

A Project Report on
**Heart Disease Identification Method using
Machine Learning**

Submitted by

Samuel Robert (Roll no. 53)

Sneha Sable (Roll no. 54)

Shobhan Akshay Giridharan (Roll no. 63)

Payal Thorat (Roll no. 68)

in partial fulfillment for the award of the degree

BACHELOR OF ENGINEERING

in

Electronics and Telecommunication Engineering

Under the Guidance of

Mrs. Pallavi Patil



St. Francis Institute of Technology, Mumbai

University of Mumbai

2021 - 2022

CERTIFICATE

This is to certify that Samuel Robert, Sneha Sable, Shobhan Akshay Giridharan, Payal Thorat are the bonafide students of St.Francis Institute of Technology, Mumbai. They have successfully carried out the project titled “Heart Disease Identification Method using Machine Learning” in partial fulfilment of the requirement of B. E. Degree in Electronics and Telecommunication Engineering of Mumbai University during the academic year 2021-2022. The work has not been presented elsewhere for the award of any other degree or diploma prior to this.

(Pallavi Patil)

(Dr. Gautam Shah)
EXTC HOD

(Dr. Sincy George)
Principal

Project Report Approval for B.E.

This project entitled '*Heart Disease Identification Method using Machine Learning*' by **Samuel Robert, Sneha Sable, Shobhan Akshay Giridharan, Payal Thorat** is approved for the degree of Bachelor of Engineering in Electronics and Telecommunication from University of Mumbai.

Examiners

1. - - - - -

2. - - - - -

Date:

Place:

ACKNOWLEDGEMENT

We are thankful to a number of individuals who have contributed towards our final year project and without their help; it would not have been possible. Firstly, we offer our sincere thanks to our project guide, Mrs. Pallavi Patil for her constant and timely help and guidance throughout our preparation.

We are grateful to all project co-ordinators for their valuable inputs to our project. We are also grateful to the college authorities and the entire faculty for their support in providing us with the facilities required throughout this semester.

We are also highly grateful to Dr. Gautam A. Shah, Head of Department (EXTC), Principal, Dr. Sincy George, and Director Bro. Jose Thuruthiyil for providing the facilities, a conducive environment and encouragement.

Signatures of all the students in the group

(Samuel Robert)

(Sneha Sable)

(Shobhan Akshay Giridharan)

(Payal Thorat)

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included; we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in this submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Signatures of all the students in the group

(Samuel Robert)

(Sneha Sable)

(Shobhan Akshay Giridharan)

(Payal Thorat)

Abstract

Heart disease, alternatively known as cardiovascular disease, encases various conditions that impact the heart and is the primary basis of death worldwide over the span of the past few decades. Machine Learning is used across many ranges around the world. The healthcare industry is no exclusion. Machine Learning can play an essential role in predicting presence/absence of locomotors disorders, Heart diseases and more. Such information, if predicted well in advance, can provide important intuitions to doctors who can then adapt their diagnosis and dealing per patient basis. We work on predicting possible Heart Diseases in people using Machine Learning algorithms. In this project we perform the comparative analysis of classifiers like Decision Tree, Naive Bayes, Logistic Regression, Support Vector Machine and Random Forest. It uses Heart Statlog Cleaveland Hungary Dataset of UCI repository. In this Dataset there are 1190 instances and 12 attributes. This research paper aims to envision the probability of developing heart disease in the patients. The results portray that the highest accuracy score is achieved with Random Forest.

Keywords: Heart Disease Prediction, Decision Tree, Naive Bayes, Random Forest, Logistic Regression, Support Vector Machine, Machine Learning, Python Programming, Confusion Matrix.

Contents

Acknowledgement	iii
Declaration	iv
Abstract	v
List of Figures	vii
List of Tables	viii
Nomenclature	ix
1 Introduction	1
1.1 Motivation	2
1.2 Problem Statement	2
1.3 Methodology	3
1.3.1 Collection of dataset	3
1.3.2 Selection of attributes	3
1.3.3 Training and Testing of Model	4
1.3.4 Prediction of Disease	4
1.3.5 Data Flow Model	5
1.4 Organization of Project Report	6
2 Literature Review	7
3 Software Hardware Support	13
3.1 Hardware requirements:	13
3.2 Software requirements:	13

3.3	Python:	13
3.3.1	Python Libraries:	14
3.3.2	Jupyter notebook:	16
4	Heart Disease Prediction Using Machine Learning Techniques	17
4.1	DATA SET:	17
4.2	MACHINE LEARNING ALORITHMS:	18
4.2.1	Support Vector Machine (SVM):	18
4.2.2	Logistic Regression:	20
4.2.3	Decision Trees:	22
4.2.4	Random Forest:	23
4.2.5	Naive Bayes:	25
4.3	VOTING METHOD:	27
4.4	USER INTERFACE:	27
5	ANALYSIS AND DISCUSSIONS OF EXPERIMENTAL RESULTS	29
6	Conclusion and Future work	33
	Appendix-I:Timeline of the project	37

List of Figures

1.1	Flow Chart of the project	5
4.1	Decision tree Flowchart[15]	23
4.2	Random forest working[14]	25
5.1	Streamlit output 1	30
5.2	Streamlit output 2	30
5.3	Streamlit output 3	31
5.4	Streamlit output 4	31
5.5	Confusion matrix	32
6.1	Timeline of project (SEM VII)	38
6.2	Timeline of project (SEM VIII)	39

List of Tables

4.1	Attributes of Dataset	18
5.1	Model Wise Accuracy	29
5.2	Model Wise Confusion Matrix Report	29

Nomenclature

AI	Artificial Intelligence
DT	Decision Tree
LR	Logistic Regression
MLA	Machine Learning Algorithm
ML	Machine Learning
RF	Random Forest
SVM	Support Vector Machine
UCI	University of California, Irvine

Chapter 1

Introduction

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease[16]. Heart disease is one of the biggest causes of morbidity and mortality among the population of the world[16]. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of data analysis. The load of cardiovascular disease is rapidly increasing all over the world from the past few years[16]. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms[16]. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduces the complications.

Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the health care industry. This project aims to predict future Heart Disease by analyzing data of patients which classifies whether they have heart disease or not using machine-learning algorithm. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not.

1.1 Motivation

According to the World Health Organization more than 10 million die due to Heart diseases every single year around the world. Heart disease is the second leading cause of death and a major cause of disability worldwide. The main challenge in today's healthcare is provision of best quality services and effective accurate diagnosis. Even if heart diseases are found as the prime source of death in the world in recent years, they are also the ones that can be controlled and managed effectively. We can use different machine learning models to diagnose the disease and classify or predict the results. Such information, if predicted well in advance, can provide important insights to doctors who can then adapt their diagnosis and treatment per patient basis. This is the main motivation behind this work. The proposed work makes an attempt to detect these heart diseases at early stage with good precision to avoid disastrous consequences.

1.2 Problem Statement

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either it are expensive or are not efficient to calculate chance of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients everyday in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. Since we have a good amount of data in today's world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medical data.

1.3 Methodology

The working of the system starts with the collection of data and selecting the important attributes. The data is then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data. This system is implemented using the following modules.

- 1.) Collection of Dataset
- 2.) Selection of attributes
- 3.) Balancing of Data
- 4.) Disease Prediction

1.3.1 Collection of dataset

Initially, we collect a dataset for our heart disease prediction system. After the collection of the dataset, we split the dataset into training data and testing data. The training dataset is used for prediction model learning and testing data is used for evaluating the prediction model. For this project, 80 of training data is used and 20 of data is used for testing. The dataset used for this project is Heart Disease UCI. The dataset consists of 12 attributes.

1.3.2 Selection of attributes

Attributes selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system. Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc are selected for the prediction.

1.3.3 Training and Testing of Model

for this section, we have used Sklearn module of python, which is used for training and testing of data from the dataset. The parameters used for training and testing part were test size , which is important factor which decides the size of the testing parameters from the dataset.

1.3.4 Prediction of Disease

Various machine learning algorithms like Support Vector Machine, Naive Bayes, Decision Tree, Random Tree, Logistic Regression are used for classification. Comparative analysis is performed among algorithms and the algorithm that gives the highest accuracy is used for heart disease prediction.

Further we have build a basic web application to help doctors in diagnosing heart disease by using 12 inputs for 12 features. When the user enters input data, and clicks the submit button this application predicts chances of heart disease.

1.3.5 Data Flow Model

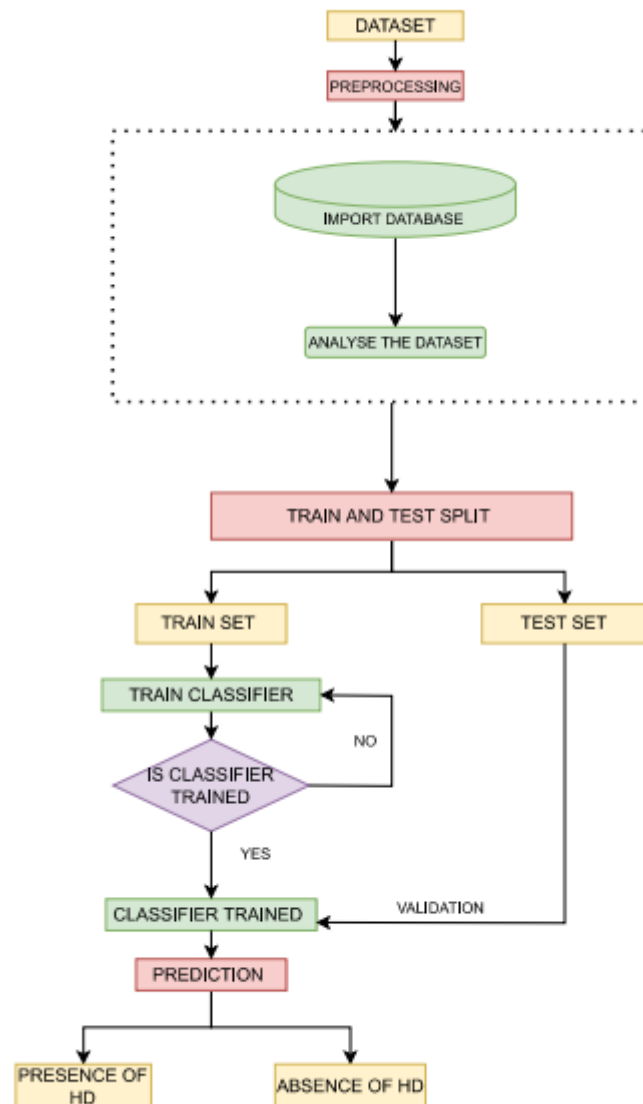


Figure 1.1: Flow Chart of the project

1.4 Organization of Project Report

This project report is organized as follows:

- **Chapter 1** presents the Introduction, Motivation, Methodology of the proposed Heart Disease Prediction Model.
- **Chapter 2** presents the literature survey on the existing work that has already been done in the field of Machine Learning.
- **Chapter 3** provides a brief explanation of Software and Hardware support that is provided to the prediction model.
- **Chapter 4** is dedicated towards detailed explanation about dataset along with the flowchart that explains the flow of the prediction model along with the machine learning models that are used.
- **Chapter 5** shows analysis and discussions of experimental results.
- **Chapter 6** presents the conclusions and future scope for this project.

Chapter 2

Literature Review

Recursion Enhanced Random Forest With an Improved Linear Model (RERF-ILM):

- In this paper[1], the proposed excursion Enhance Random Forest with an improved linear model (RFRF-ILM) to predict heart diseases[1].
- In this work[2], information for several cardiac patients, classification algorithms, has been accumulate to predict the cardiac disorder of the person concerned[2].
- Dataset clustering is based on Decision Tree (DT) feature variables and criteria[?]. To estimate its Sensitivity and Specificity, the classifier is then applied to each data set with the rhythm classification graphical structure[2].
- The proposed RFRF-ILM method is utilized merging the features of the linear model and random forest[3]. RFRF-ILM achieves high accuracy in the prediction of heart disease. The support vector machine is utilized to enhance the performance of the algorithm[3].

- Results from the Effective Heart Disease Prediction classification using the RERF-ILM method show that the accuracy of heart disease is more precise than with other methods (GC, DLT, EHDPS, FCBF)[4]. The compared classification results of our approach with existing GC, DLT, EHDPS, and FCBF methods used for data sets to determine the performance of the proposed approach with more accuracy[4].
- Naive Bayes's model of learning employs Bayes rules utilizing separate features as analyzed by the naive Bayes model[3]. Each R instance is assigned the highest probability class subsequently with high performance[3].
- From the experimental results, this indicates that coronary artery disease develops more often in older ages and also important in this disease's outbreak is high blood pressure provides a further aspect that should be taken into consideration in the occurrence of coronary artery disease[2].

HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System.

- The proposed HDPM was applied to both datasets and showed positive results for increasing the prediction accuracy as compared to other models[5]. We selected six state-of-the-art MLAs (NB, LR, MLP, SVM, DT, and RF) that have been widely used in the research community and have a proven track record for accuracy and efficiency for comparison[5]. We performed 10-fold cross-validation for all models and collected eight performance metrics: accuracy (acc), precision (pre), recall/sensitivity/true positive rate (rec/sec/TPR), f measure (f), MCC, false positive rate (FPR), false negative rate (FNR), and true negative rate (TNR)[5]. The findings revealed that the proposed model outperformed other models by achieving acc, pre, rec/sec, f up to 95.90%, 97.14%, 94.67%, 95.35% for dataset I and 98.40%, 98.57%, 98.33%, 98.32% for dataset II, respectively. In term of MCC, the proposed HDPM achieved the highest MCC value up to 0.92 and 0.97 for datasets I and II, respectively, which confirms the superiority of our proposed model relative to other models[5].
- We further investigated the performance of the proposed HDPM using a receiver operating characteristic (ROC) curve visualization[6]. The ROC curve consists of the TP rate as the y-axis and FP rate as the x-axis with the area under the ROC curve (AUC) being calculated to show the performance of the model[6]. The best model is achieved when the value of AUC is close or equal to 1[6].
- We proposed an effective heart disease prediction model (HDPM) for heart disease diagnosis by integrating DBSCAN, SMOTE-ENN, and XGBoost-based MLA to improve prediction accuracy[7]. The DBSCAN was applied to detect and remove the outlier data, SMOTE-ENN was used to balance the unbalanced training dataset and XGBoost MLA was adopted to learn and generate the prediction model[7][8].

- The experimental results confirmed that the proposed model achieved better performance than that of state-of-the-art models and previous study results, by achieving an accuracy up to 95.90% and 98.40% for datasets I and II, respectively[8]. In addition, the statistical-based analysis result also showed the significant improvement for the proposed model as compared with the other models[8].

Novel Feature Reduction (NFR) Model With Machine Learning and Data Mining Algorithms for Effective Disease Risk Prediction.

- The NFR model employs heart datasets of Cleveland, Hungarian, Statlog, and Switzerland taken from the ML Repository of UCI[9]. The dataset consists of 76 raw attributes/features. Of these 14 features, eight features are categorical and six features are numeric[9].
- The objective of this research article is to present a Novel Feature Reduction (NFR) model that improves the accuracy and AUC by reducing the number of features without omitting relevant features[9]. The proposed NFR model is aligned with five ML classification algorithms: logistic regression (LR), support vector machine (SVM), boosted regression tree (BRT), stochastic gradient boosting (SGB), and random forest (RF) for datasets of healthcare to enhance the diagnosis capabilities by creating smaller feature subsets from a higher number of features[9].

The proposed model comprises of two approaches, where

- i) the first approach is based on a heuristic process evaluating the prediction performance by reducing features and the improvement in the AUC simultaneously with the accuracy as evaluation metrics, in order to acquire the best subset of highly contributing features[9].
 - ii) The second approach evaluates the accuracy and AUC of all individual features attained on aligning with ML classifiers and forms the subsets with the highest accuracies, AUCs and least difference between them[10]. These subsets are then combined in various combinations to achieve the best reduced set of highly contributing features in the disease risk prediction[10].
- Compared with several existing studies, the proposed model achieves better performance. On applying the two approaches of the NFR model to the four datasets, the Switzerland heart dataset yields the best results with a maximum accuracy of 95.52% and a maximum AUC of 99.2% was achieved using the BRT

algorithm with 41.67% feature reduction. A 25% of performance improvement was achieved in the run time of the algorithm[10].

Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques.

- Proposed hybrid HRFLM approach is used combining the characteristics of Random Forest (RF) and Linear Method (LM)[11]. HRFLM proved to be quite accurate in the prediction of heart disease[11]. The future course of this research can be performed with diverse mixtures of machine learning techniques to better prediction techniques[11]. Furthermore, new feature selection methods can be developed to get a broader perception of the significant features to increase the performance of heart disease prediction[11].
- The prediction models are developed using 13 features and the accuracy is calculated for modeling techniques. This compares the accuracy, classification error, precision, F-measure, sensitivity and specificity[11]. The highest accuracy is achieved by HRFLM classification method in comparison with existing methods[11].
- Identifying the processing of raw healthcare data of heart information will help in the long term saving of human lives and early detection of abnormalities in heart conditions[12]. Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards heart disease[12].

Chapter 3

Software Hardware Support

3.1 Hardware requirements:

Processor : intel i3 and above

Ram : Min 4GB

Hard Disk : Min 100GB

3.2 Software requirements:

Operating System : Windows family

Technology : Python 3.9

IDE : Jupyter notebook

App Framework : Streamlit

3.3 Python:

Python is a programming language that is preferred for programming due to its vast features, applicability, and simplicity. The Python programming language best fits machine learning due to its independent platform and its popularity in the programming community.

Python is a programming language that distinguishes itself from other programming languages by its flexibility, simplicity, and reliable tools required to create modern software.

Python is consistent and is anchored on simplicity, which makes it most appropriate for machine learning. The Python programming language best fits machine learning due to its independent platform and its popularity in the programming community.

Advantages:

- Presence of third-party modules
- Extensive support libraries (NumPy for numerical calculations, Pandas for data analytics etc.)
- Open source and community development
- Versatile, Easy to read, learn and write
- User-friendly data structures
- Object-oriented language
- Portable and Interactive

3.3.1 Python Libraries:

NumPy:

NumPy is a very popular python library for large multi-dimensional array and matrix processing, with the help of a large collection of high-level mathematical functions. It is very useful for fundamental scientific computations in Machine Learning.

Scikit-learn:

Scikit-learn is one of the most popular ML libraries for classical ML algorithms. It is built on top of two basic Python libraries, viz., NumPy and SciPy. Scikit-learn supports most of the supervised and unsupervised learning algorithms. Scikit-learn can also be used for data-mining and data-analysis, which makes it a great tool who is starting out with Machine Learning.

Seaborn

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. Seaborn is a library in Python predominantly used for making statistical graphics. Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python. Visualization is the central part of Seaborn which helps in exploration and understanding of data.

Pandas:

Pandas is a popular Python library for data analysis. It is not directly related to Machine Learning. As we know that the dataset must be prepared before training. In this case, Pandas comes handy as it was developed specifically for data extraction and preparation. It provides high-level data structures and wide variety tools for data analysis. It provides many inbuilt methods for groping, combining and filtering data.

Matplotlib:

Matplotlib is a very popular Python library for data visualization. Like Pandas, it is not directly related to Machine Learning. It particularly comes in handy when a programmer wants to visualize the patterns in the data. It is a 2D plotting library used for creating 2D graphs and plots. A module named pyplot makes it easy for programmers for plotting as it provides features to control line styles, font properties, formatting axes, etc. It provides various kinds of graphs and plots for data visualization, viz., histogram, error charts, bar chats, etc.

Streamlit:

Streamlit is a free, open-source, all-python framework that enables data scientists to quickly build interactive dashboards and machine learning web apps with no front-end web development experience required. If we know python, then we are all equipped to use Streamlit to create and share our web apps, within hours, not weeks.

3.3.2 Jupyter notebook:

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. Its uses include data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more. It allows the user to construct the content in a mixture of Markdown, an extended version of Markdown called MyST, Maths Equations using MathJax, Jupyter Notebooks, reStructuredText, the output of running Jupyter Notebooks at build time. Multiple output formats can be produced (currently single files, multipage HTML web pages and PDF files).

Installation and Execution:

The best method of installing the Jupyter Notebooks is by the installation of the Anaconda package. The Jupyter Notebook and the Jupyter Lab comes pre-installed in the Anaconda package, and you don't have to install this on your own.

One of the requirements here is Python, either Python v3.7 or greater. Here simply install the Jupyter Notebook the Pythonic way, then proceed to issue the following command i.e. `pip install jupyter` in the command prompt or PowerShell.

With this installation of the Jupyter Notebook is completed, you should be able to access it successfully. Just type the `jupyter notebook` command in the command prompt or PowerShell. This should launch your Jupyter Notebook in your selected web browser with a local hosting URL.

Chapter 4

Heart Disease Prediction Using Machine Learning Techniques

4.1 DATA SET:

The dataset used was the Heart disease Dataset which is taken from Heart Statlog Cleaveland Hungary Dataset of UCI repository. In this Dataset there are 1190 instances and 12 attributes. It consists of attributes mainly named as age, sex, chest pain type, resting bps, cholesterol, fasting blood sugar, resting ecg, max heart rate, exercise angina, oldpeak, ST slope and target. The shape of the dataset is (1190 x 12). We have used the already processed UCI Heart Statlog Cleaveland Hungary Dataset available in the Kaggle website for our analysis. The complete description of the 14 attributes used in the proposed work is mentioned in Table 4.1 shown below.

Feature Name	Feature Code	Description
Age	AGE	Age in years
Sex	SEX	Male=1,Female=0
Chest pain	CPT	Atypical angina=1 Typical angina=2 Asymptotic=3 Non-anginal pain=4
Resting blood pressure	RBP	mm hg, hospitalized
serum cholesterol	SCH	In mg/dl
fasting blood sugar>120 mg/dl	FBS	fasting blood sugar>120 mg/dl (T=1) (F=0)
resting electrographic	RES	Normal=0 ST T=1 Hypertrophy=2
Maximum Heart Rate	MHR	-
exercise induced angina	EIA	yes=1 no=0
old peak=ST depression induced by exercise relative to rest	OPK	-
The slope of peak exercise ST segment	PES	Up sloping=1 Flat=2 Down sloping=3
Target	Target	negative=0, positive=1

Table 4.1: Attributes of Dataset

4.2 MACHINE LEARNING ALORITHMS:

4.2.1 Support Vector Machine (SVM):

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems[21]. However, primarily, it is used for Classification problems in Machine Learning[21]. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future[21]. This best decision boundary is called a hyperplane.SVM chooses the extreme points/vectors that help in creating the hyperplane[21]. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine[21].

Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression[21]. But generally, they are used in classification problems[21]. In the 1960s, SVMs were first introduced but later they got refined in 1990. SVMs have their unique way of implementation as compared to other machine learning algorithms[21]. Lately,

they are extremely popular because of their ability to handle multiple continuous and categorical variables.

The following are important concepts in SVM -

Support Vectors - Data Points that are closest to the hyperplane are called support vectors. Separating line will be defined with the help of these data points[21].

Hyperplane - As we can see in the above diagram, it is a decision plane or space which is divided between a set of objects having different classes[21].

Margin - It may be defined as the gap between two lines on the closest data points of different classes[21]. It can be calculated as the perpendicular distance from the line to the support vectors[21]. Large margin is considered as a good margin and small margin is considered as a bad margin[21].

Types of SVM:

SVM can be of two types:

- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier[21].
- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier[21]. The objective of the support vector machine algorithm is to find a hyperplane in an N dimensional space (N - the number of features) that distinctly classifies the data points[21].

The advantages of support vector machines are:

- Effective in high dimensional spaces[21].
- Still effective in cases where the number of dimensions is greater than the number of samples[21].
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient[21].

- Versatile: different kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels[21].

The disadvantages of support vector machines are:

- If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial[21]. SVM do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation[21].

4.2.2 Logistic Regression:

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique[18]. It is used for predicting the categorical dependent variable using a given set of independent variables[18].

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value[18]. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1[18].

Logistic Regression is much similar to the Linear Regression except that how they are used[18]. Linear Regression is used for solving Regression problems, whereas logistic regression is used for solving the classification problems[18].

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1)[18]. The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc[18].

Advantages:

Logistic Regression is one of the simplest machine learning algorithms and is easy

to implement yet provides great training efficiency in some cases[18]. Also due to these reasons, training a model with this algorithm doesn't require high computation power[18].

The predicted parameters (trained weights) give inference about the importance of each feature. The direction of association i.e. positive or negative is also given[18]. So we can use Logistic Regression to find out the relationship between the features[18].

This algorithm allows models to be updated easily to reflect new data, unlike Decision Tree or Support Vector Machine. The update can be done using stochastic gradient descent[18].

Logistic Regression outputs well-calibrated probabilities along with classification results[18]. This is an advantage over models that only give the final classification as results[18]. If a training example has a 95% probability for a class, and another has a 55% probability for the same class, we get an inference about which training examples are more accurate for the formulated problem[18].

Disadvantages:

Logistic Regression is a statistical analysis model that attempts to predict precise probabilistic outcomes based on independent features[18]. On high dimensional datasets, this may lead to the model being over-fit on the training set, which means overstating the accuracy of predictions on the training set and thus the model may not be able to predict accurate results on the test set[18].

This usually happens in the case when the model is trained on little training data with lots of features[18]. So on high dimensional datasets, Regularization techniques should be considered to avoid overfitting (but this makes the model complex)[18]. Very high regularization factors may even lead to the model being under-fit on the training data[18].

Non linear problems can't be solved with logistic regression since it has a linear decision surface[18]. Linearly separable data is rarely found in real world scenarios. So the transformation of non linear features is required which can be done by increasing the number of features such that the data becomes linearly separable in higher dimensions[18].

4.2.3 Decision Trees:

A decision tree is a decision support tool that uses a tree-like model of decision making process and the possible consequences[15]. It covers event outcomes, resource costs, and utility of decisions[15]. Decision Trees resemble an algorithm or a flowchart that contains only conditional control statements[15].

A decision tree is drawn upside down with the root node at top[15]. Each decision tree has 3 key parts: a root node, leaf nodes, branches[15]. In a decision tree, each internal node represents a test or an event. Say, a heads or a tail in a coin flip[15]. Each branch represents the outcome of the test and each leaf node represents a class label -a decision taken after computing all attributes. The paths from root to leaf nodes represent the classification rules[15].

Decision trees can be a powerful machine learning algorithm for classification and regression[15]. Classification tree works on the target to classify if it was a heads or a tail[15]. Regression trees are represented in a similar manner, but they predict continuous values like house prices in a neighborhood[15].

The best part about decision trees:

1. Handle both numerical and categorical data[15]
2. Handle multi-output problems
3. Decision trees require relatively less effort in data preparation[15]

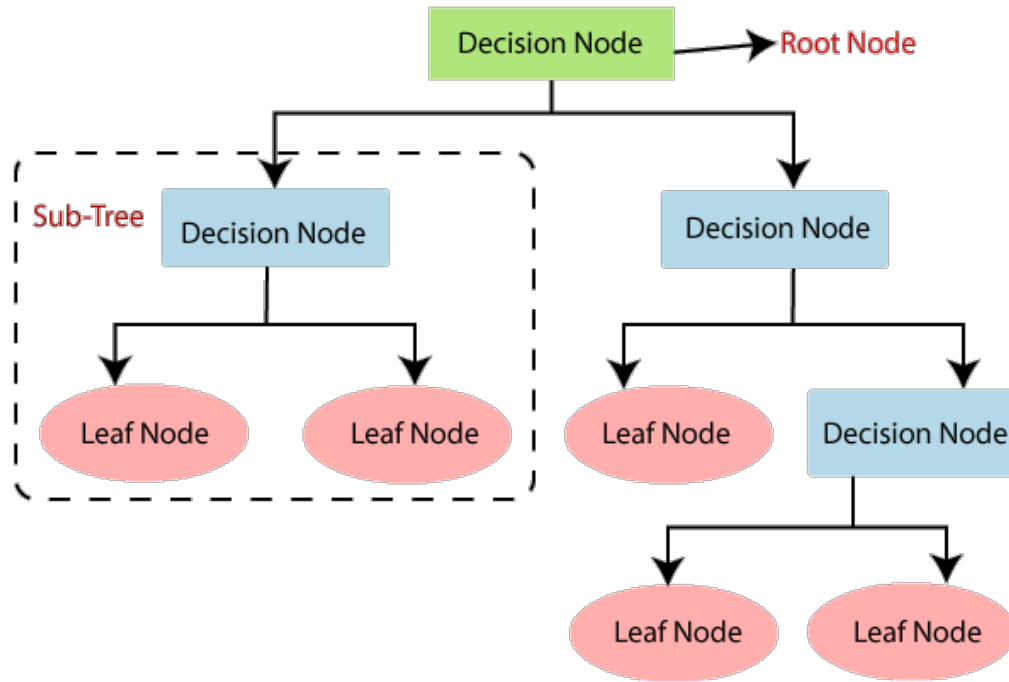


Figure 4.1: Decision tree Flowchart[15]

4. Nonlinear relationships between parameters do not affect tree performance[15]

4.2.4 Random Forest:

Random Forest is a supervised learning algorithm. It is an extension of machine learning classifiers which include the bagging to improve the performance of Decision Tree[14]. It combines tree predictors, and trees are dependent on a random vector which is independently sampled[14]. The distribution of all trees are the same[14].

Random Forests splits nodes using the best among of a predictor subset that are randomly chosen from the node itself, instead of splitting nodes based on the variables. [14].

It can be used both for classification and regression. It is also the most flexible and easy to use algorithm[14]. A forest consists of trees. It is said that the more trees it has, the more robust a forest is. Random Forests create Decision Trees on

randomly selected data samples, get predictions from each tree and select the best solution by means of voting. It also provides a pretty good indicator of the feature importance[14].

Random Forests have a variety of applications, such as recommendation engines, image classification and feature selection. It can be used to classify loyal loan applicants, identify fraudulent activity and predict diseases[14]. It lies at the base of the Boruta algorithm, which selects important features in a dataset[14].

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique[14]. It can be used for both Classification and Regression problems in ML[14]. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model[14].

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset[14]." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output[14].

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting[14].

Assumptions: Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not[14]. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result[14].
- The predictions from each tree must have very low correlations[14].

Random Forest Classifier

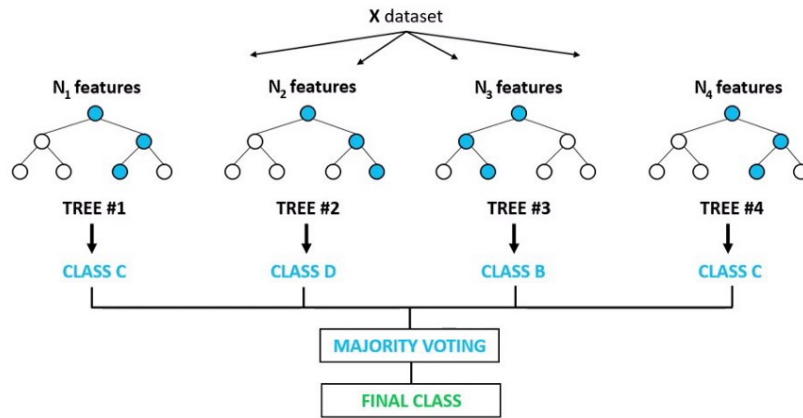


Figure 4.2: Random forest working[14]

Advantages:

- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality[14].
- It enhances the accuracy of the model and prevents the overfitting issue[14].

4.2.5 Naive Bayes:

Naive Bayes algorithm could be a supervised learning algorithm, which relies on Bayes theorem and used for solving classification problems. It is mainly utilized in text classification that has a high-dimensional training dataset[20].

Naive Bayes Classifier is one among the easy and handiest Classification algorithms which helps in building the fast machine learning models which will make quick predictions[20].

It is a probabilistic classifier, which suggests it predicts on the concept of the probability of an object[20]. Some popular samples of Naive Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles[20]. It is a

classification technique supported Bayes Theorem with an assumption of independence among predictors[20].

In simple terms, a Naive Bayes classifier assumes that the presence of a specific feature in an exceedingly class may be unrelated to the presence of any other feature[20].

The Naive Bayes model is easy to build and particularly useful for very large data sets[20]. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods[20].

The Naive Bayes algorithm is comprised of two words Naive and Bayes, Which can be described as:

- **Naive:** It is called Naive because it assumes that the occurrence of a certain feature is independent of the occurrence of other features[20]. Such as if the fruit is identified on the basis of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other[20].
- **Bayes:**
It is called Bayes because it depends on the principle of Bayes' Theorem[20].

Bayes theorem:

Bayes theorem is also known as Bayes Rule or Bayes law, which is used to determine the probability of a hypothesis with prior knowledge[20]. It depends on the conditional probability. The Gaussian model assumes that features follow a normal distribution[20]. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution[20].

4.3 VOTING METHOD:

For voting of libraries , we accessed any random patient's from the dataset and predicted that data using all the models which further predicted whether the person has heart disease or not on basis of all the parameters in the dataset. Trained models which were used were mainly Support Vector Machines (SVM), Naive Bayes (NB), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT). All the trained models predicted the Heart Disease of the particular patient with best accuracy.

4.4 USER INTERFACE:

For having a good interactive interface, we have used Streamlit Application where using python one can build interactive Interfaces. Using one such platform , we have created an interface. In this Application, user is supposed to enter his/her health and other factors such as

- **Name:** Name to be entered in Charecters
- **Age:** Age to entered in numbers / integers
- **Gender:** Choose 0: Female, 1: Male
- **Chest Pain Type:** To Be Entered in Selection Format 0, 1, 2, 3, 4
- **Blood Pressure:** Blood Pressure to be entered in numbers
- **Cholestrol:** Enter Inputs in numbers / integers
- **Sugar:** Choose between 0: No Sugar, 1: Sugar
- **ECG Normality whether Normal and Abnormal ECG:** Choose between 0: Normal, 1: AbNormal
- **Maximum Heart Rate Achieved:** Maximum Heart Rate to be entered in numbers / integers

- **Angina Details (Whether person suffers from Oxygen Problems while performing certain exercise):** 0: No, 1: Yes
- **ECG Levels:** ST Depression to be entered in float/ decimal inputs
- **Peak Exercise levels:** Choose between 0: Upslopping, 1: Flat, 2: Downslopping

Chapter 5

ANALYSIS AND DISCUSSIONS OF EXPERIMENTAL RESULTS

Algorithm	Training Accuracy	Testing Accuracy
Decision Tree	100%	92.8%
Logistic Regression	82.56%	89.09%
Support Vector Machine	70.8%	73.53%
Naive Bayes	84.03%	82.77%
Random Forest	100%	92.86%

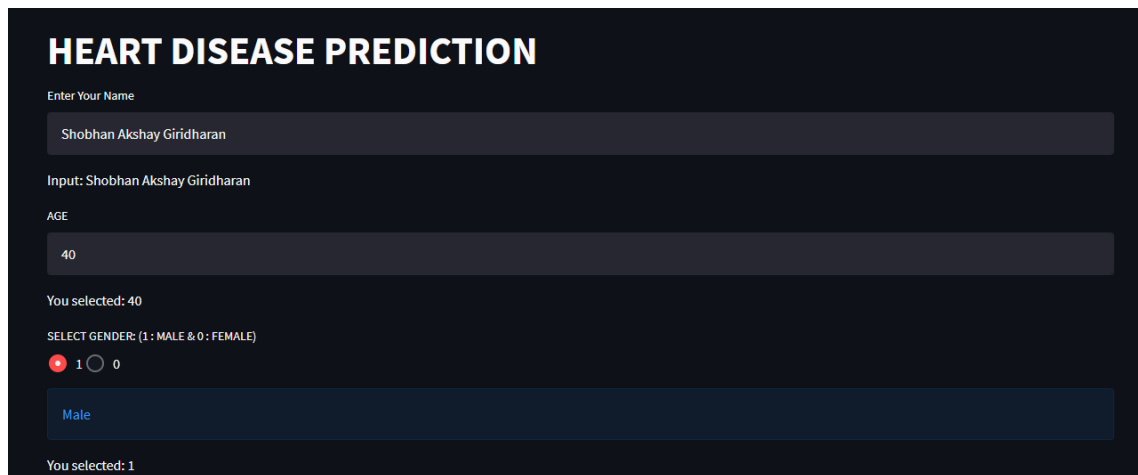
Table 5.1: Model Wise Accuracy

Table 4.2 shows the accuracy achieved from by training and testing the 5 proposed models.

Algorithm	True positive	False positive	False Negative	True Negative
Decision Tree	85	27	36	90
Logistic Regression	97	97	97	97
Support Vector Machine	85	27	36	90
Naive Bayes	85	27	36	90
Random Forest	85	27	36	90

Table 5.2: Model Wise Confusion Matrix Report

Table 4.3 shows the confusion matrix report achieved from the models which depicts the truthfulness of the classification report achieved from each model through Figure 5.5 which is confusion matrix.



HEART DISEASE PREDICTION

Enter Your Name

Shobhan Akshay Giridharan

Input: Shobhan Akshay Giridharan

AGE

40

You selected: 40

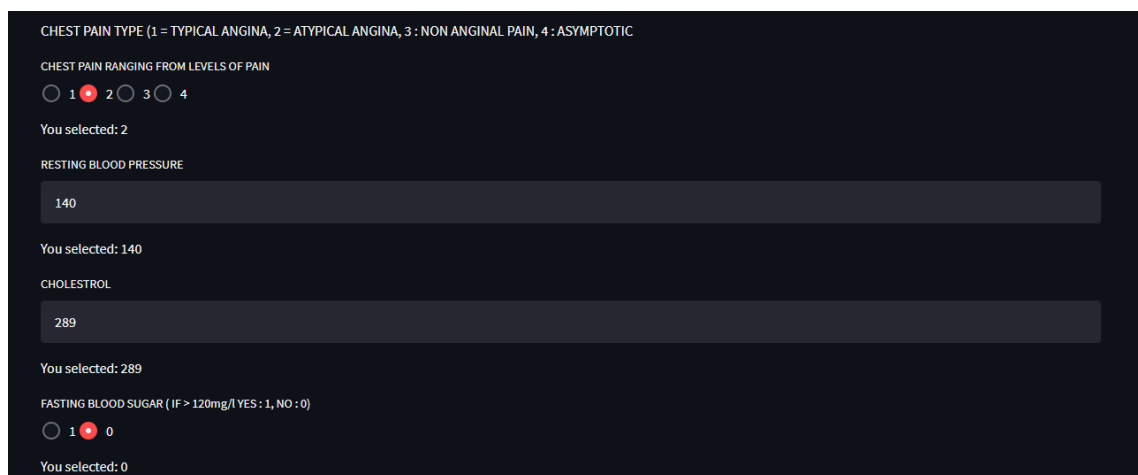
SELECT GENDER: (1 : MALE & 0 : FEMALE)

☒ 1 ☐ 0

Male

You selected: 1

Figure 5.1: Streamlit output 1



CHEST PAIN TYPE (1 = TYPICAL ANGINA, 2 = ATYPICAL ANGINA, 3 : NON ANGINAL PAIN, 4 : ASYMPTOTIC)

CHEST PAIN RANGING FROM LEVELS OF PAIN

☐ 1 ☒ 2 ☐ 3 ☐ 4

You selected: 2

RESTING BLOOD PRESSURE

140

You selected: 140

CHOLESTROL

289

You selected: 289

FASTING BLOOD SUGAR (IF > 120mg/l YES : 1, NO : 0)

☐ 1 ☒ 0

You selected: 0

Figure 5.2: Streamlit output 2

Figure 5.1, 5.2, 5.3, 5.4 shows us the snippets of the user interface based predictor made on streamlit, which will enable the user to enter his attribute values which are asked. Then after entering the values the software will predict whether the user has heart disease or not.

RESTING ECG(0 : NORMAL, 1 : ABNORMAL)

☒ 0 ☐ 1

You selected: 0

MAXIMUM HEART RATE ACHIEVED

172

You selected: 172

DO YOU GET OXYGEN PROBLEM DURING EXERCISE, IF YES? PROCEED ELSE SELECT NO

EXERCISE INDUCED ANGINA (1 : YES, 0 : NO)

☐ 1 ☒ 0

You selected: 0

ECG LEVELS ON PLOT

ST DEPRESSION INDUCED BY EXERCISE REALTIVE TO REST / OLDPEAK

0.0

You selected: 0.0

[Manage app](#)

Figure 5.3: Streamlit output 3

PEAK EXERCISE ST SEGMENT (0 : UPSLOPPING, 1 : FLAT, 2 : DOWNSLOPPING)

1

You selected: 1

PLEASE REVIEW YOUR FILLED DATA

	Name	Age	Sex	Chest Pain Type	Resting Blood Pressure	Cholestrol	Fasting Blood Sugar	Resting ECG	Max Heart R
values	Shobhan Akshay Giridharan	40	1	2	140	289	0	0	172

PREDICT

This person has less chance of heart attack

Figure 5.4: Streamlit output 4

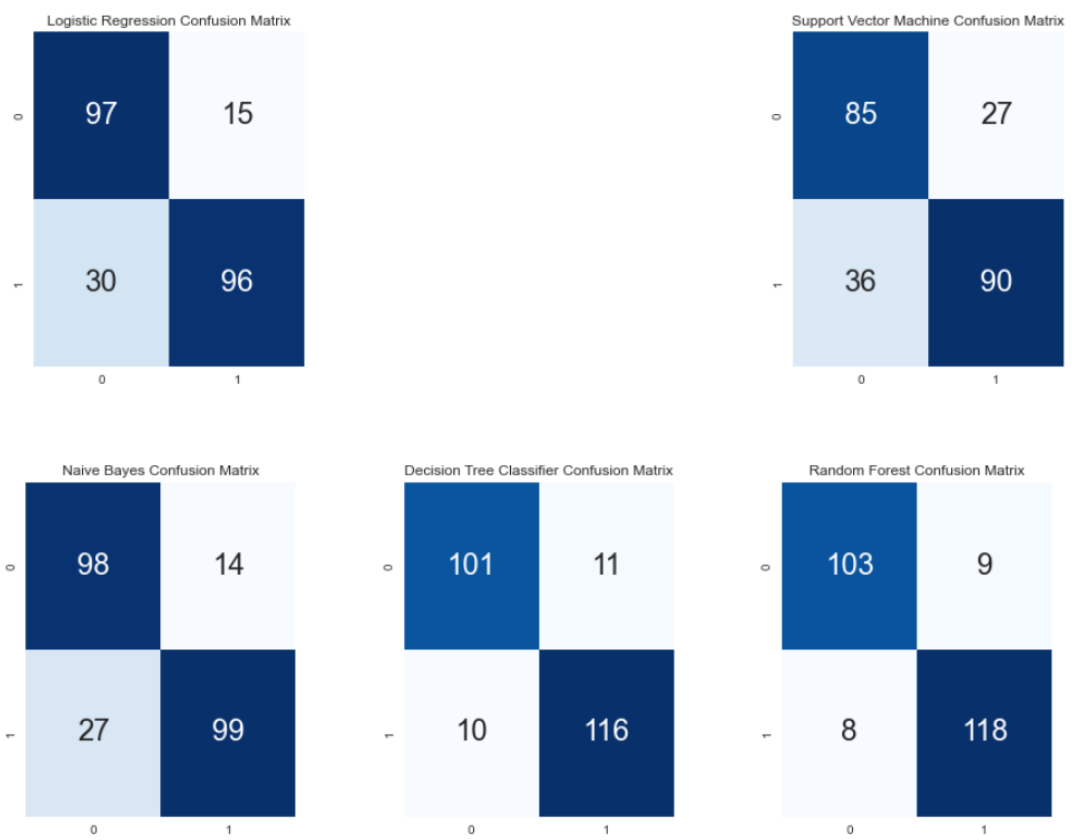


Figure 5.5: Confusion matrix

Chapter 6

Conclusion and Future work

With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. The result of this study indicates that the Random Forest Algorithm is the most efficient algorithm with accuracy score of 92.86% for prediction of heart disease. A web user interface based on the Random Forest algorithm is also developed for the user's simplicity to check and predict the accuracy of presence of Heart Disease in a person.

Future scope :

To make the study more useful and effective the following suggestion have been proposed for further improvements in this area:

- To develop improved algorithms to reduce the level of failure.
- To work on optimization of the code, so that the software can run in real time applications.
- To update the User Interface according to the latest conditions which will affect the heart disease prediction.

Bibliography

- [1] M. Kavousi, S. Elias-Smale, J. H. W. Rutten, M. J. G. Leening, R. Vliegenthart, G. C. Verwoert, G. P. Krestin, M. Oudkerk, M. P. M. de Maat, F. W. G. Leebeek, F. U. S. Mattace-Raso, J. Lindemans, A. Hofman, E. W. Steyerberg, A. van der Lugt, A. H. van den Meiracker, and J. C. M. Witteman, *Evaluation of newer risk markers for coronary heart disease risk classification: A cohort study*, " *Ann. Internal Med.*, vol. 156, no. 6, pp. 438-444, Mar. 2012.
- [2] H. Tada, O. Melander, J. Z. Louie, J. J. Catanese, C. M. Rowland, J. J. Devlin, S. Kathiresan, and D. Shiffman, *Risk prediction by genetic risk scores for coronary heart disease is independent of self-reported family history*, *Eur. Heart J.*, vol. 37, no. 6, pp. 561-567, Feb. 2016.
- [3] A. Junejo, Y. Shen, A. A. Laghari, X. Zhang, and H. Luo, Molecular diagnostic and using deep learning techniques for predict functional recovery of patients treated with cardiovascular disease, *IEEE Access*, vol. 7, pp. 120315-120325, 2019.
- [4] C. Krittanawong, H. Zhang, Z. Wang, M. Aydar, and T. Kitai, Artificial intelligence in precision cardiovascular medicine, *J. Amer. College Cardiol.*, vol. 69, no. 21, pp. 2657-2664, 2017.
- [5] P. Gomathi, S. Baskar, and P. M. Shakeel, Identifying brain abnormalities from electroencephalogram using evolutionary gravitational neocognitron neural network, *Multimedia Tools Appl.*, vol. 2019, pp. 1-20, Feb. 2019, doi: 10.1007/s11042-019.
- [6] Zhang W, Cheng B, Lin Y Driver drowsiness recognition based on computer vision technology. *Tsinghua Sci Technology* 17(3):354-362, 2012

- [7] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, D. S. Rajput, R. Kaluri, and G. Srivastava, Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis, *Evol. Intell.*, vol. 13, no. 2, pp. 185-196, Jun. 2020
- [8] B. K. Sarkar, Hybrid model for prediction of heart disease, *Soft Comput.*, vol. 24, no. 3, pp. 1903-1925, Feb. 2020.
- [9] M. Tanveer, C. Gautam, and P. N. Suganthan, Comprehensive evaluation of twin SVM based classifiers on UCI datasets, *Appl. Soft Comput.*, vol. 83, Oct. 2019, Art. no. 105617.
- [10] B. R. Kirkwood, J. A. C. Sterne, and B. R. Kirkwood, *Essential Medical Statistics*, 2nd ed. Malden, MA, USA: Blackwell Science, 2003.
- [11] M. Xu, D. Fralick, J. Z. Zheng, B. Wang, X. M. Tu, and C. Feng, The differences and similarities between two-sample T-test and paired T-test, *Shanghai Arch, Psychiatry*, vol. 29, no. 3, pp. 184-188, Jun. 2017
- [12] A. K. Dwivedi, Performance evaluation of different machine learning techniques for prediction of heart disease, *Neural Comput. Appl.*, vol. 29, no. 10, pp. 685-693.
- [13] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in *Proc. 14th Int. Joint Conf. Artif. Intell. (IJCAI)*, Montreal, QC, Canada, vol. 2, Aug. 1995, pp. 1137-1145.
- [14] <https://medium.com/analytics-vidhya/random-forest-classifier-and-its-hyperparameters-8467bec755f6>
- [15] <https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm>
- [16] https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1
- [17] <https://stackoverflow.com/questions/39120942/difference-between-standardscaler-and-normalizer-in-sklearn-preprocessing>

- [18] <https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>
- [19] <https://www.javatpoint.com/logistic-regression-in-machine-learning>
- [20] <https://www.javatpoint.com/machine-learning-naive-bayes-classifier>
- [21] <https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>

Appendix-I : Timeline of the project

TIMELINE CHART FOR SEMESTER VII																	
MONTH	JULY				AUGUST					SEPTEMBER				OCTOBER			
WEEK NO.	W1	W2	W3	W4	W1	W2	W3	W4	W5	W1	W2	W3	W4	W1	W2	W3	W4
WORK TASKS																	
1.PROBLEM DEFINITION																	
Search for topics																	
Identify the goal of the project																	
2.PREPARATION																	
Search for the IEEE Access Research papers																	
Study of related IEEE papers and Methods																	
Study of Machine Learning Models																	
Study of WEKA Tool																	
3.PLANNING																	
Dataset Analysis																	
Selection of Feature Selection Methods																	
Introduce New Methodology																	
4.EXECUTION OF THE PROJECT																	
Accuracy of Model																	

Figure 6.1: Timeline of project (SEM VII)

TIMELINE CHART FOR SEMESTER VIII																
MONTH	JANUARY				FEBRUARY				MARCH				APRIL			
WEEK NO	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4	W1	W2	W3	W4
WORK TASK																
Execution of Project																
OUTPUT AND RESULTS																
User Interference																
POSTER MAKING AND BLACK BOOK																

Figure 6.2: Timeline of project (SEM VIII)