Impact Factor (SJIF): 5.301

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 5, Issue 5, May-2018

”Classication and Prediction of Heart Disease Risk in Data Mining Techniques”.

Supriya Hangarge[1], Shweta Narwade[2], Darshana Karnawat[3], Sarswati Admane[4]. Prof. A.A.Bamanikar5

[1,2,3,4]*Student of Department of Computer Engineering, PDEA College of Engg(Manjari,bk), Hadapsar, Pune, Maharashtra, India*

[5]*Assit. Prof. Department of Computer Engineering, PDEA College of Engg(Manjari,bk), Hadapsar, Pune. Maharashtra, India*

Abstract — currently a days, health un-wellness square measure increasing day by day because of life vogue, hereditary. Especially, heart disease has become additional common recently .i.e. lifetime of folks is in danger. Every individual has completely different values for force per unit area, sterol and vital sign. But in step with medically well-tried results the traditional values of force per unit area is 120/90, sterol is and vital sign is seventy two. This paper offers the survey concerning completely different classification techniques used for predicting the chance level of every person supported age, gender, force per unit area, sterol, pulse rate. The patient risk level is classed mistreatment data mining classification techniques like Naïve mathematician, KNN, call Tree algorithmic program, and Neural Network. etc., Accuracy of the chance level is high once mistreatment additional variety of attributes.

Keywords- Classification Techniques, decision Tree algorithmic program, cardiopathy, kNN, Naïve mathematician, Neural Network, Risk level.

I. INTRODUCTION

Heart disease is that the biggest cause of death these days. Pressure level, sterol, pulse are the key reason for the guts malady. Some non-modifiable factors also are there. Like smoking, drinking conjointly reason for heart condition. The guts are AN OS of our build. If the operate of heart isn't done properly suggests that, it'll have an effect on other build half conjointly. Some risk factors of cardiopathy are case history, High vital sign, sterol, Age, Poor diet, Smoking. Once blood vessels are overstretched, the chance level of the blood vessels is augmented. This results in the vital sign. Vital sign is often measured in terms of heartbeat and pulsation. Pulse indicates the pressure within the arteries once the guts muscle contracts and pulsation indicates the pressure within the arteries once the guts muscle is in resting state. The amount of lipids or fats augmented within the blood are causes the guts malady. The lipids are in the arteries thence the arteries become slim and blood flow is additionally become slowly. Age is that the non-modifiable risk issues which conjointly a reason for heart condition. Smoking is that the reason for four-hundredth of the death of heart diseases. As a result of it limits the atomic number 8 level within the blood then it harms and tightens the blood vessels. Numerous data processing techniques like Naïve Thomas Bayes, KNN rule, call tree, are accustomed predict the chance of heart condition. The KNN rule uses the K user outlined worth to search out the values of the factors of heart condition. Call tree rule is employed to produce the classified report for the guts malady. The Naïve Thomas Bayes technique is employed to predict the guts malady through likelihood. In all this on top of mentioned techniques the patient records are classified and foreseen ceaselessly. The patient activity is monitored ceaselessly, if there's any changes occur, so the chance level of malady is hip to the patient and doctor. The doctors are ready to predict heart diseases at AN earlier stage as a result of machine learning algorithms and with the assistance of technology. This paper provides AN insight concerning KNN data processing technique accustomed predict heart diseases.

II. PROBLEM STATEMENT

We gift a heart disease prediction system supported based on naïve bayes' algorithmic rule. This system is convenient, effective and offers smart prediction of diseases to users. Exhibits the analysis of varied data processing techniques which can be helpful for medical analysts or practitioners for correct heart condition identification.

III. LITERATURE REVIEW

1. “ Web based health care detection”

Author: S.Indhumathi.etl

It has suggested a prediction of high risk heart disease using a Naïve Bayes algorithm. The preprocessed data has been considered as the training set. Two phase namely classification and prediction was discussed in that work. Preprocessing is done in the classification phase. The preprocessing includes cleaning of data, normalization and reduction of data, etc. In the prediction phase the disease types are classified and predicted, i.e. a training set is formed based on the disease

type and the test set is formed based on the questions. The predicted results are sent to the doctor. ANN, often just called a "neural network", is a mathematical model or computational model used for a biological purpose. In other words, it is an emulation of biological neural system.

2. The prediction method for heart disease using Neural Network

Author: Chaitrali S.Dangare.etl

It has mainly three layers, i.e. the input layer, hidden layer and the output layer. The input is given to the input layer and the result is obtained in the output layer. Then the actual output and the expected output are compared. The back propagation has been applied to find the error and to adjust the weight between the output and the previous hidden layers. Once, the back propagation is completed, and then the forward process is started and continued until the error is minimized. KNN is a non-parametric method which is used for classification and regression. Compared to other machine learning algorithm KNN is the simplest algorithm. This algorithm consist K-closet training examples in the feature space. In this algorithm K is a user defined constant. The test data are classified by assigning a constant value which is most chronic among the K-training samples nearest to the point. Literature shows the KNN has the strong consistency result. Decision tree builds classification models in the form of a tree structure. It breaks the dataset into smaller subset while at the same time an associated decision tree is incrementally developed. The decision tree uses a top-down approach method. The root of the decision tree is the data set and the leaf is the subset of the data set.

3. The risk level of heart disease prediction through hybrid algorithm

Author: Shovon K.Pramanik.etl.

Hybrid Algorithm is the combination of KNN algorithm and ID3. These algorithms are used for heart disease prediction. The KNN algorithm is used to preprocess the data; it is called as preprocessed algorithm. The preprocessed data are considered as training set and then the data has been classified into a tree structure. The ID3 algorithm is applied for the classifier to predict the heart disease. The incorrect values are classified through KNN Algorithm.

4. " Using Data Mining Technique in Heart Disease Diagnosis and Treatment"

Author: Mai Shouman, Tim Turner and Rob Stocker

Various single and hybrid data mining techniques in heart disease diagnosis. Using single data mining technique for heart disease diagnosis has been thoroughly investigated showing the considerable levels of accuracy. Recent investigation shows that for hybridizing more than one technique, will obtain enhanced result in diagnosis. This paper identifies gap in the diagnosis of heart disease and treatment require for it and proposes a model which close those gaps to discover if applying hybrid and single data mining techniques in heart disease treatment, data can provide reliable performance. Here author can apply different data mining techniques like multilayer perceptron, naïve bayes decision tree, neural network and kernel density on different heart disease datasets and measures the accuracy of each technique. Then applying hybrid data mining techniques on different heart disease datasets shows the different accuracies

5. A classifier approaches for heart disease prediction

Author: Dhanashree S. Madhekar, Mayur P. Bote, Shruti D.Deshmukh

A classifier approaches for heart disease prediction and shows how naive bayes classification can be used for this purpose. The proposed system will categorized medical data into five distinct categories namely no, low, average, high and very high. Also the system will predict the class label of different unknown samples, if any and for this prediction the two basic functions namely classification (training) and prediction (testing) will be performed. The accuracy of the system will depend on different algorithms, techniques applied on different databases.

IV. EXISTING SYSTEM

There was no specific existing system. User presupposed to enter hospital to notice the heart malady and diabetic' s prediction manually. When millions of effort check user get the result that required more time.

V. PROPOSED SYSTEM

Heart disease may be a general name for a range of diseases, conditions and disorders that have an effect on the guts and also the blood vessels. Symptoms of Heart Disease vary counting on the precise sort of cardiopathy. Innate Heart Disease refers to a haul with the heart's structure and performance as a result of abnormal heart development before birth. Symptom Heart Disease is once the guts doesn't pump adequate blood to the opposite organs within the body. Coronary cardiopathy or in its medical term ischemic Heart Disease is that the most frequent sort of heart downside. Coronary

Heart Disease may be a term that refers to break to the guts that happens as a result of its blood provide is shriveled, it ends up in the fatty deposits build a fait the linings of the blood vessels that offer the guts muscles with blood, leading to them narrowing The paper identifies the chance factors for the various forms of heart diseases. Pressure, steroid alcohol, pulse area unit the key reason for the guts illness. Some non-modifiable factors are there. Like smoking, drinking conjointly reason for Heart Disease. The guts are associate degree software system of our soma. If the operate of heart isn't done properly means that, it'll have an effect on different soma half conjointly. Some risk factors of Heart Disease area unit case history, High pressure, steroid alcohol, Age, Poor diet, Smoking.

Advantages:

- Easy to predict heart disease as basic level.
- Diabetic' s prediction will be easy to manipulate.

VI. ALGORITHM

KNN:

A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function.

If K = 1, then the case is simply assigned to the class of its nearest neighbor.

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

$$\sum_{i=1}^{k}|x_i - y_i|$$

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$

It should also be noted that all three distance measures are only valid for continuous variables. In the instance of categorical variables the Hamming distance must be used. It also brings up the issue of standardization of the numerical variables between 0 and 1 when there is a mixture of numerical and categorical variables in the dataset.

$$D_H = \sum_{i=1}^{k}|x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

Choosing the optimal value for K is best done by first inspecting the data. In general, a large K value is more precise as it reduces the overall noise but there is no guarantee. Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value. Historically, the optimal K for most datasets has been between 3-10. That produces much better results than 1NN.

Naïve Bayes:

R and L are conditionally independent given M if for all x,y,z in {T,F}:

P(R=x □ M=y ^ L=z) = P(R=x □ M=y)

More generally:

Let S1 and S2 and S3 be sets of variables.

Set-of-variables S1 and set-of-variables S2 are conditionally independent given S3 if for all Assignments of values to the variables in the sets, P(S1's assignments | S2' s assignments & S3' s assignments) = P (S1' s assignments | S3' s assignments)

P(A|B) = P(A ^B)/P(B)

Therefore P(A^B) = P(A|B).P(B) – also known as Chain Rule

Also P(A^B) = P(B|A).P(A)

Therefore P(A|B) = P(B|A).P(A)/P(B)

P(A,B|C) = P(A^B^C)/P(C)
= P(A|B,C).P(B^C)/P(C) – applying chain rule
= P(A|B,C).P(B|C)
= P(A|C).P(B|C) , If A and B are conditionally independent given C.
This can be extended for n values as P(A1,A2..An|C) = P(A1|C).P(A2|C)..P(An|C) if A1, A2..An are conditionally independent has given C.

A. **BLOCK DEIAGRAM OF SYSTEM**

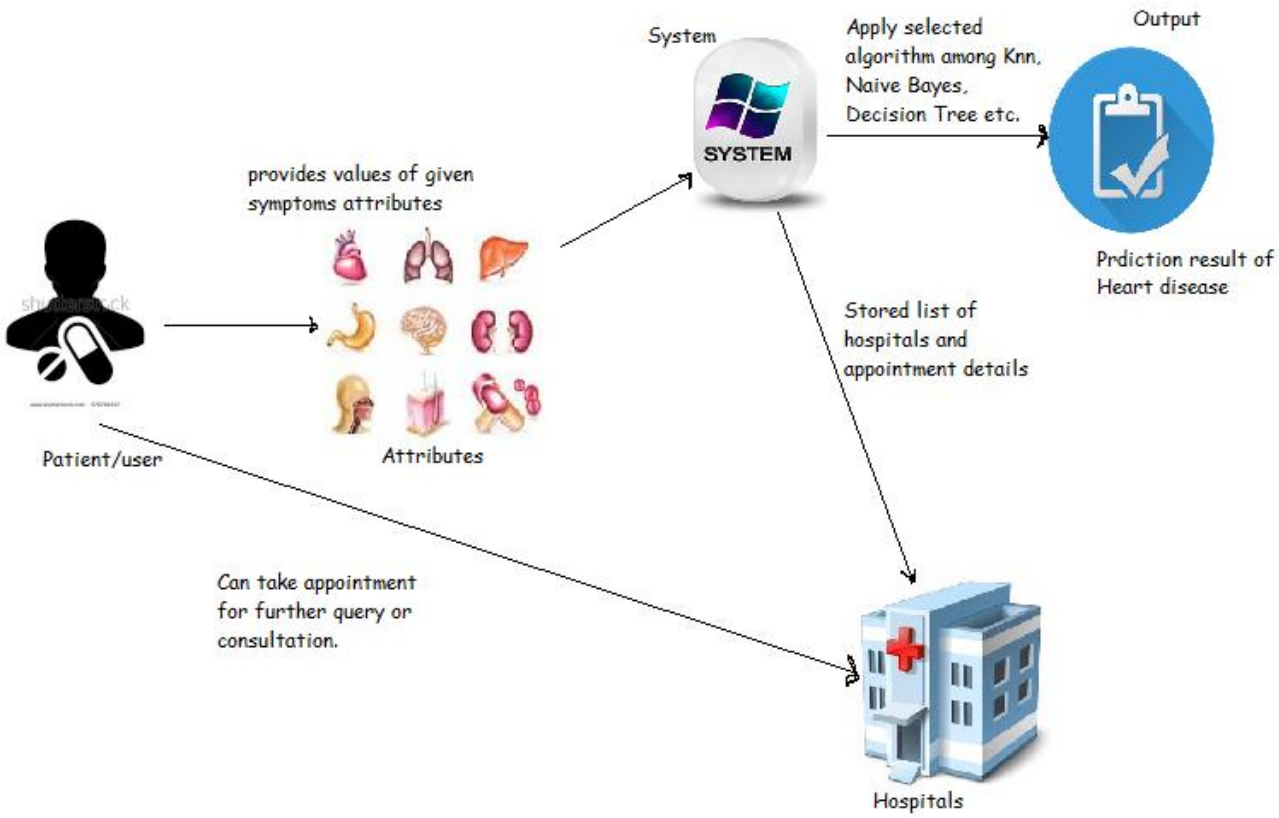


Figure6.1: Block diagram of system

Heart disease could be a general name for a variety of diseases, conditions and disorders that have an effect on the heart and also the blood vessels. Symptoms of cardiopathy vary looking on the specific sort of cardiopathy. Non inheritable cardiopathy refers to a haul with the heart's structure and function thanks to abnormal heart development before birth. Symptom coronary failure is once the centre does not pump adequate blood to the opposite organs in the body. Coronary cardiopathy or in its medical term anaemia cardiopathy is the most frequent sort of heart problem. Coronary cardiopathy is a term that refers to wreck to the centre that happens as a result of its blood offer is minimized, it results in the fatty deposits build a fait the linings of the blood vessels that offer the centre muscles with blood, resulting in them narrowing The paper identifies the risk factors for the various styles of heart diseases. Pressure, steroid alcohol, vital sign square measure the key reason for the centre malady. Some non-modifiable factors also are there. Like smoking, drinking additionally reason for cardiopathy. The centre is Associate in nursing software system of our body. If the perform of heart isn't done properly means that, it'll have an effect on different body half additionally. Some risk factors of cardiopathy square measure family history, High pressure, steroid alcohol, Age, Poor diet, Smoking.

B. **SYSTEM REQUIREMENT**
Hardware Requirements:

- System : Pentium IV 2.4 GHz. And above
- Hard Disk : 40 GB.
- Floppy Drive: 1.44 Mb.

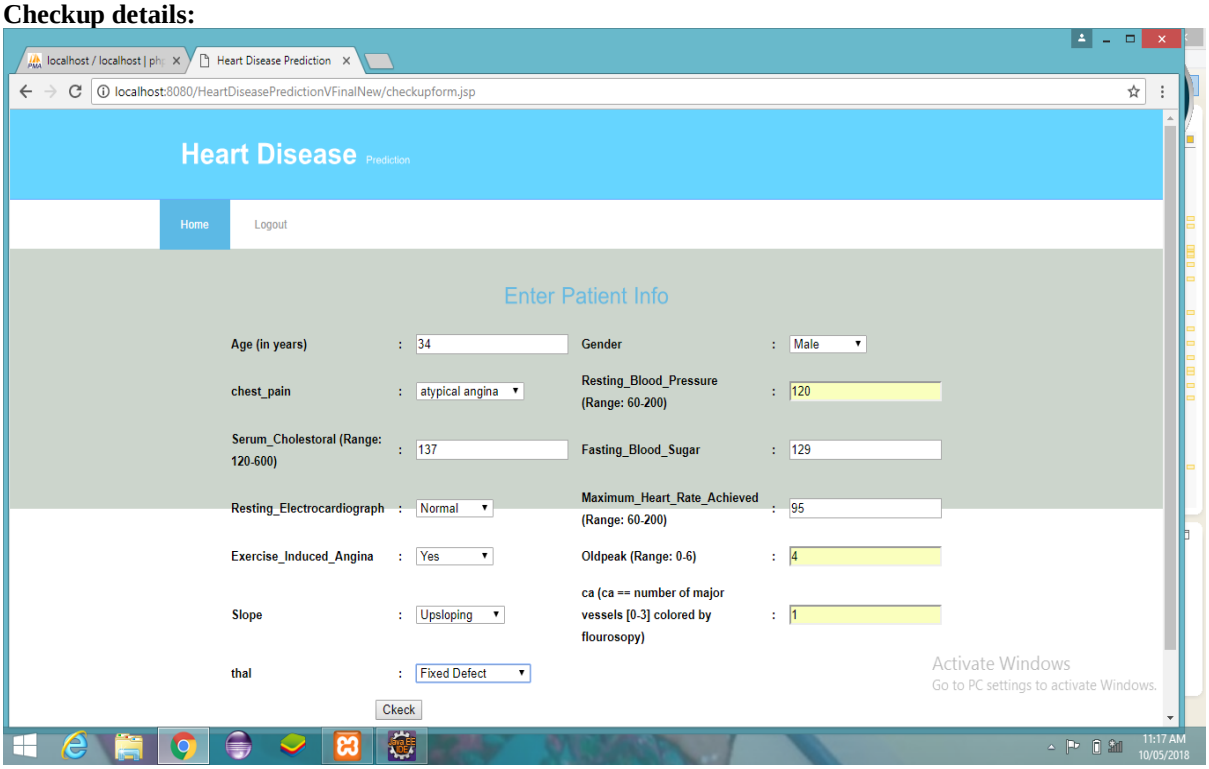
- Monitor : 15 VGA Colour.
- Mouse : Logitech.
- Ram : 512 Mb.

Software Requirements:

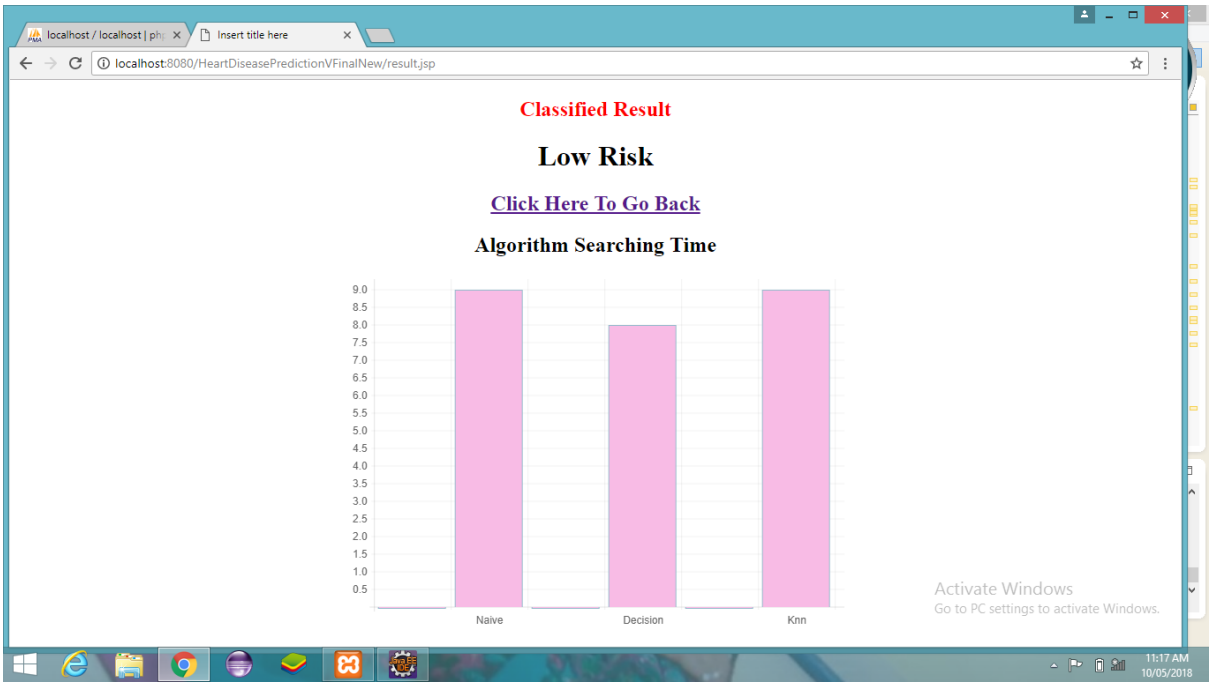
- Operating system : Windows XP/7/LINUX.
- Front End : Java/J2SE
- Back End : MySQL 5.5
- Tool/IDE : Eclipse kepler
- Server : Apache Tomcat

VII. RESULTS

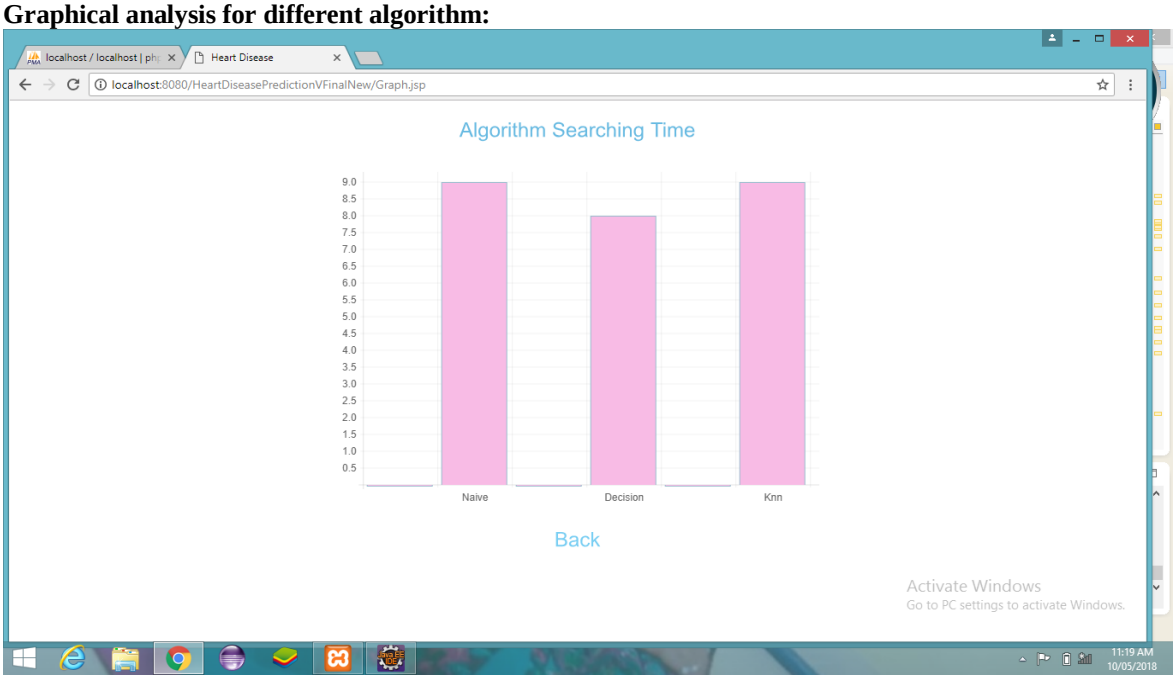
1. **Checkup details:**



2. **Disease predication and solution:**



3. Graphical analysis for different algorithm:



VIII. APPLICATION

- Any heart diseases detection site.
- To verify seriousness of your heart problem.
- Help the medical practitioners to understand the root causes of disease in depth.

IX. CONCLUSION AND FUTURE SCOPE

A conclusion is created that neural network is best among all the classification techniques once we state prediction or classification of a nonlinear data. BP formula that is that the best classifier of Artificial Neural Network which uses the change technique of weights by propagating the errors backward is used. however it has downside of being stuck in a native minima answer therefore to resolve this downside, we can use associate degree economical optimizing technique to any improve its accuracy and apply in the predictions of assorted applications. During this paper, we have a tendency to develop a cardiopathy prediction system that may assist medical professionals in evaluating a patient' s cardiopathy supported the clinical information of the patient. Our approaches embody 3 steps. Firstly, we have a tendency to choose

necessary clinical options, i.e., age, sex, pain kind, sterol, abstinence glucose, resting cardiogram, GHB vital sign, exercise induced angina, old peak, slope, number of vessels color, and thal. Secondly, we have a tendency to develop a synthetic neural network formula for classifying cardiopathy based on these clinical options. The accuracy of prediction is near eighty per. Finally, we tend to develop easy cardiopathy predict system; that generates prediction results victimization artificial neural network (ANN), Decision Tree, Naive mathematician Classification techniques. The HDPS system may be a computer-aided system developed from C and C sharp setting. Hopefully, this technique is employed in the classification of cardiopathy.

ACKNOWLEDGMENT

Authors want to acknowledge Principal, Head of department and guide of their project for all the support and help rendered. To express profound feeling of appreciation to their regarded guardians for giving the motivation required to the finishing of paper.

REFERENCES

[1] K. Sudhakar, Dr. M. Manimekalai, " Study of Heart Disease Prediction using Data Mining" , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 1, pp.1157-60, January 2014.

[2] S. U. Amin, K. Agarwal, and R. Beg, " Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors," ,IEEE Conference on Information and Communication Technologies (ICT 2013), 2013.

[3] Miss. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte, " A Data mining approach for prediction of heart disease using neural network' s" , International Journal of Computer Engineering & Technology(IJCET), Volume 3, Issue 3, October – December (2012), pp. 30-40.

[4] S.Indhumathi, Mr.G.Vijaybaskar, " Web based health care detection using naive Bayes algorithm" , International Journal of Advanced Research in Computer Engineering & Technology(IJARCET**)**, Volume 4 Issue 9, pp.3532-36, September 2015.

[5] G. Purusothaman, P. Krishnakumari, " A Survey of Data Mining Techniques on Risk Prediction: Heart Disease" , Indian Journal of Science and Technology, Vol 8(12), June 2015.