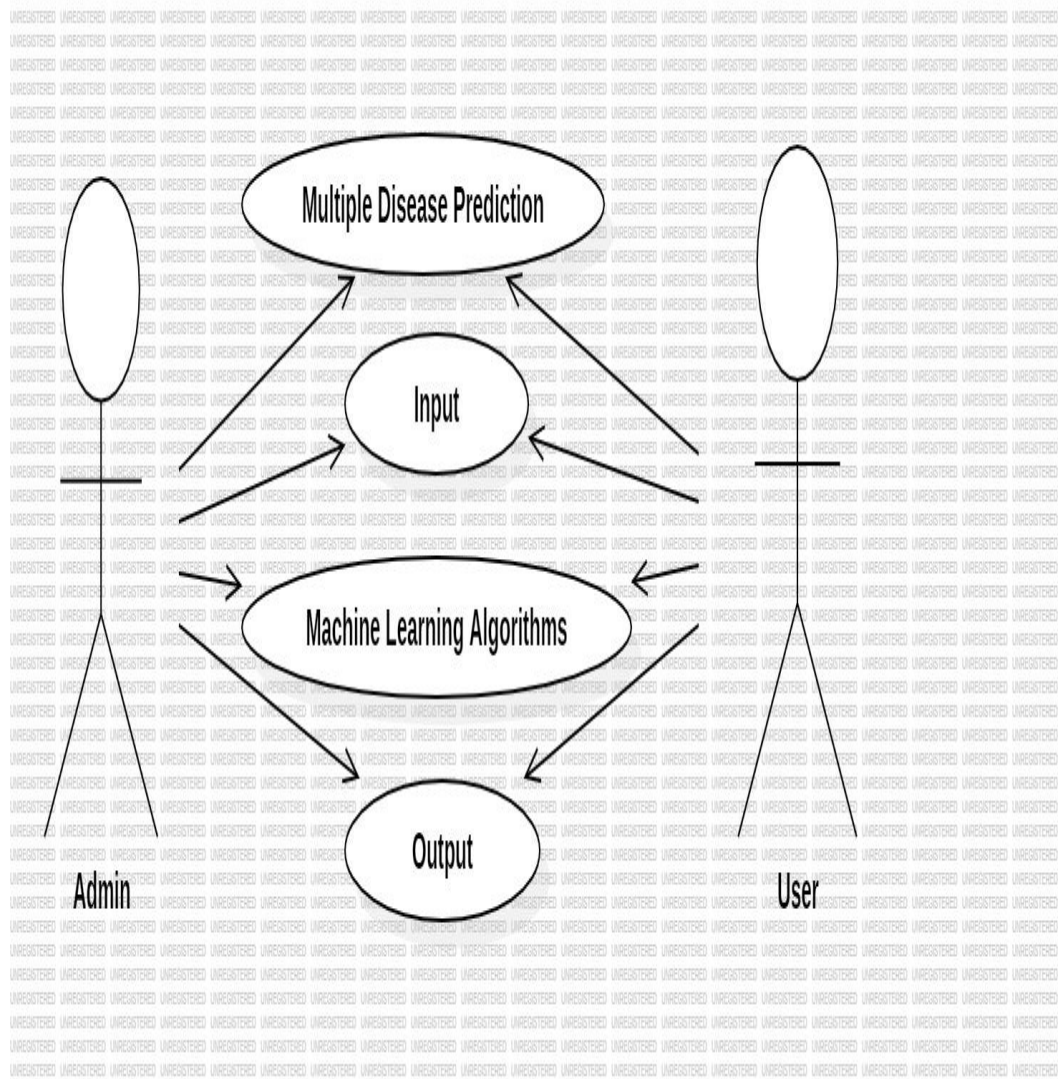
4. The DiseasePrediction component receives the setPatient(patient) message and sets the patient attribute accordingly.
5. The user interface component sends the trainModel(trainingData) message to the DiseasePrediction component, passing the training data as a parameter.
6. The DiseasePrediction component receives the trainModel(trainingData) message and sets the training data for the prediction model.
7. The DiseasePrediction component sends the predictDisease() message to the PredictionModel component.
8. The PredictionModel component receives the predictDisease() message and performs the disease prediction based on the symptoms and patient information.
9. The PredictionModel component sends the predicted disease back to the DiseasePrediction component.
10. The DiseasePrediction component receives the predicted disease and sends it to the user interface component using the displayResult(disease) message.
11. The user interface component receives the predicted disease and displays the result to the user.
12. This sequence diagram illustrates the flow of messages and interaction between the components involved in the disease prediction process, from setting symptoms and patient information to training the model and predicting the disease.



### 5.2.3 USECASE DIAGRAM

In this use case diagram, we have the following actors and use cases:

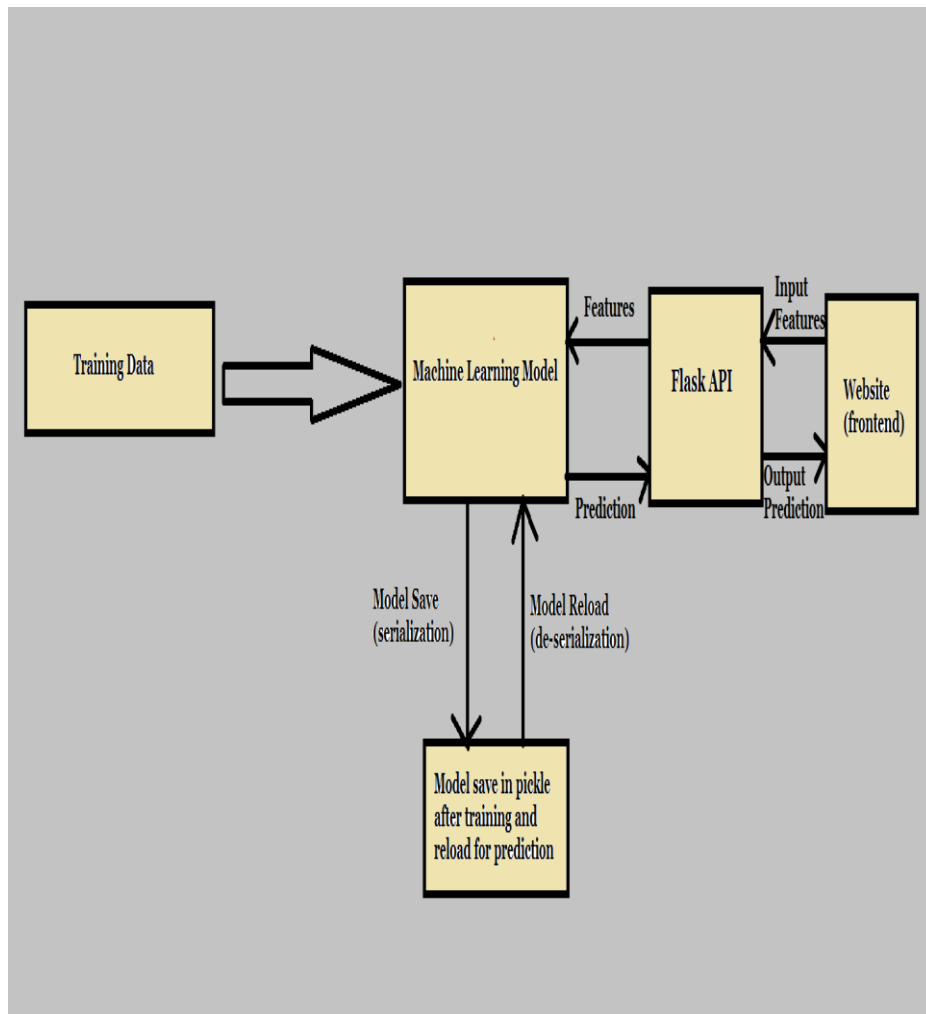
1. User: Represents the user of the system who interacts with it.
2. Enter Symptoms: The user can enter the symptoms of the patient.
3. Enter Patient Details: The user can enter the details of the patient (e.g., name, age, gender).
4. Train Model: The user can initiate the training process for the prediction model.
5. Predict Disease: The user can request the system to predict the disease based on the entered symptoms and patient details.
6. View Prediction: The user can view the predicted disease.
7. The use case diagram illustrates the various actions that a user can perform in the disease prediction system. The user can enter symptoms and patient details, train the model, request disease prediction, and view the predicted disease.



## **CHAPTER 6: SYSTEM IMPLEMENTATION**

### **6.1 SYSTEM PROCESS**

In our current system implementation, a typical user interaction with the proposed scheme follows a straightforward process. The system utilizes a FLASK API to enable multiple disease prediction. While many existing machine learning models for healthcare analysis concentrate on a single disease per analysis, our system allows for the analysis of various diseases, including diabetes, diabetes retinopathy, heart disease, and breast cancer. Moreover, there is potential to expand the system's capabilities to incorporate additional diseases in the future, such as skin diseases and fever analysis.



**Figure:** System Design for Multiple Disease Prediction

To facilitate the analysis of multiple diseases, the implementation utilizes machine learning algorithms, TensorFlow, and a Flask API. The behavior of the model is saved as a Python pickle file, allowing for easy loading whenever needed using Python unpickling.

The article emphasizes the importance of incorporating all relevant parameters that may contribute to the development of a disease in order to accurately assess its potential effects. For example, in the analysis of diabetes, various factors such as age, sex, BMI, insulin, glucose, blood pressure, diabetes pedigree function, pregnancies, serum creatinine, potassium levels, Glasgow Coma Scale, heart rate, respiration rate, body temperature, LDL cholesterol, HDL cholesterol, and triglyceride levels are considered.

The final model behavior is saved as a Python pickle file, and a Flask API is developed. Users can access this API by providing the disease parameters and name. The Flask API then calls the corresponding model and returns the patient's status. The system is designed to aid in monitoring patients' conditions and provide timely warnings to help reduce mortality rates.

## **6.2 MODULE DESCRIPTION**

The process of predicting multiple diseases can be broken down into several steps:

### **6.2.1 Data Collection**

The first step is to gather data from various sources such as electronic health records, medical literature, and patient surveys. This data should include information such as patient demographics, medical history, symptoms, lab results, and imaging findings.

### **6.2.2 Data Preprocessing**

The collected data needs to be preprocessed to remove any irrelevant or incomplete information, correct errors, and standardize the format. This step is crucial for ensuring the quality and accuracy of the data.

### **6.2.3 Feature extraction and selection**

Next, relevant features need to be extracted from the preprocessed data, which can help in predicting the diseases. This process can involve statistical methods, machine learning algorithms, or expert knowledge. Once features have been extracted, a feature selection algorithm can be used to identify the most important features for predicting multiple diseases.



#### **6.2.4 Model selection and training**

After feature selection, a suitable machine learning model needs to be chosen to predict the multiple diseases. This can involve techniques such as logistic regression, decision trees, random forests, or deep learning. The chosen model needs to be trained on the preprocessed data to optimize its performance.

#### **6.2.5 Model evaluation**

Once the model has been trained, it needs to be evaluated using appropriate metrics such as accuracy, precision, recall, and F1 score. This evaluation can help to identify any weaknesses or areas for improvement in the model.

#### **6.2.6 Deployment**

Finally, the trained model can be deployed in a clinical setting to predict multiple diseases for patients. This can involve integrating the model into an electronic health record system or other clinical decision support tools. It is important to regularly monitor and update the model to ensure its accuracy and relevance over time.

## **CHAPTER 7: SOURCE CODE**

```
# Diabetes Prediction

## **Exploratory Data Analysis**

# Importing the packages

import numpy as np

import pandas as pd

import statsmodels.api as sm

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.preprocessing import scale, StandardScaler

from sklearn.model_selection import train_test_split, GridSearchCV, cross_val_score

from sklearn.metrics import confusion_matrix, accuracy_score, mean_squared_error,

r2_score, roc_auc_score, roc_curve, classification_report

from sklearn.linear_model import LogisticRegression

from sklearn.neighbors import KNeighborsClassifier

from sklearn.svm import SVC

from sklearn.neural_network import MLPClassifier

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.ensemble import GradientBoostingClassifier

from sklearn.model_selection import KFold

import warnings

warnings.simplefilter(action = "ignore")
```

```

sns.set()

plt.style.use('ggplot')

%matplotlib inline

df = pd.read_csv('../data/diabetes.csv')

df.head()

df.info()

df.describe()

print("Dataset shape:", df.shape)

df["Outcome"].value_counts()*100/len(df)

50df.Outcome.value_counts()

plt.figure(figsize=(8,7))

plt.xlabel('Age', fontsize=12)

plt.ylabel('Count', fontsize=12)

df["Age"].hist(edgecolor = "black");

print("Max Age: " + str(df["Age"].max()) + ', ' + " Min Age: " + str(df["Age"].min()))

fig, ax = plt.subplots(4,2, figsize=(20,20))

sns.distplot(df.Age, bins = 20, ax=ax[0,0], color="red")

sns.distplot(df.Pregnancies, bins = 20, ax=ax[0,1], color="red")

sns.distplot(df.Glucose, bins = 20, ax=ax[1,0], color="red")

sns.distplot(df.BloodPressure, bins = 20, ax=ax[1,1], color="red")

sns.distplot(df.SkinThickness, bins = 20, ax=ax[2,0], color="red")

sns.distplot(df.Insulin, bins = 20, ax=ax[2,1], color="red")

sns.distplot(df.DiabetesPedigreeFunction, bins = 20, ax=ax[3,0], color="red")

sns.distplot(df.BMI, bins = 20, ax=ax[3,1], color="red")

```

```

df.groupby("Outcome").agg({"Pregnancies":"mean"})

df.groupby("Outcome").agg({"Age":"mean"})

df.groupby("Outcome").agg({"Age":"max"})

df.groupby("Outcome").agg({"Insulin": "mean"})

df.groupby("Outcome").agg({"Insulin": "max"})

df.groupby("Outcome").agg({"Glucose": "mean"})

df.groupby("Outcome").agg({"Glucose": "max"})

df.groupby("Outcome").agg({"BMI": "mean"})

# Visualize the distribution of the outcome variable in the data -> 0 - Healthy, 1 -
Diabetic

f,ax=plt.subplots(1,2,figsize=(18,8))

df['Outcome'].value_counts().plot.pie(explode=[0,0.1],autopct='%1.1f%%',ax=ax[0],s
hadow=True)

ax[0].set_title('target')

ax[0].set_ylabel("")

sns.countplot(df['Outcome'],ax=ax[1])

ax[1].set_title('Outcome')

plt.show()

# corr() is used to find the pairwise correlation of all columns in the dataframe

5152

df.corr()

# Correlation matrix of the data set

f, ax = plt.subplots(figsize= [20,15])

sns.heatmap(df.corr(), annot=True, fmt=".2f", ax=ax, cmap ='magma' )

```

```

ax.set_title("Correlation Matrix", fontsize=20)

plt.savefig("corr.png", dpi=400)

plt.show()

## **Data Preprocessing**

## Missing Observation Analysis

We saw on df.head() that some features contain 0, it doesn't make sense here and this
indicates missing value. Below we replace 0 value by NaN:

df[['Glucose','BloodPressure','SkinThickness','Insulin','BMI']]
=

df[['Glucose','BloodPressure','SkinThickness','Insulin','BMI']].replace(0,np.NaN)

df.head()

# Now, we can look at where are missing values

df.isnull().sum()

# Visualizing the missing observations using the missingno library

import missingno as msno

msno.bar(df, color="orange");

# The missing values will be filled with the median values of each variable

def median_target(var):

temp = df[df[var].notnull()]

temp = temp[[var, 'Outcome']].groupby(['Outcome'])[var].median().reset_index()

return temp

# The values to be given for incomplete observations are given the median value of
people who are not sick and the median values of people who are sick.

columns = df.columns

```

```

columns = columns.drop("Outcome")

for i in columns:

    median_target(i)

    df.loc[(df['Outcome'] == 0 ) & (df[i].isnull()), i] = median_target(i)[i][0]

    df.loc[(df['Outcome'] == 1 ) & (df[i].isnull()), i] = median_target(i)[i][1]

df.head()

# Number of missing valuesdf.isnull().sum()

#### Pair plot for clean data

```

The pairs plot builds on two basic figures, the histogram and the scatter plot. The histogram on the diagonal allows us to see the distribution of a single variable while the scatter plots on the upper and lower triangles show the relationship between two variables.

```

p=sns.pairplot(df, hue = 'Outcome')

## Outlier Observation Analysis

for feature in df:

    Q1 = df[feature].quantile(0.25)

    Q3 = df[feature].quantile(0.75)

    IQR = Q3-Q1

    lower = Q1- 1.5*IQR

    upper = Q3 + 1.5*IQR

    if df[(df[feature] > upper)].any(axis=None):

        print(feature,"yes")

    else:

        print(feature, "no")

```

```

# Outlier observation of Insulin

import seaborn as sns

plt.figure(figsize=(8,7))

sns.boxplot(x = df["Insulin"], color="red");

# Conducting a stand alone observation review for the Insulin variable

# Suppressing contradictory values

Q1 = df.Insulin.quantile(0.25)

Q3 = df.Insulin.quantile(0.75)

IQR = Q3-Q1

lower = Q1 - 1.5*IQR

upper = Q3 + 1.5*IQR

df.loc[df["Insulin"] > upper,"Insulin"] = upper

import seaborn as sns

plt.figure(figsize=(8,7))

sns.boxplot(x = df["Insulin"], color="red");

## Local Outlier Factor (LOF)

# Determining the outliers between all variables with the LOF method

from sklearn.neighbors import LocalOutlierFactor

lof =LocalOutlierFactor(n_neighbors= 10)

lof.fit_predict(df)

df_scores = lof.negative_outlier_factor_

np.sort(df_scores)[0:30]

# Choosing the threshold value according to lof scores

threshold = np.sort(df_scores)[7]

```

```

threshold

# Deleting those that are higher than the threshold

outlier = df_scores > threshold

df = df[outlier]

# Examining the size of the data.

df.shape

## Feature Engineering

Creating new variables is important for models. But we need to create a logical new
variable. For this data set, some new variables were created according to BMI, Insulin
and glucose variables.

# The variables in the data set are an effective factor in increasing the performance of
the models by standardization.

from sklearn.preprocessing import RobustScaler
transformer = RobustScaler().fit(X)

X = transformer.transform(X)

X = pd.DataFrame(X, columns = cols, index = index)

X.head()

X = pd.concat([X, categorical_df], axis = 1)

X.head()

y.head()

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.30, random_state
= 0)

# scaling data

from sklearn.preprocessing import StandardScaler

```



```

scaler = StandardScaler()

X_train = scaler.fit_transform(X_train)

X_test = scaler.transform(X_test)

# LR

# fitting data to model

from sklearn.linear_model import LogisticRegression

log_reg = LogisticRegression()

log_reg.fit(X_train, y_train)

# model predictions

y_pred = log_reg.predict(X_test)

# accuracy score

from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

56print(accuracy_score(y_train, log_reg.predict(X_train)))

log_reg_acc = accuracy_score(y_test, log_reg.predict(X_test))

print(log_reg_acc)

# confusion matrix

print(confusion_matrix(y_test, y_pred))

# classification report

print(classification_report(y_test, y_pred))

# KNN

from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier()

knn.fit(X_train, y_train)

# model predictions

```

```

y_pred = knn.predict(X_test)

# accuracy score

print(accuracy_score(y_train, knn.predict(X_train)))

knn_acc = accuracy_score(y_test, knn.predict(X_test))

print(knn_acc)

# confusion matrix

print(confusion_matrix(y_test, y_pred))

# classification report

print(classification_report(y_test, y_pred))

# SVM

from sklearn.svm import SVC

from sklearn.model_selection import GridSearchCV

svc = SVC(probability=True)

parameters = {

'gamma' : [0.0001, 0.001, 0.01, 0.1],

'C' : [0.01, 0.05, 0.5, 0.1, 1, 10, 15, 20]

}

grid_search = GridSearchCV(svc, parameters)

grid_search.fit(X_train, y_train)

# best parameters

grid_search.best_params_

# best score

grid_search.best_score_

svc = SVC(C = 1, gamma = 0.1, probability=True)

```

```

svc.fit(X_train, y_train)

# model predictions

y_pred = svc.predict(X_test)

# accuracy score

print(accuracy_score(y_train, svc.predict(X_train)))

svc_acc = accuracy_score(y_test, svc.predict(X_test))

print(svc_acc)

# confusion matrix

print(confusion_matrix(y_test, y_pred))

# classification report

print(classification_report(y_test, y_pred))

# DT

from sklearn.tree import DecisionTreeClassifier

58dtc = DecisionTreeClassifier()

dtc.fit(X_train, y_train)

# accuracy score, confusion matrix and classification report of decision tree

dtc_acc = accuracy_score(y_test, dtc.predict(X_test))

print(f"Training Accuracy of Decision Tree Classifier is {accuracy_score(y_train,
dtc.predict(X_train))}")

print(f"Test Accuracy of Decision Tree Classifier is {dtc_acc} \n")

print(f"Confusion Matrix :- \n{confusion_matrix(y_test, dtc.predict(X_test))}\n")

print(f"Classification Report :- \n {classification_report(y_test, dtc.predict(X_test))}")

# hyper parameter tuning of decision tree

from sklearn.model_selection import GridSearchCV

```

```

grid_param = {
    'criterion' : ['gini', 'entropy'],
    'max_depth' : [3, 5, 7, 10],
    'splitter' : ['best', 'random'],
    'min_samples_leaf' : [1, 2, 3, 5, 7],
    'min_samples_split' : [1, 2, 3, 5, 7],
    'max_features' : ['auto', 'sqrt', 'log2']
}

grid_search_dtc = GridSearchCV(dtc, grid_param, cv = 50, n_jobs = -1, verbose = 1)

grid_search_dtc.fit(X_train, y_train)

# best parameters and best score
print(grid_search_dtc.best_params_)
print(grid_search_dtc.best_score_)

# best estimator
dtc = grid_search_dtc.best_estimator_

59# accuracy score, confusion matrix and classification report of decision tree
dtc_acc = accuracy_score(y_test, dtc.predict(X_test))

print(f"Training Accuracy of Decision Tree Classifier is {accuracy_score(y_train,
dtc.predict(X_train))}")

print(f"Test Accuracy of Decision Tree Classifier is {dtc_acc} \n")

print(f"Confusion Matrix :- \n{confusion_matrix(y_test, dtc.predict(X_test))}\n")

print(f"Classification Report :- \n {classification_report(y_test, dtc.predict(X_test))}")

# RF

from sklearn.ensemble import RandomForestClassifier

```

```

rand_clf = RandomForestClassifier(criterion = 'entropy', max_depth = 15,
max_features = 'auto', min_samples_leaf = 2, min_samples_split = 3, n_estimators =
130)

rand_clf.fit(X_train, y_train)

y_pred = rand_clf.predict(X_test)

# accuracy score

print(accuracy_score(y_train, rand_clf.predict(X_train)))

ran_clf_acc = accuracy_score(y_test, y_pred)

print(ran_clf_acc)

# confusion matrix

print(confusion_matrix(y_test, y_pred))

# classification report

print(classification_report(y_test, y_pred))

# GBDT

from sklearn.ensemble import GradientBoostingClassifier

gbc = GradientBoostingClassifier()

parameters = {

'loss': ['deviance', 'exponential'],

'learning_rate': [0.001, 0.1, 1, 10],

'n_estimators': [100, 150, 180, 200]

}

grid_search_gbc = GridSearchCV(gbc, parameters, cv = 10, n_jobs = -1, verbose = 1)

grid_search_gbc.fit(X_train, y_train)

# best parameters

```

```

grid_search_gbc.best_params_

# best score

grid_search_gbc.best_score_

gbc = GradientBoostingClassifier(learning_rate = 0.1, loss = 'deviance', n_estimators
= 180)

gbc.fit(X_train, y_train)

y_pred = gbc.predict(X_test)

# accuracy score

print(accuracy_score(y_train, gbc.predict(X_train)))

gbc_acc = accuracy_score(y_test, y_pred)

print(gbc_acc)

# confusion matrix

print(confusion_matrix(y_test, y_pred))

# classification report

print(classification_report(y_test, y_pred))

# XGBoost

from xgboost import XGBClassifier

6162

xgb = XGBClassifier(objective = 'binary:logistic', learning_rate = 0.01, max_depth =
10, n_estimators = 180)

xgb.fit(X_train, y_train)

y_pred = xgb.predict(X_test)

# accuracy score

print(accuracy_score(y_train, xgb.predict(X_train)))

```

```
xgb_acc = accuracy_score(y_test, y_pred)

print(xgb_acc)

# confusion matrix

print(confusion_matrix(y_test, y_pred))

# classification report

print(classification_report(y_test, y_pred))
```

## CHAPTER 8: RESULTS

Fig 8.1 represents the accuracy of Liver prediction model performed, given by ROC curve, among all the models SVM gave an accuracy of 71%.

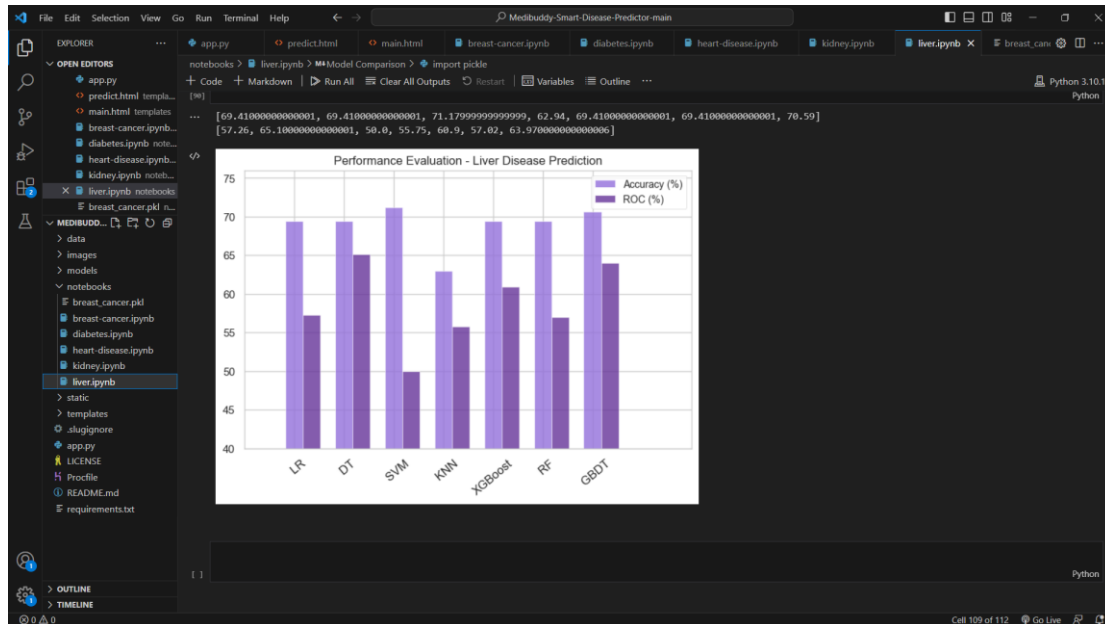


Figure 8.1: Liver Disease Performance Evaluation

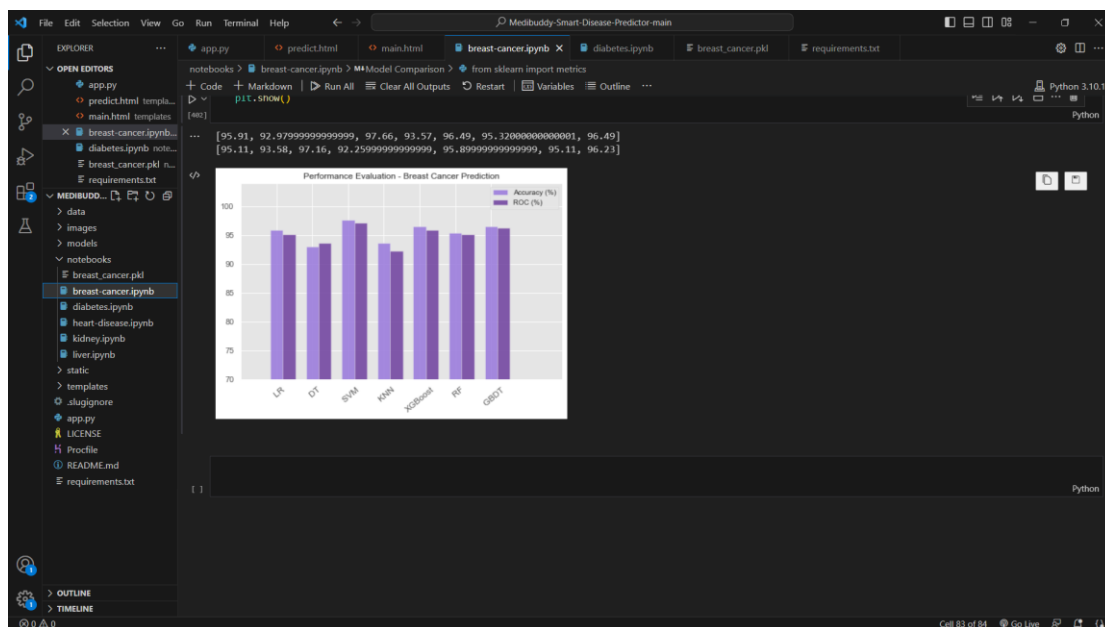
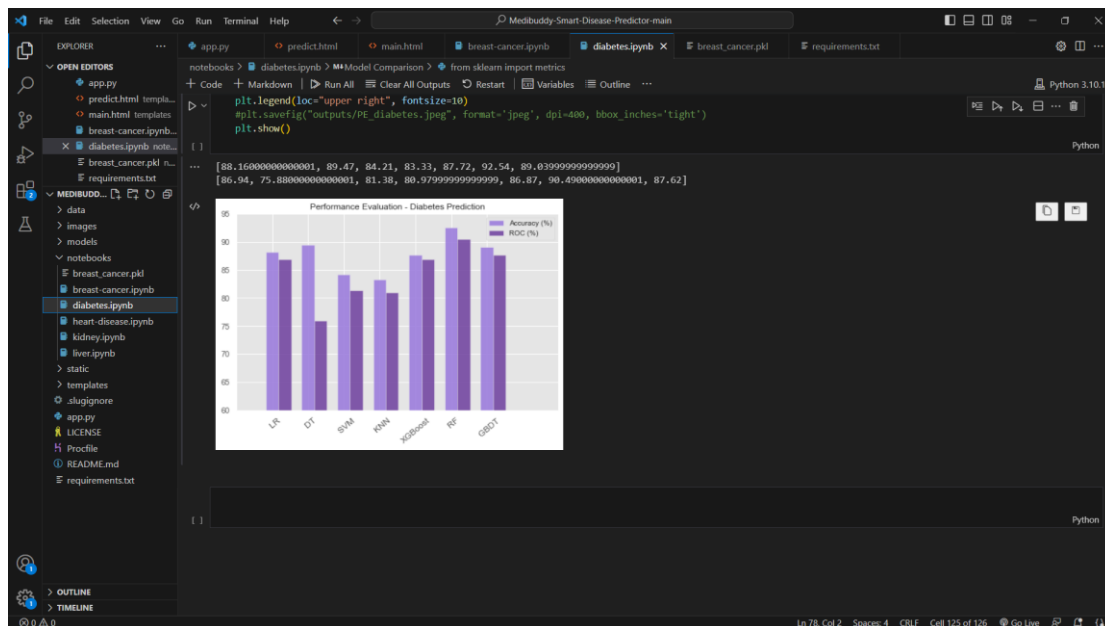


Figure 8.2: Breast Cancer Performance Evaluation

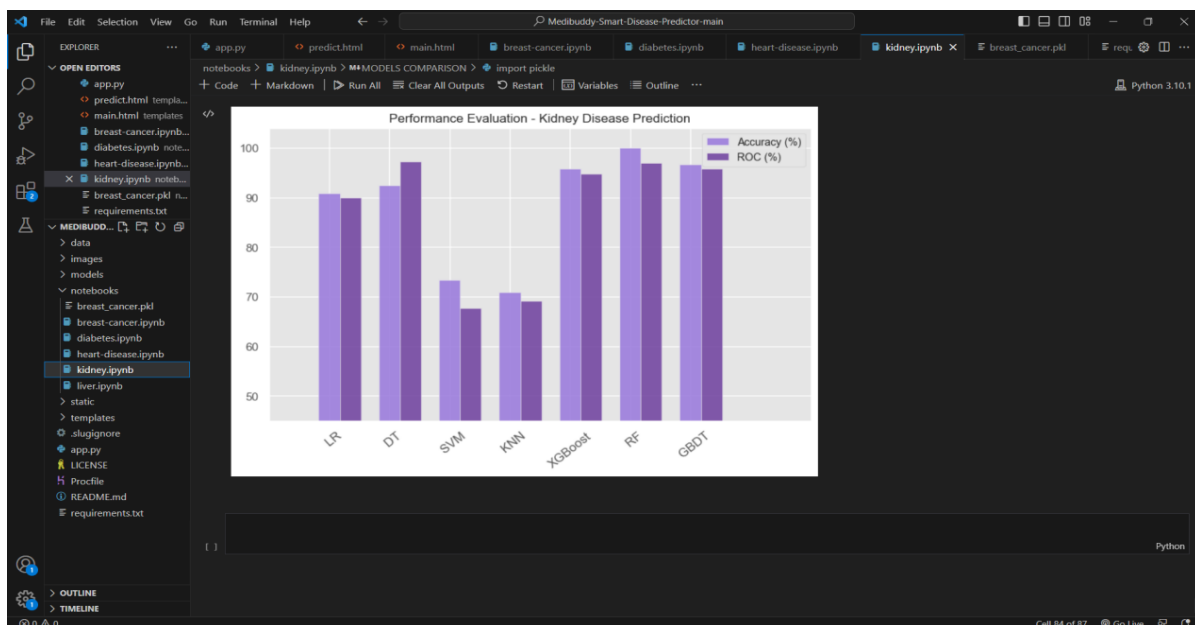


Fig 8.2 represents the accuracy of Breast Cancer prediction model performed , given by ROC curve, among all the models SVM gave an accuracy 97.7%.



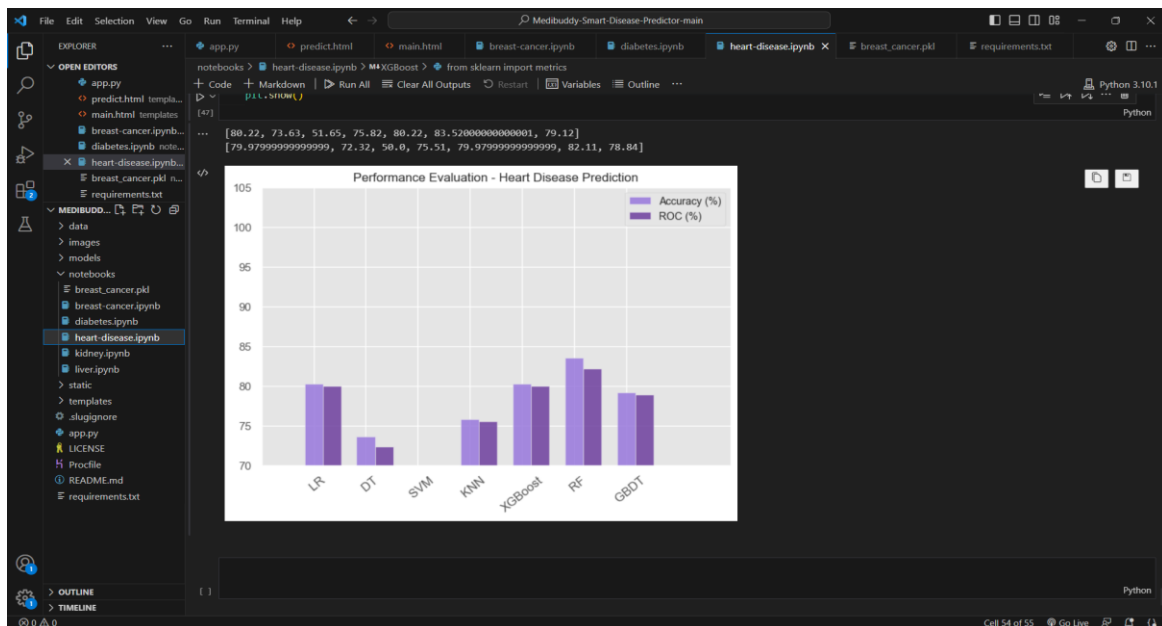
**Figure 8.3: Diabetes Performance Evaluation**

The accuracy of Diabetes prediction model performed is given by ROC curve, among all the models Random Forest gave an accuracy of 92.5%, the above fig shows comparison if different models.



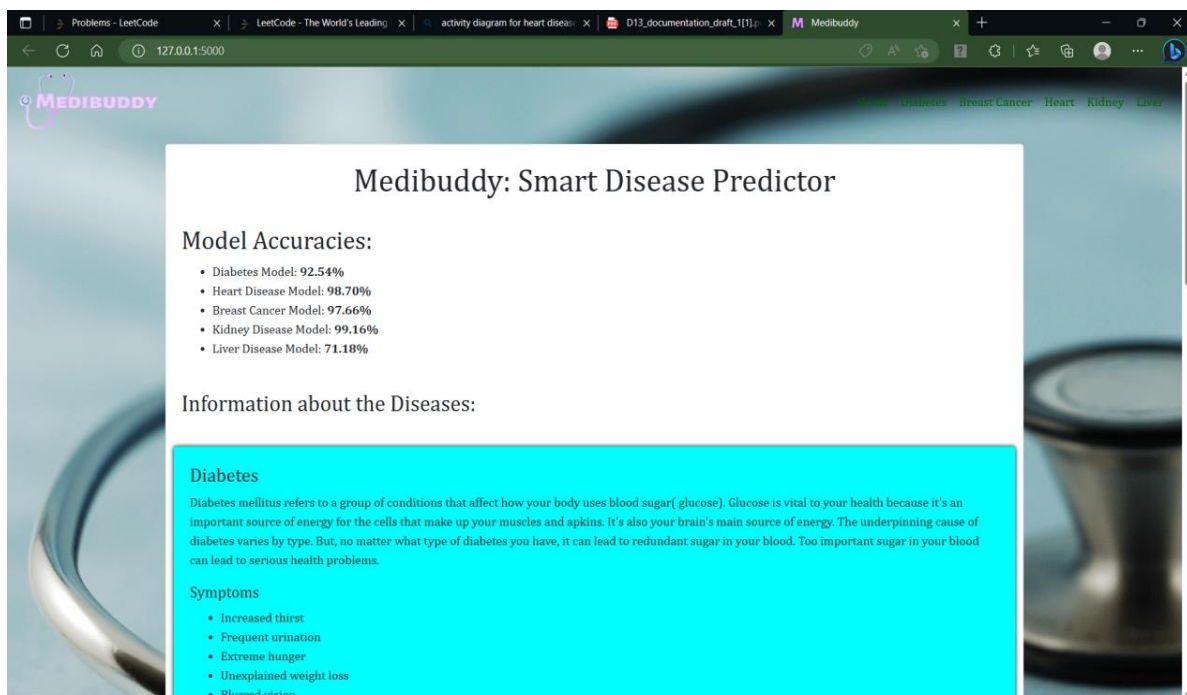
**Figure 8.4: Kidney Disease Performance Evaluation**

The accuracy of Kidney prediction model performed is given by ROC curve, among all the models Random Forest gave an accuracy of 99.16%.



**Figure 8.5: Heart Disease Performance Evaluation**

The accuracy of Heart prediction model performed is given by ROC curve, among all the models Random Forest gave an accuracy of 98%.



**Figure 8.6: Home Page**

The fig 8.6 represents the actual web site that shows details about multiple disease, the web page include details about the disease and symptoms related to disease.

The patient now can choose his preferred disease at top right and start entering values in the text boxes the fig 8.7 represents diabetes web page where user have to enter values in the text boxes.

Diabetes Predictor

No. of Pregnancies

Glucose

Blood Pressure (in mmHg)

Skin Thickness (in mm)

Insulin (in  $\mu\text{U/ml}$ )

BMI

Diabetes Pedigree Function

Age (in years)

Predict

**Figure 8.7 : Diabetes Page**

The fig 4.8 represents breast cancer web page where user have to enter values in the text boxes. The user has to enter detailed values in the text boxes, the user then clicked predict to get the results. It will redirect to result page.

**Breast Cancer Predictor**

Texture Mean	Smoothness Mean	Compactness Mean
Concave Points Mean	Symmetry Mean	Fractal Dimension Mean
Texture Standard Error	Area Standard Error	Smoothness Standard Error
Compactness Standard Error	Concavity Standard Error	Concave Points Standard Error
Symmetry Standard Error	Fractal Dimension Standard Error	Texture Worst
Area Worst	Smoothness Worst	Compactness Worst
Concavity Worst	Concave Points Worst	Symmetry Worst
	Fractal Dimension Worst	

**Predict**

**Figure 8.8: Breast Cancer**

The fig 4.9 represents heart web page where user have to enter values in the text boxes. The user has to enter detailed values in the text boxes, the user then clicked predict to get the results. It will redirect to result page.

**Heart Disease Predictor**

Age (in years)	Sex (1 = Male; 0 = Female)
Chest Pain Type	Resting Blood Pressure (in mm Hg)
Serum Cholesterol (in mg/dl)	Fasting Blood Sugar > 120 mg/dl (1 = True; 0 = False)
Resting Electrocardiograph Results	Maximum Heart Rate Achieved
Exercise Induced Angina (1 = Yes; 0 = No)	ST Depression Induced by Exercise Relative to Rest
The Slope of the Peak Exercise ST Segment	Number of Major Vessels (0-3) Colored by Fluoroscopy
Thal: 1 = Normal; 2 = Fixed Defect; 3 = Reversible Defect	

**Predict**

**Figure 8.9: Heart Disease**

The fig 8.10 represents kidney web page where user have to enter values in the text boxes. The user has to enter detailed values in the text boxes, the user then clicked predict to get the results. It will redirect to result page.

Age (in years)	Blood Pressure (in mm/Hg)	Specific Gravity
Albumin (0, 1, 2, 3, 4, 5)	Sugar (0, 1, 2, 3, 4, 5)	Red Blood Cells (0: Abnormal; 1: Normal)
Pus Cell (0: Abnormal; 1: Normal)	Pus Cell Clumps (0: Not Present; 1: Present)	Bacteria (0: Not Present; 1: Present)
Blood Glucose Random (in mgs/dl)	Blood Urea (in mgs/dl)	Serum Creatinine (in mgs/dl)
Sodium (in mEq/L)	Potassium (in mEq/L)	Haemoglobin (in gms)
Packed Cell Volume	White Blood Cell Count (in cells/cumm)	Red Blood Cell Count (in millions/cmm)
Hypertension (0: No; 1: Yes)	Diabetes Mellitus (0: No; 1: Yes)	Coronary Artery Disease (0: No; 1: Yes)
Appetite (0: Good; 1: Poor)	Pedal Edema (0: No; 1: Yes)	Anemia (0: No; 1: Yes)
Predict		

**Figure 8.10: Kidney Disease**

The fig 8.11 represents liver web page where user have to enter values in the text boxes. The user has to enter detailed values in the text boxes, the user then clicked predict to get the results. It will redirect to result page.

Age (in years)	Gender (0: Male; 1: Female)
Total Bilirubin (in mg/dL)	Conjugated Bilirubin (in mg/dL)
Alkaline Phosphatase (in IU/L)	Alanine Aminotransferase (in IU/L)
Aspartate Aminotransferase (in IU/L)	Total Proteins (in g/dL)
Albumin (in g/dL)	Albumin and Globulin Ratio
Predict	



**Figure 8.11: Liver Disease**

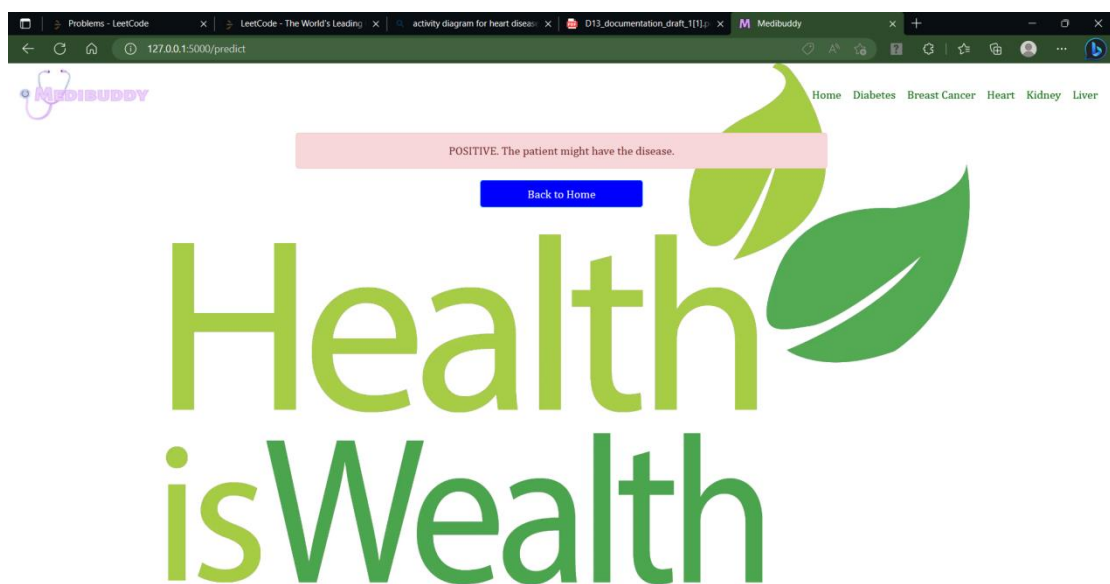
The fig 8.12 represents diabetes web page where user have to enter values in the text boxes. The user has to enter detailed values in the text boxes, the user then clicked predict to get the results. It will redirect to result page. The values are displayed as shown in the fig are entered by user.

Diabetes Predictor

0
85
25
2
2
26
0.351
35

Predict

**Figure 8.12: Insert Page**



**Figure 8.13: Home Page**

The fig 8.13 represents results web page where user will get the results if the user is facing any problem then the result page will display positive, the patient may be facing disease else negative the patient is safe and also displays home button below.

The above fig 8.13 represents the result page for diabetes. In this way for every page the result page will be displayed as shown above.

## **CHAPTER 9: CONCLUSION**

The primary objective of this project is to accurately predict diseases based on patient-reported symptoms by implementing Machine Learning algorithms. In this study, four Machine Learning algorithms were utilized, achieving a mean accuracy of over 95%. This signifies significant improvement and higher accuracy compared to previous works, making the system more reliable and satisfying for users.

**Improved Accuracy** Machine Learning models excel in processing large amounts of data and identifying complex patterns that may not be discernible to humans. This capability enhances the accuracy of disease prediction by considering various factors such as symptoms, medical history, genetics, lifestyle choices, and environmental factors.

**Early Detection** Machine Learning algorithms can analyze data from multiple sources to detect subtle indicators or early warning signs of diseases. Early detection enables timely intervention, leading to improved treatment outcomes and potentially saving lives.

**Personalized Medicine** Machine Learning models can incorporate individual patient characteristics to generate personalized predictions or treatment recommendations. This approach allows healthcare providers to tailor interventions based on each patient's unique needs, resulting in more effective and targeted healthcare strategies.

**Efficient Resource Allocation** Accurate disease prediction by Machine Learning models optimizes resource allocation in healthcare settings. This includes streamlining workflows, optimizing staff schedules, and effectively managing medical equipment and supplies, leading to cost savings and improved healthcare delivery.

**Data-Driven Insights** Machine Learning algorithms reveal hidden relationships and insights within large datasets. This assists researchers and healthcare professionals in identifying new risk factors, understanding disease mechanisms, and developing innovative treatments or interventions.

**Continuous Learning and Improvement** Machine Learning models continually learn and improve with more data exposure. This adaptability allows models to incorporate new knowledge, refine predictions, and adapt to evolving healthcare trends and patterns.



**Challenges and Ethical Considerations** While Machine Learning offers promising opportunities, there are challenges and ethical considerations to address. These include ensuring data privacy and security, addressing biases in training data, maintaining transparency and interpretability of models, and responsibly using predictive models in healthcare decision-making.

Overall, the integration of Machine Learning in disease prediction has the potential to revolutionize healthcare by improving prediction accuracy, enabling early interventions, facilitating personalized medicine, optimizing resource allocation, and generating data-driven insights. Continued research, collaboration between healthcare professionals and data scientists, and ethical considerations are essential for harnessing the full potential of Machine Learning in disease prediction and providing enhanced healthcare solutions for patients. Additionally, this system is user-friendly and accessible to a wide range of users without any specific threshold.

## **CHAPTER 10: REFERENCES**

- [1] In a research paper titled "Disease Prediction System," authored by Khurana, S., Jain, A., Kataria, S., Bhasin, K., Arora, S., and Gupta, A.D., published in the International Research Journal of Engineering and Technology in 2019, the authors discuss a system for predicting diseases. The paper is found in Volume 6, Issue 5, with page numbers 5178-5184.
- [2] In a paper titled "Heart Disease Prediction with Machine Learning Approaches," Kamboj (2020) explores the use of machine learning techniques for predicting heart disease. The research was published in the International Journal of Science and Research, Volume 9, Issue 7, with pages 1454-1458.
- [3] Ware, Rakesh, and Choudhary (2020) present a study on heart attack prediction using machine learning techniques. The research was published in the International Journal of Recent Technology and Engineering, Volume 8, Issue 5, with pages 1577-1580. Additionally, it was also featured in the Turkish Journal of Computer and Mathematics Education, Research Article Vol.12 No.6 (2021), with pages 4013-4023.
- [4] Shirsath and Patil (2018) discuss disease prediction using machine learning methods applied to big data. The study is published in the International Journal of Innovative Research in Science and Technology, Volume 7, Issue 6, with pages 6752-6757.
- [5] Marimuthu et al. (2018) conduct a review on heart disease prediction utilizing machine learning and data analytics approaches. The research is featured in the International Journal of Computer Application, Volume 181, Issue 18, with pages 20-25.
- [6] Battineni et al. (2020) explore the application of machine learning predictive models in chronic diseases. The study is published in the International Journal of Personalised Medicine, Volume 10, Issue 21, with pages 1-11.
- [7] Ardabili et al. (2020) present a study on COVID-19 outbreak prediction using machine learning. The research is featured in the Journal of Algorithms, Volume 13, Issue 249, with pages 1-36.
- [8] Shrestha and Chatterjee (2019) propose a heart disease prediction system using machine learning techniques. The research is published in the LBEF Research Journal of Science Technology and Management, Volume 1, Issue 2, with pages 115-132.

- [9] Magar et al. (2020) investigate heart disease prediction using machine learning methods. The study is featured in the Journal of Emerging Technologies and Innovative Research, Volume 7, Issue 6, with pages 2081-2085.
- [10] Alotaibi (2019) implements a machine learning model for predicting heart failure disease. The research is published in the International Journal of Advanced Computer Science and Application, Volume 10, Issue 6, with pages 261-268.
- [11] Godse et al. (2019) compare different machine learning algorithms for multiple disease prediction. The study is featured in the International Journal of Advance Research in Computer and Communication Engineering, Volume 8, Issue 12, with pages 50-52.
- [12] Anitha and Sridevi (2019) explore heart disease prediction using data mining techniques. The research is published in the Journal of Analysis and Computation, Volume 13, Issue 2, with pages 48-55.
- [13] Bindhika et al. (2020) investigate heart disease prediction using machine learning techniques. The study is featured in the International Research Journal of Engineering and Technology, Volume 7, Issue 4, with pages 5272-5276.
- [14] Pingale et al. (2019) propose a disease prediction model based on machine learning. The research is published in the International Research Journal of Engineering and Technology, Volume 6, Issue 12, with pages 2810-2813.
- [15] Chauhan et al. (2020) conduct a study on disease prediction.