HEALTHICA HEALTH ANALYSIS APP

A MINI PROJECT REPORT

submitted by

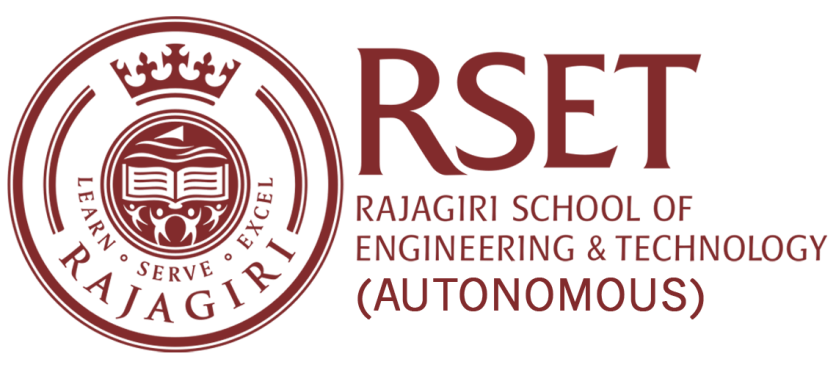
**Abhijith Suresh  
Anand CA**

**Devu Balakrishnan**

**Ephram Devlal Palathingal**

to

the APJ Abdul Kalam Technological University  
in partial fulfillment of the requirements for the award of the Degree  
of  
Bachelor of Technology  
In  
*Information Technology*



**Department of Information Technology**

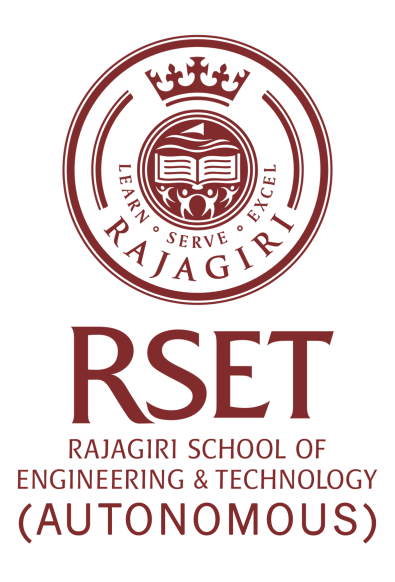
Rajagiri School of Engineering & Technology(Autonomous)  
Rajagiri Valley, Kakkanad, Kochi-39

JULY 2022

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**RAJAGIRI SCHOOL OF ENGINEERING & TECHNOLOGY (AUTONOMOUS)**

**RAJAGIRIVALLEY, KAKKANAD, KOCHI- 39**



**CERTIFICATE**

This is to certify that the report entitled **HEALTHICA HEALTH ANALYSIS APP** submitted by **Abhijith Suresh, Anand CA**, **Devu Balakrishnan**, **Ephram Devlal Palathingal** with Register number: **(RET19IT002),**  **(RET19IT011)**, **(RET19IT023)**, **(RET19IT026)** during the academic year 2021-22 to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Information Technology is a bonafide record of the project work carried out by him/her under my/our guidance and supervision. This report in any form has not been submitted to any other University or Institute for any purpose.

Mr. Tinku Soman Jacob Mr. Binu A Dr. Neeba E A

Project GuideProject Coordinator Head of the Department

Assistant Professor Assistant Professor Assistant Professor

Dept. of IT Dept. of IT Dept. of IT

RSET RSET RSET

Submitted for the practical examination conducted on………………

**Internal Examiner**

ACKNOWLEDGEMENT

I am extremely thankful to the Principal **Dr. P.S. Sreejith** for giving me his consent for this mini project. I express my thanks to the Head of the Department **Dr. Neeba E A** for her valuable suggestions and support. I am indebted to the mini project coordinators **Mr. Binu A**, Asst. Professor, Dept. of Information Technology and **Mr. Tinku Soman Jacob**, Asst. Professor, Dept. of Information Technology for their constant help and support throughout the presentation of the project.

I extend my sincere and heartfelt thanks to my guide, **Mr. Tinku Soman Jacob**, Asst. Professor, Dept. of Information Technology, for helping me in my presentation and providing timely advice and guidance.

I thank God almighty for all the blessings received during this endeavor. Last, but not least I thank all my friends for the support and encouragement they have given me during the course of my work.

ABSTRACT

It might have happened so many times that you or someone of yours, need a doctor’s help immediately, but they are not available. Most people may not be able to visit doctors or clinics every time they become sick or show any symptoms. Many of them might want to check themselves to find out if they're free of all the deadly diseases. They might even wish to get updated about their checkup reports and medications virtually without going to the hospital every time they get sick. They might also not be able to track their health and avail prolonged care.

The project showcases the utilization of a web-based application to analyze health for any possible illness and ailments. Using disease predictors to find acute and/or chronic diseases such as diabetes and other heart diseases. Machine learning (ML) predictions are cross-verified instantly by one of the many doctors without any consultation fee. With the Symptoms Check predictor, the model collates and analyzes symptoms provided by a given patient to hypothesize possible ailments and diseases. Using this disease predictor one can identify the possibility of any given disease by just entering in the symptoms they experience. Intelligent machine learning techniques are implemented to predict Diabetes, Myocardial infarction, Abnormal Heartbeat, and History of Myocardial infarction accurately.

It is essential to convert biomedical signals, such as electro-cardiograms (ECG) into digital form, for such computerized investigation hence for detecting heart diseases an electrocardiogram (ECG) a simple test that can be used to check your heart's rhythm and electrical activity is used as an input in envisioning its chance. Diabetes predictor uses the pedigree function, glucose level of the person, and BMI values to predict the probability of it.

It is also vital for Individuals to keep track of their health and checkups periodically, getting checkup reports and keeping track of them can be made easy through “Healthica”.

Using this web application, patients, doctors and hospitals can have easy access to the patient’s Database. Users can also track their medical prescription, set reminders, and view and track their health status through the Dashboard personalized for each patient. This Dashboard can be accessed through the web app and one will periodically receive notifications, reminders, and updates about health-related doubts and queries.

The salient features of the web application are a wealth of health and nutritional information, one tap disease predictor system, a detailed Health Report generation, and accurate diabetes as well as heart disease prediction.

Contents

[ACKNOWLEDGEMENT 3](#_Toc108968293)

[ABSTRACT 4](#_Toc108968294)

[LIST OF FIGURES 6](#_Toc108968296)

[CHAPTER 1: INTRODUCTION 7](#_Toc108968297)

[1.1 General Background 7](#_Toc108968298)

[1.2 Objective 8](#_Toc108968299)

[1.3 Objective 8](#_Toc108968300)

[CHAPTER 2: LITERATURE SURVEY 9](#_Toc108968301)

[2.1 Heart Disease Prediction Models 9](#_Toc108968302)

[2.2 Diabetes Mellitus Prediction Models 11](#_Toc108968303)

[CHAPTER 3: DESIGN OF THE PROPOSED SYSTEM 16](#_Toc108968304)

[CHAPTER 4: SYSTEM ARCHITECTURE 19](#_Toc108968305)

[4.1 SYSTEM ARCHITECTURE DIAGRAM 19](#_Toc108968306)

[4.2 ER DIAGRAM 22](#_Toc108968307)

[4.3 DATABASE DESIGN 23](#_Toc108968308)

[4.4 MODULES INVOLVED 23](#_Toc108968309)

[CHAPTER 5: METHODOLOGIES AND ML ALGORITHMS 24](#_Toc108968310)

[5.1 Disease Prediction: 24](#_Toc108968311)

[5.2 Diabetes mellitus prediction: 26](#_Toc108968312)

[5.3](#_Toc108968312) Heart Disease Prediction:…………………………………………………………………………………………………………………………………31

[CHAPTER 6: TOOLS AND TECHNIQUES 42](#_Toc108968313)

[6.1 Programming Languages 42](#_Toc108968314)

[6.2 Libraries 43](#_Toc108968315)

[6.3 Data Sets 45](#_Toc108968316)

[CHAPTER 7: RESULTS 47](#_Toc108968317)

[7.1 EXPERIMENTAL RESULTS 47](#_Toc108968318)

[7.2 PERFORMANCE ANALYSIS 48](#_Toc108968319)

[CHAPTER 8: CONCLUSION 50](#_Toc108968320)

[REFERENCES 51](#_Toc108968321)

LIST OF FIGURES

1 [Figure 3.1 Block Diagram 17](#_Toc108976770)

2 [Figure 3.2:Comparison between existing and proposed systems for heart disease prediction 18](#_Toc108976771)

3 [Figure 4.1: System Architecture 19](#_Toc108976772)

4 [Figure 4.2: Dataflow Diagram Level 0 20](#_Toc108976773)

5 [Figure 4.3:Dataflow Diagram Level 1 21](#_Toc108976774)

6 [Figure 4.4: ER Diagram 22](#_Toc108976775)

7 [Figure 4.5: Database Design 23](#_Toc108976776)

8 [Figure 5.1: Symptoms Disease Prediction Design 24](#_Toc108976777)

9 [Figure 5.2: Cross Validation Accuracy 25](#_Toc108976778)

10 [Figure 5.3: Comparison Bar Graph 26](#_Toc108976779)

11 [Figure 5.4: Diabetes Pedigree Function Generation Model 27](file:////Users/ephramdevlal/Downloads/Project%20Draft%20Report_V2%20(1).docx#_Toc108976780)

12 [Figure 5.5: Support Vector Machine 28](#_Toc108976781)

13 [Figure5.6: Work Flow of Diabetes Prediction Model 31](#_Toc108976782)

14 [Figure 5.7: Input Image of ECG 32](#_Toc108976783)

15 [Figure 5.8: Leads of ECG 33](#_Toc108976784)

16 [Figure 5.9: Preprocessed Leads of ECG 34](#_Toc108976785)

17 [Figure 5.10: Contour Images of ECG 35](#_Toc108976786)

18 [Figure 5.11: Normalized 1D Signal 35](#_Toc108976787)

19 [Figure 5.12: Accuracy of K-Nearest Neighbour 36](#_Toc108976788)

20 [Figure 5.13: Accuracy of Logistic Regression 37](#_Toc108976789)

21 [Figure 5.14: Accuracy of SVM 38](#_Toc108976790)

22 [Figure 5.15: Accuracy after using XGBoost 38](#_Toc108976791)

23 [Figure 5.16: Final accuracy after Ensembling 39](#_Toc108976792)

24 [Figure 5.17: Hyper-Parameter Tuning using GridSearchCV 39](#_Toc108976793)

25 [Figure 7.1: Accuracy Comparison Chart 48](#_Toc108976794)

26 [Figure 7.2: Heart Disease Model accuracy using Ensembling 49](#_Toc108976795)

27 [Figure 7.3: Accuracy Comparison Chart 2 49](#_Toc108976796)

CHAPTER 1

INTRODUCTION

Healthica Health Analysis App was designed to offer users an avenue that is always available to quench their health concerns. To provide users with a service that ensures that even if they were to wake up in the middle of the night when healthcare may not always be readily available, they will always have the Healthica Health Analysis App to rely on and seek immediate medical knowledge to any questions they may have.

1.1 General Background

Expensive consultation fees, long waiting times, and exposure to a plethora of diseases and infections have always been a deterrent to visiting the doctors’ office. The costs of medication on top of the consultation fee and the various additional expenses that are imposed on hapless patients may not make medical care an easily affordable option for many. The need to be reassured of one’s health, to know and understand the changes happening to our bodies or that of our loved ones have often led us to the doctor’s doorstep. Some would begrudgingly make their way to the clinic or hospital for fear of contracting some other illness other than the one that they may have visited the doctor for. This was especially witnessed during the emergence of Covid-19 when hospitals mainly housed Covid-19 patients and the general populace was much more averse to visiting the hospital or clinic for the common cold or non-life-threatening diseases. Most people were not able to visit doctors or clinics every time they had become sick or showed signs of any symptoms. As a result, they were unable to track their health and avail prolonged care. They might have even wished to get updated about their checkup reports and medications virtually without having to go to the hospital to collect them.

The medical practitioner asks various questions to gauge the possible symptoms faced by the patient and offers the most likely possibility as a diagnosis for what the symptoms indicate.

It might have happened many times that you or someone of yours, need a doctor’s help immediately, but they are not available. Most people may not be able to visit doctors or clinics every time they become sick or show any symptoms. They might even wish to get updated about their checkup reports and medications virtually without going to the hospital every time they get sick. They might also not be able to track their health and avail prolonged care.

1.2 Objective

The project aims to offer a web-based alternative that has been trained with real-life data sets to offer a similar diagnosis to that of a doctor’s based on the symptoms experienced. Users are able to receive a medical diagnosis in the comfort of their own homes.

The project showcases the utilization of a web-based application to analyze health for any possible illness and ailments using disease predictors to find acute and/or chronic diseases ranging from normal viral fever and flu to cancers, pneumonia, diabetes, and other heart diseases.

* Able to predict Symptoms
* Able to predict Diseases
* Able to view and track medical prescription
* Able to set reminders for medical prescription
* Personalized Dashboard
* Access to medical knowledge
* Analyze health for any possible illness and ailments
* Means to provide an understanding of one’s medical status
* Self-diagnosis
* Independent of doctor's schedule
* Information

1.3 Scope

The scope of the project mainly focuses on the following: a machine learning model which can be used to identify if a patient is likely to have heart disease based on input variables, another machine learning model to predict a patients’ likelihood of getting diabetes, a functional symptoms predictor which is able to predict some of the symptoms they are facing, a functional disease predictor to learn about possible diseases that a patient may have. The web app should also have features that allow users to access a personalized dashboard where they are able to view and track their medical prescriptions or set reminders so that they can consume their medicines on time. Most importantly, the application should have information regarding the various diseases so that users can acquire more knowledge regarding their illnesses or ailments.

CHAPTER 2

LITERATURE SURVEY

Several works have been done related to disease prediction systems using different machine learning algorithms in Medical Centers.

Bo Jin, Chao Che et al. (2018) proposed a “Predicting the Risk of Heart Failure With EHR Sequential Data Modeling” model designed by applying neural networks. This paper used the electronic health record (EHR) data from real-world datasets related to congestive heart disease to perform the experiment and predict the heart disease before itself. We tend to use one-hot encryption and word vectors to model the diagnosing events and foretold coronary failure events victimization the essential principles of an extended memory network model. By analyzing the results, we tend to reveal the importance of respecting the sequential nature of clinical records [1]

2.1 Heart Disease Prediction Models

Senthil Kumar Mohan et al,[1] proposed Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques in which strategy that objective is to find critical includes by applying Machine Learning bringing about improving the exactness in the expectation of cardiovascular malady. The expectation model is created with various blends of highlights and a few known arrangement strategies. We produce an improved exhibition level with a precision level of 88.7% through the prediction model for heart disease with hybrid random forest with a linear model (HRFLM) they likewise educated about Diverse data mining approaches and expectation techniques, Such as, KNN, LR, SVM, NN, and Vote have been fairly famous of late to distinguish and predict heart disease.

Lakshmana Rao et al,[5] Machine Learning Techniques for Heart Disease Prediction in which the contributing elements for heart disease are more (circulatory strain, diabetes, current smoker, high cholesterol, etc..). So, it is difficult to distinguish heart disease. Different systems in data mining and neural systems have been utilized to discover the seriousness of heart disease among people. The idea of CHD ailment is bewildering, in addition, in this manner, the disease must be dealt with warily. Not doing early identification, may impact the heart or cause sudden passing. The perspective of therapeutic science furthermore, data burrowing is used for finding various sorts of metabolic machine learning, a procedure that causes the framework to gain from past information tests, models without being expressly customized. Machine learning makes rationale dependent on chronicled information.

Abhay Kishore et al,[4] developed Heart Attack Prediction Using Deep Learning in which This paper proposes a heart attack prediction system using Deep learning procedures, explicitly Recurrent Neural System to predict the probable prospects of heart-related infections of the patient. Recurrent Neural Network is a very ground-breaking characterization calculation that utilizes Deep Learning approach in Artificial Neural Network. The paper talks about in detail the significant modules of the framework alongside the related hypothesis. The proposed model deep learning and data mining to give the precise outcomes least blunders. This paper gives a bearing and point of reference for the advancement of another type of heart attack prediction platform. Prediction stage.

Mr. Santhana Krishnan.J and Dr. Geetha.S, [6] Prediction of heart disease using machine learning algorithm This Paper predicts heart disease for Male Patient using Classification Techniques. The detailed information about Coronary Heart diseases such as its Facts, Common Types, and Risk Factors has been explained in this paper. The Data Mining tool used in WEKA (Waikato Environment for Knowledge Analysis), a good Data Mining Tool for Bioinformatics Fields. All three available Interfaces in WEKA are used here; Naive Bayes, Artificial Neural Networks and Decision Tree are Main Data Mining Techniques and through these techniques heart disease is predicted in this System. The main Methodology used for prediction is Decision Trees like CART, C4.5, CHAID, J48, ID3 Algorithms, and Naive Bayes Techniques.

Avinash Golande et al,[7] proposed Heart Disease Prediction Using Effective Machine Learning Techniques in which Specialists utilize a few data mining strategies that are available to support the authorities or doctors distinguish the heart disease. Usually, methodology utilized are decision trees, k- closest and Naive Bayes. Other unique characterization-based strategies utilized are packing calculation, Part thickness, consecutive negligible streamlining and neural systems, straight Kernel self- arranging guide and SVM (Bolster Vector Machine). The following area obviously gives subtleties of systems that were utilized in the examination.

V.V. Ramalingam et Al,[8] proposed Heart disease prediction using machine learning techniques in which Machine Learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data. Many researchers, in recent times, have been using several machine learning techniques to help the healthcare industry and the professionals in the diagnosis of heart-related diseases. This paper presents a survey of various models based on such algorithms and techniques and analyzes their performance. Models based on supervised learning algorithms such as Support Vector Machines (SVM), K- Nearest Neighbour (KNN), NaÃ¯ve Bayes, Decision Trees (DT), Random Forest (RF) and ensemble models are found very popular among the researchers and systems have been applied to different clinical datasets to robotize the investigation of huge and complex information. Numerous scientists, as of late, have been utilizing a few Machine Learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data.

2.2 Diabetes Mellitus Prediction Models

Diabetes mellitus is amongst the most significant severe problems in the medical profession. Classification is amongst the most significant decision-making methods in today’s practical circumstances. The primary goal is to categorize the data as diabetes or nondiabetic and increase the classification accuracy. Machine learning in the diagnosis of diabetes is mostly about understanding patterns from the diabetes dataset which would be given. Machine learning in recent times has always been the developing, dependable, and supportive technology in the medical sector. This study is focused on the identification of diabetes types of patients based on personal and clinical information utilizing machine learning classifiers. This section contains a summary of the works suggested by different researchers during the last decade. It is beneficial to identify the shortcomings of suggested works in the field of diabetic patients’ treatment regimen machine learning classifiers. Diagnosis of diabetes is a growing area of study.

Sonu Kumari and Archana Singh proposed [3] an intelligent and effective methodology for the automated detection of Diabetes Mellitus using Neural Network. The paper [4] approached the aim of diagnoses by using ANNs and demonstrated the need for preprocessing and replacing missing values in the dataset being considered. Through the Modified training set, a better accuracy was achieved with lesser time required for training the set. Sajida[8] by using CPCSSN(Canadian primary care sentinel surveillance Network ) dataset and three machine learning methods to predict Diabetes Diseases (DD) in the early stages to save human life from an early death. In this study Bagging, Adaboost and Decision tree(J48) were used to predict diabetes and the researcher was compare the result of those methods and concluded that Adaboost method provided effective and better accuracy than the other methods in data mining tools. Sadri [20] used Naive Bayes, RBF Network and J48 data mining algorithms for diagnosing type II diabetes. They used the WEKA tool. Finally, they found Naive Bayes, having an accuracy rate of 76.96% more than other algorithms. In this paper[27], Prediction of diabetes is done using ensemble voting classifiers for Pima Indian diabetes dataset, in comparison with different classification algorithms, the highest accuracy of 80% and 81% is achieved for the data set by using 10-fold cross-validation and by spitting data into 30% testing and 70% training. J. Pradeep Kandhasamy, S. Balamurali [58] This research study compares the performance of algorithms that are used to predict diabetes using data mining techniques. Also authors classifiers J48 Decision Tree, KNearest Neighbors, and Random Forest, Support Vector Machines to classify patients with diabetes mellitus. Authors compared four prediction models for predicting diabetes mellitus using 8 important attributes under two different situations. One is before pre-processing the dataset. Here the studies conclude that the decision tree J48 classifier achieves higher accuracy of 73.82 % than other three classifiers. After pre-processing, the dataset is given more accuracy.

Pradeep & Dr.Naveen [10] in this paper, the performance of machine learning techniques were compared and measured based on their accuracy. The accuracy of the technique varies from before preprocessing and after preprocessing as they identified in this study. This indicates that in the prediction of diseases the pre-processing of data sets has its own impact on the performance and accuracy of the prediction. Song [6] describes and explains different classification Algorithms using different parameters such as Glucose, Blood Pressure, Skin Thickness, insulin, BMI, Diabetes Pedigree, and age. The research did not include pregnancy parameters to predict diabetes disease (DD). In this research, the researchers were using only small sample data for the prediction of Diabetes. The algorithms used in this paper were five different algorithms GMM, ANN, SVM, EM, and Logistic regression. Finally, The researchers conclude that ANN (Artificial Neural Network) was providing High accuracy for the prediction of Diabetes. P. Chen[19] in their work has performed statistical testing on medical measurement index results of both patients with diabetes and without diabetes. They have further used boosting algorithms to give an excellent classification of diabetes models based on the given medical data. In this study [39]medical bioinformatics analyses have been accomplished to predict diabetes. The WEKA software was employed as a mining tool for diagnosing diabetes. The Pima Indian diabetes database was acquired from a UCI repository used for analysis. The dataset was studied and analyzed to build an effective model that predicts and diagnoses diabetes disease. In this study, we aim to apply the bootstrapping resampling technique to enhance the accuracy and then apply Naïve Bayes, Decision Trees and (KNN) and compare their performance.

Yunsheng[9] in his study was the new approach that used the KNN algorithm by removing the outlier/OOB(out of bag) using DISKR(decrease the size of the training set for K-nearest neighbor .and also in this study the storage space was minimized. Therefore, the space complexity becomes less efficient after removing parameters or instances which have less effect or factor the researchers got better accuracy. V. Kumar and L. Velide,[21] used Data mining Approach for Prediction and Treatment Of Diabetes Disease. The techniques they used were Naïve Bayes, JRip, J48 (4.5), DT, NN .They used the WEKA tool for implementation. They got 68.5% of accuracy for the J48 algorithm. The research paper [46] elaborates on a detailed review of existing data mining methods used for prediction of diabetes. It also gives about the types of diabetes disease Type1, type2, and type3. The aim of the diabetes predictor is to predict diabetes with the help of Data mining methods such as the K-Nearest Neighbor Algorithm, Bayesian Classifier, Naive Bayesian Classifier, Bayesian Network, all the methods are used for prediction of diabetes. This paper also mentions the effects of diabetes on patients. In this paper[54,]the proposed methodology aims at providing an efficient hybrid classification framework for predicting and monitoring Diabetes disease. The main aim of this research is to identify and construct models that would assist medical practitioners in an efficient way by the way benefiting the people to attain longer life in this world.

Byoung Geol Choi, Seung-Woon Rha, Sung Wook Kim, Jun Hyuk Kang, Ji Young Park, and Yung-Kyun Noh[26] successfully developed and verified a T2DM prediction system using machine learning and an EMR database, and it predicted the 5-year occurrence of T2DM similarly to with a traditional prediction model. In Kemal Polat’s work[29], Discussing Generalized Discriminant Analysis (GDA) and Least Square Support Vector Machine (LS-SVM) are used for the diagnosis of diabetes disease. Also, proposed a new cascade learning system based on Generalized Discriminant Analysis and Least Square Support Vector Machine. The proposed system includes two stages. In the first stage, they used Generalized Discriminant Analysis to discriminate feature variables between healthy and patient (diabetes) data as a pre-processing step. In the second stage, they used LS-SVM for classification of diabetes dataset. While LS-SVM obtained 78.21% classification accuracy using 10-fold cross-validation, the proposed system called GDA-LS-SVM obtained 82.05% classification accuracy using 10-fold cross-validation and it is very promising compared to the previously reported classification techniques.

Hasan Temurtas [30] a multilayer neural network structure, trained by Levenberg–Marquardt (LM) algorithm and a probabilistic neural network structure were used. The results of the study were compared with the results of the previous studies that also focused on diabetes disease diagnosis and by using the same UCI machine learning database obtains 79.62% accuracy. The classification accuracy of MLNN with LM obtained by this study using correct training was comparatively better than those obtained by other studies except the classification accuracies by Polat and Gunes. Santi Wulan [31] Implemented MKS-SSVM technique to improve accuracy of the result has been developed by many researchers. It is called Multiple Knot Spline SSVM (MKS-SSVM). Implement an experiment on Pima Indian diabetes dataset to evaluate the effectiveness of our method. The accuracy of previous results of this data is still below 80% using SSVM, which is a smooth support vector machine. Then, the proposed MKS-SSVM showed better performance in classifying diabetes disease diagnosis with an accuracy of 93.2% which is better than previously reported results. It can be concluded from the study[63] that hybrid deep learning provides the most satisfactory results for prediction of diabetes. Least error rate and highest area under the ROC curve, accuracy and precision values provide evidence of better performance as compared to pure SVM and pure Deep Learning models.

In Preeti Verma, Inderpreet Kaur, Jaspreet Kaur’s paper[36] work, the performance of this method is evaluated using 10-fold cross-validation accuracy, confusion matrix. The obtained classification accuracy using 10-fold cross-validation is 96.58% in comparison with other spline SSVM techniques. The results of this study showed that the modified spline SSVM was effective to detect diabetes disease diagnosis and this is a very promising result compared to the previously reported results. In Dr. B .L. Shivkumar and S Thiyagarajan’s work[23], an effective machine learning algorithm is proposed for the classification of type dm patients. This machine learning algorithm used for classification will find the optimal hyper-plane which divides the various classes. Sneha and Tarun[24] proposed a method that aims to focus on selecting the attributes that fail in early detection of Diabetes Mellitus using Predictive analysis. The result shows the decision tree algorithm and the Random forest has the highest specificity of 98.20% and 98.00%, respectively holds best for the analysis of diabetic data. Naïve Bayesian outcome states the best accuracy of 82.30%. The research also generalizes the selection of optimal features from the dataset to improve the classification accuracy. This paper[47] focuses on the use of data mining algorithms that can be very helpful in early prediction and in consequence early precautions before the diagnosis of disease. The main goal of this paper is to provide a comparison and suggest the best algorithm which can be used for pattern recognition or prediction in healthcare fields. After the implementations of these algorithms, it can be said that for PID dataset Decision Tree gives the best accuracy 75.65%. The tool used for testing and validation is Rapid Miner while all algorithms worked with a 70:30 ratio for training and testing.

The key limitation of the existing works and methods is that all of the diabetes prediction models work on user input pedigree function data without providing any functionality for the user to know their pedigree function which varies from individual to individual .ie a person who has no idea about his function would never be able to get through the prediction correctly In this project, an individual through the pedigree analysis using darts theory and the equation can easily find their probability value making the prediction more accurate and exact.

At present, there are plenty of algorithms available that could detect and predict heart anomalies from clinical reports. However, in this project, the focus is more on discovering and extracting patterns from Electrocardiogram (ECG or EKG) image reports. By digitizing ECG records, the need for time consuming manual intervention for comprehending the report can be eliminated. With digitization, the automation of diagnosis and analysis can be achieved more quickly. Many papers related to cardiovascular prediction focused on other features that included diet, age, gender, and many other dimensions, and then predicted cardiovascular diseases based on these features. Our work is more on predicting diseases by providing the ECG chart to our model

Another main advantage of our System is that we have incorporated the 3 different prediction models in a single Health app which is an expansion of all the existing web/mobile based web applications so far.

CHAPTER 3

DESIGN OF THE PROPOSED SYSTEM

The healthcare domain is one of the prominent research fields in the current scenario with the rapid improvement of technology and data. It is difficult to handle the huge amount of data of the patients. It is easier to handle this data through Big Data Analytics. There are a lot of procedures for the treatment of multiple diseases across the world. Machine Learning is an emerging approach that helps in the prediction and diagnosis of a disease. The project depicts the prediction of disease based on symptoms as well as the prediction of Diabetes using different attributes like glucose level, Insulin,BMI, Pedigree function, age. Heart diseases are predicted using the ECG paper records into a 1-D signal. This can be achieved by extracting the P, QRS, and T waves that exist in ECG signals to demonstrate the electrical activity of the heart using various techniques by Machine Learning algorithms such as Support vector machines, Decision Tree, K nearest-neighbors(KNN) and Random Forest are employed on the provided dataset to predict the disease. Its implementation is done through the python programming language. The Project demonstrates the best algorithm based on their accuracy. The accuracy of an algorithm is determined by the performance on the given dataset.

****

Figure 3. 1 Block Diagram

The web application system generates a checkup report every time the patient uses any of the symptoms and disease prediction models which is then cross-verified and reviewed by the doctors registered inside the application. It also allows a user to receive medicine reminders enabled as notifications on the screen at the particular time they need to take it. The system fetches all these data into the dashboard of the patient/doctor currently logged in to the system.

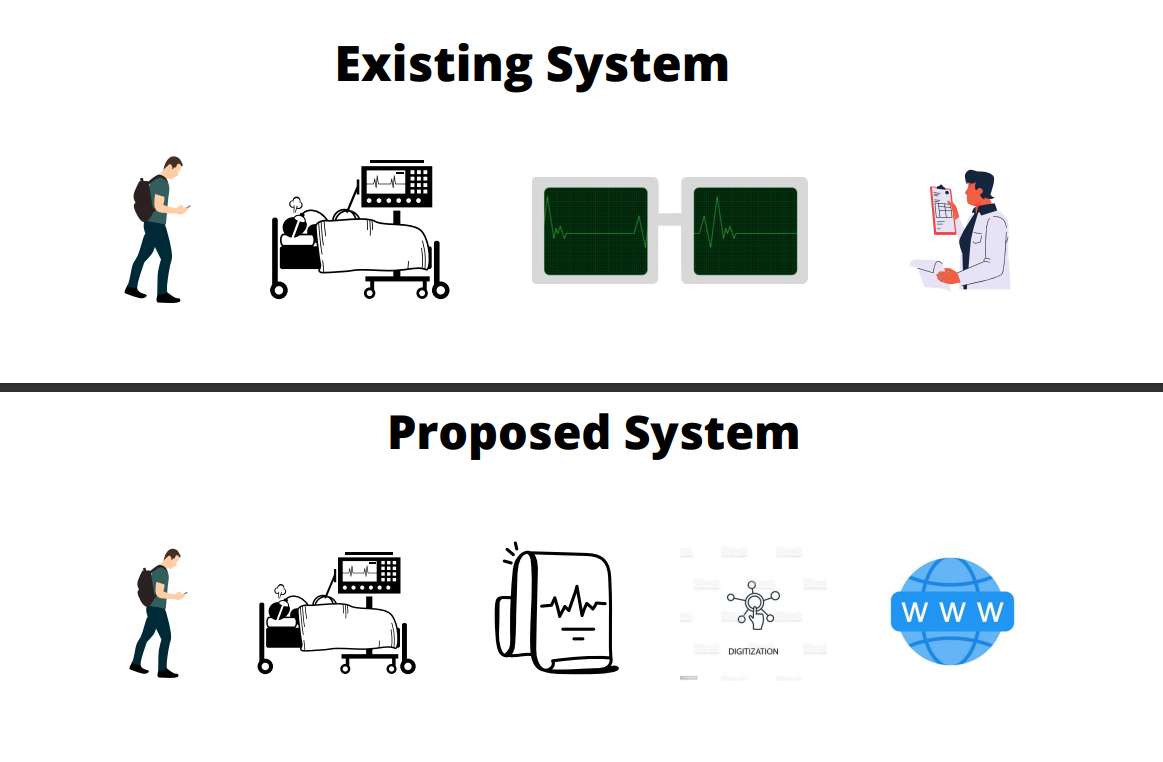


Figure 3. 2:Comparison between existing and proposed systems for heart disease prediction

CHAPTER 4

SYSTEM ARCHITECTURE

System architecture is the conceptual model that defines the structure, behavior, and more of a view of how the system works. The project aims at developing a machine learning algorithm to predict diseases that the user might possibly have. Three different ML models are used to predict a wide range of diseases based on various inputs for example the symptoms, the ECG image scan and from the details filled in by the user while logging into the system. These are given as inputs to different ML models to produce a reliable output with maximum accuracy, and the result is finally displayed in the dashboard.

4.1 SYSTEM ARCHITECTURE DIAGRAM



Figure 4.1: System Architecture

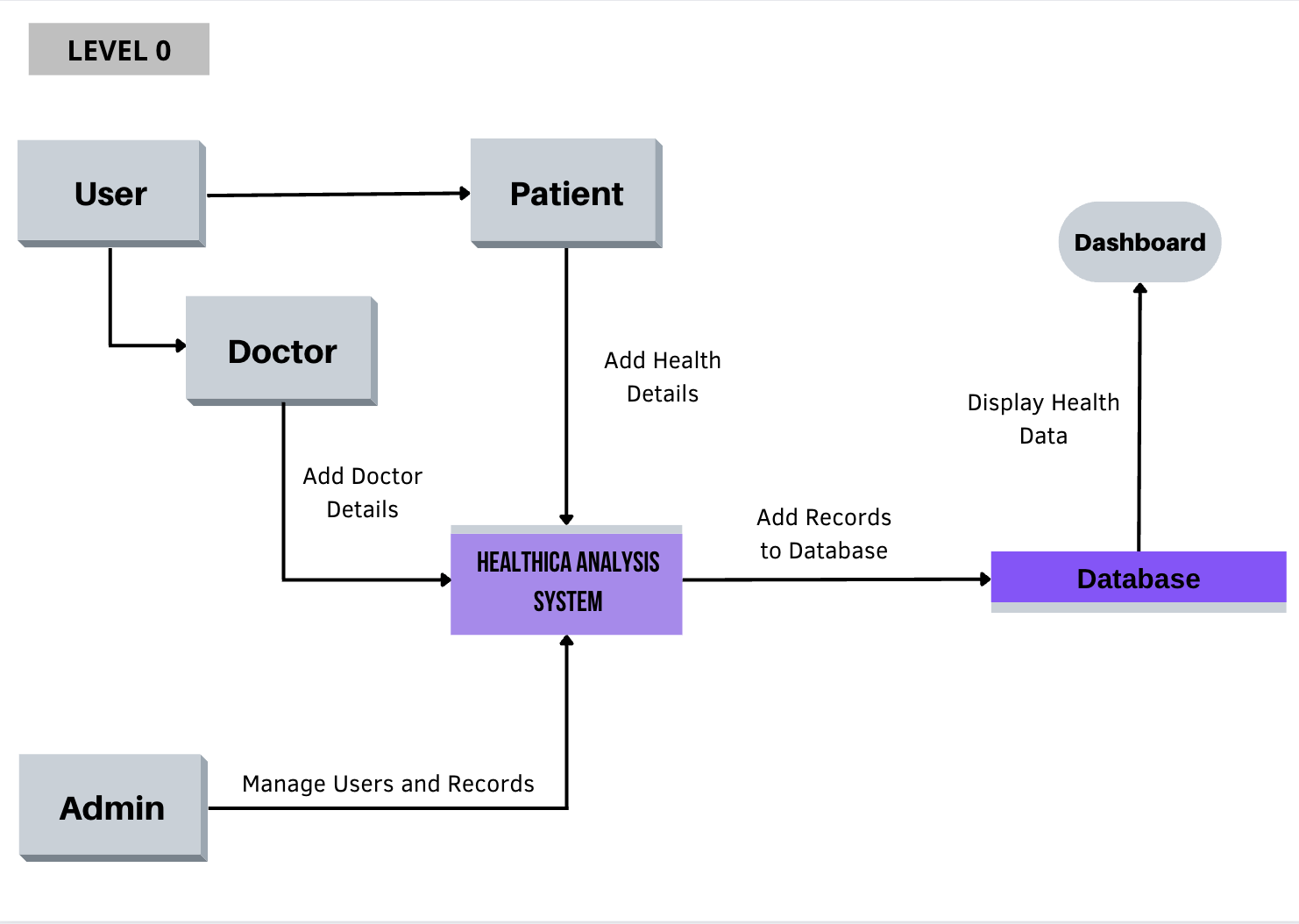
****

Figure 3.2: Dataflow Diagram Level 0

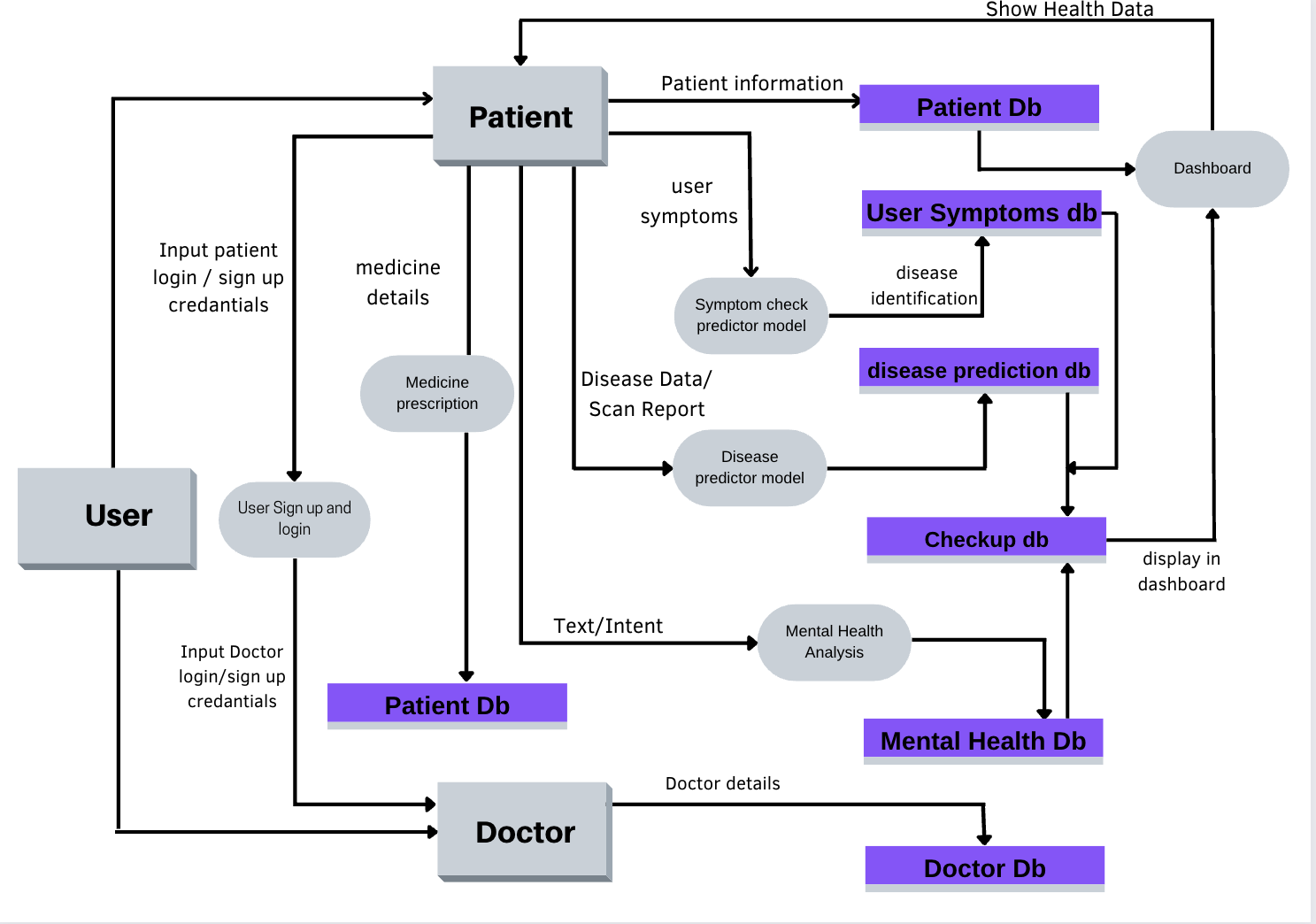


Figure 4.3:Dataflow Diagram Level 1

4.2 ER DIAGRAM

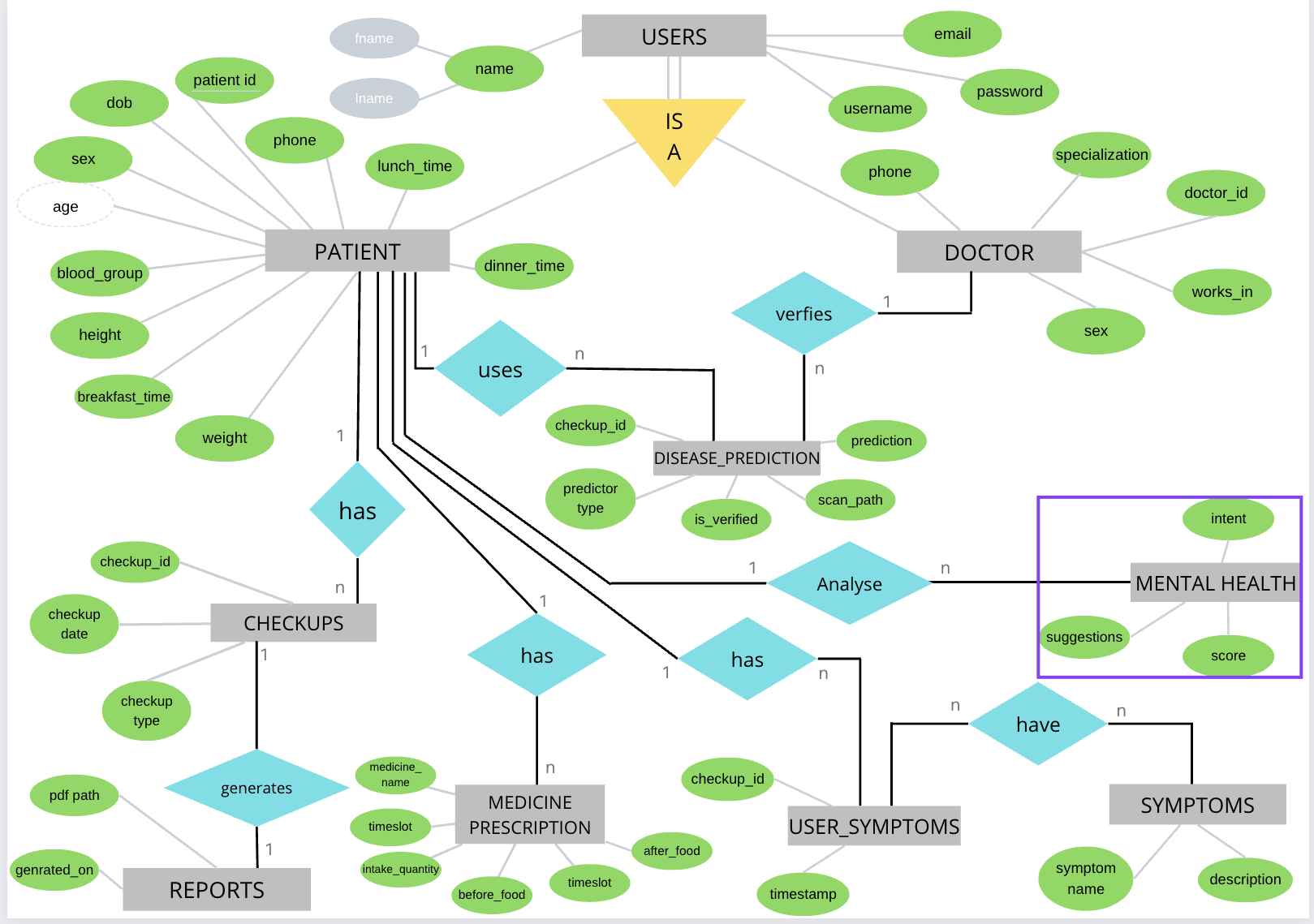


Figure 4.4: ER Diagram

4.3 DATABASE DESIGN

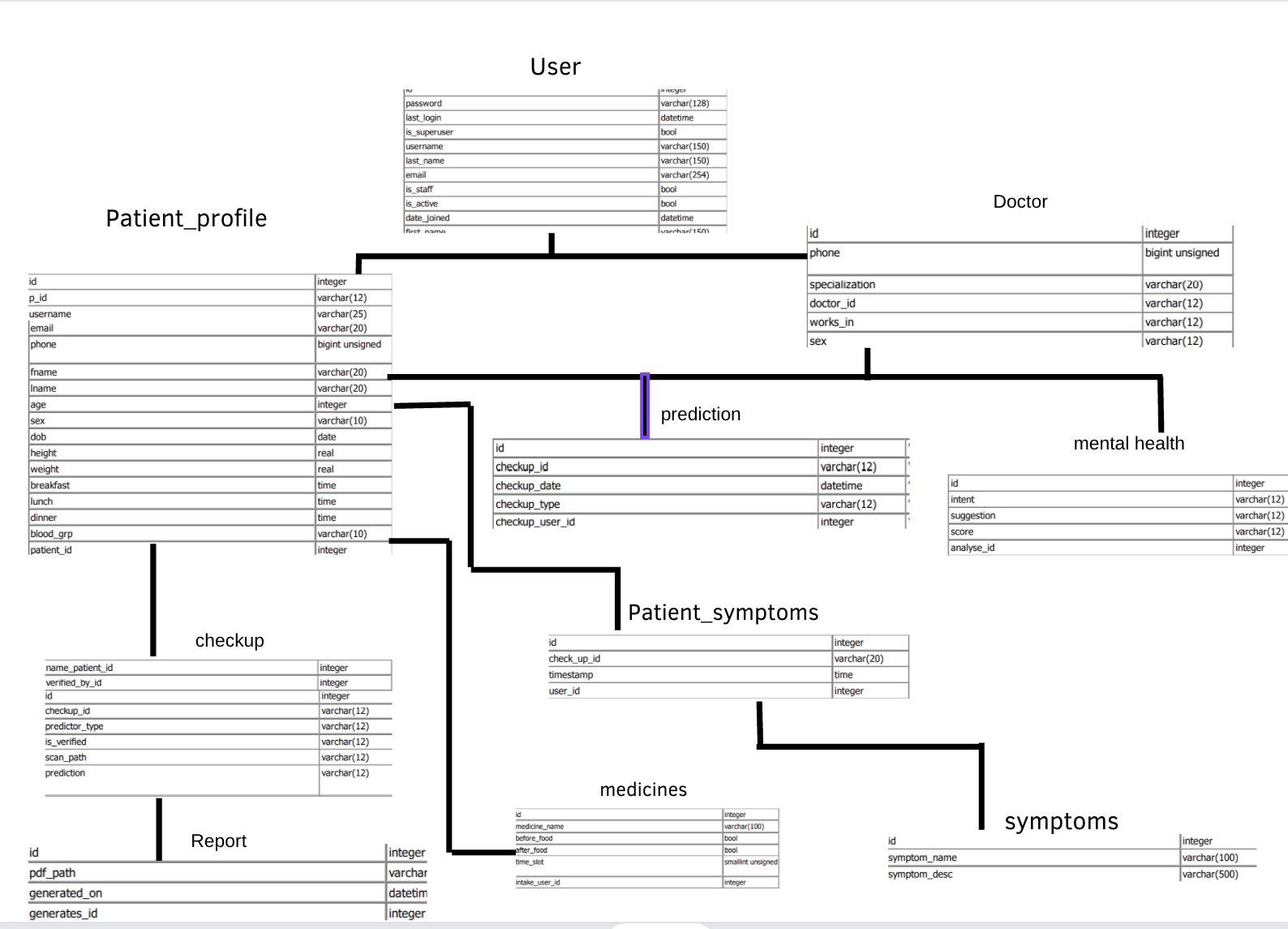


Figure 4.5: Database Design

4.4 MODULES INVOLVED

1.User login/sign up module

2. Symptoms -disease predictor module

3. Diabetes predictor module

4. Heart Disease prediction module

5. Health report/checkup dashboard

6. Medicine reminder details

7. Medicine reminder notification

CHAPTER 5

METHODOLOGIES AND ML ALGORITHMS

5.1 Disease Prediction:

5.1.1 Symptoms -disease prediction:

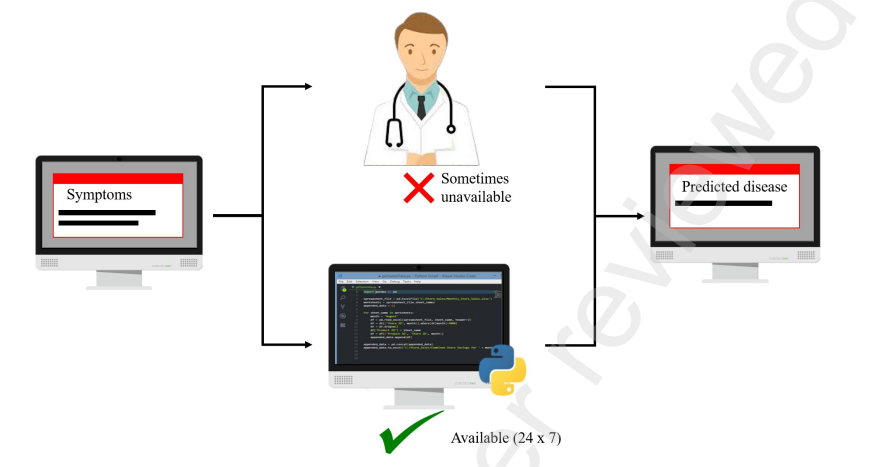
****

Figure 5.1: Symptoms Disease Prediction Design

5.1.2 Data scraping

Web scraping is a technique to fetch data from websites. While surfing on the web, many websites don’t allow the user to save data for personal use. One way is to manually copy-paste the data, which is both tedious and time-consuming. Web Scraping is the automation of the data extraction process from websites. This event is done with the help of web scraping software known as web scrapers. They automatically load and extract data from the websites based on user requirements. These can be custom built to work for one site or can be configured to work with any website.

The Disease Symptom dataset was created in a separate python program, dataset scrapping was done using the NHP website and Wikipedia data.

Beautiful Sou**p** is a Python library that is used for web scraping purposes to pull the data out of HTML and XML files. It creates a parse tree from page source code that can be used to extract data

5.1.3 Cleaning the Data

Cleaning is the most important step in a machine learning project. The quality of our data determines the quality of our machine learning model. So it is always necessary to clean the data before feeding it to the model for training. In our dataset all the columns are numerical, the target column i.e. prognosis is a string type and is encoded to numerical form using a label encoder.

5.1.4 Model Building

After gathering and cleaning the data, the data is ready and can be used to train a machine learning model. We will be using this cleaned data to train the KNN, Support Vector Classifier, Logistic regression classifier, Decision Tree Classifier, I Bayes Classifier, and Random Forest Classifier, Multilayer Perceptron Classifier. We performed Cross Validation Accuracy on different classifiers with 5 different splits.

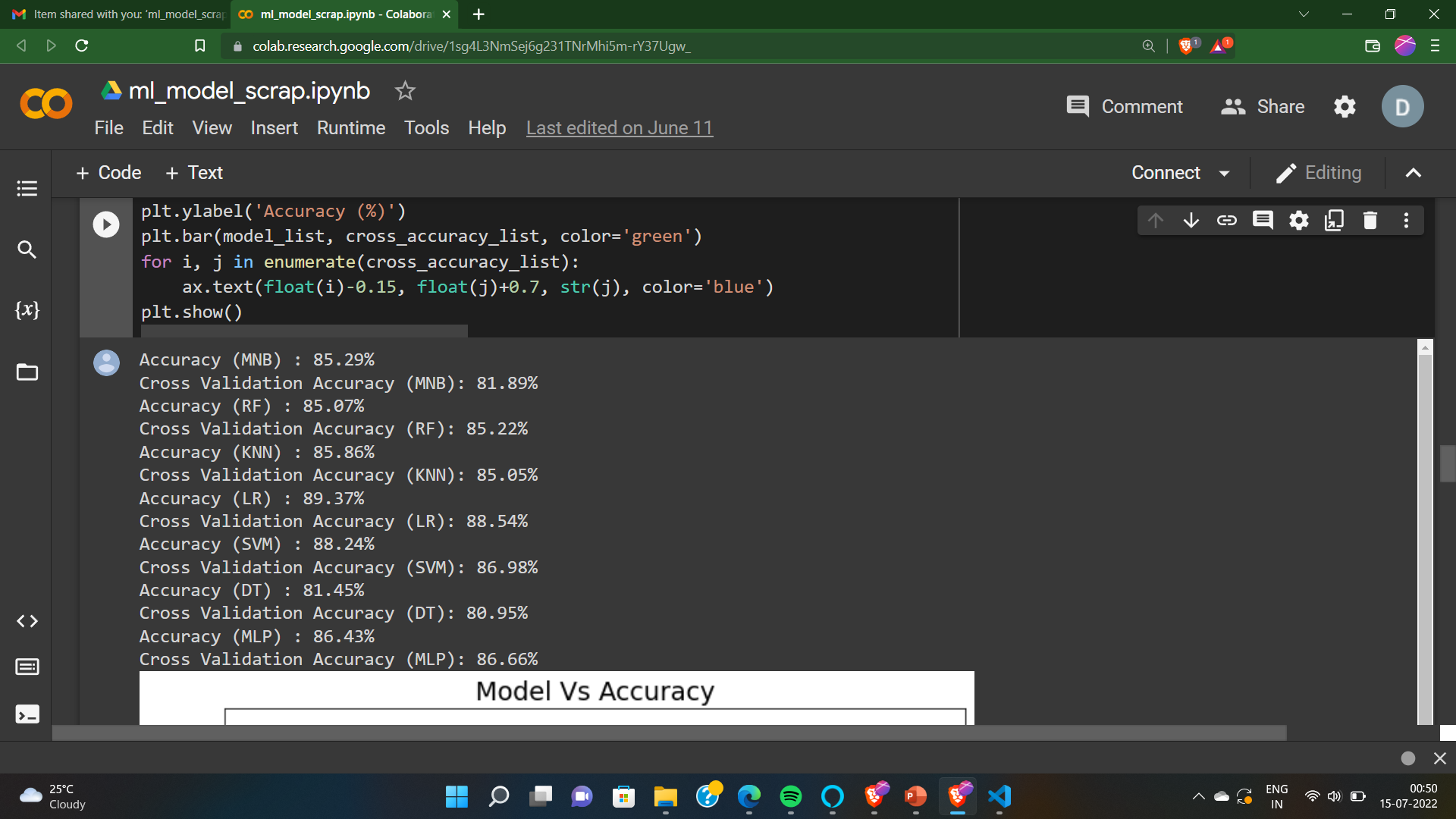


Figure 5.2: Cross Validation Accuracy

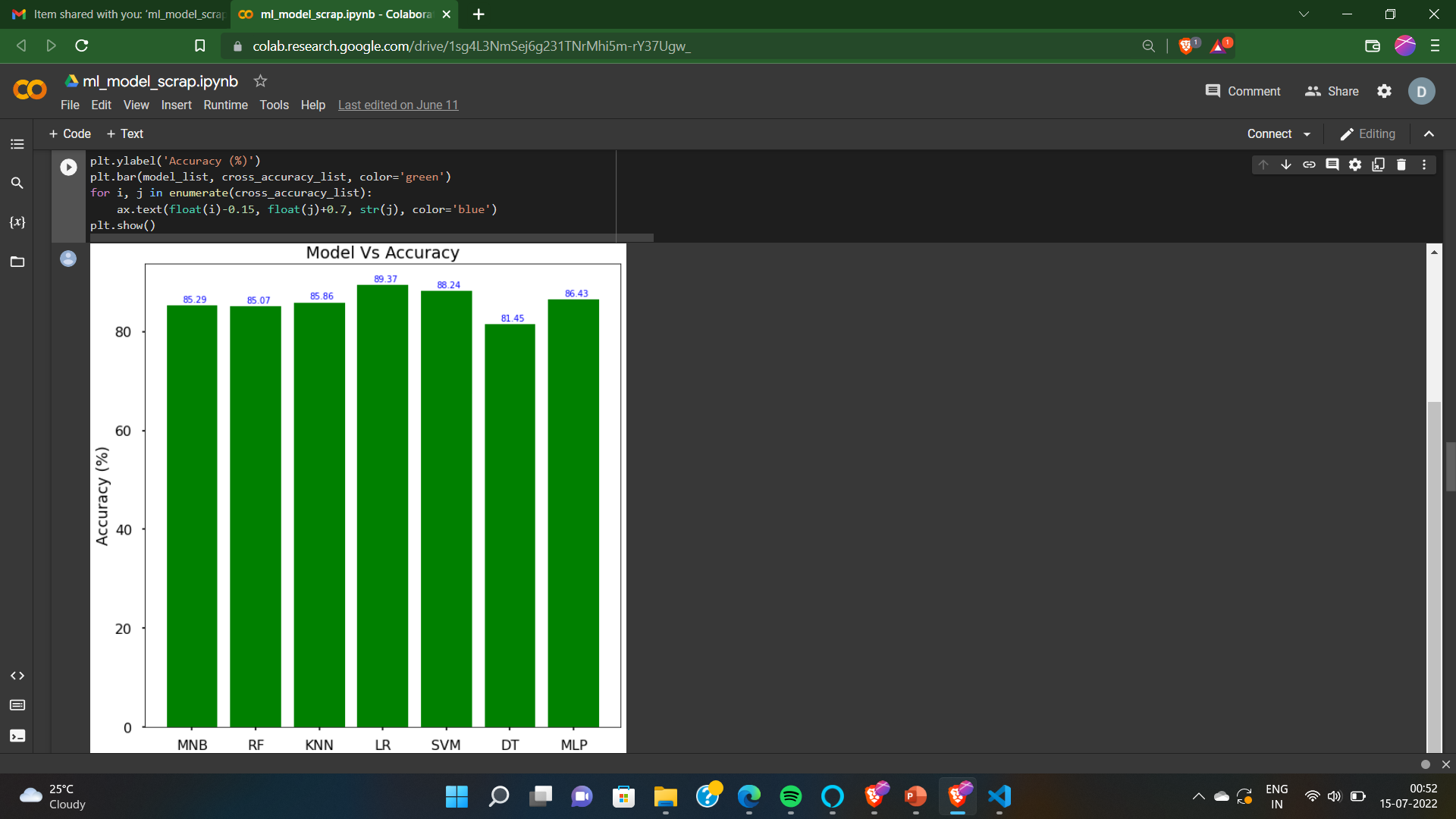
****

Figure 5.3: Comparison Bar Graph

5.2 Diabetes mellitus prediction:

5.2.1 Data Preparation

This dataset is used to predict whether a patient is likely to get diabetes based on the input parameters like Age, Glucose, Blood pressure, Insulin, BMI, diabetes pedigree function, etc. Each row in the data provides relevant information about the patient.

5.2.2 Data Preprocessing

Data preprocessing is the most important process. Mostly healthcare related data contains missing values and other impurities that can affect the effectiveness of data. To improve quality and effectiveness obtained after the mining process, Data preprocessing is done. To use Machine Learning Techniques on the dataset effectively this process is essential for accurate results and successful prediction. For this diabetes dataset, we need to perform pre-processing in two steps.

Missing Values removal: Remove all the instances that have zero (0) as worth. Having zero as worth is not possible. Therefore, this instance is eliminated. By eliminating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces dimensionality of data and helps to work faster.

Splitting of data: After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train algorithms on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Basically, the aim of normalization is to bring all the attributes under the same scale.

Removing less valued attributes: Pregnancies, Skin thickness

5.2.3 Diabetes pedigree function generation

Diagram

Description automatically generated

Figure 5.4: Diabetes Pedigree Function Generation Model

5.2.4 Apply Machine Learning

When data has been ready, we apply Machine Learning Technique. We use different classification and ensemble techniques, to predict diabetes. The methods were applied on the Pima Indians’ diabetes dataset. The main objective is to apply Machine Learning Techniques to analyze the performance of these methods and find their accuracy and also to be able to figure out the responsible/important feature which plays a major role in prediction.

5.2.5 Support Vector Machine

Support Vector Machine is also known as supervised machine learning algorithm(SVM). SVM is the most popular classification technique. SVM creates a hyperplane that separates two classes. It can create a hyperplane or set of hyperplanes in high dimensional space. This hyperplane can be used for classification or regression also. SVM differentiates instances in specific classes and can also classify the entities which are not supported by data. Separation is done by a hyperplane performs the separation to the closest training point of any class.

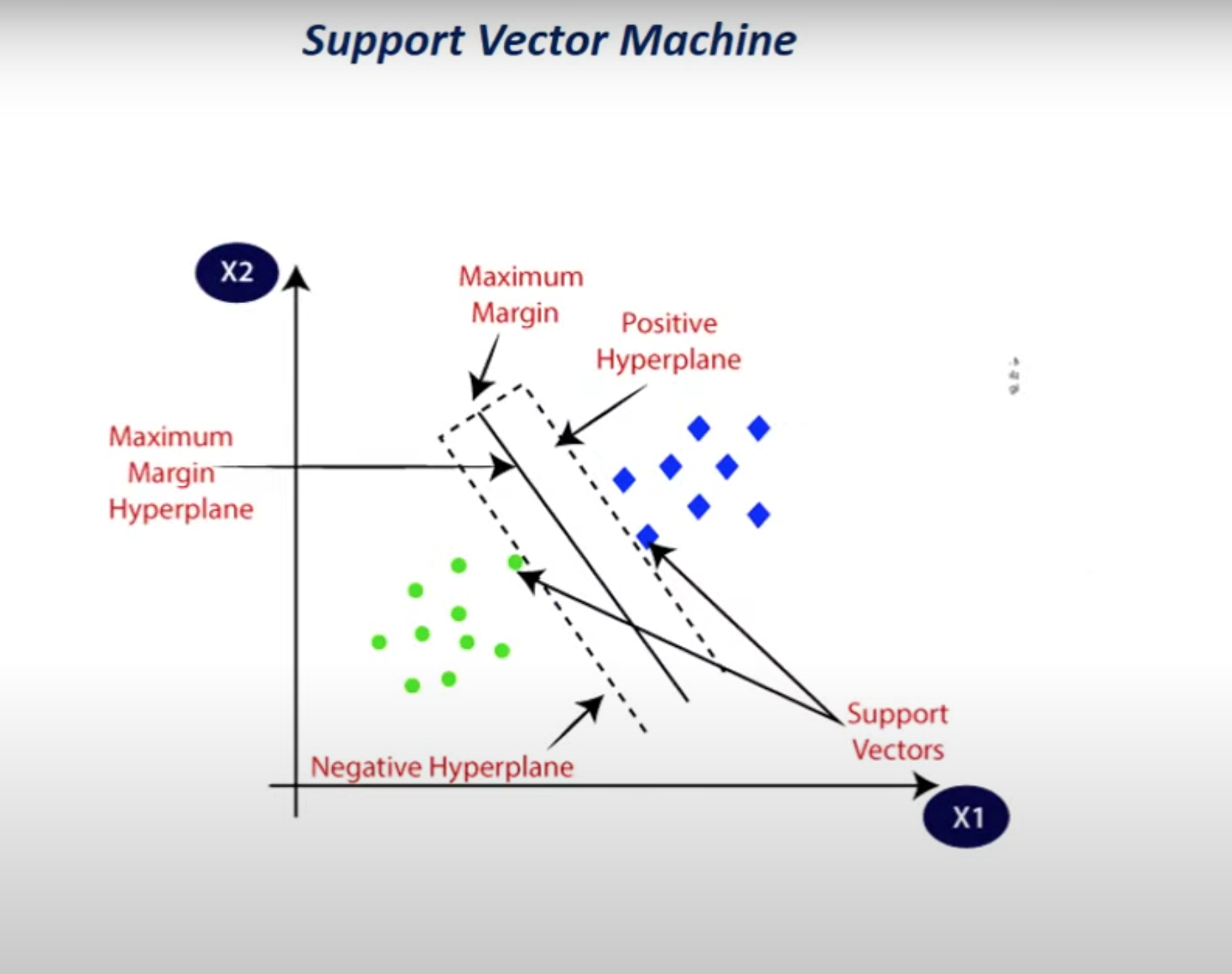


Figure 5.5: Support Vector Machine

5.2.5 K-Nearest Neighbor

KNN is also a supervised machine learning algorithm. KNN helps to solve both the classification and regression problems. KNN is a lazy prediction technique. KNN assumes that similar things are near to each other. Many times, data points which are similar are very near to each other. KNN helps to group new work based on similarity measures. KNN algorithm records all the records and classifies them according to their similarity measure. For finding the distance between the points uses a tree-like structure. To make a prediction for a new data point, the algorithm finds the closest data points in the training data set to its nearest neighbors. Here K= Number of nearby neighbors, it’s always a positive integer. The Neighbor’s value is chosen from the set of classes.

5.2.6 Decision Tree

Decision tree is a basic classification method. It is a supervised learning method. Decision tree is used when the response variable is categorical. Decision tree has a tree-like structure based model which describes classification process based on input features. Input variables are any types like graph, text, discrete, continuous, etc.

5.2.7 Ensembling

Ensembling is a machine learning technique Ensemble means using multiple learning algorithms together for some task. It provides better prediction than any other individual model I why it is used. The main cause of error is noise bias and variance, ensemble methods help to reduce or minimize these errors. There are two popular ensemble methods such as Bagging, Boosting, ada-boosting, Gradient boosting, voting, averaging, etc. Here In this work, we have used Bagging (Random forest) and Gradient boosting ensemble methods for predicting diabetes.

5.2.8 Random Forest

It is a type of ensemble learning method and is also used for classification and regression tasks. The accuracy it gives is greater than compared other models. This method can easily handle large datasets. Random Forest is developed by Leo Bremen. It is a popular ensemble Learning Method. Random Forest Improve Performance of Decision Tree by reducing variance. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the individual trees.

5.2.9 Model building

This is the most important phase which includes model building for prediction of diabetes. In this, we have implemented various machine learning algorithms which are discussed above for diabetes prediction.

Procedure of Proposed Methodology-

Step1: Import required libraries, Import diabetes dataset.

Step2: Pre-process data to remove missing data.

Step3: Perform a percentage split of 80% to divide the dataset as a Training set and 20% to Test set.

Step4: Select the machine learning algorithm i.e. K- Nearest Neighbor, Support Vector Machine, Decision Tree, Logistic regression, Random Forest and Gradient boosting algorithm.

Step5: Build the classifier model for the mentioned machine learning algorithm based on the training set.

Step6: Test the Classifier model for the mentioned machine learning algorithm based on the test set.

Step7: Perform Comparison Evaluation of the experimental performance results obtained for each classifier.

Step8: After analyzing based on various measures conclude the best performing algorithm.

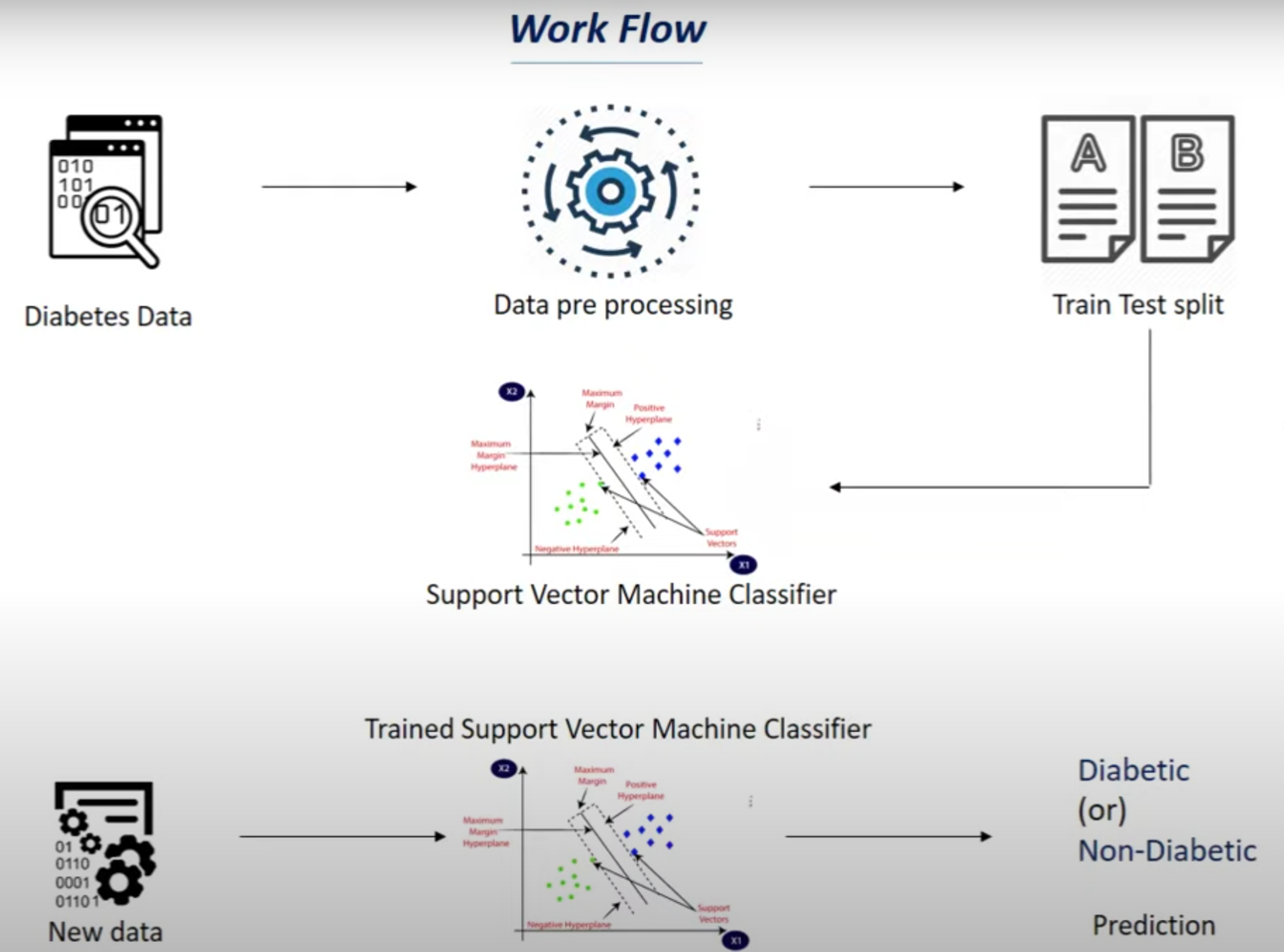


Figure5.6: Work Flow of Diabetes Prediction Model

5.3 Heart Disease Prediction:

Here, we have used 4 categories for image classification for our ECG images.

* Normal
* Myocardial infarction
* Abnormal Heartbeat
* History of Myocardial infarction

5.3.1 Data preparation

The dataset contains ECG image signals from both healthy individuals and persons with cardiovascular problems. All the ECG images from the above Mendeley dataset are parsed and transformed per our business requirements. This is the most important phase of our application (data preparation, data cleaning, data engineering and feature extraction). Each one of the images goes through various processes to extract the data in the resultant format

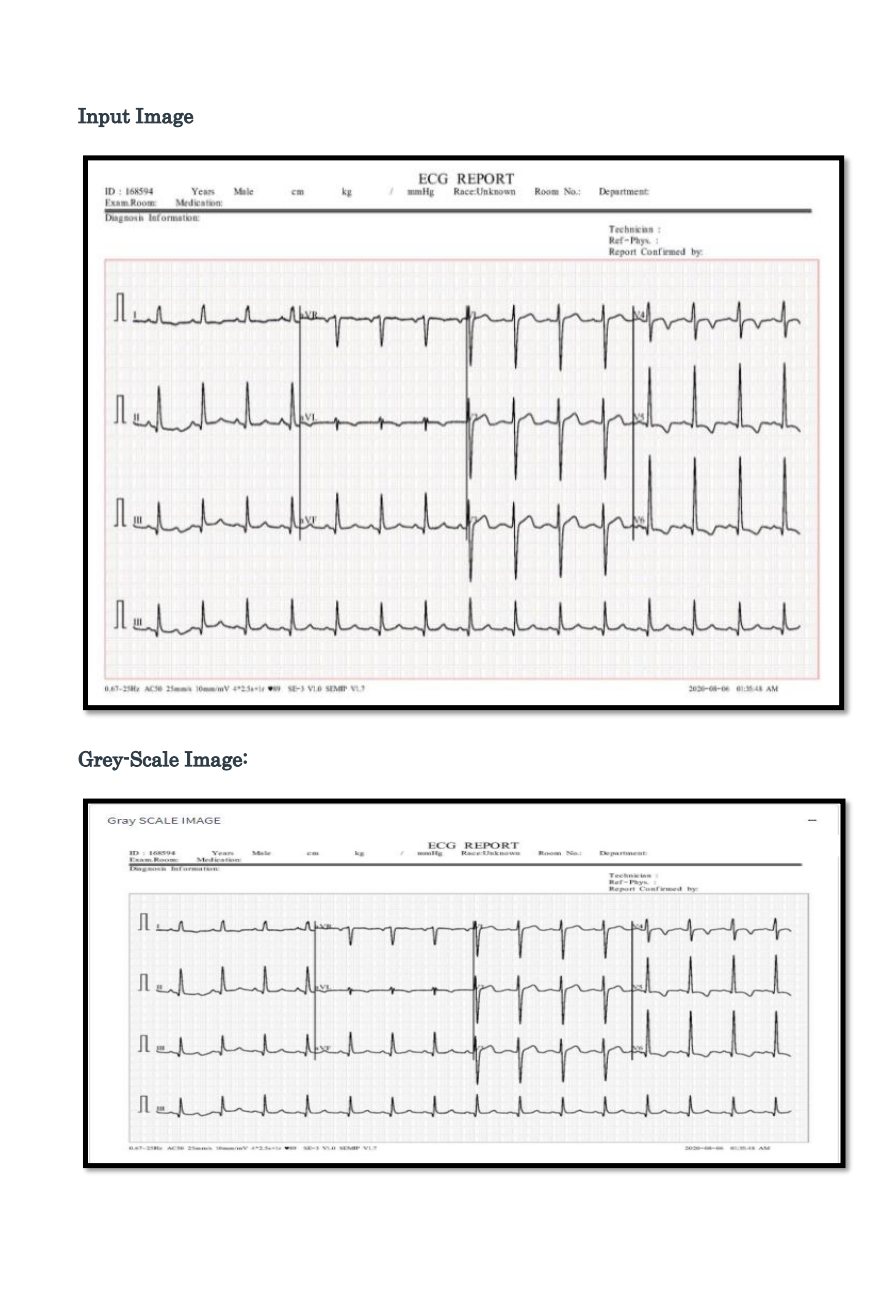


Figure 5.7: Input Image of ECG

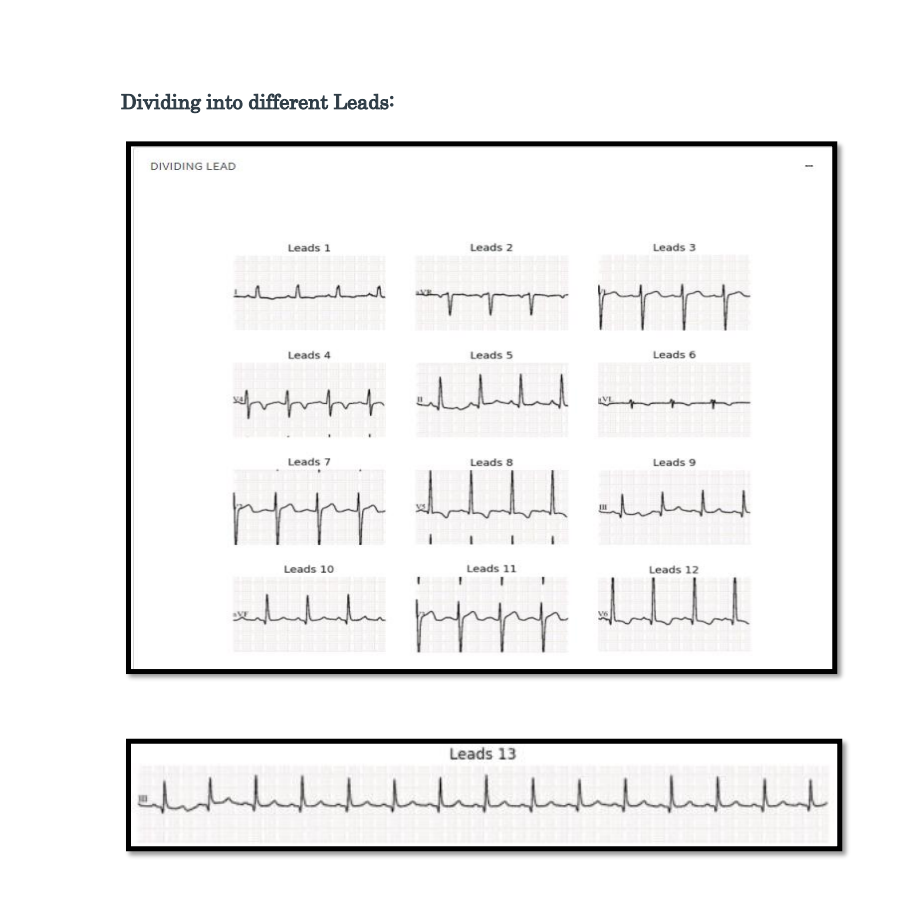


Figure 5.8: Leads of ECG

5.3.2 Data Cleaning & Feature Transformation

To prepare Leads (1-13) for further processing, each individual lead image is transformed by removing Gridlines, converting to Grayscale, applying Gaussian filtering, and performing Thresholding to convert to binary image.

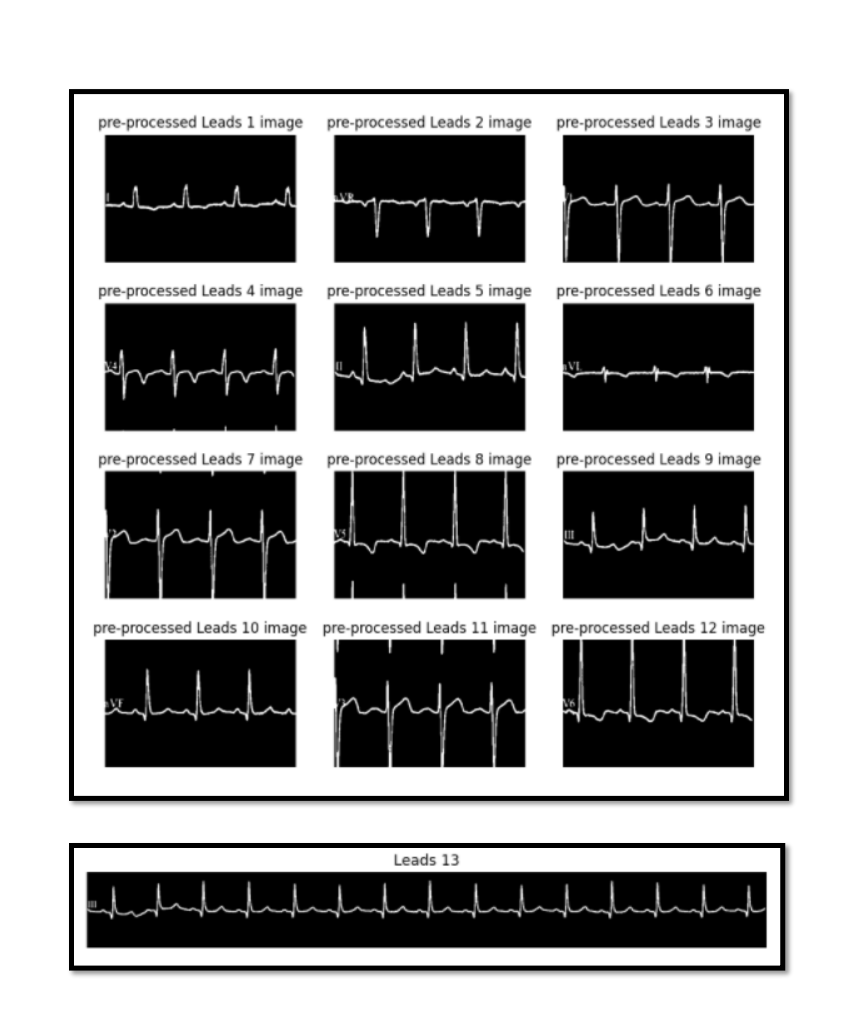


Figure 5.9: Preprocessed Leads of ECG

The transformed image is traced to extract only the signals from the image using the contour technique, and the values are scaled using the MinMax Scalar. The normalized output is saved in CSV format as a 2D signal.

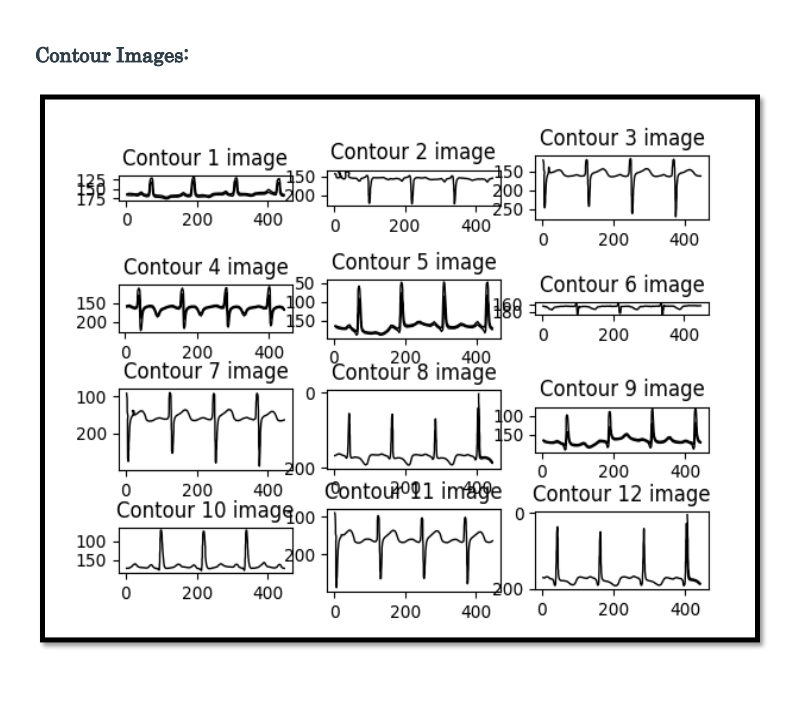


Figure 5.10: Contour Images of ECG

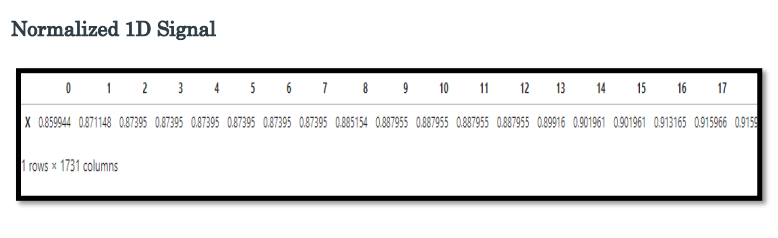


Figure 5.11: Normalized 1D Signal

We transformed all the 1D rows into columns using transpose. With both 1D and 2D CSV files and cropped 1 to 13 lead images, we perform different Supervised classification algorithms: k-nearest neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), and Voting Based Ensemble Classifier based on CSV DATA.

5.3.3 Data Engineering

Before performing data modeling, ECG images categorically belonging to four categories of patients i.e., patients with Myocardial Infarction, Abnormal Heartbeat, Myocardial Infarction History, and Good Health are combined on the lead level (from 1 to 12) and then convert target column with array ([‘No’, ‘HB’, ‘MI’, ‘PM’] into numeric using groups encoder. Post dimension reduction technique like Principal component Analysis is applied to understand the data and validate the variance explained is under an acceptable limit. Here, in this case, Total Variance Explained: 99.5. Post Dimension reduction, following data mining techniques, are applied on 12 leads combined:

* **K-Nearest Neighbors (KNN)**:

K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm that can be used for both classifications as well as regression predictive problems. However, it is mainly used for the classification of predictive problems in the industry. The K-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

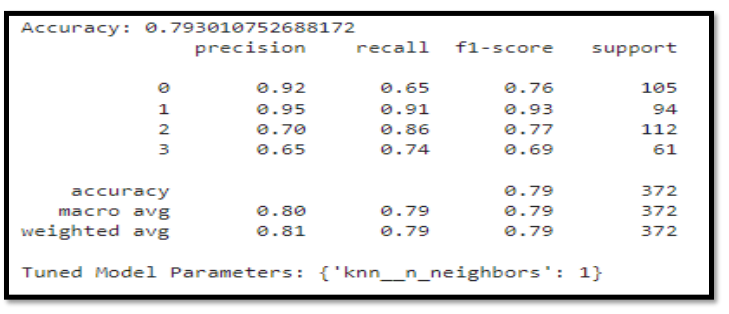


Figure 5.12: Accuracy of K-Nearest Neighbour

* **Logistic Regression**

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). This type of statistical model (also known as the logit model) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn’t vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied to the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds or the natural logarithm of odds,

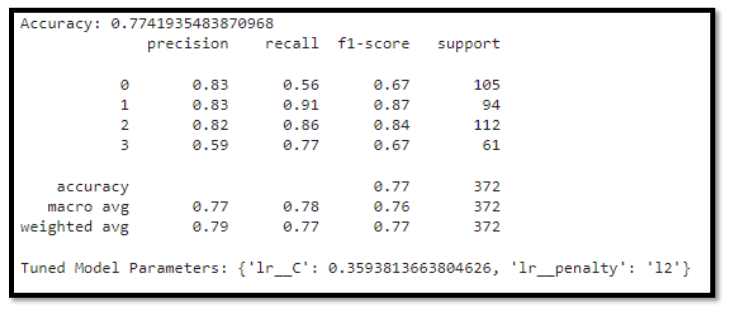


Figure 5.13: Accuracy of Logistic Regression

* **Support Vector Machine (SVM)**

Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms that are used both for classification and regression. But generally, they are used in classification problems. SVMs have their unique way of implementation as compared to other machine learning algorithms.

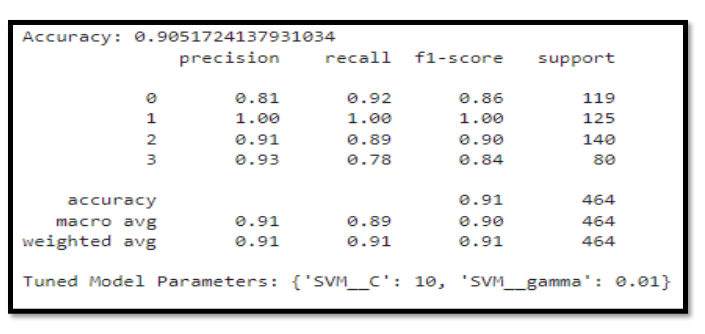


Figure 5.14: Accuracy of SVM

* **XGBoost**

XGBoost is an implementation of gradient boosted decision trees designed for speed and performance that is dominative competitive machine learning.XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems. It’s vital in the understanding of XGBoost to first grasp the machine learning concepts and algorithms that XGBoost builds upon supervised machine learning, decision trees, ensemble learning, and gradient boosting.

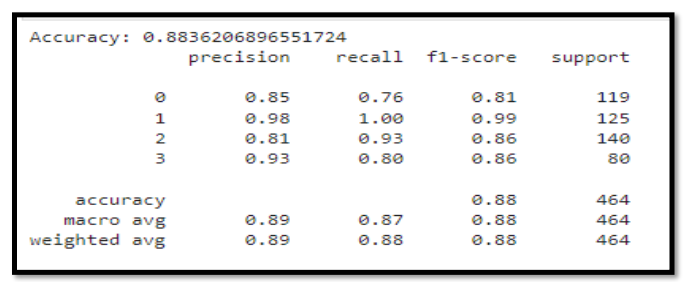


Figure 5.15: Accuracy after using XGBoost

* **Voting Based Ensemble Classifier with GridSearchCV:**

Under Voting-based Ensemble classification, three Machine learning models like k-nearest neighbors (KNN), Support Vector Machine (SVM) and Random Forest Classifier are stacked and voted to pick one model which gives the highest accuracy. For tuning the hyperparameters we have used GridSearchCV. Based on the voting, the classification report is printed as shown below:

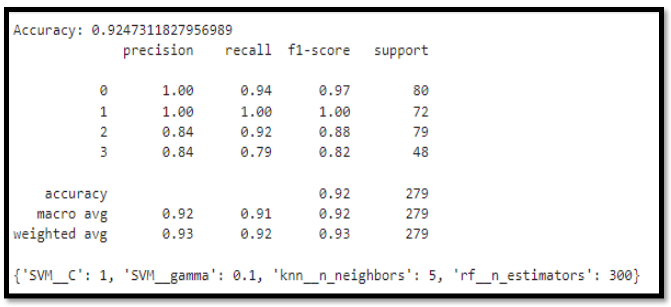


Figure 5.16: Final accuracy after Ensembling

Once the model is acceptable, we will pickle the model for future use and prediction. In this case, the model is pickled into model\_test.pkl

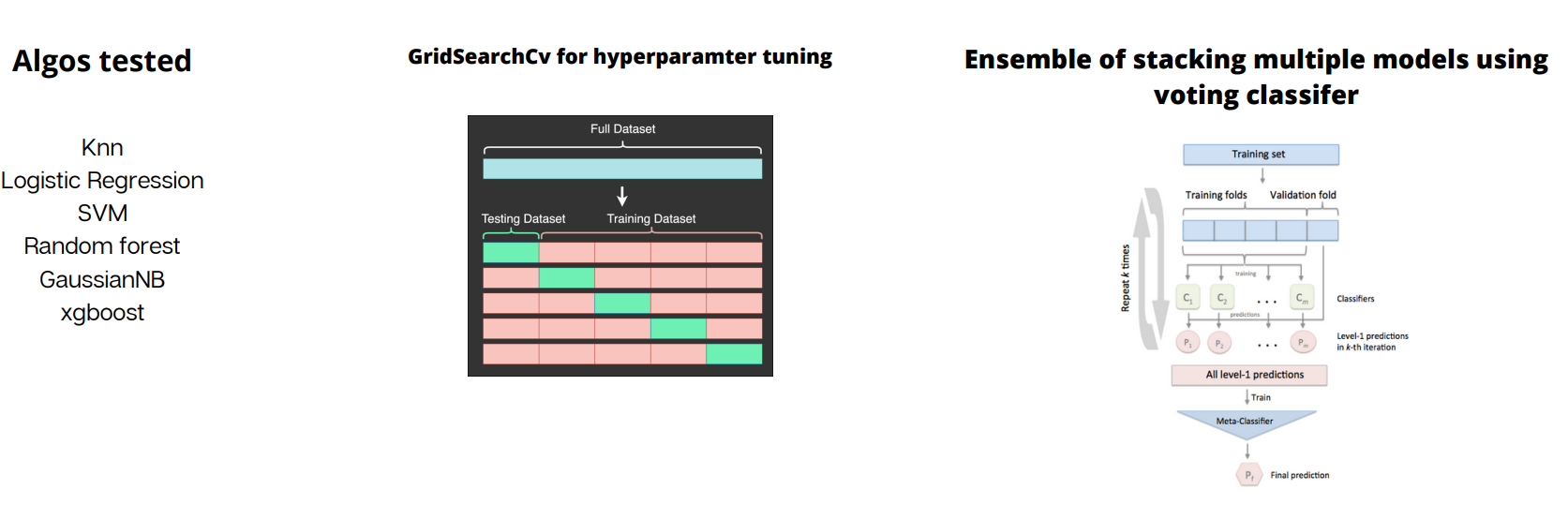


Figure 5.17: Hyper-Parameter Tuning using GridSearchCV

5.3.4 Data Scraping using Beautiful Soup

Beautiful Soup is a Python library designed for quick turnaround projects like screen-scraping. Three features make it powerful:

1. Beautiful Soup provides a few simple methods and Pythonic idioms for navigating, searching, and modifying a parse tree: a toolkit for dissecting a document and extracting what you need. It doesn't take much code to write an application
2. Beautiful Soup automatically converts incoming documents to Unicode and outgoing documents to UTF-8. You don't have to think about encodings unless the document doesn't specify an encoding and Beautiful Soup can't detect one. Then you just have to specify the original encoding.
3. Beautiful Soup sits on top of popular Python parsers like [lxml](http://lxml.de/) and [html5lib](http://code.google.com/p/html5lib/), allowing you to try out different parsing strategies or trade speed for flexibility.

5.3.5 Python Views and Features

Django Views are one of the vital participants of the MVT Structure of Django. as per Django Documentation, a view function is a python function that takes a Web request and returns a Web response. This response can be the HTML contents of a Web page, a redirect, a 404 error, an XML document, an image, or anything that a web browser can display. Django views are part of the user interface, they usually render the HTML/CSS/JavaScript in your Template files into what you see in your browser when you render a web page.

5.3.6 Django SMTP for Emails

Django comes with a ready and easy-to-use light engine to send e-mail. Similar Python libraries are imported, here smtplib. In Django, you just need to import Django.core.mail. The function send\_mail uses a separate connection for each message. The function send\_mass\_mail opens a single connection to the mail server and is mostly intended to handle mass emailing. EMAIL\_BACKEND is specified as Django.core.mail.backends.SMTP.EmailBackend. It is the default configuration that uses an SMTP server for email delivery. Defined email settings will be passed as matching arguments to EmailBackend.

5.3.7 Ajax for Real-Time Web Data

​​The primary method for facilitating asynchronous communication over the Web is asynchronous JavaScript and XML (AJAX). Although AJAX provides the Web with warranted real-time functionality, unconventional techniques of programming are required at the cost of extensive use of resources. Implementing asynchronous communication over the Web, WebSockets, which is an emerging protocol, has the potential to address many challenges. There has been no independent study analyzing AJAX and WebSockets quantitatively. This paper, therefore, provides the following contributions to integrating Web technologies in real-time systems. First, it quantitatively compares AJAX and Websocket performance in terms of connection time by varying the number of requests over the web. Second, it compares the data bytes transferred by varying the number of requests. Third, it quantitatively compares network bandwidth consumption between AJAX and WebSocket

5.3.8 Javascript Events

The change in the state of an object is known as an Event. In HTML, there are various events which represent that some activity is performed by the user or by the browser. When javascript code is included in HTML, js react over these events and allow the execution. This process of reacting to the events is called Event Handling. Thus, js handles the HTML events via Event Handlers.

5.3.9 Django Channels for Notifications

Django Channels is a project that takes Django and extends its abilities beyond HTTP, to handle WebSockets, chat protocols, IoT protocols, and more. It’s built on a Python specification called ASGI. Notifications may be sent in two ways: using the notify method of the Notifiable mixin or using the NotificationSender.Adding Notifiable mixin in your Profile model will allow you to easily send notifications to profiles using notify method. The notify method expects to receive a notification instance. Alternatively, you may send notifications via the NotificationSender. This is useful primarily when you need to send a notification to multiple notifiable entities such as a list of profiles. To send notifications using the NotificationSender, pass all of the notifiable entities and the notification instance to the send method.

CHAPTER 6

TOOLS AND TECHNIQUES

6.1 Programming Languages

6.1.1 Python

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse.

6.1.2 Django

Django is an advanced Web framework written in Python that makes use of the model view controller (MVC) architectural pattern. Django was created in a fast-moving newsroom environment, and its key objective is to ease the development of complicated, database-driven websites. This Web framework was initially developed for The World Company for managing some of their news-oriented sites. In July 2005, it was publicly released under a BSD license. Django uses Python extensively to create files, settings and

data models. It is designed to address two main challenges: the rigorous requirements of highly experienced Web developers and the intense deadlines of a newsroom.

6.1.3 Celery

Django-celery provides Celery integration for Django; Using the Django ORM and cache backend for storing results, auto-discovery of task modules for applications listed in INSTALLED\_APPS, and more. Celery is a task queue/job queue based on distributed message passing. It is focused on real-time operation but supports scheduling as well. The execution units, called tasks, are executed concurrently on a single or more worker servers. Tasks can execute asynchronously (in the background) or synchronously (wait until ready). Celery is already used in production to process millions of tasks a day. Celery is written in Python, but the protocol can be implemented in any language. It can also operate with other languages using Webhooks.

6.1.4 Html

HTML is a markup language that is used for creating attractive web pages with the help of styling, and which looks in a nice format on a web browser. An HTML document is made of many HTML tags and each HTML tag contains different content.HTML5 is the newest version of the HyperText Markup language. The first draft of this version was announced in January 2008. It is a very easy and simple language. It can be easily understood and modified. It is very easy to make an effective presentation with HTML because it has a lot of formatting tags.

6.1.5 Javascript

JavaScript, on the other hand, is a dynamic programming language that supports Math calculations, allows you to dynamically add HTML contents to the DOM, creates dynamic style declarations, fetches contents from another website, and lots more. JavaScript is an interpreted language. Thus, it doesn't need to be compiled. JavaScript renders web pages interactively and dynamically. This allows the pages to react to events, exhibit special effects, accept variable text, validate data, create cookies, detect a user’s browser, etc.

6.2 Libraries

6.2.1 Numpy

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more. It also has functions for working in the domain of linear algebra, Fourier transform, and matrices. At the core of the NumPy package, is the *ndarray* object. This encapsulates *n*-dimensional arrays of homogeneous data types, with many operations being performed in compiled code for performance.

6.2.2 Pandas

Pandas is an open-source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named Numpy, which provides support for multi-dimensional arrays. As one of the most popular data wrangling packages, Pandas works well with many other data science modules inside the Python ecosystem, and is typically included in every Python distribution, from those that come with your operating system to commercial vendor distributions like ActiveState’s ActivePythonPandas makes it simple to do many of the time consuming, repetitive tasks associated with working with data, including Data cleansing, Data fill, Data normalization, Data visualization, Statistical analysis, Data inspection and Loading and saving data.

6.2.3 SciPy

SciPy in Python is an open-source library used for solving mathematical, scientific, engineering, and technical problems. It allows users to manipulate the data and visualize the data using a wide range of high-level Python commands. SciPy is built on the Python NumPy extension. SciPy is built on top of NumPy. SciPy module in Python is a fully-featured version of Linear Algebra while Numpy contains only a few features. Most new Data Science features are available in Scipy rather than Numpy**.**

6.2.4 Matplotlib

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. It was introduced by John Hunter in 2002. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram, etc. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK.

6.2.5 SciKit-Learn

Scikit-learn is a key library for the Python programming language that is typically used in machine learning projects. Scikit-learn is focused on machine learning tools including mathematical, statistical and general purpose algorithms that form the basis for many machine learning technologies. As a free tool, Scikit-learn is tremendously important in many different types of algorithm development for machine learning and related technologies. It features various classification, regression and clustering algorithms including support-vector-machines, random forests, gradient boosting, *k-means* and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

6.2.6 Beautiful Soup

Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work. It is very fast, extremely lenient, parses pages the same way a browser does ”prettify()” is a built-in function provided by the Beautiful Soup module, it gives the visual representation of the parsed URL Source code. i.e. it arranges all the tags in a parse-tree manner with better readability.

6.3 Data Sets

6.3.1 Diabetes Dataset

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. The datasets consist of several medical predictor variables and one target variable, Outcome. Predictor variables include their diabetes pedigree function, BMI, insulin level, age, and so on.

6.3.2 Heart Disease - ECG Dataset

ECG images dataset of Cardiac Patients created under the auspices of Ch. Pervaiz Elahi Institute of Cardiology Multan, Pakistan that aims to help the scientific community in conducting research for Cardiovascular diseases. The 4 categories for image classification for ECG images are:

* Normal
* Myocardial infarction
* Abnormal Heartbeat
* History of Myocardial Infarction

ECG images: https://data.mendeley.com/datasets/gwbz3fsgp8/2 The above dataset contains ECG image signals from both healthy individuals and persons with cardiovascular problems.

6.3.3 Symptoms and Corresponding Diseases Dataset

This dataset was not readily available on the internet to use, hence we scrapped the information of each disease from the nph.gov diseases site and their corresponding symptoms were scrapped from Wikipedia to create a symptoms-disease dataset.

CHAPTER 7

RESULTS

7.1 EXPERIMENTAL RESULTS

Different machine learning models were used to examine the prediction of disease using symptoms for the available input dataset. We used 7 different ML models for the prediction. Out of the 11 models we managed to get 85 % or above accuracy for all the 7 models. As shown in Figure 4, among all the models, we gained the highest accuracy for the logistic regression model at 89.5%. The accuracy is high because this value changed according to our dataset i.e. it was small and large for the training set. Due to this variation, it proved to be the most accurate model as compared to the other ML algorithms.

Another important real-world medical problem is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of diabetes. During this work, 3 machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 81.17 % using the Simple Vector Machine algorithm. In the future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithm

The Heart Disease experimental results determine the adequacy of the designed system with an achieved accuracy of 81.17 % using ensemble technique where the 3 different models are stacked and combined to provide improved accuracy and performance of 92.4%, the 3 models were KNN, Logistic regression, SVM. Here the user uploads ECG images to our web app. From there, image conversion techniques such as rgb2gray conversion, denoising, Gaussian Filtering, thresholding and contouring are implemented to extract signals without the grid lines. The signal is then dimensionally reduced, and the necessary waves (P, QRS, T) are extracted using segmentation and fed into our pre-trained model from the analysis. Once the model finishes the analysis, it returns the results to the user based on the findings.

7.2 PERFORMANCE ANALYSIS

7.2.1 Symptom disease model accuracy

Chart, bar chart, histogram

Description automatically generated

Figure 7.1: Accuracy Comparison Chart

7.2.2 Heart disease model accuracy using ensembling

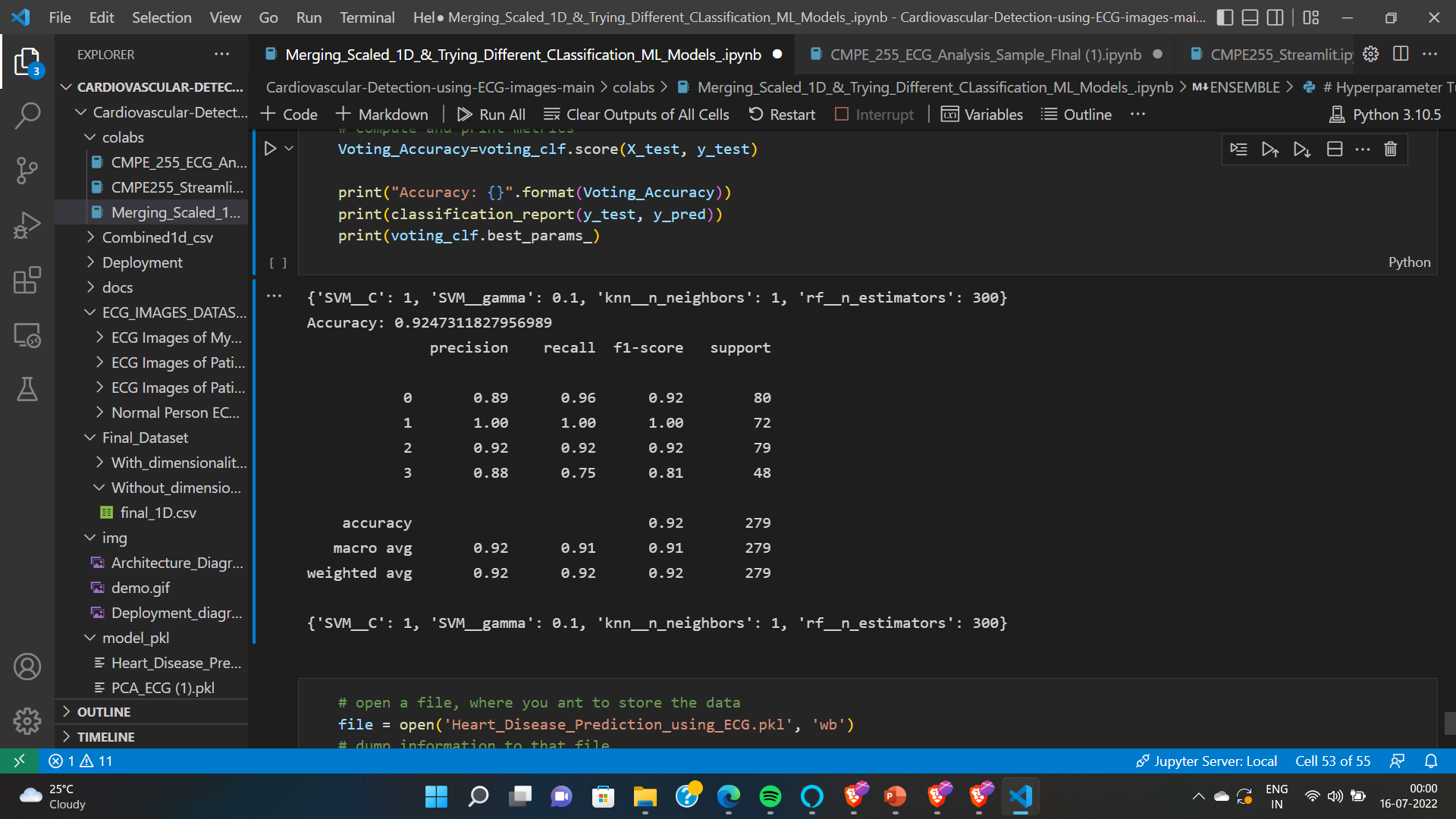
****

Figure 7.2: Heart Disease Model accuracy using Ensembling

7.3.3 Diabetes model accuracy

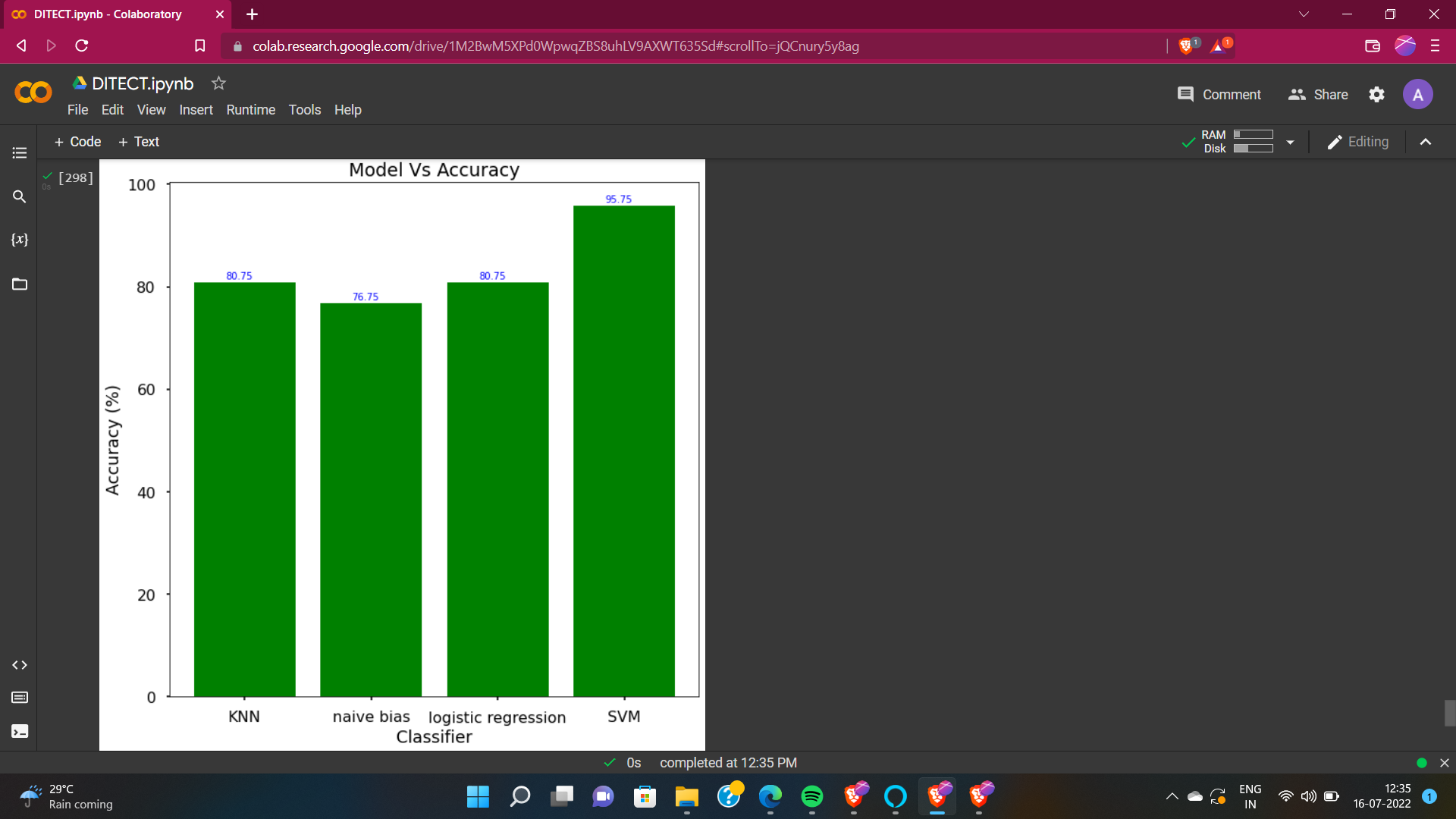


Figure 4.3: Accuracy Comparison Chart 2

CHAPTER 8

CONCLUSION

One of the important real-world medical problems is the detection of diseases at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of diseases like diabetes, heart diseases and other common illnesses based on symptoms. During this work, 3 machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on 3 Databases. Experimental results determine the adequacy of the designed system with an achieved accuracy of 85-90% using logistic regression, KNN and SVM. algorithms. In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of disease analysis including some other machine learning algorithms

REFERENCES

[1] <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

[2]"Heat Illness: Prevention, Symptoms & Treatment", *Cleveland Clinic*, 2022. [Online]. Available: <https://my.clevelandclinic.org/health/diseases/16425-heat-illness> .

[3]"Diabetes Symptoms, Causes, & Treatment | ADA", *Diabetes.org*, 2022. [Online]. Available: <https://www.diabetes.org/diabetes>

[4]"XGBoost - GeeksforGeeks", *GeeksforGeeks*, 2022. [Online]. Available: <https://www.geeksforgeeks.org/xgboost/>

[5]*Opendiabetesjournal.com*, 2022. [Online]. Available: <https://opendiabetesjournal.com/contents/volumes/V7/TODIAJ-7-5/TODIAJ-7-5.pdf>