

HEART DISEASE DETECTION USING ARTIFICIAL INTELLIGENCE

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

by

KOTI SAHITHI (Reg. No. 37110375)

PODILA SHIVANI (Reg. No. 37110570)



**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
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

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 **KOTI SAHITHI (Reg. No. 37110375)** and **PODILA SHIVANI (Reg. No.37110570)** who carried out the project entitled **"HEART DISEASE DETECTION USING ARTIFICIAL INTELLIGENCE"** under my supervision from **November 2020** to **March 2021**.

Internal Guide

Dr. N.Srinivasan, M.E., Ph.D.,

Head of the Department

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I **PODILA SHIVANI** hereby declare that the Project Report entitled “**HEART DISEASE DETECTION USING ARTIFICIAL INTELLIGENCE**” is done by me under the guidance of **Dr.N.Srinivasan, M.E., Ph.D.**, Department of Computer Science and Engineering at Sathyabama Institute of Science and Technology is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering.

DATE:

PLACE: CHENNAI

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** 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, **Dr. S. Vigneswari, M.E., Ph.D.,** and **Dr. L. Lakshmanan, 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.N.Srinivasan, M.E., Ph.D.,** for her 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. Artificial Intelligence 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 project is to summarize some of the current research on predicting heart diseases using data mining techniques analyses 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.

TABLE OF CONTENTS

ABSTRACT	iv
LIST OF FIGURES	vii
LIST OF ABBREVIATIONS	viii

CHAPTER NO.	TITLE	PAGE NO
1.	INTRODUCTION	1
	1.1 INTRODUCTION	1
	1.2 TYPES OF HEART DISEASES	2
	1.3 PREVALENCE OF HEART DISEASES	3
	1.4 APPLICATION OF AI IN HEART DISEASE	3
2.	LITERATURE SURVEY	6
3.	METHODOLOGY	12
	3.1 SYSTEM SPECIFICATION	12
	3.1.1 <i>HARDWARE SPECIFICATION</i>	12
	3.1.2 <i>SOFTWARE SPECIFICATION</i>	12
	3.2 SYSTEM IMPLEMENTATION	13
	3.2.1 <i>DESIGN STRUCTURE</i>	13
	3.2.1.1 <i>INPUT DESIGN</i>	13
	3.2.1.2 <i>OUTPUT DESIGN</i>	14
	3.3 PROBLEM STATEMENT	14
	3.4 MODULES	15
	3.4.1 <i>DATA PREPROCESSING</i>	15
	3.4.2 <i>MODEL EVALUATION</i>	16
	3.4.3 <i>CLASSIFICATION</i>	17
	3.4.4 <i>PREDICTION</i>	18
	3.5 ARCHITECTURE DIAGRAM	19
	3.6 FLOW DIAGRAM	22
	3.6.1 <i>FLOW CHART</i>	23

	3.7 ALGORITHMS USED	24
	3.7.1 SVM ALGORTIHM	24
	3.7.1.1 FEAURES OF SVM	24
	3.7.1.2 WORKING OF SVM	25
	3.7.2 RANDOM FOREST ALGORITHM	27
	3.8 FINDING ACCURACY	29
4.	RESULTS AND DISCUSSION	32
5.	CONCLUSION AND FUTURE WORK	33
	REFERENCES	35
	APPENDICES	36
	A.SAMPLE CODE	36
	B.SCREENSHOTS	40
	C.PUBLICATION WITH PLAGIARISM REPORT	43

LIST OF FIGURES

FIGURE NO.	FIGURE NAME	PAGE NO.
3.1	ARCHITECTURE DIAGRAM OF HEART DISEASE PREDICTION MODEL	19
3.2	FLOW DIAGRAM	22
3.3	CLASS OF DATASETS	23
3.4	GRAPH OF HYPER PLANES	26
3.5	RANDOM FOREST MODEL	28
(a)	HEART DISEASE DETECTION HOME PAGE	40
(b)	NO HEART DISEASE	40
(c)	STAGE 1 HEART DISEASE	41
(d)	STAGE 2 HEART DISEASE	41
(e)	STAGE 3 HEART DISEASE	42

LIST OF ABBREVIATIONS

- i. SVM - SUPPORT VECTOR MACHINE
- ii. RF - RANDOM FOREST
- iii. AI - ARTIFICIAL INTELLIGENCE
- iv. CVD - CARDIOVASCULAR DISEASE

CHAPTER 1

INTRODUCTION

1.1 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.

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 are as follows.

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.

Artificial intelligence (AI)-based applications have found widespread applications in many fields of science, technology, and medicine. The use of enhanced computing power of machines in clinical medicine and diagnostics has been under exploration since the 1960s.

More recently, with the advent of advances in computing, algorithms enabling machine learning, especially deep learning networks that mimic the human brain in function, there has been renewed interest to use them in clinical medicine. In cardiovascular medicine, AI-based systems have found new applications in cardiovascular imaging, cardiovascular risk prediction, and newer drug targets. This article aims to describe different AI applications including machine learning and deep learning and their applications in cardiovascular medicine.

AI-based applications have enhanced our understanding of different phenotypes of heart failure and congenital heart disease. These applications have led to newer treatment strategies for different types of cardiovascular diseases, newer approach to cardiovascular drug therapy and post marketing survey of prescription drugs. However, there are several challenges in the clinical use of AI-based applications and interpretation of the results including data privacy, poorly selected/outdated data, selection bias, and unintentional continuance of historical biases/stereotypes in the data which can lead to erroneous conclusions. Still, AI is a transformative technology and has immense potential in health care.

1.2 TYPES OF HEART DISEASES

There are five types of cardiovascular diseases. It includes chest pain, Nausea, Indigestion, Heartburn and stomach pain, pain in the arms, feeling dizzy and light headed ,fatigue.

Chest pain 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.

Nausea, Indigestion, Heartburn and Stomach Pain are some of the often overlooked symptoms of heart attack. Women tend to show these symptoms more than men.

Pain in the Arms is the pain often starts in the chest and then moves towards the arms, especially in the left side. Feeling Dizzy and Light Headed are the things that lead to the loss of balance and Fatigue is the simple chore which begin to set a feeling of tiredness should not be ignored.

1.3 PREVALENCE OF HEART 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.

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. Stroke happens when something slows or blocks blood flow to your brain.

Your brain can't get the oxygen and nutrients it needs, and brain cells start to die. When blood can't get to the part of your brain that controls a certain function, your body doesn't work like it should. A stroke can happen because of a blocked artery or a leaking or burst blood vessel. It needs immediate treatment to limit brain damage and other complications. Stroke is the leading cause of disability and one of the top causes of death in the United States.

1.4 APPLICATION OF AI IN HEART DISEASE

In general, governments and authorities around the world have developed and implemented laws and regulations to establish a better medical environment for patients. The current medical system still has many shortcomings, including uneven distribution of senior clinicians, high rate of misdiagnosis of primary clinicians, long training period for clinicians, shortage of clinicians in undeveloped areas, and high medical expenses for patients.

However, with the development and advancement of machine learning technology in recent years, AI has gradually changed from theory to practicality. The multiple applications of AI in medicine are being proven. Additionally, AI technology has become an important factor that may affect the development of the medical industry and may improve the level of medical services.

AI has several applications in the medical field. Firstly, AI can help clinicians diagnose disease and optimize treatment processes. After being applied to traditional medical procedures, AI can reduce the rate of misdiagnosis and improve diagnostic efficiency.

Then, with the advent of deep learning, AI has the ability to recognize medical images and provide clinicians with more reliable imaging diagnostic information. Thirdly, by using big AI algorithms can often provide more accurate results for patient prediction. AI can also help support drug research and improve the efficiency of new drug development.

Finally, the combination of AI and surgical robots will improve the accuracy of many complex and difficult operations. With the development of AI, big data analysis and cloud computing technologies, AI can provide patients with high quality medical services.

Besides, AI will enhance the advancement of smart medical and precision medicine, which will reduce patient waiting time and cost, and obtain safe, convenient and high quality medical services. Therefore, the implementation of deep learning of AI will benefit cardiovascular medicine.

At present, the application of AI in cardiovascular medicine is very close to reality. Nowadays, getting personal medical advice from virtual assistants such as Alexa or Siri is no longer a far-fetched attachment. However, some researchers are concerned about the rapid development of AI, and they believe that guidelines should be created in artificial-intelligence research to protect human rights.

Besides, there are still several unsolved technical problems in the clinical application of AI. The human brain can solve highly abstract reasoning problems by using physical neural networks. However, the mechanism of the network is only partially understood; therefore, artificial networks can only provide narrow insights.

A key factor in preventing the implementation of machine learning in medicine is that the training of a model requires massive datasets, which are not easily accessible. Additionally, without a large amount of high-quality data, the model can likely be influenced, which will lead to biases and inevitably cause the failure of the model when it makes predictions about future data.

Furthermore, the difficulty of AI in making differential diagnoses is another major problem, which can be attributed to the relatively slow development of AI use in medicine. Ultimately, the superficial knowledge of the human brain's processing, the poor data quality, the complex computational processes, the difficulty in measuring anti diastole and the inaccessibility to traditional clinical workflow have led to controversy in the clinical application of AI.

CHAPTER 2

LITERATURE SURVEY

There are few research papers that explore the computational methods to predict heart diseases. The summaries of them have been presented in a nutshell.

1. 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. 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 having 13 neurons, 6 neurons and 2 neurons respectively. Learning Vector Quantization was used in this study. It is a special case of an artificial neural network that applies a prototype-based supervised classification algorithm. C programming

was used as a tool to implement heart disease classification and prediction trained via artificial neural network.

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.

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 :

MA. 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 casualties 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.

CHAPTER 3

METHODOLOGY

3.1 SYSTEM SPECIFICATION

3.1.1 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

3.1.2 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
- Pandas 19.0.1
- Sea born 0.7.1
- Scipy 0.18.1
- Sci-kit learn 0.18.1
- Matplotlib

3.2 SYSTEM IMPLEMENTATION

Design is a multi- step that focuses on data structure software architecture, procedural details, algorithm etc... and interface between modules. The design process also translates 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.

3.2.1. DESIGN STRUCTURE

3.2.1.1 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.

3.2.1.2 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.

3.3 PROBLEM STATEMENT

Heart disease can be managed effectively with a combination of lifestyle changes, medicine and in some cases, surgery. With the right treatment, the symptoms of heart disease can be reduced and the functioning of the heart improved. The predicted results can be used to prevent and thus reduce cost for surgical treatment and other expensive.

The overall objective of my work will be to predict accurately with few tests and attributes the presence of heart disease. Attributes considered form the primary basis for tests and give accurate results more or less. Many more input attributes can be taken but our goal is to predict with few attributes and faster efficiency the risk of having heart disease. Decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the data set and databases.

This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs.

According to the opportunities to improve care and reduce costs concurrently could apply to as much as 30% of overall healthcare spending. The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors.

3.4 MODULES

Here there are four modules involved. Following are the four modules used:

- 1) Data pre-processing
- 2) Model evaluation
- 3) Classification
- 4) Prediction

3.4.1 Data preprocessing

Here the raw data in the heart disease data is cleaned and the metadata is appended 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 metadata into this and then this metadata will be attached to the data and replace the converted data with metadata.

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.

The classifier is unable to process the raw dataset because of some of its symbolic features. Thus, pre-processing is essential, in which non-numeric or symbolic features are eliminated or replaced, because they do not indicate vital participation in intrusion detection.

However, this process generates overhead including more training time; the classifier's architecture becomes complex and wastes memory and computing resources. Therefore, the non-numeric features are excluded from the raw dataset for improved performance of intrusion detection systems.

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 artificial intelligence 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.

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 and if it is functioning in its optimum, a technique named parameter tuning was used. Scikit-learn library contains the class called Grid Search CV which performs the parameter tuning.

3.4.2 Model evaluation

We apply the algorithm for testing part and get the accuracy of this model. So, we find accuracy of the model by collecting the random samples from the patients. From the accuracy we will notice which algorithm works better. Here we used SVM (support vector machine) algorithm and RF (random forest) algorithm. These algorithms are used for testing part which comes under model evaluation.

Results obtained for heart disease data sets with SVM and Random Forest using different kernel functions like linear, polynomial, radial basis and sigmoid are tabulated. It is noticed that there is a varying accuracy of classification with different probabilistic estimate with different kernel function.

Results are observed to be much better with Radial basis function with SVM and with certain set of parameters and these are comparable with Random Forest technique.

The random forest algorithm possesses a number of appealing properties making it well-suited for classification of microarray data:

- (i) It is applicable when there are more predictors than observations
- (ii) It performs embedded gene selection and it is relatively insensitive to the large number of irrelevant genes
- (iii) It incorporates interactions between predictors,
- (iv) It is based on the theory of ensemble learning that allows the algorithm to learn accurately both simple and complex classification functions,
- (v) It is applicable for both binary and multi category classification tasks
- (vi) According to its inventors it does not require much fine-tuning of parameters and the default parameterization often leads to excellent performance.

Recent work reported an empirical evaluation of random forests in the cancer microarray gene expression domain and concluded that random forest classifiers have predictive performance comparable to that of the best performing alternatives for classification of microarray gene expression data

3.4.3 Classification

Classification is a classic data mining technique based on artificial intelligence. 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.

Placing an activity into normal and intrusive categories is the core function of an intrusion detection system, which is known as an intrusive analysis engine. Thus, different classifiers have been applied as intrusive analysis engines in intrusion detection in the literature, such as multilayer perceptron, SVM, RF.

3.4.4 Prediction

The prediction as its name implied is one of a data mining 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 dataset values and finally it will predict the disease.

3.5 ARCHITECTURE DIAGRAM

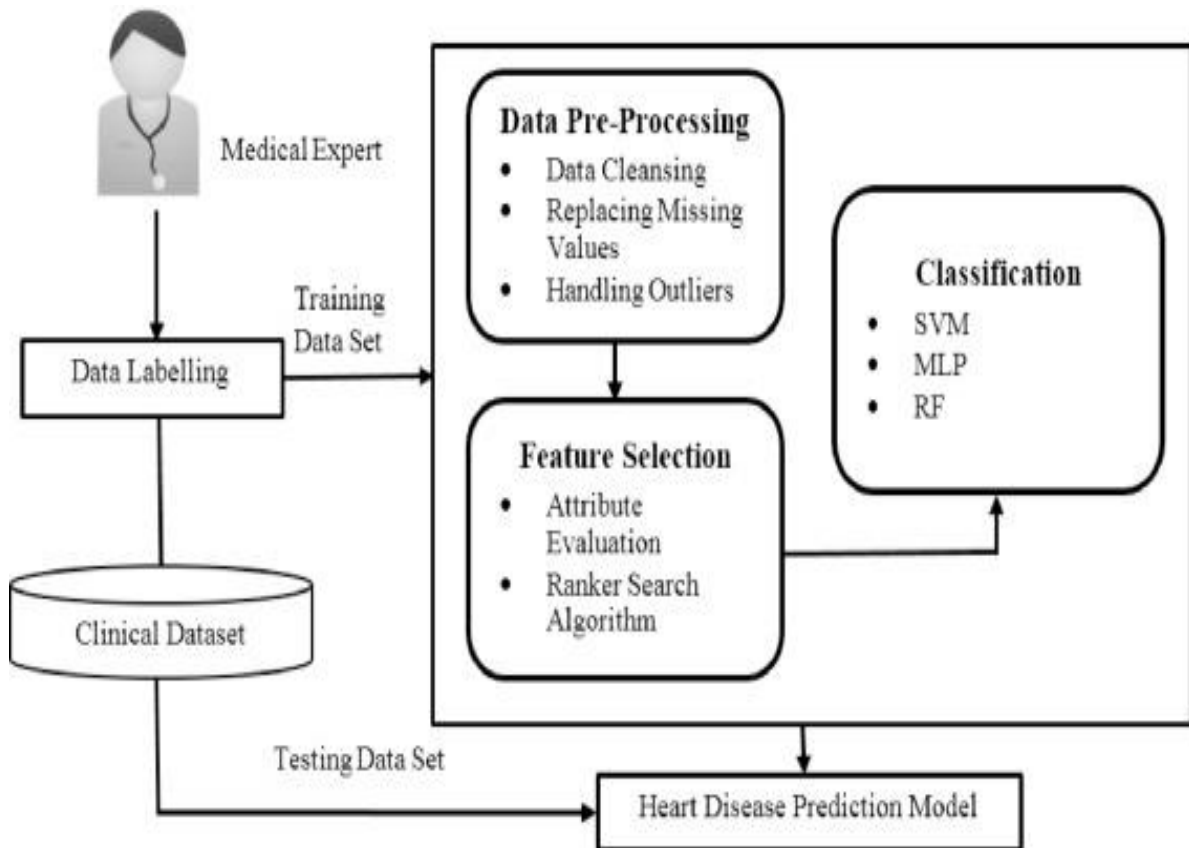


Fig 3.1 Architecture diagram of heart disease prediction model

Data set 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.

Our dataset is based on UCI heart Disease Data Set and we have 303 instances. According to UCI, "This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. "We guess too many features will bring too much noise so people has done feature extraction and reduce 76 features to 14 features.

To better understand the meaning of the features, we have the responsibility to explain some of the attributes of original dataset from UCI as follows

- age: age in years
- sex: sex (1 = male; 0 = female)
- cp: chest pain type -- Value 0: typical angina -- Value 1: atypical angina -- Value 2: non-anginal pain -- Value 3: asymptomatic
- trestbps: resting blood pressure (in mm Hg on admission to the hospital)
- chol: serum cholestoral in mg/dl
- fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- target: Heart disease (0 = no, 1 = yes)

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.

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.

Dataset selection for experimentation is a significant task, because the performance of the system is based on the correctness of a dataset. The dataset can be collected by numerous means, such as sanitized dataset, simulated dataset, tested dataset and standard dataset. However, complications occur in the application of the first three methodologies.

A real traffic method is expensive, whereas the sanitized method is unsafe. The development of a simulation system is also complex and challenging. Additionally, different types of traffic are required to model various network attacks, which is complex and costly. To overcome these difficulties, dataset is used to validate the proposed system for intrusion detection.

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.

Data set 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.

3.6 FLOW DIAGRAM

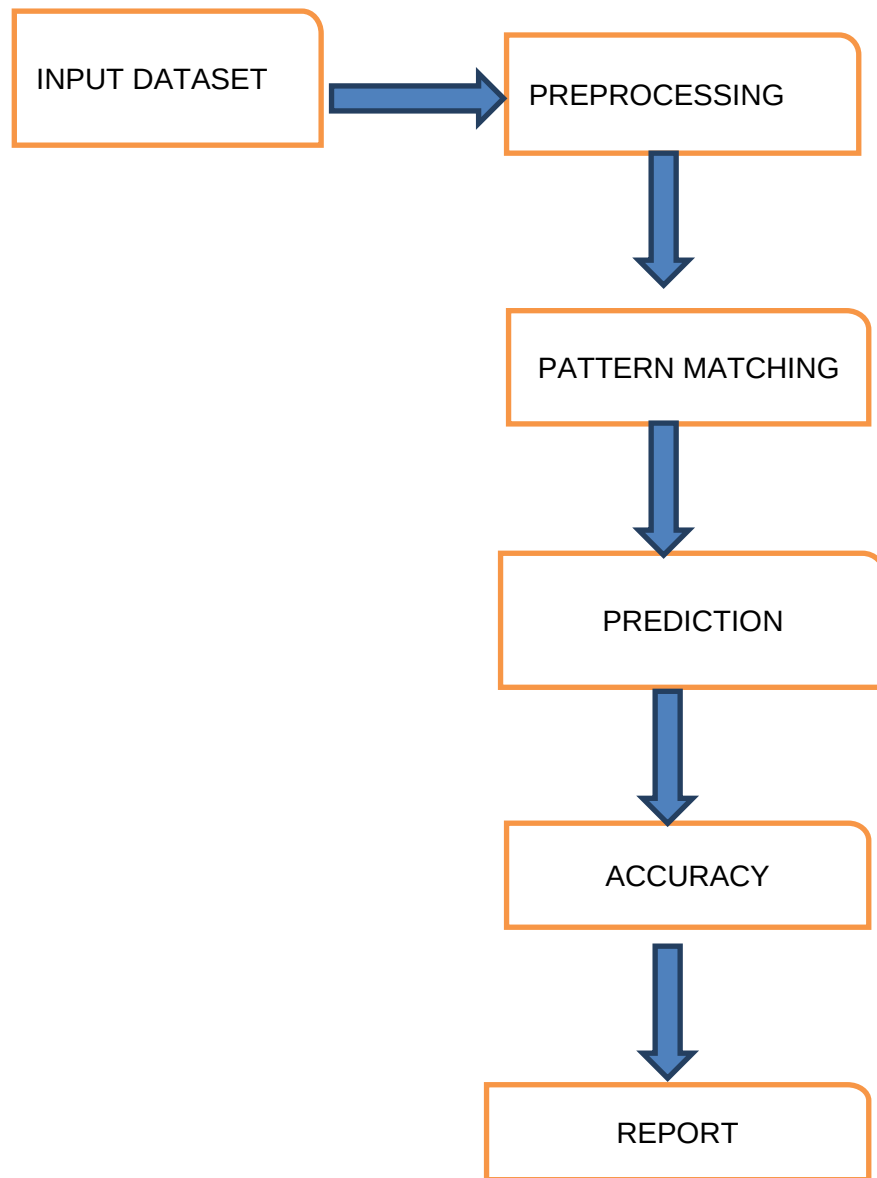


Fig 3.2 Flow diagram

3.6.1 Flow chart

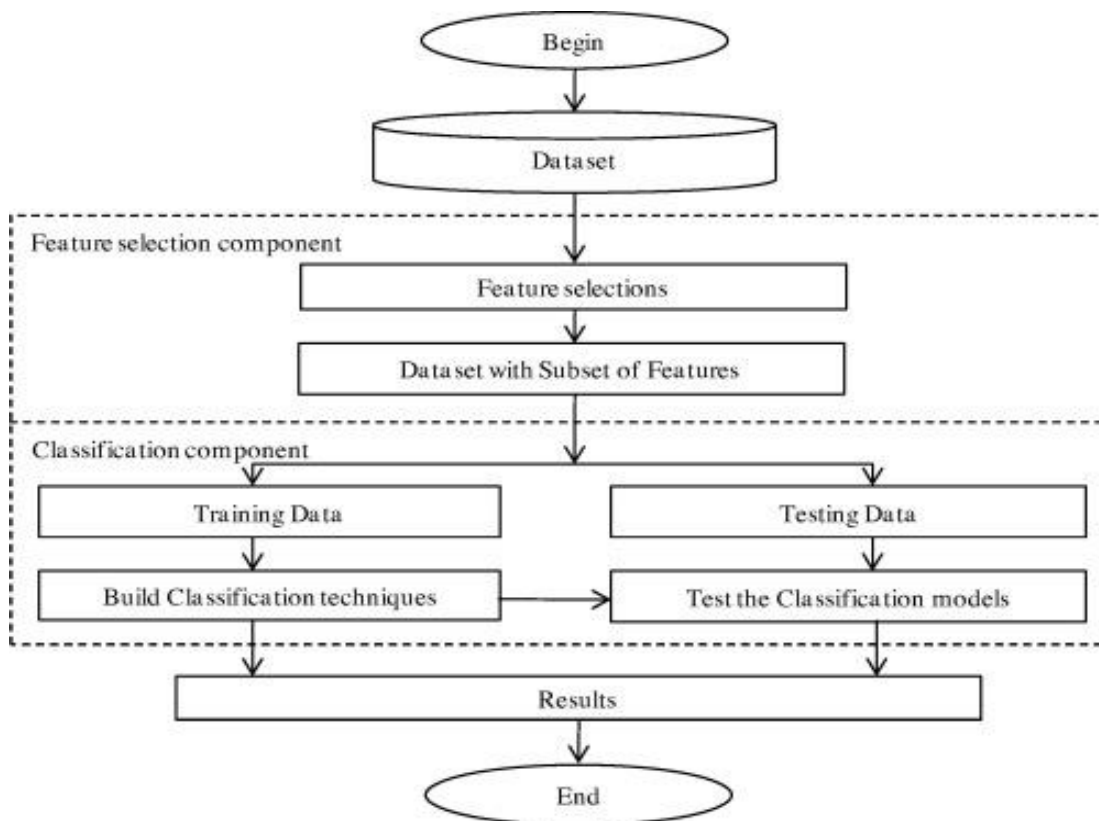


Fig 3.3 Classification of datasets

From the fig 3.3, first the raw data is collected from the samples and then after removing unwanted data from the raw data all the data is saved as dataset. Dataset is sent for feature selection; on this process dataset with subset of features is separated from the main dataset. After feature selection classification is done on splitting into training data and testing data. In training data we build the techniques for classification and in testing part we test the classification models. Now, from these methods we will obtain the result.

Here the raw data in the heart disease data is cleaned and the metadata 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 metadata into this and then this metadata will be attached to the data and replace the converted data with metadata.

3.7 ALGORITHMS USED

In this heart disease detection model we used two algorithms namely SVM algorithm and Random forest algorithm. By finding accuracy we will differentiate between these algorithms. Now, we will discuss about these two algorithms in brief.

3.7.1 SVM algorithm

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Artificial intelligence and Machine Learning. 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. This best decision boundary is called a hyper plane.

It is a supervised learning method which classifies data into two classes over a hyper plane. Support vector machine performs a similar task like C4.5 except that it doesn't use Decision trees at all. Support vector machine attempts to maximize the margin (distance between the hyper plane and the two closest data points from each respective class) to decrease any chance of misclassification. Some popular implementations of support vector machine are scikit-learn, MATLAB and of LIBSVM.

3.7.1.1 Features of SVM

- SVM is a supervised learning algorithm. This means that SVM trains on a set of labelled data.
- SVM studies the labelled training data and then classifies any new input data depending on what it learned in the training phase.
- A main advantage of SVM is that it can be used for both classification and regression problems. Though SVM is mainly known for classification, the SVR is used for regression problems.
- SVM can be used for classifying non-linear data by using the kernel trick. The kernel trick means transforming data into another dimension that has a

clear dividing margin between classes of data. After which you can easily draw a hyper plane between the various classes of data.

3.7.1.2 Working of SVM

A support vector machine is a type of model used to analyse data and discover patterns in classification and regression analysis. Support vector machine (SVM) is used when your data has exactly two classes. An SVM classifies data by finding the best hyper plane that separates all data points of one class from those of the other class. The larger margin between the two classes, the better the model is.

A margin must have no points in its interior region. The support vectors are the data points that on the boundary of the margin.

SVM is based on mathematical functions and used to model complex, and real world problems. SVM performs well on data sets that have many attributes, such as the CHDD. Support Vector Machines map the training data into kernel space.

There are many differently used kernel spaces – linear (uses dot product), quadratic, polynomial, Radial Basis Function kernel, Multilayer Perceptron kernel, etc. to name a few. In addition, there are multiple methods of implementing SVM, such as quadratic programming, sequential minimal optimization and least squares.

The challenging aspect of SVM is kernel selection and method selection such that your model is not over optimistic or pessimistic. Considering that the CHDD has a large number of instances as well as features, it is arguable whether the kernel chosen is RBF or linear.

Although the relation between the attributes and class labels are nonlinear, due to the large number of features, RBF kernel may not improve performance. It is recommended that both kernels be tested and the more efficient one be finally selected.

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. This best decision boundary is called a hyper plane.

SVM chooses the extreme points/vectors that help in creating the hyper plane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyper plane

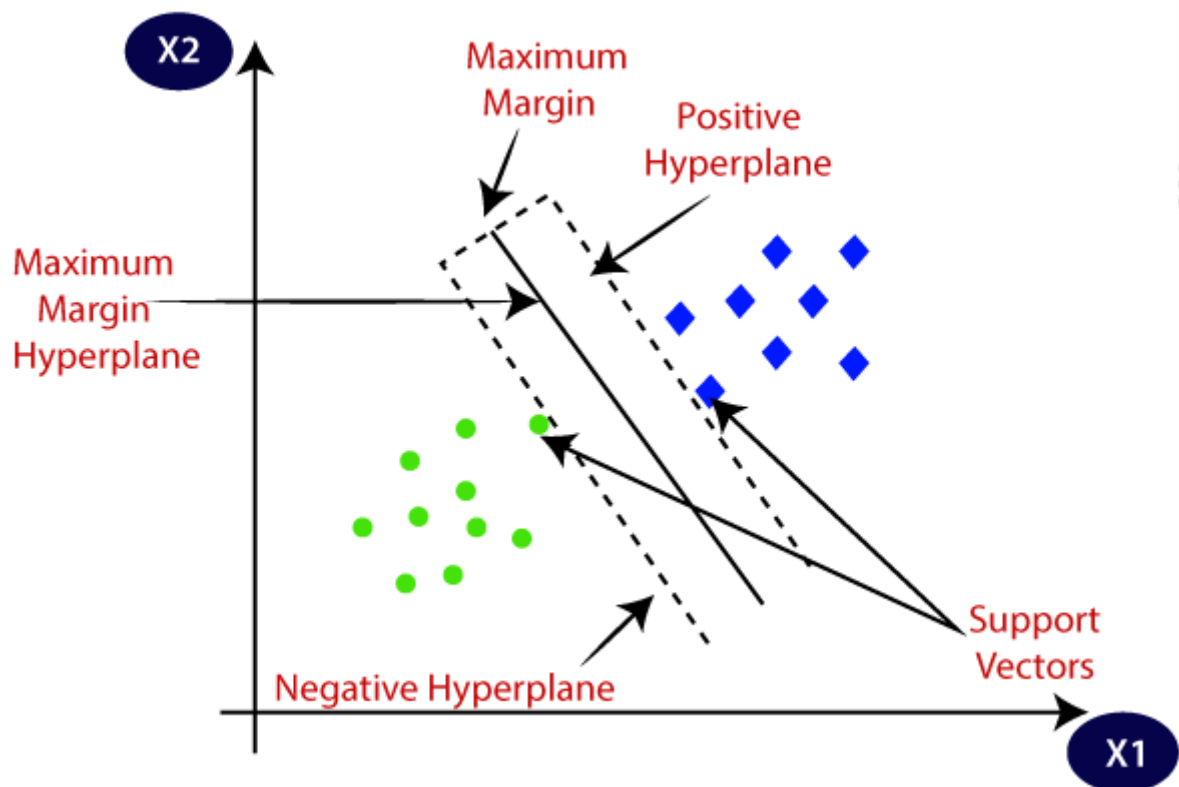


Fig 3.4 Graph of hyper planes

There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyper plane of SVM. The dimensions of the hyper plane depend on the features present in the dataset, which means if there are 2 features (as shown in image), then hyper plane will be a straight line. And if there are 3 features, then hyper plane will be a 2-dimension plane. We always create a hyper plane that has a maximum margin, which means the maximum distance between the data points.

The data points or vectors that are the closest to the hyper plane and which affect the position of the hyper plane are termed as Support Vector. Since these vectors support the hyper plane, hence called a Support vector.

3.7.2 RANDOM FOREST 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. 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

- Select random samples from a given dataset.
- Construct a decision tree for each sample and get a prediction result from
- Perform a vote for each predicted result.
- Select the prediction result with the most votes as the final prediction.
- Perform a vote for each predicted result.
- Select the prediction result with the most votes as the final prediction.

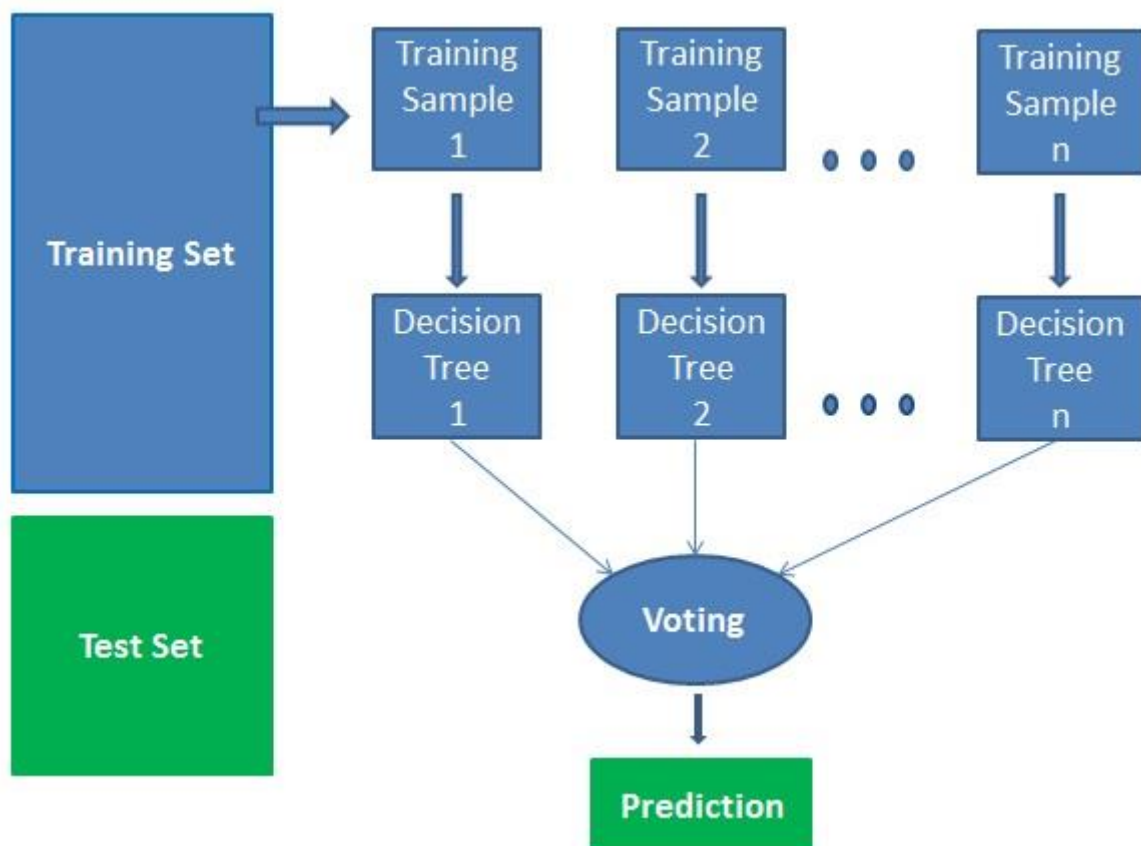


Fig 3.5 Random forest model

Advantages

- Random forest 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 over fitting 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.

3.8 FINDING ACCURACY

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:

Accuracy=Number of correct predictions/Total number of predictions

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

Accuracy=(TP+TN)/(TP+TN+FP+FN)

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Let's try calculating accuracy for the following model that classified 100 tumor samples

True Positive (TP)

Reality: Malignant; AI model predicted: Malignant; Number of TP results: 1

False Positive (FP)

Reality: Benign; AI model predicted: Malignant; Number of FP results: 1

False Negative (FN)

Reality: Malignant; AI model predicted: Benign; Number of FN results: 8

True Negative (TN)

Reality: Benign; ML model predicted: Benign; Number of TN results: 90

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

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.

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 goes.

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. Accuracy can be computed by comparing actual test set values and predicted values. Well, you got a classification rate of 96.49%, considered as very good accuracy.

For further evaluation, you can also check precision and recall of model. Since our project is a classification problem, we use accuracy, precision; recall and F1 score to evaluate the models.

We would like to introduce the meaning of TP, FP, TN and FN. A true positive (TP) is a positive outcome predicted by the model correctly while a false positive (FP) is a positive outcome predicted by the model incorrectly. A true negative (TN) is a negative outcome predicted by the model correctly while a false negative (FN) is a negative outcome predicted by the model incorrectly. We did not use cross-validation because our dataset is not very sufficient. We split the dataset into 80% for training and 20% for test. Here is the table of results of different methods and we will talk about each evaluation of methods in details.

CHAPTER 4

RESULTS AND DISCUSSION

This project summarizes some of the recent works done in data mining related to cardiovascular diseases. Data mining 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.

Python is chosen in most of the research work for practical execution of the project. Weka, Tanagra, Mat lab etc. are some of the other popular tools used for data analysis. 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, Navies Bayes, and many. We used SVM and Random forest algorithm techniques.

The proposed system is GUI-based, user-friendly, scalable, reliable and an expandable system. The proposed working model can also help in reducing treatment costs by providing Initial diagnostics in time. General physicians can utilize this tool for initial diagnosis of cardio-patients.

There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. As we have developed a generalized system, in future we can use this system for the analysis of different data sets.

CHAPTER 5

CONCLUSION AND FUTURE WORK

The objective of this research work is to provide an insight of different data mining techniques that can be employed in automated heart disease prediction systems. Heart disease is one of the leading causes of deaths worldwide and the early prediction of heart disease is very important. The computer aided heart disease prediction system facilitates the physician as a tool for heart disease diagnosis.

Some of the Heart Disease classification systems were reviewed in this study and based on different research studies it was concluded that data mining plays a major role in heart disease classification. Neural Network with offline training is a good tool for disease prediction at early stage. The good performance of the system can be obtained by preprocessed and normalized dataset.

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 Hanna, Dec 2010, "Prediction of Heart Disease Medical Prescription Function", IEEE International Conference on Computational Intelligence, 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, 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, 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, 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", (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.

APPENDICES

A. SAMPLE CODE

```
# Import SVM from sklearn import svm# Creating a SVM Classifier

Classifier = svm

SVC(C=0.01, break_ties=False, cache_size=200, class_weight=None,
coef0=0.0, decision_function_shape='ovr', degree=1, gamma='scale',
kernel='rbf',max_iter=-1,probability=False,random_state=None,
shrinking=True,tol=0.001, verbose=False)

# Training the model

classifier.fit(X_train, y_train)

# importing required libraries

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# reading csv file and extracting class column to y.
x = pd.read_csv("C:\...\cancer.csv")
a = np.array(x)
y = a[:,30] # classes having 0 and 1

# Import train_test_split function
from sklearn.model_selection import train_test_split

# Split dataset into training set and test set
X_train, X_test, y_train, y_test = train_test_split(heart_disease.data, heart
disease.target, test_size=0.3,random_state=109) # 70% training and 30%
test

#Import svm model
from sklearn import svm
```

```

#Create a svm Classifier
clf = svm.SVC(kernel='linear') # Linear Kernel

#Train the model using the training sets
clf.fit(X_train, y_train)

#Predict the response for test dataset
y_pred = clf.predict(X_test)

# extracting two features
x = np.column_stack((x.malignant,x.benign))

# import support vector classifier
# "Support Vector Classifier"
from sklearn.svm import SVC
clf = SVC(kernel='linear')

# fitting x samples and y classes
clf.fit(x, y)

# 569 samples and 2 features
x.shape

print (x),(y)

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'])
    gender = int(request.form['gender'])
    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, gender, 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)
```

B.SCREENSHOTS

Heart Disease Diagnosis

Enter your age: Your current age in years

Enter your Gender: Male

Resting blood pressure (in mm Hg on admission to the hospital): Your Blood Pressure

Serum Cholesterol in mg/dl: Cholesterol

Fasting blood sugar > 120mg/dl: Yes

Rest ECG results: Normal

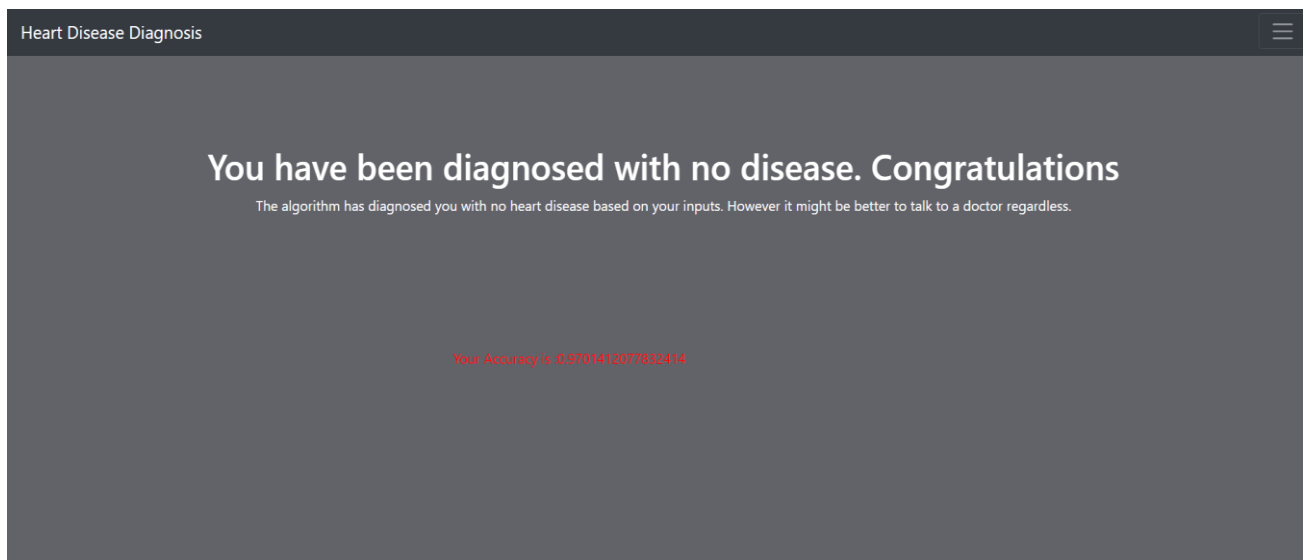
Maximum heart rate achieved during ecg: Max heart rate

Chest pain during exercise?: Yes

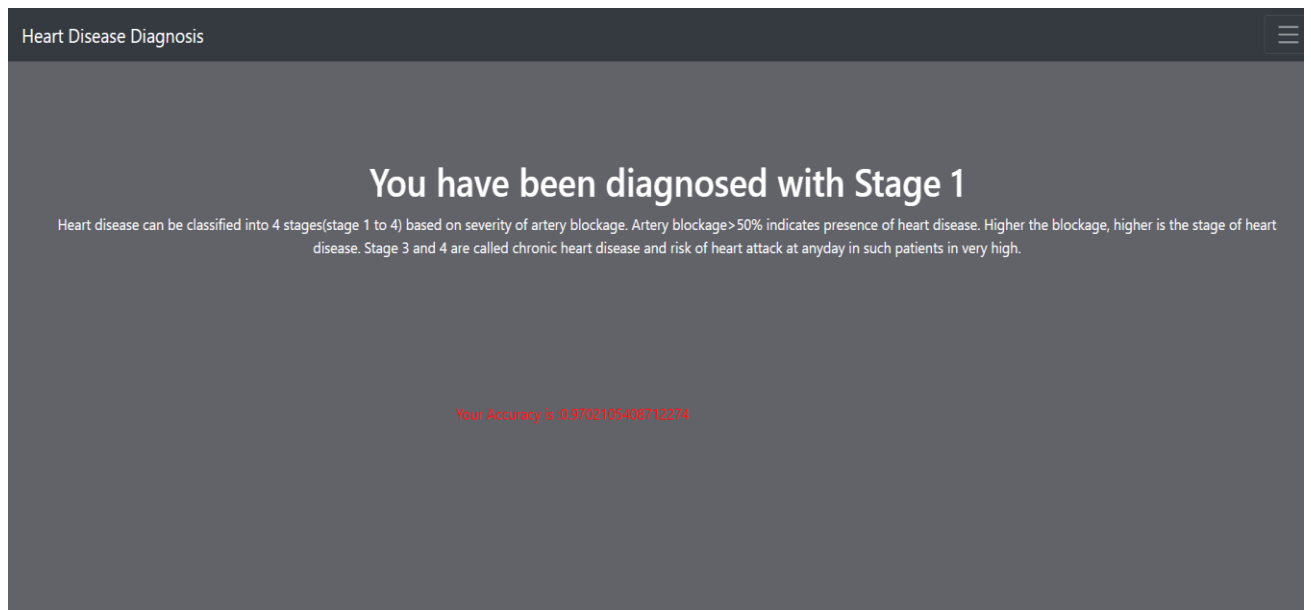
Chest pain type?: No chest pain

Submit

fig(a) Heart disease detection system home page



fig(b) No Heart disease



Fig(c) Stage 1 Heart disease

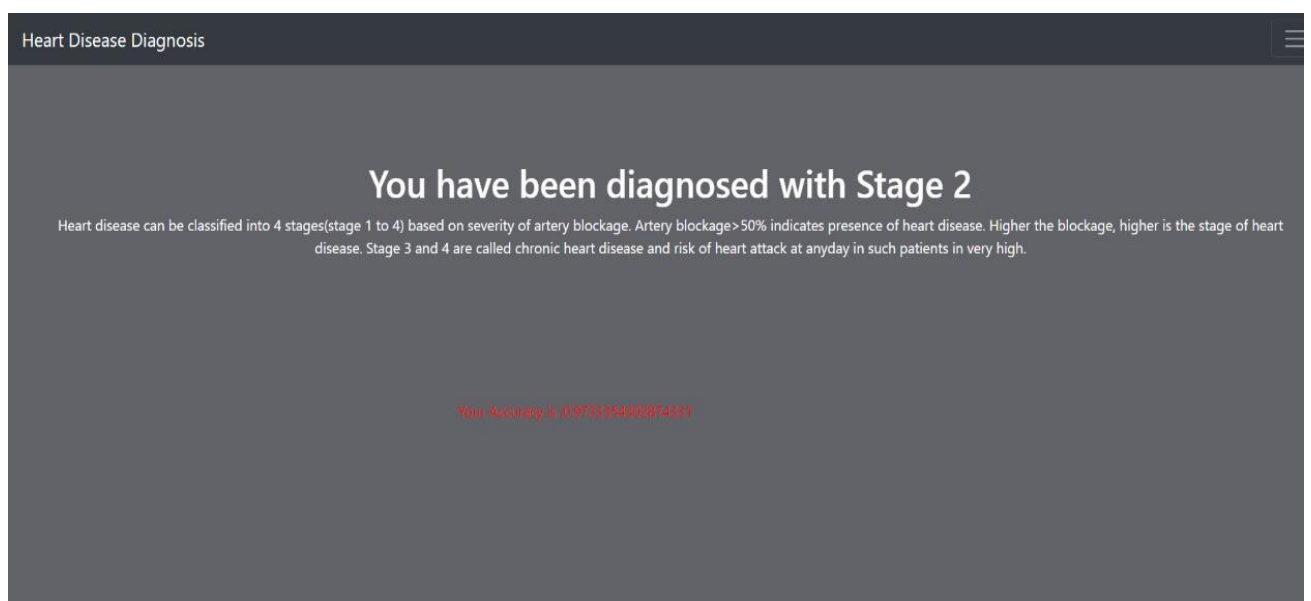


Fig (d) Stage 2 Heart disease

You have been diagnosed with Stage 3

Heart disease can be classified into 4 stages(stage 1 to 4) based on severity of artery blockage. Artery blockage>50% indicates presence of heart disease. Higher the blockage, higher is the stage of heart disease. Stage 3 and 4 are called chronic heart disease and risk of heart attack at anyday in such patients is very high.

Your Accuracy is 0.975184679309494

Fig(e) Stage 3 Heart disease

C. PLAGIARISM REPORT

Text	Paragraph	Synopsis
Submission date: 13-Mar-2021 09:23AM (UTC-0800) Submission ID: 1532080731 File name: HDD_paper.docx (22.13K) Word count: 1630 Character count: 9415		
Heart Disease Detection Using Artificial Intelligence <u>P.Shivani', K.Sahithi', N.Srinivasan',</u> <u>U.G Student' ', Professor', Department of CSE, Sathyabama Institute of Science and</u> <u>Technology, Chennai.</u> <u>podilashivani%ginail.com',sahithichowdarykoti2%gmail.com',srinis'asan.cse%sathyabama.a</u> <u>c.in'</u> ABSRTACT <u>Hean Disease illness is quite possibly the most perplexing errands in the clinical field.</u> <u>Presently, around I individual has kicked the bucket of heart disease. Since the forecast of heart</u> <u>disease is a troublesome assignment, it is important to empower the expectation interaction to</u> <u>dodge the dangers in question and to educate patients as quickly as time permits. The proposed</u>		

HDD paper.docx

ORIGINALITY REPORT

11%

SIMILARITY INDEX

10%

INTERNET SOURCES

3%

PUBLICATIONS

5%

STUDENT PAPERS

PRIMARY SOURCES

1

www.ripublication.com

Internet Source

9%

2

Submitted to Nottingham Trent University

Student Paper

1%

3

iieta.org

Internet Source

1%

4

www.ijcse.net

Internet Source

1%