IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

HEART DISEASE PREDICTION SYSEM

ADITYA MANI[1], SAKSHI SINHA[2], ANMOL[3], Prof. DR. BINDU GARG [4]

[1]STUDENT, DEPARTMENT OF COMPUTER SCIENCE ENGINEERING BHARATI VIDYAPEETH UNIVERSITY COLLEGE OF ENGINEERING PUNE, INDIA

[2]STUDENT, DEPARTMENT OF COMPUTER SCIENCE ENGINEERING BHARATI VIDYAPEETH UNIVERSITY COLLEGE OF ENGINEERING PUNE, INDIA

[3]STUDENT, DEPARTMENT OF COMPUTER SCIENCE ENGINEERING BHARATI VIDYAPEETH UNIVERSITY COLLEGE OF ENGINEERING PUNE, INDIA

[4]ASSISTANT PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE ENGINEERING BHARATI VIDYAPEETH UNIVERSITY COLLEGE OF ENGINEERING PUNE, INDIA

Abstract— Healthcare is an inevitable task to be done in humans daily life. Cardiovascular disease is a category for a range of diseases that are affecting heart and blood vessels. The early methods of forecasting the cardiovascular diseases helpful in making decisions about the changes to have occurred in high-risk patients which results in the reduction of risks. The health care industry contains medical data, therefore machine learning algorithms are required to make the decisions effectively in the prediction of heart diseases. Recently research has delved into uniting these techniques to provide hybrid machine learning algorithms. In the proposed research, data pre-processing uses techniques like the removal of noisy data, removal of missed data, filling default value if applicable and classification of attributes for prediction and decision making at different levels. The performance of the diagnosis model is obtained by using the methods like classification, accuracy, sensitivity analysis. This project proposes a prediction model to predict whether a people have any heart disease or not and to provide an diagnosis on that. This is done basically by comparing the accuracies of applying rules to the individual results of Support Vector Machine, Gradient Boosting, Random forest, Naive Bayes classifier and logistic regression on the dataset taken in a region to present an accurate model of predicting cardiovascular disease.

Keywords— *Heart Disease, Heart Diseease Prediction, Machine Learning*

I. INTRODUCTION

The heart is a muscular organ which pumps blood into the body and is the central part of the body's cardiovascular system which also contains lungs. Cardiovascular system comprises a network of blood vessels, for example, veins, arteries, and capillaries. These blood vessels delivers blood all over the body. Abnormalities in normal blood flow from the heart cause several types of heart diseases which are commonly called as cardiovascular diseases (CVD). Due to the development of advance healthcare systems, lots of patient data are available (i.e . Big Data in Electronic Health Record System) which can be used for designing predictive models for Cardiovascular diseases. Data mining or machine learning is a discovery method for analyzing big data from an assorted perspective and encapsulating it into information. "Data Mining is a non-trivial extraction of implicit, previously unknown and potentially useful information about the data". Nowadays, a huge data pertaining to disease diagnosis, patients etc. are generated by healthcare industries. Data mining provides a number of techniques which discover hidden patterns or similarities from the data. Therefore, in this paper, a machine learning algorithm is proposed for the implementation of a heart disease prediction system which has been validated on two open access heart disease prediction datasets.

The heart pumps blood with a rhythm determined by a group of pacemaking cells into the sinoatrial node. These generate a current that causes contraction of heart, travelling through the atrioventricular node and along the conduction system of the heart. The heart receives blood low in oxygen from the systemic circulation, which enters the right atrium from the superior and inferior venae cavae and then passes to the right ventricle. From here it is pumped into the pulmonary circulation, through the lungs and where it receives oxygen and gives off carbon dioxide. Oxygenated blood then returns to the left atrium, passes through the left ventricle and is pumped out through the aorta to the systemic circulation−where the oxygen is used and metabolized to carbon dioxide. The heart beats at a resting rate that is close to 72 beats per minute. Exercise temporarily increases the rate, but lowers resting heart rate in the long term, and is good for the heart health.

Problem Statement

Cardiovascular diseases are the most common cause of death over the last few decades in the developed as well as underdeveloped and developing countries. Early detection of cardiac diseases and continuous supervision of the clinicians can reduce the mortality rate. However, accurate detection of heart diseases in all cases and consultation of a patient for 24 hours by a doctor is not available since it requires more time and expertise.

Literature Survey

There are numerous works has been done related to disease prediction systems using different data mining techniques and machine learning algorithms in medical centres.

Review of Existing Models, Approaches, Problems:-

Different researchers have been contributed for the development of this field. Predication of heart disease based upon machine learning algorithm is always curious case for researchers recently there is a wave of papers and

research material on this area. Our goal in this chapter is to bring out all state of art work by different authors and researchers.

 K. Polaraju et al, proposed Prediction of Heart Disease using Multiple Regression Model and it proves that Multiple Linear Regression is appropriate for predicting heart disease chance. The work is performed using training data set consists of 3000 instances with 13 different attributes which has mentioned earlier. The data set is divided into two parts that is 70 percent of the data are used for training and 30 percent used for testing. Based on the results, it is clear that the classification accuracy of Regression algorithm is better compared to other algorithms.

 Marjia et al, developed the heart disease prediction using KStar, j48, SMO, and Bayes Net and Multilayer perception using WEKA software tools. Based on performance from different factor SMO and Bayes Net achieve optimum performance than KStar, Multilayer perception and the J48 techniques using k-fold cross validation. The accuracy performances achieved by those algorithms are still not satisfactory. Therefore, the accuracy's performance is improved more to give better decision to diagnosis disease

 S. Seema et al, focuses on the techniques that can predict chronic disease by mining the data containing in historical health records using Naïve Bayes, Decision tree, Support Vector Machine(SVM) and Artificial Neural Network(ANN). A comparative study is performed on the classifiers to measure the better performance on an accurate rate. From this experiment, SVM gives the highest accuracy rate, whereas for diabetes Naïve Bayes gives the highest accuracy.

Requirement Analysis-

Functional Requirement -

a. Predict disease with the given symptom.

b. Compare the given symptoms with the input dataset

Non-functional requirements -

a. Display the list of symptoms where user can select the symptom.

b. ALGORITHM is used to classify the data set.

User Requirements:-

Age : Age is the most important risk factor in developing cardiovascular diseases, with approximately a tripling of risk with each decade of life. Coronary fatty streaks that can begin to form in adolescence. It is estimated that 82 percent of people who die of coronary heart disease are 65 and above. Simultaneously, the risk of stroke doubles every decade after age of 55.

Sex : Men are at greater risks of heart disease than pre-menopausal women. Once past menopause, it has argued that a woman's risk is similar to a man's although more recent data from the WHO and UN disputes this. If a female has diabetes, she is more likely to develop heart disease or cardiovascular disease than a male with diabetes.

Angina (Chest Pain) : Angina is chest pain or discomfort caused when your heart muscle does not get enough oxygen-rich blood. It may feel like pressure in your chest. The discomfort also occur in your shoulders, arms, neck, jaw, or back. Angina pain may even feel like indigestion.

Serum Cholestrol : A high level of low-density lipoprotein (LDL) cholesterol is most likely to narrow arteries. A high level of triglycerides, a type of blood fat related to your diet, also ups your risk of heart attack. However, a high level of high-density lipoprotein (HDL) cholesterol (the "good" cholesterol) lowers your risks of heart attack.

Fasting Blood Sugar : Not producing enough of a hormone secreted by your pancreas (insulin) or not responding to insulin properly causes your body's blood sugar levels to rise, increasing your risks of heart attack.

Exercise induced angina : The pain or discomfort associated with angina usually feels tight, squeezing, and can vary from mild to severe. Angina is usually felt in the centre of your chest, but may spread to either or both of your shoulders, or your back, neck, arm. It can even be felt in hands. o Types of Angina a.) Stable Angina / Angina Pectoris b.) Unstable Angina c.)Variant (Prinzmetal) Angina d.) Microvascular Angina.

Peak exercise ST segment : A treadmill ECG stress test is considered abnormal when there is a horizontal or down-sloping ST-segment depression ≥ 1 mm at 60–80 ms after J point. Exercise ECGs with up-sloping ST-segment depressions are typically reported as an 'equivocal' test. In general, the occurrence of horizontal or down-sloping ST-segment depression at a lower workload (calculated in METs) or heart rate indicates a worse prognosis and higher likelihood of multi-vessel diseases. The duration of ST-segment depression is also important, as prolonged recovery after peak stress is consistent with a positive treadmill ECG stress tests. Another finding that is highly indicative of significant CAD is the occurrence of ST-segment elevation > 1 mm; these patients has been frequently referred urgently for coronary angiography.

II. SENSOR

Arduino Uno: The Arduino Uno is a microcontroller board based on the ATmega328. Arduino is an open-source, prototyping platform and its simplicity makes it ideal for hobbyists to. The Arduino Uno has 14 digital input/output pins (of which 6 can be used as PWM outputs), 6 analog inputs, a 16 MHz crystal oscillator, a USB connection, a power jack, an ICSP header, and a reset button. It contains everything needed to support the microcontroller; simply connect it to a computer with a USB cable .

The Arduino Uno differs from all the preceding boards in that it does not use the FTDI USB-to-serial driver chip. Instead, it features the Atmega8U2 microcontroller chip programmed as a USB-to-serial converters.

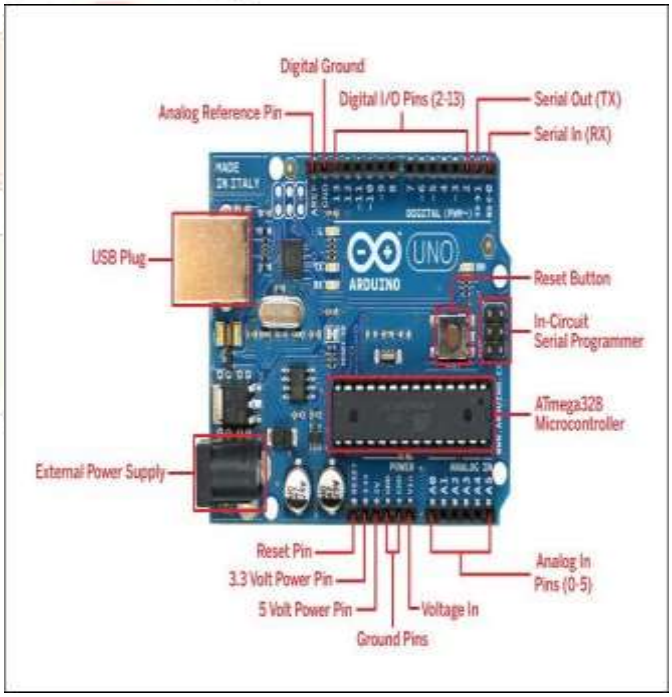


Fig.1 Arduino Uno

Heartbeat Sensor : The heartbeat sensor has emerged on the postulation of light modulation on blood flow through the finger in each pulse. Any change of light intensity through that organ (a vascular region) is predicted with the rate of heart pulses and since light is also absorbed by blood, those signal pulses are equivalent to the heartbeat pulses.

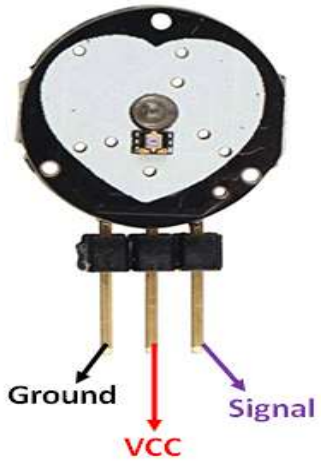


Fig-2 Heartbeat Sensor

Pin Configuration-

Pin Number	Pin Name	Wire Colour	Description
1	Ground	Black	Connected to the ground of the system
2	Vcc	Red	Connect to +5V or +3.3V supply voltage
3	Signal	Purple	Pulsating output signal.

Blood Pressure Sensor -

Blood Pressure & Pulse reading are shown on display with serial out for external projects of embedded circuit processing and display. Shows Systolic, Diastolic and Pulse Readings. Compact design fits over your wrist like a watch. Easy to use wrist style eliminates pumping.

Fig-3 Blood Pressure Sensor

AD8232 ECG Sensor –

This sensor is a cost-effective board used to measure the electrical activity of the heart. This electrical activity can be charted as an ECG or Electrocardiogram and output as an analog reading. ECGs can be extremely noisy, the AD8232 Single Lead Heart Rate Monitor acts as an op amp to help obtain a clear signal from the PR and QT Intervals easily.

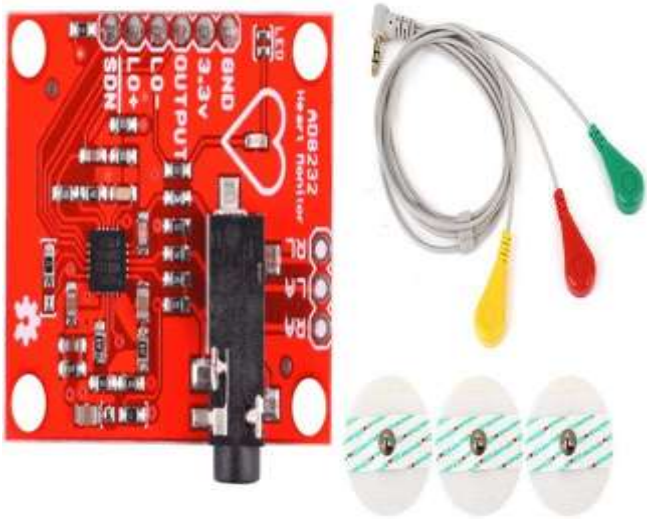


Fig- 4 AD8232 ECG Sensor

Machine Learning Algorithms

1**) Naive Bayes :** Naive Bayes is a surprisingly powerful algorithm for predictive modeling. It is a statistical classifier which assumes no dependency between attributes attempting to maximize the posterior probability in determining the class. Theoretically, this classifier has the minimum error rate, but may not be the case always. Inaccuracies are caused by assumptions due to class conditional independence and the lack of available probability data. This model is associated with two types of probabilities which can be calculated from the training dataset directly:
a) The probability of every class.
b) The conditional probability of each class with each x value.
According to Bayesian theorem P(A | B) = P(A)∗ P(B A) P(B) , where P(B | A) = P(A□ B) P(A) . Bayesian classifier calculates conditional probability of an instance belonging to each class, based on the above formula, and based on such conditional probability data, the instance is classified as the class with the highest conditional probability. If these probabilities are calculated, then the probabilistic model can be implemented to make predictions with new data usS. Nashif et al. DOI: 10.4236/wjet.2018.64057 862 World Journal of Engineering and Technology ing Naïve Bayes Theorem. When the data is real-valued it is likely to assume a Gaussian distribution (bell curve). Thus, these probabilities can easily be estimated. Naive Bayes is called naive because of assuming each input variable independent. This classifier algorithm uses conditional independence, means it assumes that an attribute value on a given class is independent of the values of other attributes.

2) **Artificial Neural Networks :** Artificial neural networks also called Multilayer Perceptron are known as biologically inspired and it is capable of modeling extremely complex non-linear functions. ANNs are one of the major tools used in machine learning. As the name “neural” suggests, they are brain-oriented systems that are intended to duplicate the way how humans learn. Neural networks consist of 3 layers of input, output and hidden layer. In most cases, a hidden layer comprises units that transform the input to a pattern that the output layer manipulates. ANN’s are excellent tools for the purpose of finding patterns that are so complex or ambiguous to a human programmer to extract and teach the machine how to recognize. Neural networks are in use since the 1940s and over the last decades they have become an important part of artificial intelligence because of the arrival of a new technique, which is called “backpropagation”, that allows networks know how to adjust their hidden layers of neurons in cases where the outcomes don’t match with the creator’s expectation. In Figure 2 the interconnection between the layers is shown.

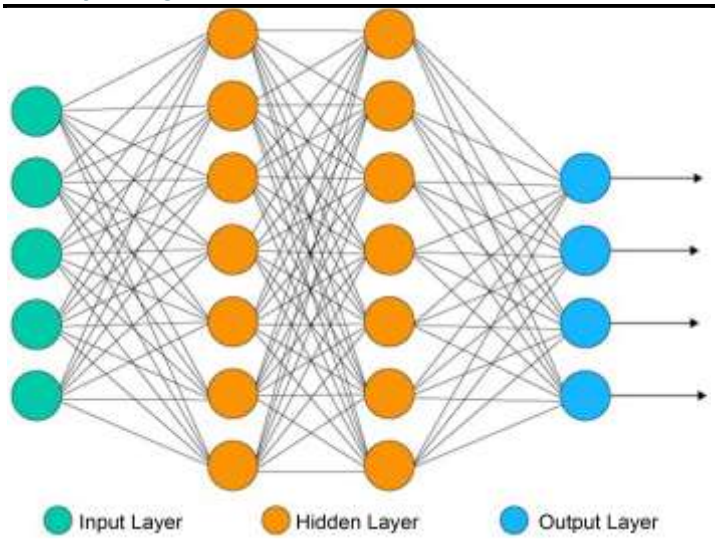
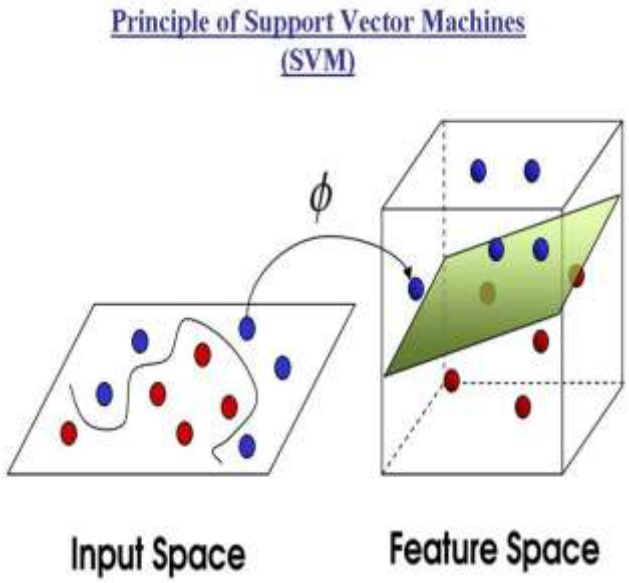


Fig-5 Artificial Neural Network

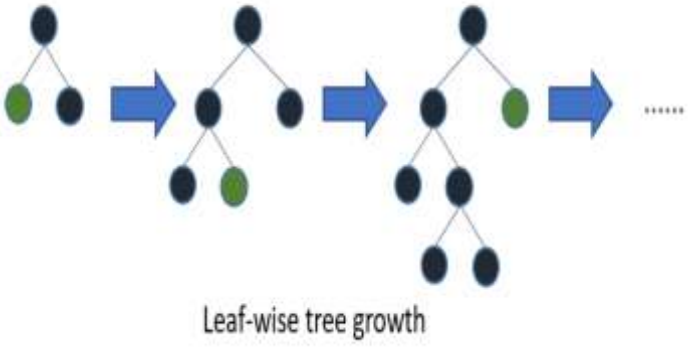
3) Support Vector Machine : SVM is a technique for ramification of both linear and non-linear data. It applies a non-linear mapping method so that it can transform the training data into a higher dimension. A hyperplane is a kind of line which separates the input variable space in SVM. The hyperplane can separate the points in the input variable space containing their class that is either 0 or 1. In two-dimensions, one can visualize this as a line and it is assumed that each input points can be completely separated by this line. The distance between the hyperplane and adjacent data coordinates is called margin. The line which has the largest margin can distinguish between the two classes is known as the optimal hyperplane. These points are called support vectors, as they define or support the hyperplane. In practice, there is an optimization algorithm which is used to calculate the values for the parameters that maximize the margin. Figure 3 depicts the feature transformation process.



4) **Random Forest :** Random Forest is one of the most renowned and most powerful machine learning algorithms. It is one kind of machine learning algorithm that is called Bagging or Bootstrap Aggregation. In order to estimate a value from a data sample such as mean, the bootstrap is a very powerful statistical approach. Here, lots of samples of data are taken, the mean is calculated, after that all of the mean values are averaged to give a better prediction of the real mean value. In bagging, the same method is used, but instead of estimating the mean of every data sample, decision trees are generally used. Here, numerous samples of the training data are considered and models are generated for
every data sample. While a prediction for any data is needed, each model gives a prediction and these predictions are then averaged to get a better estimation of the real output value.

5**) Simple Logistic Regression**: Logistic regression is a technique of machine learning which is taken from the field of statistics. This method can be used for binary classification where values are distinguished with two classes. Logistic regression is similar to linear regression where the goal is to calculate the values of the coefficients within every input variable. Unlike linear regression, here the prediction of the output is constructed using a non-linear function which is called a logistic function. The logistic function transforms any value within the range of 0 to 1. The predictions made by logistic regression are used as the probability of a data instance concerning to either class 0 or class 1. This can be necessary for problems where more rationale for a prediction is needed. Logistic regression works better when attributes are unrelated to output variable and attributes correlated to one another are removed.

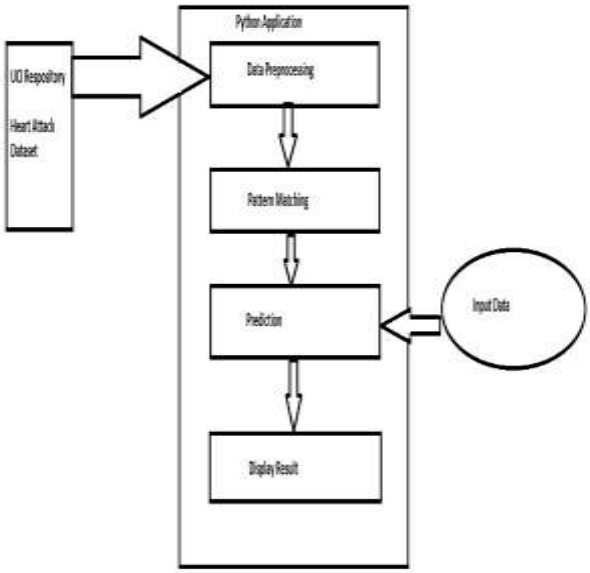
6)**LightGBM -** Light GBM is a gradient boosting framework that uses tree based learning algorithm.
Light GBM grows tree vertically while other algorithm grows trees horizontally meaning that Light GBM grows tree leaf-wise while other algorithm grows level-wise. It will choose the leaf with max delta loss to grow. When growing the same leaf, Leaf-wise algorithm can reduce more loss than a level-wise algorithm.
Below diagrams explain the implementation of LightGBM and other boosting algorithms.



PROPOSED MODEL FRAMEWORK

Proposed Project Plan. - we are proposing a heart disease prediction system by using Arduino, Heartbeat Sensor, Pressure and Temprature Sensor etc. by the help of these we will create a data record in the excel sheet directly and then we apply machine learning algorithms on that data and we can obtain desirable results. Using Arduino we take input form the patients. All the Sensors(Heart Beat Sensor,Pressure Sensor etc) are connected with the Arduino which is directly connected with the Computer. After taking the input succesfully from the patients. Arduino will represent that data in the Excel File and after that our Application(Python Based Application) will read all the data and apply the suitable machie learing algorithm and will provide the desirable outputs.

Data Flow-



The workflow of the complete system`have been mentioned below:-
 Collection and selection of different heart disease datasets in order to train various machine learning algorithms.
 Comparison of various data mining algorithm's accuracy and performance in predicting heart disease.
 Selection of the best algorithm from performance characteristics of the models to develop a intelligent heart disease prediction python based application.
 A full-fledged tentative design of an python application with respective criteria has been shown in the Analysis and Results section.
 First user have to register in the application by submitting his user id and password, after registration the main page will be displayed.
 Then user have to input the data according to given attribues by the application.
 After submitting, all the data will be processed.

 Data processing-
Data Set is taken from UCI Repository having 14 column.
1. Age: displays the age of the individual.
2. Sex: displays the gender of the individual using the following format :
1 = male
0 = female
3. Chest-pain type: displays the type of chest-pain experienced by the individual using the following format :
1 = typical angina
2 = atypical angina
3 = non — anginal pain
4 = asymptotic
4. Resting Blood Pressure: displays the resting blood pressure value of an individual in mmHg (unit)
5. Serum Cholestrol: displays the serum cholesterol in mg/dl (unit)
6. Fasting Blood Sugar: compares the fasting blood sugar value of an individual with 120mg/dl.
If fasting blood sugar > 120mg/dl then : 1 (true)
else : 0 (false)
7. Resting ECG : displays resting electrocardiographic results
0 = normal
1 = having ST-T wave abnormality
2 = left ventricular hyperthrophy
8. Max heart rate achieved : displays the max heart rate achieved by an individual.
9. Exercise induced angina :
1 = yes
0 = no
10. ST depression induced by exercise relative to rest: displays the value which is an integer or float.
11. Peak exercise ST segment :
1 = upsloping
2 = flat
3 = downsloping
12. Number of major vessels (0–3) colored by flourosopy : displays the value as integer or float.
13. Thal : displays the thalassemia :
3= normal
6= fixed defect
7 = reversible defect
14. Diagnosis of heart disease : Displays whether the individual is suffering from heart disease or not :
0 = absence
1, 2, 3, 4 = present.

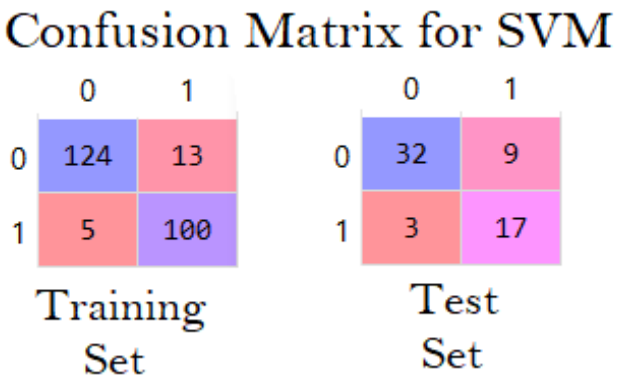
 Classification - null values either drop or will be imputed. We can impute the mean in place of the null values however one can also delete these rows entirely.
 Now let us divide the data in the test and train set.
 In this project, We have divided the data into an 80: 20 ratio. That is, the training size is 80% and testing size is 20% of the whole data.
 Different models are applied to get the results.
The evaluation metric used is the confusion matrix.

confusion matrix-
The confusion matrix displays the correctly predicted as well as incorrectly predicted values by a classifier..

		Actual Values: Positive (1)	Actual Values: Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

The sum of TP and TN, from the confusion matrix, is the number of correctly classified entries by the classifier.

SVM



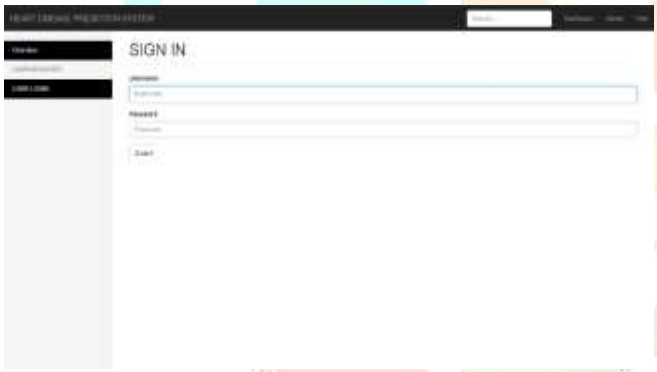
Accuracy for SVM for training set = ((124+100)/(5+13+124+100))*100 = 92.51%
Accuracy for SVM for test set = 80.32%

 By the help of confusion matrix for all the algorithms we can compare the accuracy among them.
 Algorithm with high accuracy is applied on the new data input by the user for the prediction.

Application Work Flow:
This is the Home page of the application:

After Clicking User Registeration a new page get opened. And fill all the required Fields in the showing form. After successful submission user will redirected to the login page.

This is a sign In form page. Fill this form by Valid username and password.



After Successful Sign In user will redirected to the page as shown Below. New options are poped up. As show inside a red box. You can Logout also by clicking on Logout button .

Click On Check Heart Status option and user will redirected to new page , having a form related to patient report detail. Fill up the form according to patient detail and you have a option to link patient file which should be a excel file. By clicking upload option you can upload patient file.

Click option choose file and file explorer option will popped up and choose your file from your local directory where your patient report excel file is stored.

After submitting , a patient report is generated. As shown below . You can go to home page by clicking on Dashboard option for further operation.

You can Check your Profile and you can update it by clicking update.

CONCULSION

Heart attack is crucial health problem in human society. This paper has summarised state of art techniques and available methods for predication of this disease. Deep learning an emerging area of artificial intelligence showed some promising result in other field of medical diagnose with high accuracy. It is still an open domain waiting to get implemented in heart diseasepredication.Some methods of deep learning has been discussed which can be implemented for heart disease predication, along with pioneer machine learning algorithms. An analytical comparison has been done for finding out bestavailable algorithm for medical dataset. In future our aim is to carry forward the work of temporal medical dataset, where dataset varies with time and retraining of dataset is required.

REFERENCES

[1] William Carroll; G. Edward Miller, "Disease among Elderly Americans : Estimates for the US civilian non institutionalized population, 2010," Med. Expend. Panel Surv., no. June, pp. 1–8, 2013.

[2] V. Kirubha and S. M. Priya, "Survey on Data Mining Algorithms in Disease Prediction," vol. 38, no. 3, pp. 124–128, 2016.

[3] M. A. Jabbar, P. Chandra, and B. L. Deekshatulu, "Prediction of risk score for heart disease using associative classification and hybrid feature subset selection," Int. Conf. Intell. Syst. Des. Appl. ISDA, pp. 628–634, 2012.

[4] M. Sultana, A. Haider, and M. S. Uddin, "Analysis of data mining techniques for heart disease prediction," 2016 3rd Int. Conf. Electr. Eng. Inf. Commun. Technol. iCEEiCT 2016, 2017.

[5] M. Akhil, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," Procedia Technol., vol. 10, pp. 85–94, 2013.

[6] S. Kumra, R. Saxena, and S. Mehta, "An Extensive Review on Swarm Robotics," pp. 140–145, 2009.

[7] T. M. Lakshmi, A. Martin, R. M. Begum, and V. P. Venkatesan, "An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data," Int. J. Mod. Educ. Comput. Sci., vol. 5, no. 5, pp. 18–27, 2013.

[8] P. Sharma and A. P. R. Bhartiya, "Implementation of Decision Tree Algorithm to Analysis the Performance," Int. J. Adv. Res. Comput. Commun. Eng., vol. 1, no. 10, pp. 861–864, 2012.