ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 2, Issue 2, pp: (22-26), Month: October 2015 – March 2016, Available at: www.paperpublications.org

Comparative Study of Data Mining Classification Algorithms in Heart Disease Prediction

Durgadevi. A[1], Prof. Saravanapriya.K[2]

[1,2]PG and Research Department of Computer Applications Sacred Heart College, Tirupattur, Vellore

Abstract: **This research paper intends to provide an accuracy of the best algorithm from the classification algorithms. The main objective of this research work is to predict more accurately the presence of heart disease in patients. The dataset used here is from the UCI Machine Learning Repository based Hungarian-14-heart disease with 294 instances. In this research paper, we use weka tool with five classifiers like Naïve Bayes, Logistic function, RBF Network function, Decision Table rule, SMO function and their performance on the diagnosis has been compared. From that Naïve Bayes provides 86% accurate result in 0 seconds and RBF Network provides 86% of accuracy in 0.17 seconds. The research result shows both the Naïve Bayes outperforms with a time duration of 0 seconds to build the model.**

Keywords: **CVD - Cardio Vascular Disease, Machine learning algorithm, Supervised and Unsupervised algorithms, Logistic function Algorithm, RBF - Radial Basis Function Network, SMO function -Sequential Minimal Optimization Algorithms.**

1. INTRODUCTION

Various data mining algorithms and techniques like Classification, Clustering, Artificial Intelligence, Regression, Neural Networks, Association Rules, Genetic Algorithm, Decision Trees, Nearest Neighbour method etc., are used for knowledge discovery from databases. In this paper we were discussed about the five different classifiers used to predict the heart disease with 14 attributes.

Classification algorithms: Classification is the most commonly applied data mining technique, which needs a set of pre-classified examples to develop a model that can classify the population of records at huge. Fraud detection and credit risk applications are particularly well suited to such kind of analysis. This attitude commonly employs decision tree or neural network-based classification algorithms. Mainly data classification process involves learning and classification. In learning the training data are analysed by classification algorithm. During classification test data are used to estimate the accuracy of the classification rules. In the case of accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection type of application, this would comprise complete records of both fraudulent and valid activities determined on a record-by-record basis. To determine the set of parameters required for appropriate discrimination the classifier-training algorithm uses these pre-classified examples. Then the algorithm encodes these parameters into a model called a classifier.

In ***Supervised learning*** the computer is presented inputs and their desired outputs, and the goal is to learn a general rule that maps inputs to outputs. In ***Unsupervised learning*** there is no labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself or a means towards an end. ***Bayes' Theorem*** finds the probability of an event occurring given the probability of another event that has already occurred. If B represents the dependent event and A represents the prior event, Bayes' theorem can be stated as follows.

ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 2, Issue 2, pp: (22-26), Month: October 2015 – March 2016, Available at: www.paperpublications.org

Prob (B given A) = Prob(A and B)/Prob(A)

RBF Network is a three-layer network, namely the input, the output and the hidden layer, where each hidden unit in a hidden layer implements a radial activated function. The main advantages of RBF's over feed-forward networks are its accuracy and shorter computational time. ***Sequential minimal optimization*** is an algorithm for solving the quadratic programming problem that arises during the training of support vector machines. ***Logistic function*** predicts the probability of an outcome that can only have two values. The prediction is based on the use of one or several predictors. **Decision table** is the rule which is used to make the decision. Decision tables may be defined in a variety of different ways.

2. DATASET DESCRIPTION

In this paper we have used Hungarian-14-heart disease dataset with 14 attributes. Using the WEKA tool to analyze the data and it provides the result with the accuracy of the algorithm.

Table 2.1 Hangarian-14-heart-disease attributes

NO	Attribute Name	Description
1	Age	Age in Year
2	Sex	(value 1: Male; value 0 : Female)
3	Cp	value 1: typical type 1 angina, value 2:typical type angina, value 3: non-angina pain; value 4:asymptomatic)
4	Trestbps	mm Hg on admission to the hospital
5	Chol	Serum Cholesterol (mg/dl)
6	Fbs	Fasting Blood Sugar (value 1: > 120 mg/dl; value 0: < 120 mg/dl)
7	Restecg	Resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
8	Thalach	(value 3: normal; value 6: fixed defect; value 7:reversible defect)
9	Exang	exercise induced angina (value 1: yes; value 0: no)
10	Oldpeak	ST depression induced by exercise relative to Rest
11	Slope	the slope of the peak exercise ST segment (value 1: Un sloping; value 2: flat; value 3: down sloping)
12	Ca	number of major vessels colored by floursopy (value 0– 3)
13	Thal	value 3: normal; value 6: fixed defect; value 7: reversible defect
14	Num	Diagnosis of heart disease.

Table 2.2 accuracy of different classification algorithms

Algorithm to be tested	Correctly classified instances	Incorrectly classified instances	Time taken to build model in seconds
Naïve Bayes	86%	14%	0 sec
RBFNetwork	86%	14%	0.17sec
SMO	84%	16%	0.12sec
Logistic	85%	15%	0.31sec
Decision Table	79%	21 %	0.17sec

The above table shows Hangarian-14-heart-disease dataset with 14 attributes are used to predict the heart disease in this study.

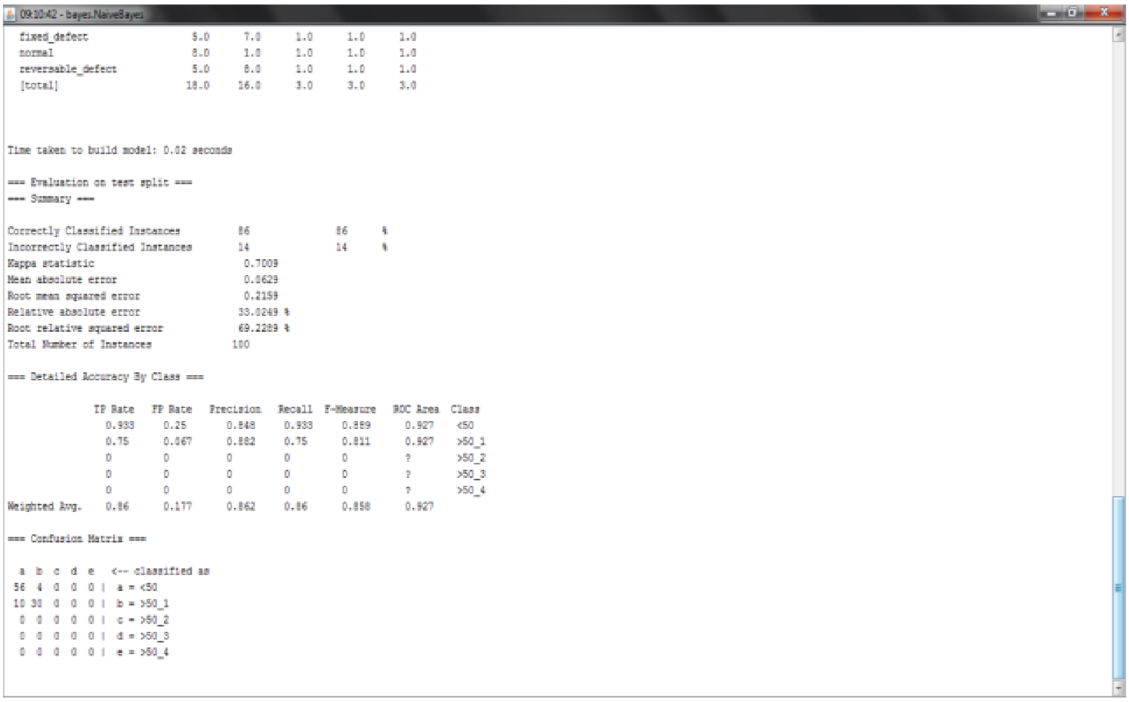
ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 2, Issue 2, pp: (22-26), Month: October 2015 – March 2016, Available at: www.paperpublications.org

3. PERFOMANCE COMPARISON OF CLASSIFICATION ALGORITHMS

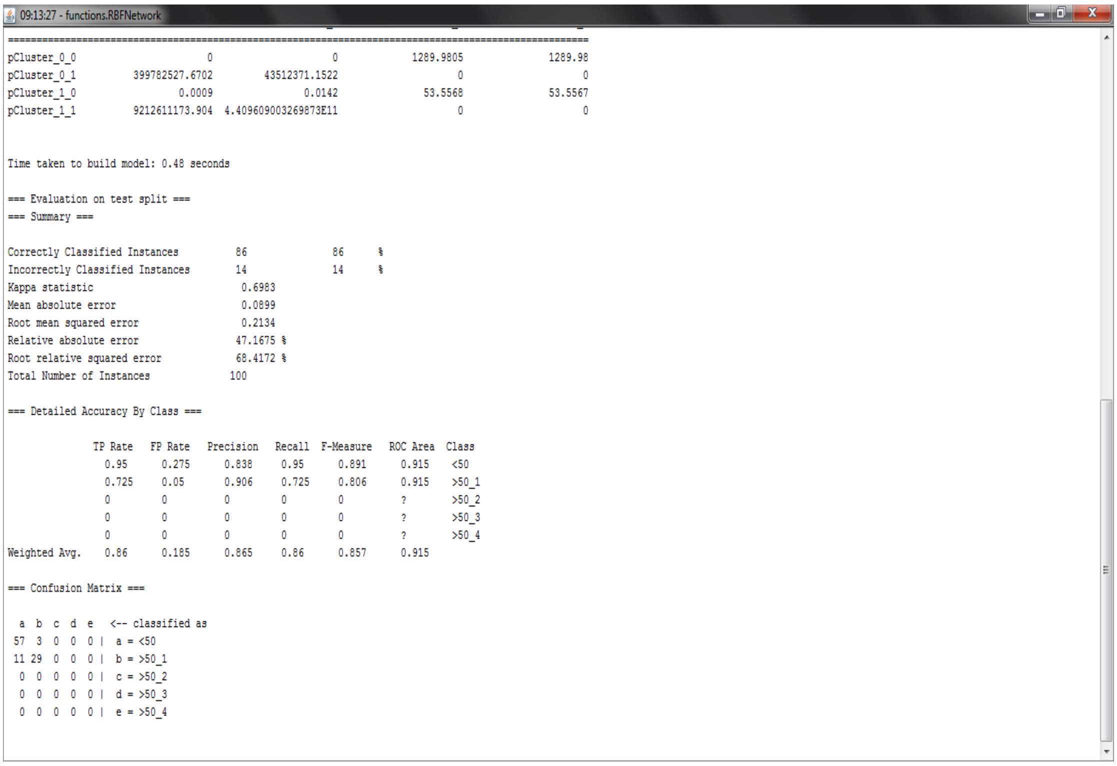
In the above table shows the accuracy with the correctly classified and incorrectly classified instances and time taken to build the model.

All the algorithm have same limitation a time when the dataset is very large. Using the split 66.0% train data test option is used to split the data.



```
09:10:42 - bayes.NaiveBayes
  fixed_defect              5.0     7.0     1.0     1.0     1.0
  normal                    8.0     1.0     1.0     1.0     1.0
  reversable_defect         5.0     8.0     1.0     1.0     1.0
  [total]                  18.0    16.0     3.0     3.0     3.0


Time taken to build model: 0.02 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances          86               86      %
Incorrectly Classified Instances        14               14      %
Kappa statistic                          0.7009
Mean absolute error                      0.0629
Root mean squared error                  0.2159
Relative absolute error                 33.0249 %
Root relative squared error             69.2289 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.933     0.25       0.848     0.933     0.889      0.927    <50
                 0.75      0.067      0.882     0.75      0.811      0.927    >50_1
                 0         0          0         0         0          ?        >50_2
                 0         0          0         0         0          ?        >50_3
                 0         0          0         0         0          ?        >50_4
Weighted Avg.    0.86      0.177      0.862     0.86      0.858      0.927

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 56  4  0  0  0 |  a = <50
 10 30  0  0  0 |  b = >50_1
  0  0  0  0  0 |  c = >50_2
  0  0  0  0  0 |  d = >50_3
  0  0  0  0  0 |  e = >50_4
```

Figure 3.1 Screenshot for accuracy of Naïve bayes with 86%

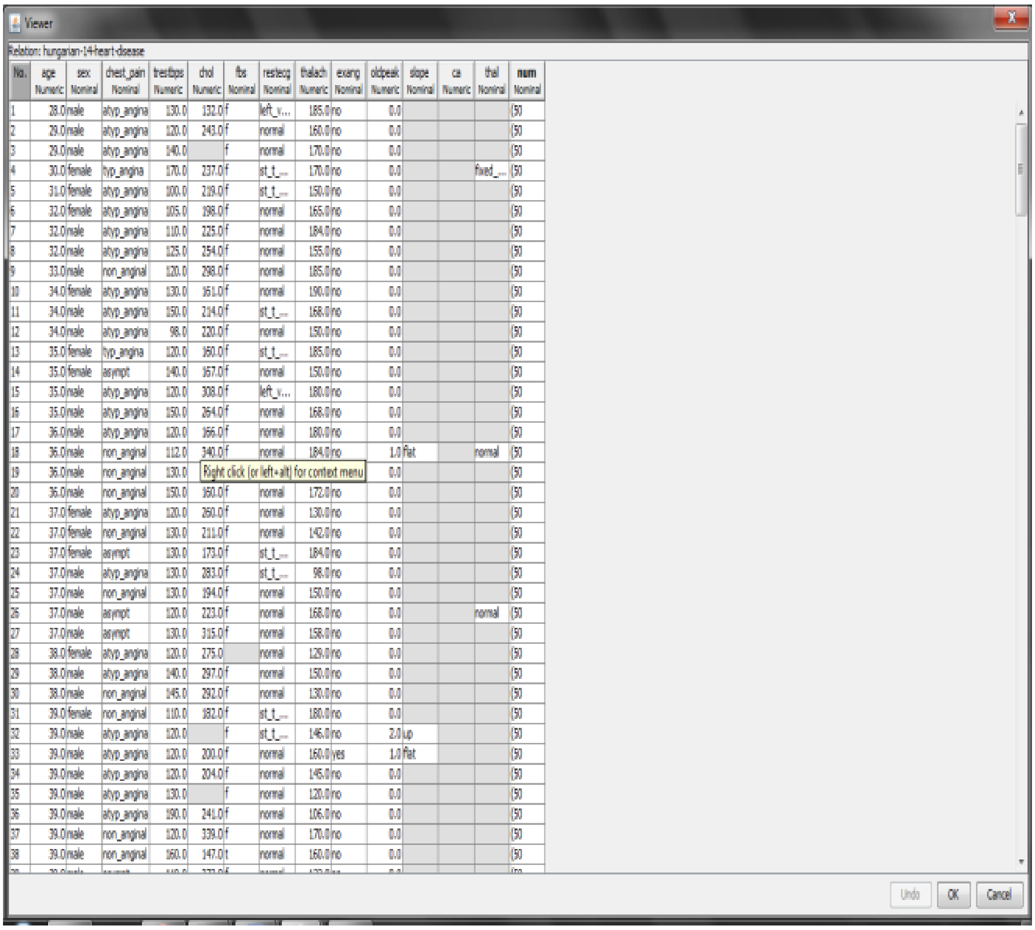


```
09:13:27 - functions.RBFNetwork
==============================================================================================
pCluster_0_0                       0                     0             1289.9805           1289.98
pCluster_0_1          399782527.6702         43512371.1522                     0                 0
pCluster_1_0                  0.0009                0.0142               53.5568           53.5567
pCluster_1_1          9212611173.904  4.409609003269873E11                     0                 0


Time taken to build model: 0.48 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances          86               86      %
Incorrectly Classified Instances        14               14      %
Kappa statistic                          0.6983
Mean absolute error                      0.0899
Root mean squared error                  0.2134
Relative absolute error                 47.1675 %
Root relative squared error             68.4172 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.95      0.275      0.838     0.95      0.891      0.915    <50
                 0.725     0.05       0.906     0.725     0.806      0.915    >50_1
                 0         0          0         0         0          ?        >50_2
                 0         0          0         0         0          ?        >50_3
                 0         0          0         0         0          ?        >50_4
Weighted Avg.    0.86      0.185      0.865     0.86      0.857      0.915

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 57  3  0  0  0 |  a = <50
 11 29  0  0  0 |  b = >50_1
  0  0  0  0  0 |  c = >50_2
  0  0  0  0  0 |  d = >50_3
  0  0  0  0  0 |  e = >50_4
```

Figure 3.2 Screenshot for accuracy of RBFNetwork function with 86%

ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 2, Issue 2, pp: (22-26), Month: October 2015 – March 2016, Available at: www.paperpublications.org



No.	age Numeric	sex Nominal	chest_pain Nominal	trestbps Numeric	chol Numeric	fbs Nominal	restecg Nominal	thalach Numeric	exang Nominal	oldpeak Numeric	slope Nominal	ca Numeric	thal Nominal	num Nominal
1	28.0	male	atyp_angina	130.0	132.0	f	left_v...	185.0	no	0.0				<50
2	29.0	male	atyp_angina	120.0	243.0	f	normal	160.0	no	0.0				<50
3	29.0	male	atyp_angina	140.0		f	normal	170.0	no	0.0				<50
4	30.0	female	typ_angina	170.0	237.0	f	st_t_...	170.0	no	0.0			fixed_...	<50
5	31.0	female	atyp_angina	100.0	219.0	f	st_t_...	150.0	no	0.0				<50
6	32.0	female	atyp_angina	105.0	198.0	f	normal	165.0	no	0.0				<50
7	32.0	male	atyp_angina	110.0	225.0	f	normal	184.0	no	0.0				<50
8	32.0	male	atyp_angina	125.0	254.0	f	normal	155.0	no	0.0				<50
9	33.0	male	non_anginal	120.0	298.0	f	normal	185.0	no	0.0				<50
10	34.0	female	atyp_angina	130.0	161.0	f	normal	190.0	no	0.0				<50
11	34.0	male	atyp_angina	150.0	214.0	f	st_t_...	168.0	no	0.0				<50
12	34.0	male	atyp_angina	98.0	220.0	f	normal	150.0	no	0.0				<50
13	35.0	female	typ_angina	120.0	160.0	f	st_t_...	185.0	no	0.0				<50
14	35.0	female	asympt	140.0	167.0	f	normal	150.0	no	0.0				<50
15	35.0	male	atyp_angina	120.0	308.0	f	left_v...	180.0	no	0.0				<50
16	35.0	male	atyp_angina	150.0	264.0	f	normal	168.0	no	0.0				<50
17	36.0	male	atyp_angina	120.0	166.0	f	normal	180.0	no	0.0				<50
18	36.0	male	non_anginal	112.0	340.0	f	normal	184.0	no	1.0	flat		normal	<50
19	36.0	male	non_anginal	130.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.0				<50
20	36.0	male	non_anginal	150.0	160.0	f	normal	172.0	no	0.0				<50
21	37.0	female	atyp_angina	120.0	260.0	f	normal	130.0	no	0.0				<50
22	37.0	female	non_anginal	130.0	211.0	f	normal	142.0	no	0.0				<50
23	37.0	female	asympt	130.0	173.0	f	st_t_...	184.0	no	0.0				<50
24	37.0	male	atyp_angina	130.0	283.0	f	st_t_...	98.0	no	0.0				<50
25	37.0	male	non_anginal	130.0	194.0	f	normal	150.0	no	0.0				<50
26	37.0	male	asympt	120.0	223.0	f	normal	168.0	no	0.0			normal	<50
27	37.0	male	asympt	130.0	315.0	f	normal	158.0	no	0.0				<50
28	38.0	female	atyp_angina	120.0	275.0		normal	129.0	no	0.0				<50
29	38.0	male	atyp_angina	140.0	297.0	f	normal	150.0	no	0.0				<50
30	38.0	male	non_anginal	145.0	292.0	f	normal	130.0	no	0.0				<50
31	39.0	female	non_anginal	110.0	182.0	f	st_t_...	180.0	no	0.0				<50
32	39.0	male	atyp_angina	120.0		f	st_t_...	146.0	no	2.0	up			<50
33	39.0	male	atyp_angina	120.0	200.0	f	normal	160.0	yes	1.0	flat			<50
34	39.0	male	atyp_angina	120.0	204.0	f	normal	145.0	no	0.0				<50
35	39.0	male	atyp_angina	130.0		f	normal	120.0	no	0.0				<50
36	39.0	male	atyp_angina	190.0	241.0	f	normal	106.0	no	0.0				<50
37	39.0	male	non_anginal	120.0	339.0	f	normal	170.0	no	0.0				<50
38	39.0	male	non_anginal	160.0	147.0	t	normal	160.0	no	0.0				<50

Figure 3.3 Screen shot of report viewer (CVD)

In figure 3.1 and 3.2 shows the naïve bayes and the RBFNetwork function results.Both the naïve bayes and the RBFNetwork function will provides the 86% of accuracy. The Naïve Bayes algorithm provides 0.7009 kappa statistics and the RBFNetwork function provides 0.6983 kappa statistics. The Naïve Bayes theorem provides the highest kappa statistics. Figure 3.3 shows the CVD report.

4. CONCLUSION AND FUTURE WORK

Naïve Bayes, Logistic function, RBF Network, Decision Table, SMO function algorithms are used for testing and the testing results show that the Naïve Bayes algorithm outperforms with 86% accuracy in 0 seconds. A comparative study is applied to determine the most effective techniques that are capable for the detection of heart valve disease with a high accuracy. To implement the combination of classification techniques to improve the performance of the algorithms.

REFERENCES

[1] Anita Devi, Abhishek Misal, “A Survey on Classifiers Used in Heart Valve Disease Detection” ,International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering Vol. 2, Issue 1, January 2013

[2] K.R. Lakshmi , M.Veera Krishna and S.PremKumar,“Performance comparison of Data Mining Techniques for Predicting of Heart Disease Survivability”, International Journal of Scientific and Research Publications, Volume 3, Issue 6, June 2013 1 ISSN 2250-3153

[3] B.Venkatalakshmi, M.V Shivsankar ,“Heart Disease Diagnosis Using Predictive Data mining” International Journal of Innovative Research in Science, Engineering and Technology Volume 3, Special Issue 3, March 2014 International Conference on Innovations in Engineering and Technology (ICIET’14)

[4] Sona Baby, Ariya T.K, "A survey paper of data mining in medical diagnosis", international journal of research in IJRCCT computer and communication technology.

[5] K.Srinivas, Dr. G. Raghavendra Rao and and Dr. A. Govardhan, "survey on prediction of heart morbidity using data mining techniques",International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.3, May 2011

[6] Beant Kaur, Williamjeet Singh," Review on Heart Disease Prediction System using Data Mining Techniques" International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 2 Issue: 10

[7] Prof. Gondkar Mayura D.1, Prof. Pawar Suvarna E, " A Survey On Data Mining Techniques To Find Out Type Of Heart Attack", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727Volume 16, Issue 1, Ver. V (Jan. 2014), PP 01-05

[8] Vikas Chaurasia,etal,Carib.j.SciTech, "Early Prediction of Heart Diseases Using Data Mining Techniques" , ,2013,Vol.1,208-217

[9] Aqueel Ahmed, Shaikh Abdul Hannan, "Data Mining Techniques to Find Out Heart Diseases: An Overview", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-1, Issue-4, September 2012

[10] Dr.Hari Ganesh S and Gajenthiran M, "Comparative study of Data Mining Approaches for prediction Heart Diseases", IOSR Journal of Engineering (IOSRJEN) www.iosrjen.org ISSN (e): 2250-3021, ISSN (p): 2278-8719 Vol. 04, Issue 07 (July. 2014), ||V3|| PP 36-39

[11] Devendra Shah, Ketan Kadam, Shubham Shinde , Akash Doiphode, "Heart Disease Prediction: A Data Mining Aspect", International Journal of Engineering Technology, Management and Applied Sciences

[12] Zarna Parekh, Avaniba Parmar, "Survey Paper on Early Diagnosis Ofcardio-Vascular Disease Using Data Mining And Neural Network", IJSRD - International Journal for Scientific Research & Development| Vol. 2, Issue 10, 2014 ISSN (online): 2321-0613

[13] Jyoti Soni, Ujma Ansari, Dipesh Sharma and Sunita Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011

[14] Vikas Chaurasia and Saurabh Pal, "Data Mining Approach to Detect Heart Dieses", International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol. 2, No. 4, 2013, Page: 56-66, ISSN: 2296-1739.