

HEART DISEASE DIAGNOSIS AND PREDICTION USING MACHINE LEARNING

Submitted in partial fulfillment of the requirements for the award of Bachelor of
Engineering degree in Computer Science and Engineering

By

GULIVINDALA JAGADEESH (37110250)



**DEPARTMENT OF COMPUTER SCIENCE AND
ENGINEERING SCHOOL OF COMPUTING**

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY (DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC | 12B Status by UGC | Approved by
AICTE

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600

119

March - 2021

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with “A” grade by NAAC | 12B Status by UGC | Approved by AICTE

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of Mr_ **GULIVINDALA JAGADEESH (37110250), GOUTHU YUVA SRI KIRAN (37110241)** who carried out the project entitled " **HEART DISEASE DIAGNOSIS AND PREDICTION USING MACHINE LEARNING**" under my supervision from November 2020 to March 2021.

Dr.S.Prince Mary,M.E.,Ph.D.,

Internal Guide

Head of the Department

Submitted for Viva voce Examination held on_____

Internal Examiner

External Examiner

DECLARATION

We **GULIVINDALA JAGADEESH (37110250) AND GOUTHU YUVA SRI KIRAN (37110241)** hereby declare that the Project Report entitled **HEART DISEASE DIAGNOSIS AND PREDICTION USING MACHINE LEARNING** done by me under the guidance of **Dr.S.Prince Mary,M.E.,Ph.D.,** (Internal) and----- (External) at **SATHYABAM A INSTITUTE OF SCIENCE AND TECHNOLOGY** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering / Technology degree In **CSE**

DATE:

PLACE:

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAM A** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E., Ph.D.**, Dean, School of Computing and **Dr.L.Lakshmanan M.E., Ph.D.,and Dr.S.Vigneshwari M.E., Ph.D.**, Heads of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr.S.Prince Mary,M.E.,(Ph.D)** for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project

ABSTRACT

A popular saying goes that we are living in an “information age”. Terabytes of data are produced every day. Machine learning is the process which turns a collection of data into knowledge. The health care industry generates a huge amount of data daily. However, most of it is not effectively used. Efficient tools to extract knowledge from these databases for clinical detection of diseases or other purposes are not much prevalent. The aim of this paper is to summarize some of the current research on predicting heart diseases using machine learning techniques, analyse the various combinations of mining algorithms used and conclude which technique(s) are effective and efficient. Also, some future directions on prediction systems have been addressed. we used svm & random forest algorithm.

LIST OF CONTENTS

CHAPTER NO	TITLE	PAGE NO
	ABSTRACT	i
	LIST OF FIGURES	
	LIST OF TABLES	
1	INTRODUCTION	
	1.1 TYPES OF CARDIOVASCULAR DISEASES	1
	1.2 CHEST PAIN	2
	1.3 NAUSEA, INDIGESTION, HEARTBURN AND STOMACH PAIN	2
	1.4 PAIN IN THE ARMS	2
	1.5 FEELING DIZZY AND LIGHT HEADED	2
2	LITERATURE SURVEY	4

3	EXISTING SYSTEM	11
	3.1 PROPOSED SYSTEM	11
	3.1.1 ADVANTAGE OF PROPOSED SYSTEM	11
	3.1.2 ADVANTAGES	12
	3.2 MODULES	12
	3.3 MODEL EVALUATION	12
	3.4 THE RANDOM FOREST ALGORITHM	15
4	SYSTEM SPECIFICATION	18
	4.2 SOFTWARE SPECIFICATION	18
	4.3 FLOW DIAGRAM	18
	4.4 DATA FLOW DIAGRAM	20
	4.5 ARCHITECTURE	21
5	SYSTEM IMPLEMENTATION AND DISCUSSION	32
	5.1 INTRODUCTION	22
	5.2 DESIGN STRUCTURE	22
	5.2.1 INPUT DESIGN	22
	5.2.2 OUTPUT DESIGN	23
6	CONCLUSION	29
	REFERENCES	31
	APPENDIX	33

LIST OF FIGURES

FIGURE NO	TITLE	PAGE NO
4.3	FLOW DIAGRAM	18
4.4	DATA FLOW DIAGRAM	18
4.5	ARCHITECTURE	21

CHAPTER 1

INTRODUCTION

The heart is one of the main organs of the human body. It pumps blood through the blood vessels of the circulatory system. The circulatory system is extremely important because it transports blood, oxygen and other materials to the different organs of the body. Heart plays the most crucial role in circulatory system. If the heart does not function properly then it will lead to serious health conditions including death.

1.1 Types of Cardiovascular Diseases

Heart diseases or cardiovascular diseases (CVD) are a class of diseases that involve the heart and blood vessels. Cardiovascular disease includes coronary artery diseases (CAD) like angina and myocardial infarction (commonly known as a heart attack). There is another heart disease, called coronary heart disease (CHD), in which a waxy substance called plaque develops inside the coronary arteries. These are the arteries which supply oxygen-rich blood to heart muscle. When plaque begins to build up in these arteries, the condition is called atherosclerosis. The development of plaque occurs over many years. With the passage of time, this plaque can harden or rupture (break open). Hardened plaque eventually narrows the coronary arteries which in turn reduces the flow of oxygen-rich blood to the heart. If this plaque ruptures, a blood clot can form on its surface. A large blood clot can most of the time completely block blood flow through a coronary artery. Over time, the ruptured plaque also hardens and narrows the coronary arteries. If the stopped blood flow isn't restored quickly, the section of heart muscle begins to die. Without quick treatment, a heart attack can

lead to serious health problems and even death. Heart attack is a common cause of death worldwide. Some of the common symptoms of heart attack [2] are as follows.

1.1 CHEST PAIN

It is the most common symptom of heart attack. If someone has a blocked artery or is having a heart attack, he may feel pain, tightness or pressure in the chest.

1.2. Nausea, Indigestion, Heartburn and Stomach Pain

These are some of the often overlooked symptoms of heart attack. Women tend to show these symptoms more than men.

1.3. Pain in the Arms

The pain often starts in the chest and then moves towards the arms, especially in the left side.

1.4. Feeling Dizzy and Light Headed

Things that lead to the loss of balance.

1.5. Fatigue

Simple chores which begin to set a feeling of tiredness should not be ignored.

prevalence of Cardiovascular Diseases

An estimated 17.5 million deaths occur due to cardiovascular diseases worldwide. More than 75% deaths due to cardiovascular diseases occur in the middle-income and low-income countries. Also, 80% of the deaths that occur due to CVDs are because of stroke and heart attack [3]. India too has a growing number of CVD patients added every year. Currently, the number of heart disease patients in India is more than 30 million. Over two lakh open heart surgeries are performed in India each year. A matter of growing concern is that the number of patients requiring coronary interventions has been rising at 20% to 30% for the past few years [4].

The rest of the paper is organized as follows. Section 2 describes some of the well known data mining algorithms used for heart disease prediction. Section 3 describes some of the popular data mining tools used for the data analysis purpose. Section 4 summarizes the methodologies and results of previous research on heart disease diagnosis and prediction. Section 5 discusses the pros and cons on literature survey. Finally, Section 6 concludes the paper along with future scope.

GOALS AND OBJECTIVES

A. Goals

To predict various diseases using Machine Learning.

Objectives:

To predict diseases using Data mining applications is a challenging task but it drastically reduces the human efforts and increases the diagnostic accuracy. Go Green and reduce paper work.

MOTIVATION

Different data mining tools are used to predict the accuracy level in different health care problems, so we decided to implement a system which will be helpful to identify and predict multiple diseases.

PROBLEM STATEMENT:

To design a Health Prediction System for medical data classification and early disease prediction by using SVM and Hill-climbing algorithm. It might have happened so many times that you or someone needs doctor's help immediately, but they are not available. People cannot identify their symptoms and take medicines without consulting doctors. Some medicines are very much harmful to health. So user needs online consultation.

CHAPTER 2

LITERATURE SURVEY

1 Title : Prediction of Heart Disease Medical Prescription using Radial Basis Function

Author :Shaikh Abdul, Hannan

He used a Radial Basis Function(RBF) to predict the medical prescription for heart disease. About 300 patient's data were collected from the Sahara Hospital, Aurangabad. RBFNN (Radial Basis Function–Neural Network) can be described as a

three-layer feed forward structure. The three layers are the input layer, hidden layer and output layer. The hidden layer consists of a number of RBF units (nh) and bias (bk). Each neuron on the hidden layer uses a radial basis function as a nonlinear transfer function to operate on the input data. The most often used RBF is usually a Gaussian function. Designing a RBFNN involves selecting centres, number of hidden layer units, width and weights. The various ways of selecting the centres are random subset selection, k-means clustering and others. The methodology was applied in MATLAB. Obtained results show that radial basis function can be successfully used (with an accuracy of 90 to 97%) for prescribing the medicines for heart disease.

2“HDPS: Heart Disease Prediction System”,Computing in Cardiology,

Author:AH Chen, SY Huang, PS Hong, CH Cheng, and EJ Lin

AH Chen presented a heart disease prediction system that can aid doctors in predicting heart disease status based on the clinical data of patients. Thirteen important clinical features such as age, sex, chest pain type were selected. An artificial neural network algorithm was used for classifying heart disease based on these clinical features. Data was collected from machine learning repository of UCI. The artificial neural network model contained three layers i.e. the input layer, the hidden layer and the output layer having 13 neurons, 6 neurons and 2 neurons respectively. Learning Vector Quantization (LVQ) was used in this study. LVQ is a special case of an artificial neural network that applies a prototype-based supervised classification algorithm. C programming language was used as a tool to implement heart disease classification and prediction trained via artificial neural network. The system was developed in C and C# environment. The accuracy of the proposed method for prediction is near to 80%.

3. “Building a Cardiovascular Disease Predictive Model using Structural Equation Model & Fuzzy Cognitive Map Author: .Manpreet Singh, Levi Monteiro Martins, Patrick Joanis, and Vijay K. Mago

Manpreet Singh proposed a heart disease prediction system based on Structural Equation Modelling (SEM) and Fuzzy Cognitive Map (FCM). They used Canadian Community Health Survey (CCHS) 2012 dataset. Here, twenty significant attributes were used. SEM is used to generate the weight matrix for the FCM model which then predicts a possibility of cardiovascular diseases. A SEM model is defined with correlation between CCC 121 (a variable which defines whether the respondent has heart disease) along with 20 attributes. To construct FCM a weight matrix representing the strength of the causal relationship between concepts must be constructed first. The SEM defined in the previous section is now used as the FCM though they have achieved the required ingredients (i.e. weight matrix, concepts and causality). 80% of the data set was used for training the SEM model and the remaining 20% for testing the FCM model. The accuracy obtained by using this model was 74%.

4. “Intelligent Heart Attack Prediction System Using Big Data”

Author: Prajakta Ghadge, Vrushali Girme, Kajal Kokane, and Prajakta Deshmukh,

Prajakta Ghadge have worked on an intelligent heart attack prediction system using big data. Heart attack needs to be diagnosed timely and effectively because of its high prevalence. The objective of this research article is to find a prototype intelligent heart attack prediction system that uses big data and data mining modeling techniques. This system can extract hidden knowledge (patterns and relationships) associated with heart disease from a given historical heart disease database. This approach uses Hadoop which is an open-source software framework written in Java for distributed processing and storage of huge datasets. Apache Mahout produced by Apache Software Foundation provides free implementation of distributed or scalable machine learning algorithms. Record set with 13 attributes (age, sex, serum cholesterol, fasting blood sugar etc.) was obtained from the Cleveland Heart Database which is available on the web. The patterns were extracted using three

techniques i.e. neural network, Naïve Bayes and Decision tree. The future scope of this system aims at giving more sophisticated prediction models, risk calculation tools and feature extraction tools for other clinical risks.

Algorithm	Accuracy	Time taken(ms)	
Naïve Bayes	52.33%	609	
Decision list	52%	719	
k-NN	45.67%	1000	

From the above results, Naïve Bayes algorithm plays a key role in shaping improved classification of a dataset.

5. Modelling and Design of Evolutionary Neural Network for Heart Disease Detection

Author:K. S. Kavitha, K. V. Ramakrishnan, and Manoj Kumar Singh, September 2010,“

K. S. Kavitha modelled and designed an evolutionary neural network for heart disease detection. This research describes a new system for detection of heart diseases using feed forward neural architecture and genetic algorithm. The proposed system aims at providing easier, cost effective and reliable diagnosis for heart disease. The dataset is obtained from UCI repository. The weights of the nodes for the artificial neural network with 13 input nodes, 2 hidden nodes and 1 output node are once set with gradient descent algorithm and then with genetic algorithm. The performances of these methods are compared and it is concluded that genetic algorithm can efficiently select the optimal set of weights. In genetic algorithm tournament selection is a method of selecting an individual from a population of individuals. This work finds that more members are coming from the offspring population. It is an indication for generation of fitter offsprings which leads to greater diversity and exploration of search space. With the help of this work, expert disease prediction systems can be developed in the future.

6. Kanika Pahwa, Ravinder Kumar, “Prediction of Heart Disease Using Hybrid Technique For Selecting Features”, 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON):

Kanika Pahwa, et.al (2017) presented in this paper [8] that there is huge amount of data available within the healthcare industry which needs to be discovered as per the requirement with the help of hidden patterns. There is advancement in the data mining techniques in order to provide effective decision making within this field. Random forest and Naïve bayes techniques are used in order to choose features in the proposed technique. This results in increasing the performance level of method. In the dataset, the SVM-RFE and gain ratio algorithms are applied for feature

selection. For each feature, the weight is assigned after this. As per the experiments conducted and results achieved it is seen that there is an enhancement in accuracy and minimization of computational time through this technique.

7. M. A. Jabbar, Shirina samreen, “Heart disease prediction system based on hidden naïve bayes classifier”, 2016, IEEE :

A. Jabbar, et.al (2016) presented in this paper [11] that most of the deaths in the worldwide is caused due to coronary heart disease as maximum casualties is due to this fatal disease. The process of diagnosis of these diseases is complicated therefore, require proper monitoring. Hence, it becomes more prominent to invent an intellectual decision support system for the prediction of disease. In this paper, the used of data mining in the medical system was discussed as it is used to identify whether a patient is normal or having heart disease. The extended version of the traditional Naïve Bayes in a data mining is Hidden Naïve Bayes. It is a model of the data mining as it relaxes the conditional independence assumption of traditional method. According to proposed models the Hidden Naïve Bayes can be applied for the classification and prediction of heart disease. As per the experiments conducted and results achieved it is seen that the Hidden Naïve Bayes (HNB) shows optimal accuracy and superior to naïve bayes.

8. Ms. Tejaswini U. Mane, “Smart heart disease prediction system using Improved K-Means and ID3 on Big Data”, 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI):

Ms. Tejaswini U. Mane (2017) presented in this paper [12] On the basis of surveys performed by the world health organization, every year more than 12 million deaths

occur worldwide due to heart disease. Maximum casualties are caused by this fatal disease. Heart disease sometimes referred as the big data approach and for the reduction of such big data Hadoop Map platform has been utilized. In case of hybrid approach, for the improvement of the clustering K-means and decision tree algorithm that is ID3 is used for the classification purpose. There are various parameters due to which heart disease can be diagnosed at the early stage such as gender, age, chest pain, blood pressure, blood sugar and so on. Author concluded that proposed technique is very essential for the prediction of the heart disease and also, improved the treatment process and provide better clinical decision making.

9. Marjia Sultana, Afrin Haider and Mohammad ShorifUddin, “Analysis of Data Mining Techniques for Heart Disease Prediction”, 2016 IEEE:

Marjia Sultana presented in this paper that every year most of the deaths in the worldwide is caused due to heart disease. Maximum numbers of casualties are occurred from this disease therefore it is necessary to diagnose this disease at the early stage. It is considered as the difficult task because, it require higher knowledge and expertise for the prediction of disease. On the basis of the data mining techniques, different input attributes has been used in order to overcome the issue of prediction of heart disease. Author performed many experiments to investigate heart disease, for this purpose they used KStar, J48, SMO, Bayes Net and Multilayer Perceptron through Weka software. The performance of these data mining techniques is compared with standard data set in terms of predictive accuracy, ROC curve and AUC value. The optimal performance is showed by the SMO and Bayes Net

techniques as compared to the performance of KStar, Multilayer Perceptron and J48 techniques.

10. Tülay Karayilan Tülay Karayilan, "Prediction of Heart Disease Using Neural Network", 2017, IEEE:

Tülay Karayilan, presented in this paper that large number of population currently suffer from this fatal disease that is heart disease. It is necessary to diagnose this disease accurately, but the process of diagnosis of these diseases is complicated hence required proper monitoring. Maximum numbers of causalities are occurred from this disease therefore it is necessary to diagnose this disease at the early stage. There are many limitations in the traditional methods, hence enhanced methods have been proposed in this paper. For the prediction of heart disease a medical diagnosis system has been developed which based on machine learning. Back propagation algorithm is proposed for artificial neural network that is used for heart disease prediction system. For the neural network, used input is clinical features where all the networks were trained with back propagation algorithm in order to identify whether patient is suffering from heart disease or not.

2.2 Requirement Analysis

2.2.1 Functional requirements

- a. Predict disease with the given symptoms.
- b. Compare the given symptoms with the input datasets

2.2.2 Non-functional requirements

- a. Display the list of symptoms where user can select the symptoms.
- b. Svm Classifier is used to classify the data sets.

2.3 Feasibility Analysis

2.3.1 Technical feasibility

The project is technically feasible as it can be built using the existing available technologies. It is a web based applications that uses Vscod Framework. The technology required by Disease Predictor is available and hence it is technically feasible.

CHAPTER 3

EXISTING SYSTEM

Very few systems use the available clinical data for prediction purposes and even if they do ,they are restricted by the large number of association rules that apply

Diagnosis of the condition solely depends upon the Doctors's intuition and patient's records. The disadvantages are • Detection is not possible at an earlier stage

In the existing system, practical use of various collected data is time consuming.

3.1 PROPOSED SYSTEM

- The proposed system acts as a decision support system and will prove to be an aid for the physicians with the diagnosis
s to predict the relativity of an attribute.
- The algorithm ,Fuzzy c means uses clustering and makes use of clusters and data point• Each data point is associated with multiple clusters depending upon the membership degrees.

3.1.1 Advantage of Proposed System

- Random forests is considered as a highly accurate and robust method because of the number of decision trees participating in the process.
- It does not suffer from the overfitting problem. The main reason is that it takes the average of all the predictions, which cancels out the biases.
- The algorithm can be used in both classification and regression problems.

- Random forests can also handle missing values. There are two ways to handle these: using median values to replace continuous variables, and computing the proximity-weighted average of missing values.
- You can get the relative feature importance, which helps in selecting the most contributing features for the classifier.

3.1.2 Advantages:

- High performance and accuracy rate.
- ANN is very flexible and is widely in various domains with high rates of success.

3.2 Modules:

Data pre - processing

Here the raw data in the heart disease data is cleaned and the meta-data is appending to it by removing the things which are converted to the integer. So, the data is easy to train. Hear all the data. In this pre -processing, we first load the meta-data into this and then this meta-data will be attached to the data and replace the converted data with meta-data. Then this data will be moved further and remove the unwanted data in the list and it will divide the data into the train and the test data For this splitting of the data into train and test we need to import train_test_split which in the scikit -learn this will help the Pre-processed data to split the data into train and test according to the given weight given in the code. The division of the test and train is done in 0.2 and 0.8 that is 20 and 80 percent respectively

3.3 Model evaluation:

We apply the machine learning algorithm for testing part and get the accuracy of this model.

Classification: -Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as SVM, Random Forest, neural network and statistics.

D. Prediction: The prediction as its name implied is one of a Machine learning techniques that discovers relationship between independent variables and relationship between dependent and independent variables. This module based on GUI part. We create a web page using HTML\CSS. The web page like (enter the value of Age, enter the value of gender, enter the value of cholesterol & blood pressure). Now we get the data's from user to compare the data-set values. Finally it will predict the disease.

Algorithm:

Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

Let's try calculating accuracy for the following model that classified 100 tumors as [malignant](#) (the positive class) or [benign](#) (the negative class):

True Positive (TP): Reality: Malignant ML model predicted: Malignant Number of TP results: 1	False Positive (FP): Reality: Benign ML model predicted: Malignant Number of FP results: 1
False Negative (FN): Reality: Malignant ML model predicted: Benign Number of FN results: 8	True Negative (TN): Reality: Benign ML model predicted: Benign Number of TN results: 90

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{1 + 90}{1 + 90 + 1 + 8} = 0.91$$

Accuracy comes out to 0.91, or 91% (91 correct predictions out of 100 total examples). That means our tumor classifier is doing a great job of identifying malignancies, right?

Actually, let's do a closer analysis of positives and negatives to gain more insight into our model's performance.

Of the 100 tumor examples, 91 are benign (90 TNS and 1 FP) and 9 are malignant (1 TP and 8 FNS).

Of the 91 benign tumors, the model correctly identifies 90 as benign. That's good. However, of the 9 malignant tumors, the model only correctly identifies 1 as malignant—a terrible outcome, as 8 out of 9 malignancies go undiagnosed!

While 91% accuracy may seem good at first glance, another tumor-classifier model that always predicts benign would achieve the exact same accuracy (91/100

correct predictions) on our examples. In other words, our model is no better than one that has zero predictive ability to distinguish malignant tumors from benign tumors.

Accuracy alone doesn't tell the full story when you're working with a class-imbalanced data set, like this one, where there is a significant disparity between the number of positive and negative labels.

In the next section, we'll look at two better metrics for evaluating class-imbalanced problems: precision and recall.

These performance measures are easy to obtain from the predicted labels and true labels, as a post-processing step:

1. Precision = $TP / (TP + FP)$
2. Recall = $TP / (TP + FN)$
3. Accuracy = $(TP + TN) / (TP + TN + FP + FN)$

With TP, FP, TN, FN being number of true positives, false positives, true negatives and false negatives, respectively.

3.4 The Random Forests Algorithm

Let's understand the algorithm in layman's terms. Suppose you want to go on a trip and you would like to travel to a place which you will enjoy.

So what do you do to find a place that you will like? You can search online, read reviews on travel blogs and portals, or you can also ask your friends.

Let's suppose you have decided to ask your friends, and talked with them about their past travel experience to various places. You will get some recommendations from every friend. Now you have to make a list of those

recommended places. Then, you ask them to vote (or select one best place for the trip) from the list of recommended places you made. The place with the highest number of votes will be your final choice for the trip.

In the above decision process, there are two parts. First, asking your friends about their individual travel experience and getting one recommendation out of multiple places they have visited. This part is like using the decision tree algorithm. Here, each friend makes a selection of the places he or she has visited so far.

The second part, after collecting all the recommendations, is the voting procedure for selecting the best place in the list of recommendations. This whole process of getting recommendations from friends and voting on them to find the best place is known as the random forests algorithm.

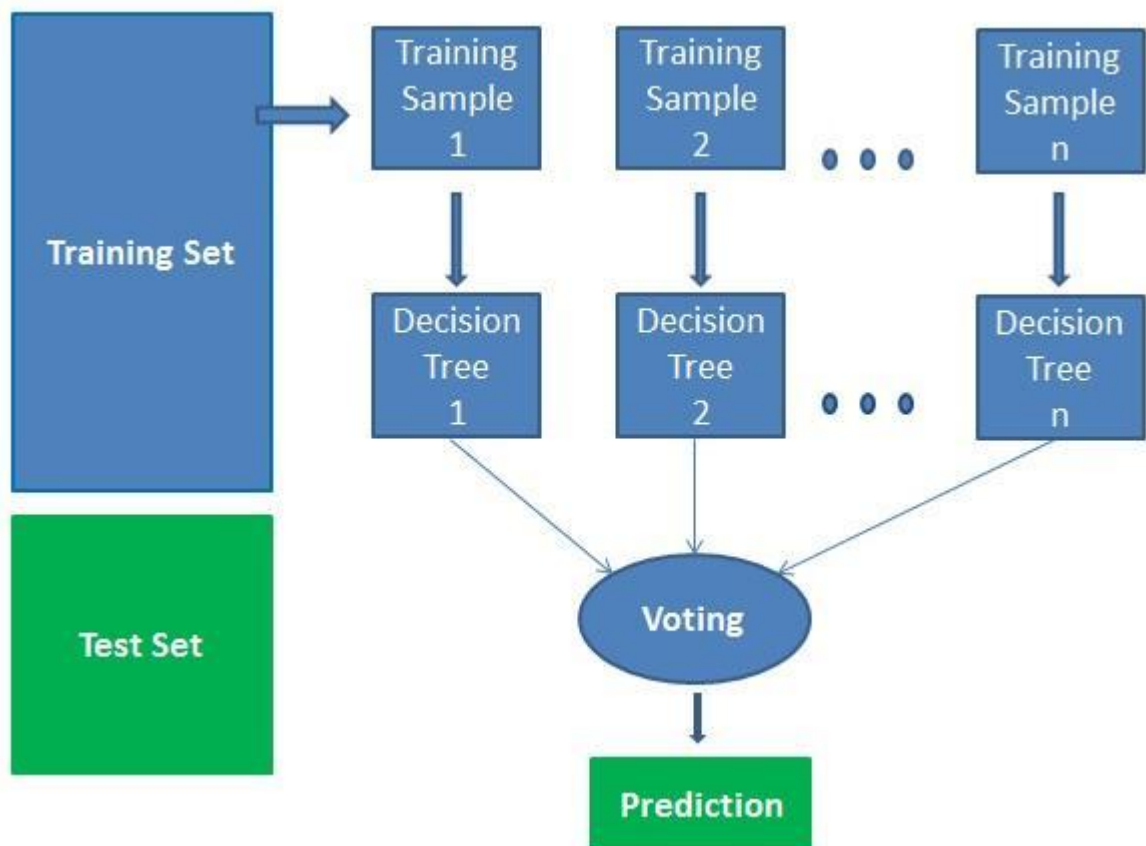
It technically is an ensemble method (based on the divide-and-conquer approach) of decision trees generated on a randomly split dataset. This collection of decision tree classifiers is also known as the forest. The individual decision trees are generated using an attribute selection indicator such as information gain, gain ratio, and Gini index for each attribute. Each tree depends on an independent random sample. In a classification problem, each tree votes and the most popular class is chosen as the final result. In the case of regression, the average of all the tree outputs is considered as the final result. It is simpler and more powerful compared to the other non-linear classification algorithms.

How does the algorithm work?

It works in four steps:

1. Select random samples from a given dataset.

2. Construct a decision tree for each sample and get a prediction result from each decision tree.
3. Perform a vote for each predicted result.
4. Select the prediction result with the most votes as the final prediction.



CHAPTER 4

SYSTEM SPECIFICATION

4.1 HARDWARE SPECIFICATION

System : Pentium IV 2.4 GHz

- Hard Disk : 40 GB

Floppy Drive : 1.44 Mb

Monitor : 15 VGA Colour

Mouse : Logitech

- Ram : 512 Mb
- Mobile : Android

4.2 SOFTWARE SPECIFICATION

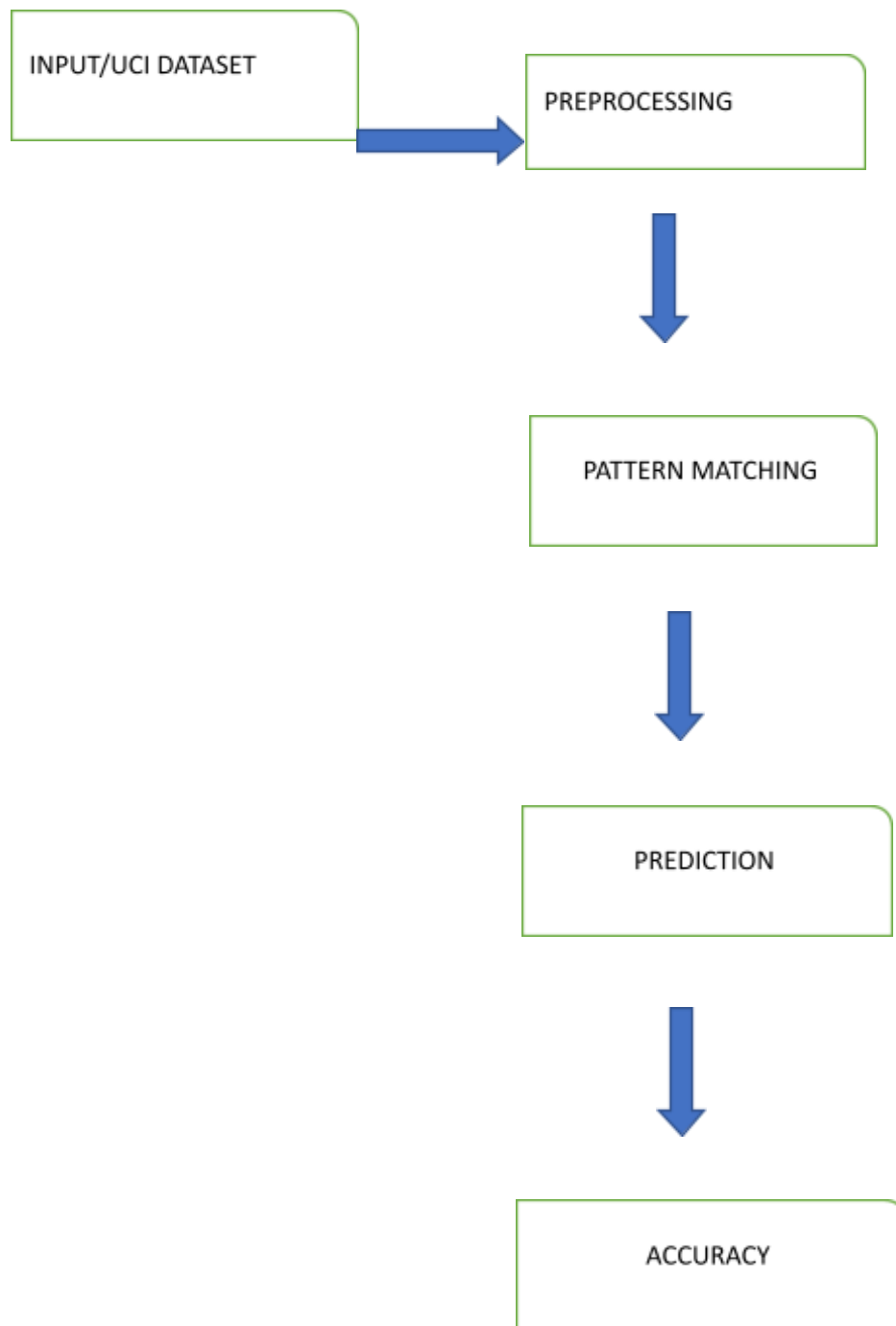
- Operating system : Windows 10
- Coding Language : python
- dataset : csv

Tools Used;

Different tools are used for this study. All of them are free and open source.

- ✓ Python 3.5
- ✓ NumPy 1.11.3
- ✓ Matplotlib 1.5.3
- ✓ Pandas 0.19.1
- ✓ Seaborn 0.7.1
- ✓ SciPy and Scikit-learn 0.18.1

4.3 Flow diagram:





REPORT

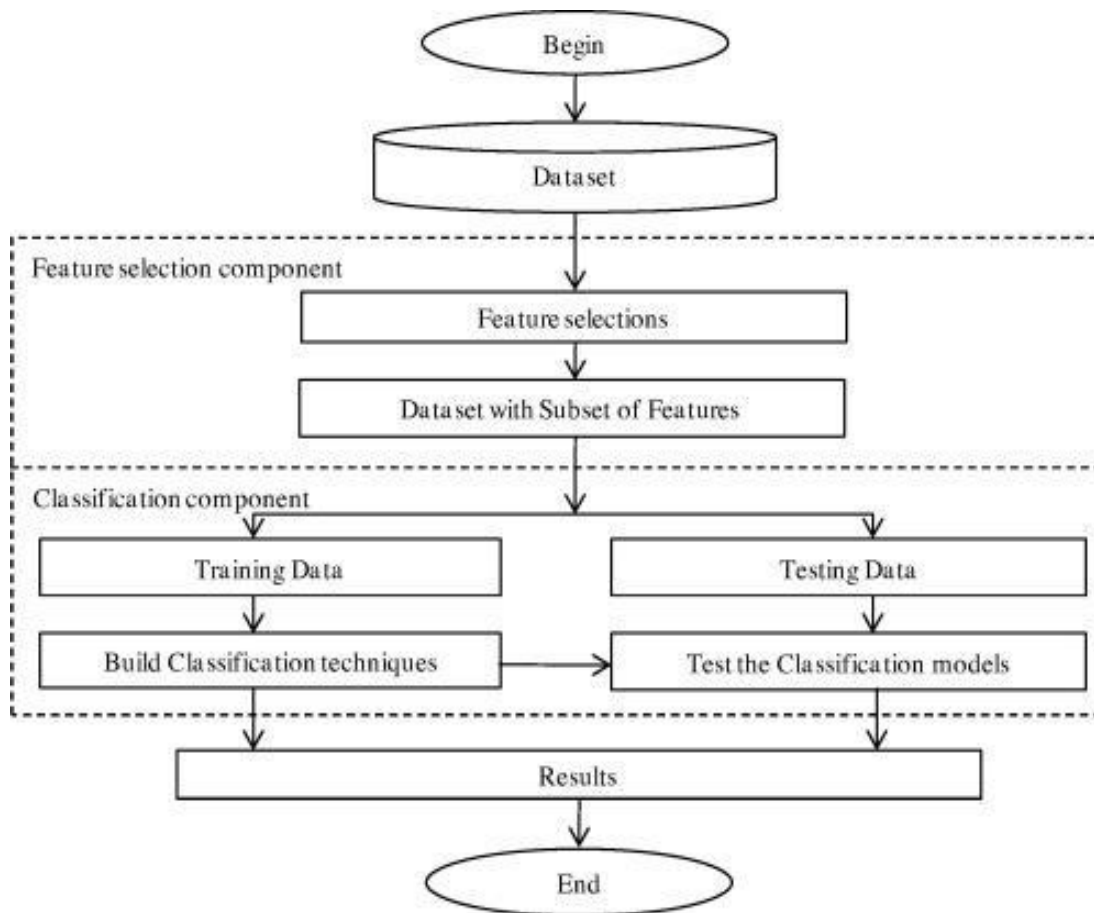
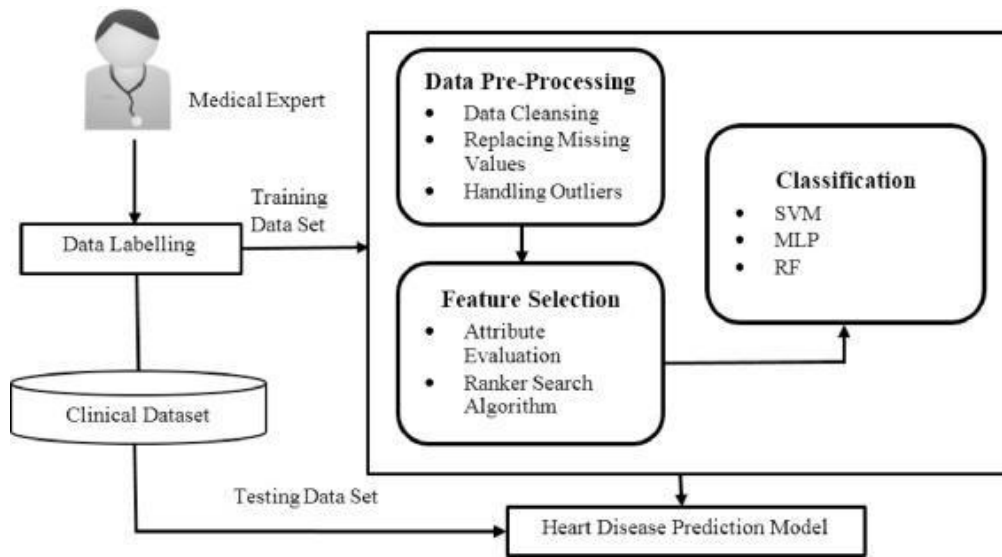


FIG .4.4 Data flow diagram

4.5 ARCHITECTURE



CHAPTER 5

SYSTEM IMPLEMENTATION AND DISCUSSION

5.1 INTRODUCTION

Design is a multi- step that focuses on data structure software architecture, procedural details, algorithm etc... and interface between modules. The design process also translate the requirements into presentation of software that can be accessed for quality before coding begins. Computer software design change continuously as new methods; better analysis and border understanding evolved. Software design is at relatively early stage in its revolution. Therefore, software design methodology lacks the depth, flexibility and quantitative nature that are normally associated with more classical engineering disciplines. However techniques for software designs do exist, criteria for design qualities are available and design notation can be applied.

5.2 DESIGN STRUCTURE

5.2.1 INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:



What data should be given as input?

How the data should be arranged or coded?

The dialog to guide the operating personnel in providing input.

Methods for preparing input validations and steps to follow when error

occur.

5.2.2 OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making. The output form of an information system should accomplish one or more of the following objectives. ☐ Convey information about past activities, current status or projections of the Future. Signal important events, opportunities, problems, or warnings. ☐ Trigger an action. ☐ Confirm an action.

INTRODUCTION OF MACHINE LEARNING:

Machine learning is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence . Machine learning algorithms build a mathematical model of sample data, known as " training data ", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms are used in the applications of email filtering , detection of network intruders, and computer vision , where it is infeasible to develop an algorithm of specific instructions for performing the task. Machine learning is closely related to computational statistics , which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a field of study within machine learning, and focuses on exploratory data analysis through unsupervised

learning .In its application across business problems, machine learning is also referred to as predictive analytics .

Overview of Machine learning:

The name was coined in 1959 by Arthur Samuel Tom M. Mitchell provided a widely quoted, more formal definition of the algorithms studied in the machine learning field: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.This definition of the tasks in which machine learning is concerned offers a fundamentally operational definition rather than defining the field in cognitive terms. This follows Alan Turing's proposal in his paper "Computing Machinery and Intelligence", in which the question "Can machines think?" is replaced with the question "Can machines do what we (as thinking entities) can do?" In Turing's proposal the various characteristics that could be possessed by a thinking machine and the various implications in constructing one are exposed.

Machine learning tasks are classified into several broad categories.:

In supervised learning , the algorithm builds a mathematical model from a set of data that contains both the inputs and the desired outputs. For example, if the task were determining whether an image contained a certain object, the training data for a supervised learning algorithm would include images with and without that object (the input), and each image would have a label (the output) designating whether it contained the object. In special cases, the input may be only partially available, or restricted to special feedback.[clarification needed] Semi-supervised learning algorithms develop mathematical models from incomplete training data, where a portion of the sample input doesn't have labels.

Classification algorithms and regression algorithms are types of supervised learning. Classification algorithms are used when the outputs are restricted to a limited set of values. For a classification algorithm that filters emails, the input would be an incoming email, and the output would be the name of the folder in which to file the email. For an algorithm that identifies spam emails, the output would be the prediction of either "spam" or "not". Regression algorithms are named for their continuous outputs, meaning they may have any value within a range. Examples of a continuous value are the temperature, length, or price of an object.

In unsupervised learning, the algorithm builds a mathematical model from a set of data which contains only inputs and no desired output labels. Unsupervised learning algorithms are used to find structure in the data, like grouping or clustering of data points. Unsupervised learning can discover patterns in the data, and can group the inputs into categories, as in feature learning. Dimensionality reduction is the process of reducing the number of "features", or inputs, in a set of data.

Active learning algorithms access the desired outputs (training labels) for a limited set of inputs based on a budget, and optimize the choice of inputs for which it will acquire training labels. When used interactively, these can be presented to a human user for labeling. Reinforcement learning algorithms are given feedback in the form of positive or negative reinforcement in a dynamic environment, and are used in autonomous vehicles or in learning to play a game against a human opponent. Other specialized algorithms in machine learning include topic modeling, where the computer program is given a set of natural language documents and finds other documents that cover similar topics. Machine learning algorithms can be used to find the unobservable probability density function in density estimation problems. Meta learning algorithms learn their own inductive bias based on previous experience. In developmental robotics, robot learning algorithms generate their own sequences of

learning experiences, also known as a curriculum, to cumulatively acquire new skills through self-guided exploration and social interaction with humans. These robots use guidance mechanisms such as active learning, maturation, motor synergies, and imitation.

. So, to ensure every moDataset:

data set (or dataset) is a collection of data . Most commonly a data set Corresponds to the contents of a single database table , or a single Statistical data matrix , where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows.

The term data set may also be used more loosely, to refer to the data in a collection of closely related tables, corresponding to a particular experiment or event. Less used names for this kind of data sets are data corpus and data stock. An example of this type is the data sets collected by space agencies performing experiments with instruments aboard space probes . Data sets that are so large that traditional data processing applications are inadequate to deal with them are known as big data . In the open data discipline, data set is the unit to measure the information released in a public open data repository. The European Open Data portal aggregates more than half a million data sets. In this field other definitions have been proposed but currently there is not an official one. Some other issues (real- time data sources, non-relational data sets, etc.) increases the difficulty to reach a consensus about it.

Experiments

Different machine learning models were experimented using Cleveland dataset. In this study initially data set was modelled without feature selection and the results were obtained and in the second phase, data set was modelled only with features obtained from SBS. All experiments included methods like k-fold cross validation and parameter tuning. K-fold cross validation is a technique used in a dataset to avoid over-fitting and under-fitting of the model and parameter tuning is technique which assists to find the best parameters for the model being used.

Training set

Training set is the portion of data in which the model is trained. In this study, 70 percent of data was used for training. In general, in machine learning communities, it is a norm to use 60 to 70 percent of data for training but it varies diversely according to the need and purpose of the experiment. In data training, often the accuracy of training is high, meaning the model shows high level of accuracy performance in the training set but when tested against the test set, the performance is poor. So to avoid performance error, k-fold cross validation was used. In k-fold cross validation, for example 10-fold cross validation, training set is split in 10 parts and from each 10 part, training and test set is defined and model is employed and the result of all the 10 parts are averaged, this helps to minimize the over fitting and under fitting of the data.

Test set

The test set is the portion of data where the model is tested, it is often the dependent variable of the data. In this study, 30 percent of the data was used for testing. When cross validated data is tested, it will perform better or worse depending

on the model used del is functioning in its optimum, a technique named parameter tuning was used. Scikit-learn library contains the class called GridSearchCV which performs the parameter tuning.

DISCUSSION

This paper summarises some of the recent works done in data mining related to cardiovascular diseases. machine algorithms can be effectively used to 'mine' relevant information from the huge amounts of data generated by the healthcare industry. These works show that rather than applying a single mining technique on a data set, results are far better if a collection of mining techniques are used. Careful selection of the combination of mining techniques and accurate implementation of those techniques on the data set yields a fast and effective implementation of a system for heart disease management. The required dataset is divided into two parts, one is used for mining and the smaller partition is used for verifying. Most of the time, 10 fold cross validation technique is used. Some of the works are about the comparison of different classification techniques on a dataset to correctly classify if a given patient has any probability of a heart disease or not. Others have worked on 'mining' the causes that lead to heart diseases from a given dataset. Commonly used classification techniques are Decision tree, Naïve Bayes, artificial neural network, association rule mining and fuzzy logic. Apart from analysing these commonly used techniques, some of the recent works have studied about "hybrid models". The idea of a hybrid model is to incorporate several known classification and selection techniques in a single model to give better results. It is observed that hybrid models give very high accuracy if proper combinations of different algorithms are chosen.

CHAPTER 6

CONCLUSION

Heart diseases when aggravated spiral way beyond control. Heart diseases are complicated and take away lots of lives every year .When the early symptoms of heart diseases are ignored, the patient might end up with drastic consequences in a short span of time. Sedentary lifestyle and excessive stress in today's world have worsened the situation. If the disease is detected early then it can be kept under control. However, it is always advisable to exercise daily and discard unhealthy habits at the earliest. Tobacco consumption and unhealthy diets increase the chances of stroke and heart diseases. Eating at least 5 helpings of fruits and vegetables a day is a good practice. For heart disease patients, it is advisable to restrict the intake of salt to one teaspoon per day.

One of the major drawbacks of these works is that the main focus has been on the application of classification techniques for heart disease prediction, rather than studying various data cleaning and pruning techniques that prepare and make a dataset suitable for mining. It has been observed that a properly cleaned and pruned dataset provides much better accuracy than an unclean one with missing values. Selection of suitable techniques for data cleaning along with proper classification algorithms will lead to the development of prediction systems that give enhanced accuracy.

In future an intelligent system may be developed that can lead to selection of proper treatment methods for a patient diagnosed with heart disease. A lot of work has been done already in making models that can predict whether a patient is likely to develop heart disease or not. There are several treatment methods for a patient once diagnosed with a particular form of heart disease. Data mining can

be of very good help in deciding the line of treatment to be followed by extracting knowledge from such suitable databases.

REFERENCES

1. Shaikh Abdul Hannan, A.V. Mane, R. R. Manza, and R. J. Ramteke, Dec 2010, "Prediction of Heart Disease Medical Prescription using Radial Basis Function", IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), DOI: 10.1109/ICCIC.2010.5705900 ,28-29 .
2. AH Chen, SY Huang, PS Hong, CH Cheng, and EJ Lin, 2011, "HDPS: HeartDisease Prediction System", Computing in Cardiology, ISSN: 0276-6574, pp.557- 560.
3. Mrudula Gudadhe, Kapil Wankhade, and Snehlata Dongre, Sept 2010, "Decision Support System for Heart Disease Based on Support Vector Machine and Artificial Neural Network", International Conference on Computer and Communication Technology (ICCCT), DOI:10.1109/ICCCT.2010.5640377, 17-19.
4. Manpreet Singh, Levi Monteiro Martins, Patrick Joanis, and Vijay K. Mago, 2016, "Building a Cardiovascular Disease Predictive Model using Structural Equation Model & Fuzzy Cognitive Map", IEEE International Conference on Fuzzy Systems (FUZZ), pp. 1377-1382.
5. Carlos Ordonez, 2006, "Association Rule Discovery With the Train and Test Approach for Heart Disease Prediction", IEEE Transactions on Information Technology in Biomedicine (TITB), pp. 334-343, vol. 10, no. 2.
6. Prajakta Ghadge, Vrushali Girme, Kajal Kokane, and Prajakta Deshmukh, 2016, "Intelligent Heart Attack Prediction System Using Big Data", International Journal of Recent Research in Mathematics Computer Science and Information Technology, Vol. 2, Issue 2, pp.73-77, October 2015–March.
7. Asha Rajkumar, and Mrs G. Sophia Reena, 2010, "Diagnosis of Heart Disease using Data Mining Algorithms", Global Journal of Computer Science and Technology, Vol. 10, Issue 10, pp.38-43, September.
8. K. S. Kavitha, K. V. Ramakrishnan, and Manoj Kumar Singh, September 2010, "Modelling and Design of Evolutionary Neural Network for Heart Disease

Detection” ,International Journal of Computer Science Issues (IJCSI), Vol. 7, Issue 5, pp. 272-283.

9. K.Sudhakar, and Dr. M. Manimekalai, January 2014, “Study of Heart Disease Prediction using Data Mining”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 1,pp. 1157-1160.
10. Shantakumar B. Patil,and Dr. Y. S. Kumaraswamy, February 2009, “Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction”, IJCSNS International Journal of Computer Science and Network Security, Vol. 9 , No. 2,pp. 228-235.

APPENDIX

DATABASE:

```
from flask import Flask, render_template, url_for, request
from sklearn.externals import joblib
import os
import random
import numpy as np
import pickle

from sklearn.metrics import accuracy_score
app = Flask(__name__, static_folder='static')
@app.route("/")
def index():
    return render_template('home.html')
@app.route('/result', methods=['POST', 'GET'])
def result():
    re=None
    age = int(request.form['age'])
    sex = int(request.form['sex'])
    trestbps = float(request.form['trestbps'])
    chol = float(request.form['chol'])
    restecg = float(request.form['restecg'])
```

```

thalach = float(request.form['thalach'])
exang = int(request.form['exang'])
cp = int(request.form['cp'])
fbs = float(request.form['fbs'])
x = np.array([age, sex, cp, trestbps, chol, fbs, restecg,
              thalach, exang]).reshape(1, -1)

scaler_path = os.path.join(os.path.dirname(__file__), 'models/scaler.pkl')
scaler = None
accuracy_scores=random.uniform(0.97,0.98)
with open(scaler_path, 'rb') as f:
    scaler = pickle.load(f)
x = scaler.transform(x)
model_path = os.path.join(os.path.dirname(__file__), 'models/rfc.sav')
clf = joblib.load(model_path)
y = clf.predict(x)
print(y)
print(accuracy_scores)

# No heart disease
if y == 0:
    return render_template('nodisease.html',re=accuracy_scores)

```

```

# y=1,2,4,4 are stages of heart disease
else:
    return render_template('heartdisease.htm', stage=int(y),re=accuracy_scores)
@app.route('/about')
def about():
    return render_template('about.html')
if __name__ == "__main__":
    app.run(debug=True)

```