VOL 1 (2017) NO 4 - 2

e-ISSN : 2549-9904
ISSN : 2549-9610

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

Comparative Analysis for Heart Disease Prediction

Sundas Naqeeb Khan[#], Nazri Mohd Nawi[#], Asim Shahzad[#], Arif Ullah[#], Muhammad Faheem Mushtaq[#], Jamaluddin Mir[*], Muhammad Aamir[#]

[#] *Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400, Johor, Malaysia*
[*] *The University of Lahore, Islamabad Campus, Pakistan*
E-mail: nazri@uthm.edu.m

Abstract— **Today, heart diseases have become one of the leading causes of deaths in nationwide. The best prevention for this disease is to have an early system that can predict the early symptoms which can save more life. Recently research in data mining had gained a lot of attention and had been used in different kind of applications including in medical. The use of data mining techniques can help researchers in predicting the probability of getting heart diseases among susceptible patients. Among prior studies, several researchers articulated their efforts for finding a best possible technique for heart disease prediction model. This study aims to draw a comparison among different algorithms used to predict heart diseases. The results of this paper will helps towards developing an understanding of the recent methodologies used for heart disease prediction models. This paper presents analysis results of significant data mining techniques that can be used in developing highly accurate and efficient prediction model which will help doctors in reducing the number of deaths cause by heart disease.**

Keywords— **Classifiers, Heart Disease Analysis.**

I. INTRODUCTION

Data mining is the process of finding formerly unknown patterns and trends in the existing data and by means of this extracted information, build predictive and usable models. The field of data mining has not been particularly useful in medical sciences, but this trend is on a fast track path of change. Today, healthcare industry delivers a broad measure of complex information with respect to, hospitals, patients, electronic patient records, disease prognosis and diagnosis and medical health care devices. This huge amount of data needs to be mined and filtered so as to enable us to extract useful information that can be useful [1].

In spite of immense technological development in healthcare sector, developing countries and all around the world, there are still in need of providing quality healthcare services at a reasonable cost that is easily affordable by their population. Although such countries have seen large scale development in terms of providing better health care facilities, yet there is still a huge demand in terms of making these facilities affordable [2].

(a) Knowledge discovery in medical databases

Data mining is a very important step in knowledge discovery. For the past few years, it has gained huge amount of interest in the field of Data Sciences. The process of knowledge discovery comprises of an iterative data cleaning, data incorporation, data assortment, configuration recognition and lastly knowledge depiction [3].

(b) Heart Disease

Heart is the blood-pumping organ that provides oxygen and other supplements to all tissues throughout the human body. Unusual heart activities can be harmful for various other organs of human body e.g. brain, kidneys etc. ceasing of cardio-vascular activity can result in an instantaneous death of an individual [4].

There exists a huge amount of research on cardio-vascular diseases and their diagnosis. Research has provided several different methodologies for the treatment of such diseases. An overview of such research is given below [5]:

Milan Kumari designed system known as Data Mining Classification strategies namely, RIPPER and decision tree using artificial neural network support vector machine (SVM), to explore the coronary disease dataset. SVM predicts coronary diseases (CVD) with least error rate and most remarkable precision [6].

Colombet et al. [8] assessed the use of CART and artificial neural networks (ANN) with the intent of predicting heart diseases in individuals. Nidhi Bhatla and Kiran Jyoti

used 15 traits in their survey for the expectation of finding and predicting coronary diseases [7]. G. Parthiban et al published a theory that they termed as "Chances of diabetic patient getting heart disease". The accuracy of this theory was verified by the researchers by applying Naïve Bayes classifier, which yields very best prediction form by means of the minimum amount of training set [8]. Jyoti Soni et al. [9] performed a large number of experiments to predict the heart disease on a particularly useful dataset. The results showed that Decision Tree performs with highest accuracy. However, they found Bayesian classification to have related truthfulness as that of decision tree method. They observed that other predictive methods like KNN, in neural network used classification that not performs on such dataset.

M. Anbarasi et al. [10] used Genetic Algorithm to determine such attributes that play a vital role in contributing towards the diagnosis of cardiac disease. The research work indirectly reduced the figure of tests needed to be taken by a patient to determine the presence of any heart disease. Decision Tree performance after incorporating subset selection was found to be quite remarkable.

Robert Detrano performed experimental results that exhibited precise classification of heart diseases, having an accuracy of nearly 77% by using logistic regression resulting discriminant purpose [10]. Zheng Yao used a novel model called R-C4.5 and was able to improve the performance of attribute selection and partitioning models. Their experiments exhibited that the rules formed by R-C4.5s could be beneficial in providing health care experts with clear and useful information regarding heart diseases [11]. Resul Das [12] then presented methodology that used SAS based software on behalf of the diagnosis of heart disease. A neural networks based technique was used by this system.

Another quite recent method is Associative classification which incorporates association rule mining and classification to a form for calculation and manages to achieve greatest accuracy. Associative classifiers are particularly suitable for applications where highest accuracy is required for prediction model [13].

This paper describes the accuracy of the different classifiers in classifying heart diseases dataset. It is organized as follows: Section 2 provides the method that had been used for performing the simulation. Section 3 performs the results in the form of graphs. Section 4 gives us the conclusion of this paper.

II. Material And Method

Due to certain resource constraints, this paper presents an analysis of a number of data mining techniques, which might be supportive for health care professionals in helping them to perform accurate analysis of heart diseases.

In this research, we used four classifiers for prediction of heart disease by using Weka version 3.6. Weka is a commonly used tool in data mining. The initial dataset comprised of 14 attributes, 303 patients record and an algorithm that was responsible for attribute selection. The algorithm was applied on dataset for pre-process. After attribute selection, certain missing values were identified that were subsequently deleted from the dataset. After deleting of missing records, 296 records were left. Out of these remaining 296 records in the data set, they were subject to highly efficient data mining techniques, namely RIPPER, Decision Tree, Artificial Neural Networks (ANNs) and Support Vector Machine (SVM).

A prominent confusion matrix was derived with the intention of calculating the sensitivity, specificity and accuracy of the results.

Below given formulae were used for calculating the parameters:

$$\text{Sensitivity} = TP / (TP + FN) \quad (1)$$

$$\text{Specificity} = TN / (TN + FP) \quad (2)$$

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN) \quad (3)$$

$$\text{True Positive rate} = TP / (TP + FN) \quad (4)$$

$$\text{False Positive rate} = FP / (FP + TN) \quad (5)$$

A Receiver Operating Characteristic (ROC) space depicts a comparative exchange between the positive and false positive.

III. Resuts And Discussion

Table 1 shows the results of sensitivity, specificity, accuracy, True Positive and False Positive meant for the different classification techniques.

TABLE 1
COMPARISON OF DATA MINING MODELS

	Sensitivity	Specificity	Accuracy	TP rate	FP rate
RIPPER	86.25%	75.82%	81.08%	0.8625	0.2410
Decision Tree C4.5	83.12%	74.26%	79.05%	0.8312	0.2573
ANN (MLP)	83.75%	75.73%	80.06%	0.8375	0.2426
SVM	90.00%	77.20%	84.12%	0.9000	0.2279

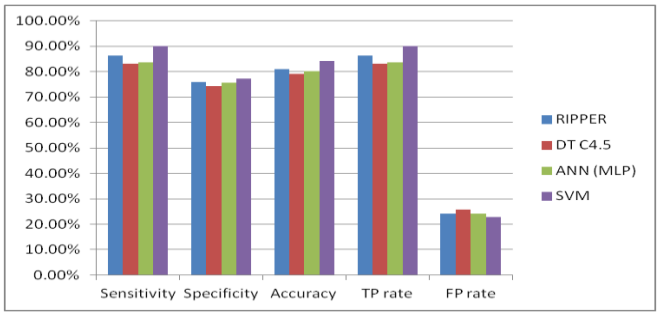

Fig. 1. Graphical representation of Data mining models with TP and FP rate

According to the results in Figure 1, data mining models SVM was found to be the best predictor of heart diseases [14].

According to research results of Nidhi Batla, a prediction system for heart disease was developed by using 15 attributes. Weka was the tool that used for the experiment. In the beginning, missing values were identified in the dataset and then they were substituted with appropriate values that use Replace Missing Values filter. The researcher then used Decision Tree, Naïve Bayes and Neural Networks for calculating the precision of the dataset.

Table 2 depicts the outcomes of this study and it shows that a neural network has in fact superior accuracy as compared to further data mining techniques.

TABLE 2
COMPARISON OF DATA MINING TECHNIQUES

Classifiers	Performance
Naïve Bayes	90.74%
Decision Tree	99.62%
Neural Networks	100%

Figure 2 shows the graphical results of diverse data mining techniques in terms of accuracy of heart disease prediction. The Neural Network based method is found to be the best classification technique as compared to others two methods [4].

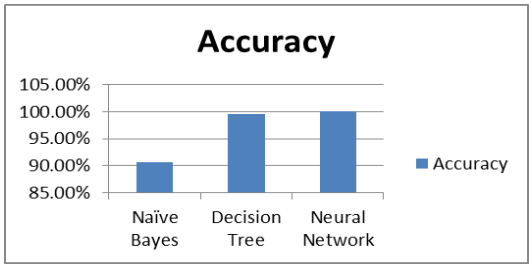

Fig. 2: Graphical Representation of Classification Techniques

It is important to note here that excluding sex and family heredity attributes, all the others attributes have numeric values. For sex, we indicate either "M" or "F" for male and female patients, respectively. For the attribute of family heredity attribute, we values like "Father", "Mother" or "Both". In such case where the patient has no record of diabetes in previous generations, the attribute value in the table is left empty.

Table 3 shows the probability of diabetic patient having heart disease by applying Naïve Bayes data mining classifier. This technique generates an most favorable prediction model by using minimum training set. Figure 3 shows the graphical form of the results that shows Naïve Bayes method is more suitable for application on a diabetic patient as there is an increased probability of the diabetic patient gets heart disease [5].

TABLE 3.
RESULTS OF CLASSIFIED INSTANCES WITH DIFFERENT EXPERIMENTS

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Incorrectly Classified Instances	0.312	0.092	0.571	0.312	0.404	0.695
Correctly Classified Instances	0.908	0.688	0.571	0.908	0.834	0.695
Weighted Average	0.74	0.52	0.714	0.74	0.712	0.695

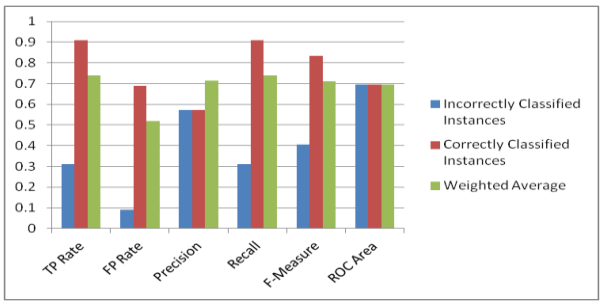

Fig. 3: Graphical Representation of results

Intelligent Heart Disease Prediction System (IHDPS) is another remarkable prediction system that uses the three commonly used data mining techniques, namely, Decision Trees, Naïve Bayes and Neural Network. IHDPS is in fact a web-based prediction that has been observed to be quite user-friendly and scalable. The traditional prediction systems lack the ability to answer "what if" questions that are answered by IHDPS. The initial data set had 909 records and 15 attributes. These were then split into two equal data sets of equal size. The training data set had 455 records and testing dataset had 454 records. It shows the results of above mentioned techniques. According to the results, Naïve Bayes method has the accuracy of 86.5% accurate predictions, followed by Neural Network (85.53% accuracy). Whereas decision Trees method proves to be most effective by having 89% accuracy [6].

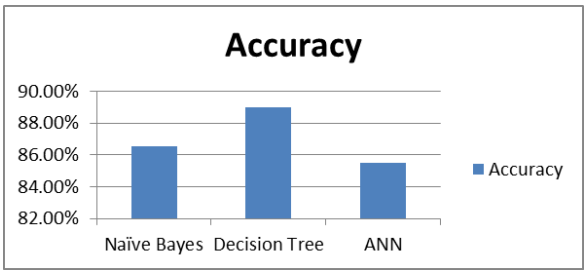

Fig. 4: Results representation obtained from IHDPS

Table 4 shows the observes that the Decision Tree data mining method performs better as compared to Naïve Bayes and ANN when it incorporates subset selection, having high model construction time. However, it must be noted that Naïve Bayes shows steadily before and after reducing attributes having same model construction time. On the other hand, cluster classification performs poorly in comparison to the further two methods [7].

TABLE 4
COMPARISON OF THREE CLASSIFIERS

Classifiers	Performance	Time of construction	MAE
Naïve Bayes	96.5	0.02	0.044
Decision Tree	99.2	0.09	0.00016
Classification via Clustering	88.3	0.06	0.117

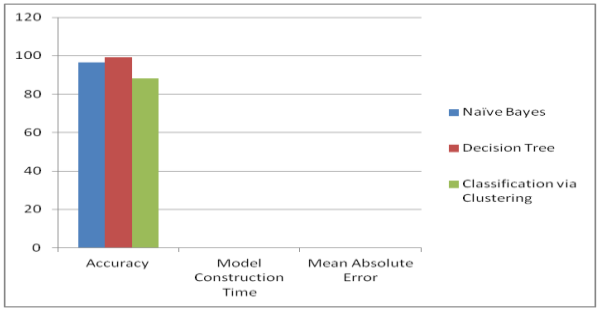

Fig. 5: Comparative Representation of three classifiers

Research done by Asha Rajkumar suggests that data classification depends on machine learning algorithms, results in higher accuracy. In this research, Tanagra was used for data classification and the data was evaluated by employing 10-fold cross validation. The results were compared after this process. The training data had 3000 instances with 14 unique attributes. These instances in the dataset were the results of different types of testing procedures that were performed on the patients to predict the occurrence of heart disease. The dataset was divided into two parts such that 70% of the data was used for training and 30% for testing.

The table 5 below contains secondary values of different classifications. These classification algorithms were compared, and it was experimental that Naive Bayes (NB) algorithm shows improved performance than the other two methods. This is primarily because it takes only few milliseconds to calculate the accuracy [18].

TABLE 5.
PERFORMANCE STUDY OF ALGORITHM

Classifiers	Performance	Time for construction
NB	52.33%	609ms
Decision Tree	52%	719ms
K-NN	45.67%	1000ms

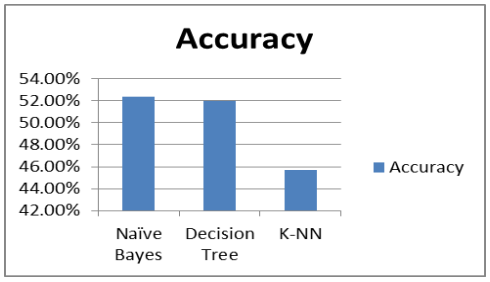

Fig.6. Predicted Accuracy of Algorithm

According to the values in table 6, the accuracy was calculated by using three main attributes, namely left ventricle hypothesis, normal and stress abnormal. Performance was determined because of accuracy comparison. Naive Bayes algorithm was observed to be having better performance [15].

TABLE 6.
ALGORITHMS PERFORMANCE ACCORDING TO RECALL AND PRECISION

Algorithm used	Values	Recall	Precision
Naïve Bayes	Left Vent hypertrophy	0.4828	0.4853
	normal	0.5705	0.4753
	St-t-abnormal	0.0000	1.0000
Decision List	Left Vent hypertrophy	0.4897	0.4855
	normal	0.5705	0.4688
	St-t-abnormal	0.0000	1.0000
K-NN	Left Vent hypertrophy	0.4552	0.5479
	normal	0.4765	0.539
	St-t-abnormal	0.0000	1.0000

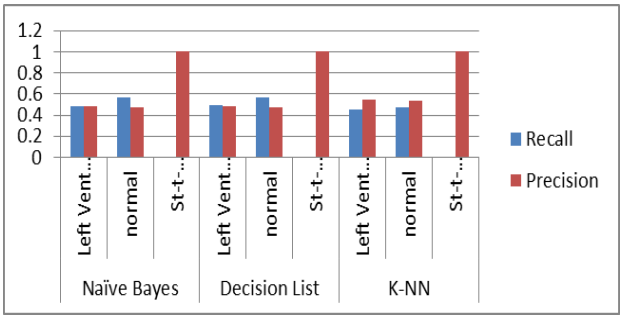

Fig. 7: Graphical representation of Algorithms

Sellapan et al used an original data with 909 records and 13 attributes. To simplify the data set, attributes were categorized for all models. By using Genetic Algorithm with Feature Subset Selection, the figure of attributes was condensed to six. The comprise data set was then given to three classification models i.e. Naïve Bayes, Decision Tree and Classification via Clustering. K-fold cross validation method was used as the test form. The analysis of attribute was the class identifier having value "buff" which indicated no cardiac disease and having value "sick" which indicated the occurrence of cardiac disease [16].

Genetic Algorithm uses natural evolution methodology to find a solution of the given problem in the unlimited search space. The search process in genetic algorithm begins with zero attributes, and an initial population, which is generated randomly. Depends upon the natural concept of survival of the fittest, new population is generated that is supposed to be the best in the current generation. Whereas the off springs created from any current population, have the best traits of their parents. Offspring are produced by the application of genetic operators i.e. cross over and mutation. The process of creating subsequent generations continues until a point is reached where it evolves a population P, every trait in P satisfying the fitness criteria. Having the initial population of 20 instances, creation of generations continued till the twenty generations, with crossover probability of 0.6 and mutation probability of 0.033. The genetic search short-listed six attributes out of thirteen.

Heart attack prediction has been presented by another export he discuss the abstraction of substantial in 2009.accoeding to this approach data warehouse is preprocessed to make for mining process .after the data process the data warehouse make the data in group using k clustering algorithm which show relevant data. And MAFIA algorithm used to mine the heart disease .which Acura like hood due to substantial age. The neural network is trained so as to enable efficient prediction of heart attack among the susceptible patients. A multi-layer Perceptron Neural Network with Back-propagation is used as training algorithm.

IV. CONCLUSION

All the above discussion showed the results of the different research papers. Results were discussed regarding the prediction of any type of heart diseases by applying Data Mining techniques with their classifiers and extension of the classifiers. In the current paper, the study pointed out different classifiers that show good results. Different research papers used different classifiers, algorithms, or techniques such as Support Vector Machine (SVM), Neural Networks (NN), and

Naïve Bayes (NB) and also its extensions, different types of Decision Trees (DT) versions, K-Nearest Neighbor (K-NN), Artificial Neural Networks (ANN), Multi-Layer Perceptron (MLP), Genetic Algorithms, and Feature Subset Selections etc. However, the NB, DT and the SVM have more accurate results as compared to other methods.

In future, we will expand this work and will apply a probabilistic approach test on these three classifiers and hope to get results that are more effective for the prediction of heart disease.

ACKNOWLEDGMENT

The authors would like to thank Universiti Tun Hussein Onn Malaysia (UTHM) Ministry of Higher Education (MOHE) Malaysia for financially supporting this Research under Trans-displinary Research Grant Scheme (TRGS) vote no. T003. This research also supported by GATES IT Solution Sdn. Bhd under its publication scheme.

REFERENCES

[1] Purusothaman, G., & Krishnakumari, P. (2015). A survey of data mining techniques on risk prediction: Heart disease. Indian Journal of Science and Technology, 8(12).

[2] Kalaiselvi, C., & Nasira, G. M. (2015). Prediction of heart diseases and cancer in diabetic patients using data mining techniques. Indian Journal of Science and Technology, 8(14).

[3] Detrano, R.; Steinbrunn, W.; Pfisterer, M., "International application of a new probability algorithm for the diagnosis of coronary artery disease". American Journal of Cardiology, Vol. 64, No. 3, 1987, pp. 304-310.

[4] Colombet, I.; Ruelland, A.; Chatellier, G.; Gueyffier, F.(2000). "Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression". Proceedings of AMIA Symp 2000, p 156-160.

[5] Bhatla, N., &Jyoti, K. (2012). An analysis of heart disease prediction using different data mining techniques. International Journal of Engineering, 1(8), 1-4.

[6] Parthiban, G., Rajesh, A., & Srivatsa, S. K. (2011). Diagnosis of heart disease for diabetic patients using naive bayes method. International Journal of Computer Applications, 24(3), 7-11.

[7] Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-48.

[8] Anbarasi, M., Anupriya, E., & Iyengar, N. C. S. N. (2010). Enhanced prediction of heart disease with feature subset selection using genetic algorithm. International Journal of Engineering Science and Technology, 2(10), 5370-5376.

[9] Detrano, R.; Steinbrunn, W.; Pfisterer, M., "International application of a new probability algorithm for the diagnosis of coronary artery disease". American Journal of Cardiology, Vol. 64, No. 3, 1987, pp. 304-310.

[10] Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE), 2(02), 250-255

[11] Das, R.; Abdul kadir, S. (2008). "Effective diagnosis of heart disease through neural networks ensembles". Elsevier, 2008.

[12] Avci, E.; Turkoglu, I., "An intelligent diagnosis system based on principle component analysis and ANFIS for the heart valve diseases". Journal of Expert Systems with Application, Vol. 2, No. 1, 2009, pp. 2873-2878.

[13] Palaniappan, S., & Awang, R. (2008, March). Intelligent heart disease prediction system using data mining techniques. In Computer Systems and Applications, 2008.AICCSA 2008. IEEE/ACS International Conference on (pp. 108-115). IEEE.

[14] Patil, S. B., & Kumaras wamy, Y. S. (2009). Intelligent and effective heart attack prediction system using data mining and artificial neural network. European Journal of Scientific Research, 31(4), 642-656.

[15] Lee, H. G., Noh, K. Y., & Ryu, K. H. (2007, May). Mining bio signal data: coronary artery disease diagnosis using linear and nonlinear features of HRV. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 218-228).Springer Berlin Heidelberg.