A Vote Share based Enhance Hybrid Classifier for Heart Disease Prediction

Ayushi Singh
Student, M.E.(CSE)
Medicaps Institute of Technology and Management
Indore, MP, India

Suneet Joshi
Professor, Dept, CSE
Medicaps Institute of Science and Technology
Indore, MP, India

ABSTRACT
The data mining is current age technology; it has a rich number of applications and domains for research and development. A number of researches are contributing in the different applications for improving the decision making, classification and other automated data analysis techniques. The proposed work is investigation of the data mining techniques for implementing with the predictive data analysis applications. Therefore a medical domain application is namely heart disease prediction system is desired to develop and implement. In observations that are found the heart disease prediction system can be implementable with the data mining based classifiers. But in most of the cases these classifiers are producing poor outcomes therefore a new technique for improving the classification performance is proposed and implemented in this work. The proposed classification technique includes the goodness of Bayesian classifier and neural network to reform the issues of single classification technique. The proposed classifier also includes a combined outcome generation technique for heart disease prediction. The combined outcomes are generated by incorporating the outcomes of both the implemented classifiers using the vote share basis. Additionally for computing the vote shares the validation outcomes are utilized with the formulation of the proposed technique. The implementation of the proposed technique is performed using the java technology and after implementation the performance study performed with respect to traditional Bayesian classification technique. For comparing the performance of both the implemented classifiers the accuracy, memory consumption, error rate and training time of the algorithms are considered as the key factor of comparison. According to the obtained performance the proposed classification technique improves the performance of traditional classification algorithms by vote share based technique. Thus the presented work is adoptable and efficient for machine learning and prediction applications where the accuracy is the key factor to achieve.

Keywords
Data Mining, Classification, Prediction, Heart Disease Prediction System, Hybrid Classifier.

1. INTRODUCTION
The data mining is a technique of analysing the data and extracting the essential patterns from the data. These patterns are used with the different applications for making decision making and prediction related task. The decision making and prediction is performed on the basis of the learning of algorithms. The data mining algorithms supports both kinds of learning supervised and unsupervised. In unsupervised learning only the data is used for performing the learning and in supervised technique the data and the class labels both are required to perform the accurate training. In supervised learning the accuracy is maintained by creating the feedbacks form the class labels and enhance the classification performance by reducing the error factors from the learning model.

The proposed work is intended to investigate these techniques in the application of the predictions. Therefore the heart disease prediction system is proposed to develop and implement. In the past decades, data mining have played an important role in heart disease research [1]. The proposed heart disease prediction system utilizes the aspects of the supervised learning for predicting the class labels of the input pattern of heart dataset samples. The proposed predictive data mining technique is being developed in the hybrid manner for predicting accurate class labels. Because the hybrid classifier includes the goodness of both kind of classifiers in the same place and improve the classification performance.

The proposed hybrid classification technique incorporates the neural network and Bayesian classification algorithms for implementing the proposed concept. In this classification model first the algorithm is trained over the training samples and a trained data model prepared by both the classification algorithms individually. In further the cross validation is performed with the same dataset as produced for training. There are two classifiers are implemented then a cross validation process direct the use of voting. Additionally using the voting process the test data is classified and their class labels are predicted.

2. PROPOSED WORK
The proposed work is intended to find the solution for the accurate classification technique development. That technique is implemented using two different classifiers contributions, namely Bayesian classifier and the back propagation neural network. The given section provides the understanding of the proposed technique development and their functional aspects. Therefore the proposed classification algorithm is prepared in two main modules training and testing. The functional architecture for the training of classifiers is provided using the figure 1 additionally there components and subcomponents are described in detail.

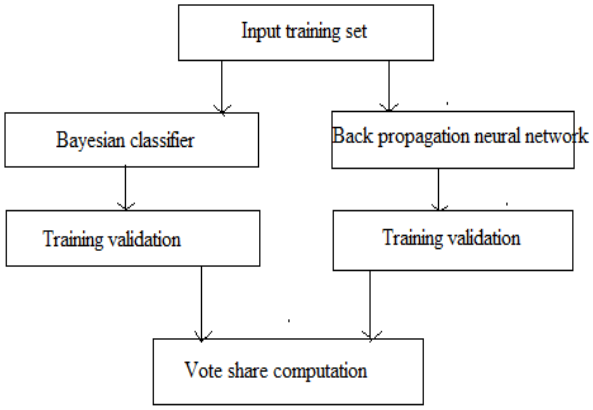

Figure 1 Training process

Input training set: The supervised classifiers are working to analyze data and for preparing the trained data model using the trained model the classification task is performed. The training process needs some pattern input or example pattern on which the classifier performs training. The training datasets are found in both the formats namely structured or unstructured format. In this presented work the structured data is utilized for training and testing of the developed classifiers. In this dataset the list of attributes are organized with their predefined class labels. During the classifiers training these class labels are used to develop accurate approximation of patterns.

Bayesian classifier: The input training samples are produced in the Bayesian classifier to perform the training. Therefore the algorithm computes the respective probability distribution for each class labels appeared on the training data samples according to their attributes. Thus the prior probability of the data according to their class labels is estimated during the Bayesian classifier's training.

Back propagation neural network: Back propagation neural network is a classical model of machine learning and classification. It usages the distributed concept of computation to learn over the pattern and frequently used in various different applications of machine learning and pattern recognition. In this training the input training samples are first encoded into binary string and then utilized for making the training. After training of the model a trained data model is prepared which can be used to distinguish the similar patterns on which the model is trained before.

Training validation: In this phase the trained data models namely Bayesian classifier and back propagation neural network is tested using the n-cross validation process. In this process the training sample is randomized and new samples using training data is prepared which is provided to the classifiers for performing the classification. According to the generated N number of test sets the mean performance is evaluated for training validation.

Vote share computation: After validation of trained classifiers the entire voting score is distributed in four parts and the vote shares are decided according to the performed training and the validation outcomes. To understand more clearly consider an example, the entire voting score is denoted by 1 and for the wining classifier one third part of voting is considered by the wining classifier and one fourth part of voting is considered for the losing classifier. For example there are two classifiers are available $C_1$and C_2. The classifier C_1 produces 80% accuracy and C_2 returns 90% accuracy during the validation process thus the vote wining classifier is C_2. Here the C_2 having the share of 0.75 and the C_1classifier having the 0.25 share for voting.

After computing the vote share of included classifiers the testing is performed, therefore the testing operation can be defined using the given figure 2. The participating components of the testing module are described as:

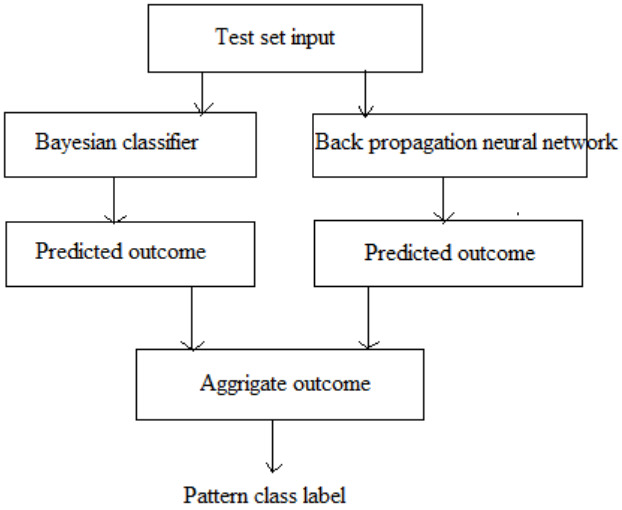

Figure 2 Testing model

Test input: After training of the classifiers the models are enabled to classify the input patterns based on their performed training. Thus in order to test the prepared trained model a set of input samples are produced to the trained classifiers for predicting the class labels of the test datasets. That can be prepared by manual entry or by producing the set of samples using a CSV file.

Bayesian classifier: The input test samples are produced to the Bayesian classifier for classification and prediction of their class labels. The Bayesian classifier accepts the test dataset and classifies each sample and for each samples a class label is produced.

Back propagation neural network: That is functioning similar as the trained Bayesian classifier. The neural network also accepts the input test samples and generates the class labels for the entire input test set instances.

Predicted outcome: These outcomes are class labels that produced by both the algorithms namely Bayesian classifier and the back propagation neural network. Here in this experiment these outcomes are treated as raw outcomes of the classifier and need to aggregate the final outcomes.

Aggregated outcome: The given component accept the outcomes of both the classifiers and responsible to generate a combined outcome for the hybrid concept. Thus first the wining classifier is decided according to the previous validation performance. Finally the outcome of the system is aggregated using the following formula.

$$class_{out} = V_1^c * P_{c1} + V_2^c * P_{c2}$$

Where

$class_{out}$ = the final predictable outcome of the proposed classifier

V_1^c=vote share of first classifier (Bayesian classifier)

V_2^c= vote share of second classifier (back propagation neural network)

P_{c1}= predicted class label by Bayesian classifier

P_{c2}= predicted class label by back propagation neural network

Pattern class labels: The combined outcome of both the classifier are computed and the computed outcome of the classifier is produced as the classification class label of the input pattern which is denoted as $class_{out}$.

Proposed algorithm

The given section introduces the summarized step of both the processes (training and testing) in terms of algorithm steps both the algorithms are demonstrated as:

Table 1 Training algorithm

Training algorithm
Input: training dataset D
Process: 1. initialize the classifiers 2. read the training samples D 3. $[V_{bays}, TM_{bays}] = train\_bayesian(D)$ 4. $[V_{BPN}, TM_{BPN}] = train\_BPN(D)$ 5. if $V_{BPN} > V_{bays}$ then 6. $V_{BPN}^c = 0.75$ 7. $V_{bays}^c = 0.25$ 8. else 9. $V_{BPN}^c = 0.25$ 10. $V_{bays}^c = 0.75$ 11. end if

The process of the training is given using table 1 and for testing process the table 2 helps to understand the process involved.

Table 2 Testing algorithm

Testing algorithm
Input: test dataset DT, V_{bays}^c , V_{BPN}^c , TM_{bays} ,TM_{BPN}
Process: 1. read the input test samples DT 2. $P_{bays}^c = TM_{bays}.classify(DT)$ 3. $P_{BPN}^c = TM_{BPN}.classify(DT)$ 4. compute the aggregate outcome $class_{out} = V_{bays}^c * P_{bays}^c + P_{BPN}^c * V_{BPN}^c$ 5. return $class_{out}$

3. RESULT ANALYSIS

The given section provides the study about the proposed classification algorithm and the comparative performance study among the implemented classifiers in different performance factors. The performance outcomes and the estimated analysis are provided in this chapter.

3.1 Accuracy

The accuracy is a measurement of the data model for finding the amount of correctly classified data using the input samples. The performance of the algorithm in terms of accuracy can be evaluated using the following formula.

$$accuracy\ \% = \frac{total\ correctly\ classified\ data}{total\ input\ datasets} X100$$

The performance of the proposed hybrid classifier and the traditional Bayesian classifier is compared using the figure 3 and the table 3.

Table 3 Accuracy

Dataset size	Hybrid classification	Bayesian classifier
50	93.82	100
100	94.66	98.52
150	95.29	95.33
200	96.28	93.52
300	98.36	92.14
400	98.69	91.42
500	99.68	90.72

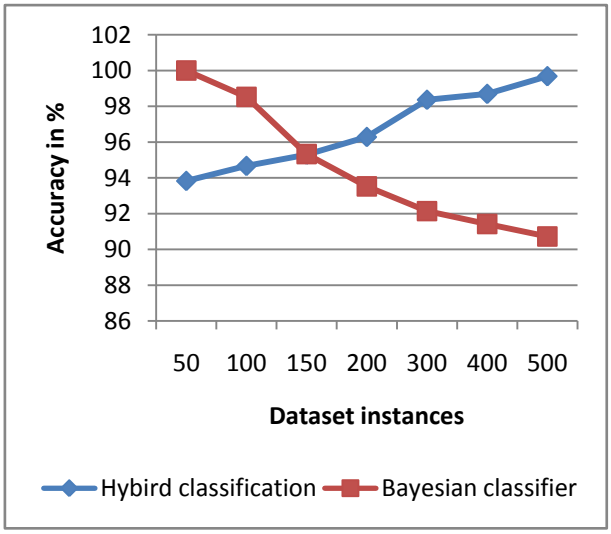

Figure 3 Accuracy

In this diagram the X axis shows the training samples in the dataset and the Y axis shows the obtained accuracy in terms of percentage. The results of both the classifiers are demonstrating the different behavior of classification aspects,

in the traditionally implemented classifier the performance of the classification is reduces as the amount of training instances are increases. On the other hand the performance of the proposed classification technique is increases as the amount of training samples are increase. Thus the proposed classifier performs more effectively as compared to traditional manner of classification. For analyzing the results in the statistical manner the mean accuracy of both the classifiers are computed and their difference in performance is reported using the figure 4. In this diagram the mean performance of both the method in terms of accuracy is demonstrated using the Y axis and the X axis contains the implemented methods for making comparative performance study. According to the obtained performance the proposed classifier is producing approximately 96% of accurate results and the traditional classifier produces the 94 % of accurate results. Thus the proposed classification technique is much efficient and accurate as compared to the traditional technique of data classification.

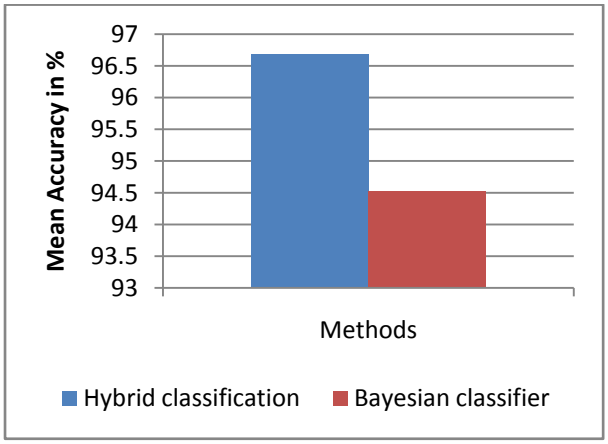

Figure 4 Mean accuracy

3.2 Error rate

The error rate of the classifier provides the estimation about the misclassified samples during the testing of the trained classifier. The evaluation of error rate can be performed using the following formula.

$$error\ rate\ \% = \frac{total\ misclassified\ samples}{total\ input\ samples} X100$$

Or

$$error\ rate\ \% = 100 - accuracy\ \%$$

The comparative error rate of the proposed and traditional classification technique is provided using the table 4 and the figure 5. The given figure includes the X axis to show the size of training samples and the Y axis shows the amount of misclassified patterns in terms of percentage. According to the demonstrated results the error rate of the proposed classifier is reduces as the amount of training instances are increases in the database. On the other hand the error rate of the traditional scheme is increases as the amount of data for learning is increases. Thus the proposed classifier is improving the outcomes of the classification with increasing the learning patterns.

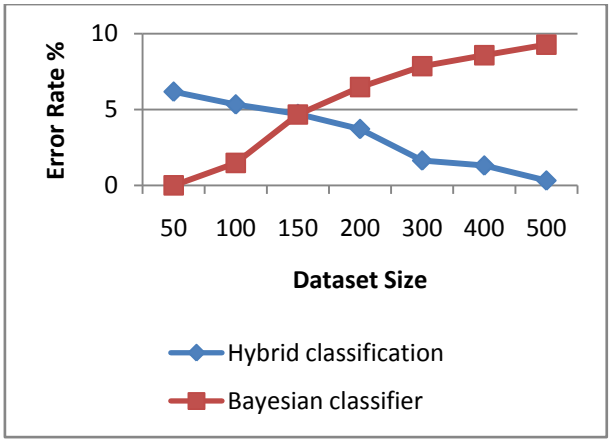

Figure 5 Error rate

Table 4 Error rate

Dataset size	Hybrid classification	Bayesian classifier
50	6.18	0
100	5.34	1.48
150	4.71	4.67
200	3.72	6.48
300	1.64	7.86
400	1.31	8.58
500	0.32	9.28

In order to understand the performance of the classification more clearly the mean error rate percentage is evaluated and reported using the figure 6. In this figure the amount of error rate produced by the algorithms are demonstrated using Y axis and X axis shows the methods implemented with the system.

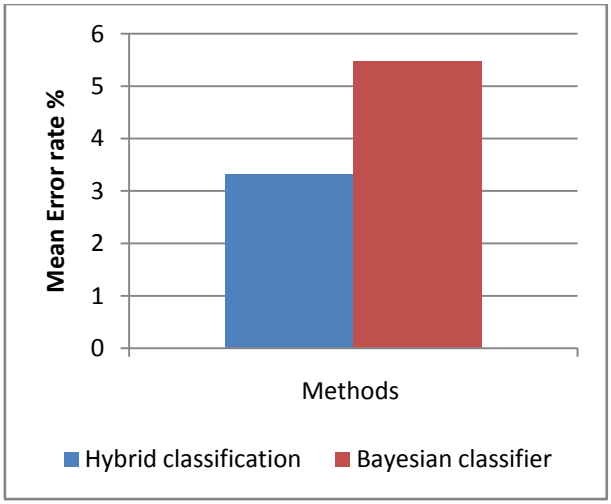

Figure 6 Mean error rate

According to the obtained results the from the mean error rate percentage the proposed hybrid classifier produces more effective and improving performance as compared to the traditional classification technique.

3.3 Memory usages

The amount of main memory required to successfully execute the algorithms is known as the memory consumption of the algorithms. The given figure 7 and the table 5 show the comparative performance of both the implemented classifiers. In the given diagram the X axis shows the number of training input samples produced for the training to the data models and the Y axis shows the amount of main memory consumed by the implemented algorithms. According to the obtained results the amount of memory consumption in the proposed data modeling is higher as compared to traditional technique because the proposed classifier needs to process the data using both the classifiers.

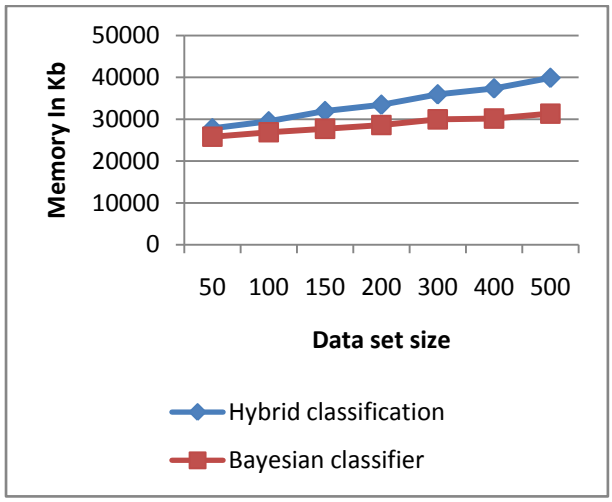

Figure 7 Memory usage

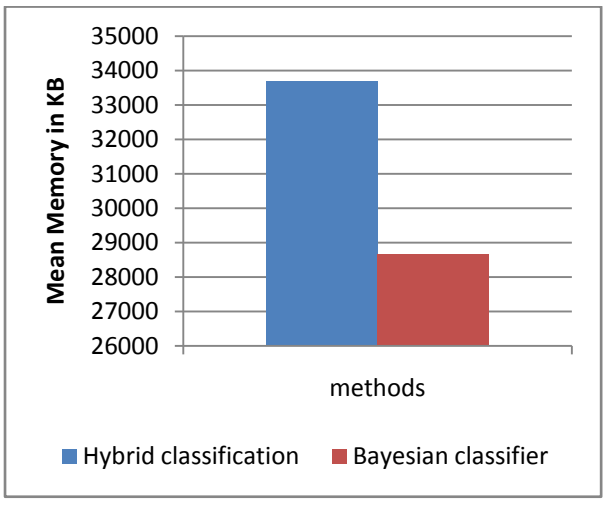

Figure 8 Mean memory consumption

Table 5 Memory usage

Dataset size	Hybrid classification	Bayesian classifier
50	27817	25818
100	29488	26891
150	31938	27716
200	33462	28612
300	35938	29981
400	37362	30164
500	39882	31332

In order to understand the memory usage difference among both the classification technique the mean memory consumption is demonstrated using the figure 8 in this diagram the X axis shows the amount of instances of the data used for training and the Y axis shows the amount of main memory consumed during evaluation of data.

3.4 Time consumption

The amount of time required to process the data using the proposed algorithm is termed here as the time consumption of the system. The comparative time consumption of both the data models during the training is demonstrated using table 6 and figure 9. In this diagram the X axis contains the amount of data used for training and the Y axis shows the amount of time required to process the data samples. According to the obtained results the proposed technique consumes higher time as compared to the traditional classifier. The proposed scheme utilizes the back propagation neural network and learning of this algorithm is an iterative process thus the amount of time is higher as compared to the Bayesian classifier.

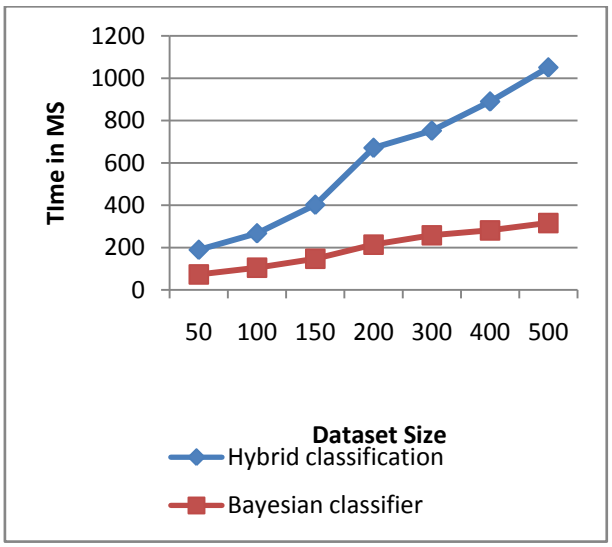

Figure 9 Time consumption

Table 6 Time consumption

Dataset size	Hybrid classification	Bayesian classifier
50	189	73
100	267	105
150	402	147
200	671	214
300	752	258
400	890	281
500	1051	316

In order to understand the difference among both the technique's performance the mean time consumption of both the algorithms are computed. According to the obtained performance the proposed technique consumes more time as compared to the traditional technique. Thus the proposed model is a time consuming model for training time.

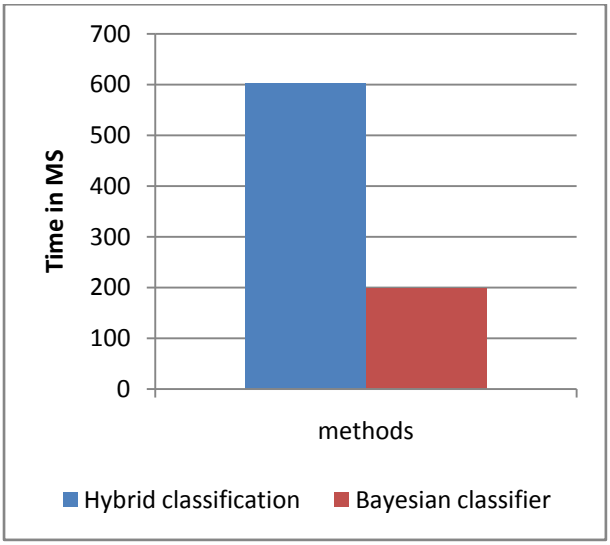

Figure 10 Mean time consumption

4. CONCLUSIONS

The key aim of designing the hybrid classification technique to improve the predictive accuracy is designed and implemented. Based on the experiments some essentials facts are obtained that are discussed in this chapter. In addition of that the limitations and the future extension of the work is also included.

4.1 Conclusion

The data mining is a technique by which the computer based algorithm analyzes the data and provides the meaningful outcomes for utilizing with the real world applications for problem solving. In this context the data mining domain provide ease in various domains of engineering, science and other industries. The use of data mining is also performed for finding the trends in data and pattern for consuming with the predictive data analysis. In this presented work the data mining techniques are investigated for obtaining the application in predictive manner.

Therefore the key work is focused on finding the predictive model for heart disease prediction. The heart disease prediction is performed by the learning of patterns available in old trends. A number of data models for predicting the class labels of input patterns are available but most of them are not much accurate for medical data analysis. Therefore a hybrid concept for analyzing the data is proposed in this work to enhance the learning of model and produce more accurate outcomes on the basis of the previous patterns.

The proposed data model includes the implementation of two different classifiers and using the weight based classification technique the model is used to predict the class labels. The key advantages with the hybrid classifiers are, these classifiers are able to combine the goodness of the implemented classifiers. The utilized algorithms for combining the outcomes are bays classifier and the neural network classifier. Both the models are taking training individually and produce the combined results during the predictive data analysis.

The implementation of the proposed technique is performed with the help of JAVA technology and their performance is evaluated. The obtained performance is estimated in terms of memory consumption, training time required, accuracy and the error rate of the classifier. According to the obtained results the performance summary is prepared and demonstrated using the table 7.

Table 7 Performance summary

S. No.	Parameters	Hybrid classifier	Bayesian classifier
1	Accuracy	High	Low
2	Error rate	Low	High
3	Memory usage	High	Low
4	Time consumption	High	Low

According to the performance summary the accuracy of the proposed classier is more effective as compared to the traditional technique. But the performance of the proposed classifier is lacked in terms of resource consumption. Therefore the proposed model is suggested to implement in those cases where the accuracy is more desired as compared to the time and the memory.

4.2 Future work

The proposed work for enhancing the classification performance is design and implemented, the experimental results demonstrate the effective performance of the proposed classifier. But the performance is not much significant due to the lack in memory and time consumption. Therefore in near future that is required to work in both the domain of performance improvement as compared to the accuracy of the classifier. Thus the model is extendable for limiting the memory and time consumption.

5. REFERENCES

[1] AH Chen, SY Huang, PS Hong, CH Cheng, EJ Lin, "HDPS: Heart Disease Prediction System", Computing in Cardiology 2011; 38:557-560.

[2] Data Mining: What is Data Mining?, http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[3] Data Mining - Applications & Trends, http://www.tutorialspoint.com/data_mining/dm_applications_trends.htm

[4] MahakChowdhary, ShrutikaSuri and MansiBhutani, "Comparative Study of Intrusion Detection System", 2014, IJCSE All Rights Reserved, Volume-2, Issue-4

[5] Mrs. PradnyaMuley, Dr. Anniruddha Joshi, "Application of Data Mining Techniques for Customer Segmentation in Real Time Business Intelligence", International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163, Issue 4, Volume 2 (April 2015)

[6] GhazalehKhodabandelou, Charlotte Hug, Rebecca Deneckere, Camille Salinesi, "Supervised vs. Unsupervised Learning for Intentional Process Model Discovery", Business Process Modeling, Development, and Support (BPMDS), Jun 2014, Thessalonique, Greece. pp.1-15, 2014

[7] Importance of Predictive Analytics in Business, http://www.orchestrate.com/blog/importance-of-predictive-analytics-in-business/

[8] David A. Dickey, N. Carolina State U., Raleigh, NC, "Introduction to Predictive Modeling with Examples", Statistics and Data Analysis, SAS Global Forum 2012

[9] Hand, Manilla, & Smyth, "Descriptive Modeling", http://www.stat.columbia.edu/~madigan/DM08/descriptive.ppt.pdf

[10] K.Jayavani, "STATISTICAL CLASSIFICATION IN MACHINE INTELLEGENT", ISRJournals and Publications, Volume: 1 Issue: 1 18-Jul-2014, I

[11] Chaitrali S. Dangare, Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012

[12] JyotiSoni, Ujma Ansari, Dipesh Sharma, SunitaSoni, "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

[13] Shadab Adam Pattekari and AsmaParveen, "Prediction System for Heart Disease Using Naive Bayes", International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294

[14] N. AdityaSundar, P. PushpaLatha, M. Rama Chandra, "Performance Analysis of Classification Data Mining Techniques Over Heart Disease Data Base", International Journal of Engineering Science & Advanced Technology, Volume-2, Issue-3, 470 – 478

[15] R. Thanigaivel, Dr. K. Ramesh Kumar, "Review on Heart Disease Prediction System using Data Mining Techniques", Asian Journal of Computer Science and Technology (AJCST) Vol.3.No.1 2015 pp 68-74

[16] M.I. López, J.M Luna, C. Romero, S. Ventura, "Classification via clustering for predicting final marks based on student participation in forums", Proceedings of the 5th International Conference on Educational Data Mining

[17] Neeraj Shah, Valay Parikh, Nileshkumar Patel, Nilay Patel, ApurvaBadheka, AbhishekDeshmukh, AnkitRathod, James Lafferty, "Neutrophil lymphocyte ratio significantly improves the Framingham risk score in prediction of coronary heart disease mortality: Insights from the National Health and Nutrition Examination Survey-III", International Journal of Cardiology, 2013 Elsevier Ireland Ltd. All rights reserved.

[18] P.K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules", Journal of King Saud University – Computer and Information Sciences (2012) 24, 27–40

[19] Nicholas P. Tatonetti, Patrick P. Ye, Roxana Daneshjou, and Russ B. Altman, "Data-Driven Prediction of Drug Effects and Interactions", Published in final edited form as: SciTransl Med. 2012 March 14; 4(125): 125ra31. doi:10.1126/scitranslmed.3003377.

[20] Peter C. Austin, Jack V. Tu, Jennifer E. Ho, Daniel Levy, and Douglas S. Lee, "Using methods from the data mining and machine learning literature for disease classification and prediction: A case study examining classification of heart failure sub-types", Published in final edited form as: J ClinEpidemiol. 2013 April ; 66(4): 398–407. doi:10.1016/j.jclinepi.2012.11.008

[21] SumanBala, Krishan Kumar, "A Literature Review on Kidney Disease Prediction using Data Mining Classification Technique", IJCSMC, Vol. 3, Issue. 7, July 2014, pg.960 – 967

[22] A. J. M. Abu Afza, Dewan Md. Farid, and ChowdhuryMofizurRahman, "A Hybrid Classifier using Boosting, Clustering, and Naïve Bayesian Classifier", World of Computer Science and Information Technology Journal (WCSIT), ISSN: 2221-0741, Vol. 1, No. 3, 105-109, 2011

[23] ShwetaPandey, Prof. Megha Mishra, "Cryptanalysis of Feistel cipher using Back propagation Neural Network", International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 3, March 2012)

[24] Pratik Gite, Sanjay Thakur, "An Effective Intrusion Detection System for Routing Attacks in MANET using Machine Learning Technique", International Journal of Computer Applications (0975 – 8887) Volume 113 – No. 9, March 2015