Available online **www.ijpras.com**

International Journal of Pharmaceutical Research & Allied Sciences, 2019, 8(2):104-110

Research Article

ISSN : 2277-3657
CODEN(USA) : IJPRPM

Implementation of Artificial Neural Networks and Decision Tree Algorithms for Heart Disease Diagnosis

Gul Mohamed Rasitha Banu[1]*, Thani Babikar[2], Illham Bashier[2], N. Sasikala[3]

[1] *Department of HI, FPHTM, Jazan University, KSA.*
[2] *Department of HE, FPHTM, Jazan University, KSA.*
[3] *Department of Computer Science, Md. Sathak College, India.*

***Email:** rashidabanu76 @ gmail.com*

ABSTRACT

An abnormal condition which troubles a living organism is called a disease. Nowadays, the most common problems the people are affected by are the heart problems. Several times, they lead to death in most cases due to the lack of correct diagnosis. The volume of data has been increasing rapidly in the area of health care. Predicting the heart problems is very difficult for the physicians. It is intractable to find the interesting patterns among enormous volumes of data. To find those, pattern recognition can be used, and to discover the hidden knowledge, data mining can be used. There have been a large number of medical data sets available in the market. Among all types of heart diseases, Cardio Vascular Disease is a type. So, many researchers carried out their works in heart disease dataset with 13 attributes, and 15 attributes with various data mining methods. In this study, ranking method was used in preprocessing a stage with total of 17 attributes for strengthening the rate of accuracy. The Zero R and J48 algorithms from NN and Multilayer Perceptron & decision tree were applied respectively on the dataset. The classifiers' performance was analyzed by error rate and time complexity with accuracy. In this research, Multilayer perceptron classifier showed high accuracy results with 13 attributes. Out of these three classifiers, J48 classifier gave high accuracy, minimum error rate and less time while using 17 attributes. Hence, these approaches can be very useful to the physicians to take decisions at the proper time. This research work was entirely carried out by WEKA (Waikato Environment Knowledge Analysis) data mining tool.

Key words: *Data Mining, Heart Disease, Artificial Neural Networks, Multilayer Perceptron (MLP), Decision Trees, WEKA, Accuracy.*

INTRODUCTION

"Data Mining is the task of extracting knowledge and discovering hidden patterns from huge volume of data" [1]. It is used in many fields such as medical domain, telecom, airlines, education, banking and education. There are many techniques namely classification, clustering and association rule mining which are available. According to the health care sector, the quality of service is very important. It means correct diagnosis gives better treatment at proper time to safeguard human life. Cardio Vascular Diseases (CVD), including coronary heart diseases, stroke, and peripheral vascular diseases constitute major public health problems worldwide [2]. As per WHO's global status report, 17.3 million people have died because of CVDs [2]. A report by Lim et al., (2012) pointed out that 16.5% of all deaths each year (9.4 million deaths) are due to the hypertension, 51% of death are due to stroke, and 45% of death are due to coronary heart diseases [3]. The number of death due to the CVDs will reach to 23.3 million by the year 2030. Most of the CVDs will be prevented by tackling associated risk factors like the lack of exercise, tobacco usage, uncontrollable blood sugar, chunk food, obesity, high cholesterol, and high BP [4]. In 2004, in Saudi Arabia, CVD was a major public health problem, with an overall prevalence of 5.5% and 11.7% amongst the people aged 45 years and older; respectively [5]. For betterment of

the human life, data mining technique plays an important role in diagnosing diseases. And by correct diagnosis, doctors can give proper treatment at a proper time.

Human body consists of many types of organs. One of the vital part in human body is heart. People will suffer from heart diseases if the organ does not function properly. If it stops its function, no one can stay alive. Generally, men suffer from heart diseases comparing with women. Due to obesity, physical inactivity, and smoking, lack of exercise, alcohol intake, excess of stress, and depression, the possibility of the heart diseases can be increased.

Different data mining techniques are used to predict the symptoms of heart diseases [6, 7]. S. B. Patil and Y. S. Kumaraswamy used K-means clustering algorithm and Maximal frequent Item set Algorithm to predict heart diseases [8]. M. Jabbar et.al proposed Cluster Based Association Rule mining for heart disease predictions [9]. A. K. Sen et.al proposed Neuro Fuzzy Integrated Approach to predict heart diseases and reduced memory requirements [10]. In the present work, to improve the classifier accuracy, only 18 relevant attributes were used, and artificial neural network algorithms such as Multilayer Perceptron and decision tree classifiers namely J48 and Zero R on the dataset were applied to predict heart diseases.

OBJECTIVES

The aims of the study were:

a. To implement data mining techniques to predict heart diseases.
b. To develop a predictive model using classifiers
c. To compare the performance of classification algorithms to identify which classifier predict the disease correctly with the high accuracy in terms of confusion matrix, and help the physicians to predict heart diseases and take decisions at a proper time.

DATA COLLECTION

The statlog heart disease database was collected from UCI machine repository [11]. It consists of 270 instances and 73 attributes. But, in this research work, to develop the system, 18 attributes were used. The attributes and their descriptions were also taken from statlog database [11].

Table1: Data set Description

S.No	Attribute	Description
1	age	Patient's Age
2	sex	Gender of Patient like Male or female
3	cp	Type of Chest pain
4	thestbps	Level of Resting blood pressure
5	chol	Blood Serum cholesterol
6	Restecg	Resting electrographic results
7	fbs	Fasting blood sugar
8	thalach	Maximum heart rate achieved
9	exang	Exercise induced agina
10	oldpeak	ST depression induced by exercise
11	solpe	Slope of the peak exercise ST segment
12	ca	Major vessels colored by floursopy
13	thal	Defect type
14	class	Represents whether heart disease is present or not

14 input attributes including class for prediction of heart disease which have been listed in table 1, were taken. To get more appropriate results, four important attributes i.e. Physical inactivity, alcohol, family history and smoking were added. The newly added attributes descriptions have been given below in table 2.

Table 2: New added Attributes' descriptions

S. No	Attribute	Description	values
15	Phy. Inactivity	Physical Inactivity	1=true 0=false
16	Alc	Alcohol intake	1=true

			0=false
17	Smoke	smoking	1=smoking 0= not smoking
18	FH	Family History	1=yes 0=no

METHODOLOGY

a. Preprocessing

Preprocessing is one of the data mining techniques. It is the task of transforming the format of data, removing irrelevant attributes, filling the missing values and so on. In this research, the preprocessing technique was applied to select relevant attributes using the ranking method.

b. Classification

Classification is one of the data mining techniques. In classification, predefined sample is used.

c. Decision tree

Decision tree is one of the classification techniques in data mining. It is a tree-like graph [12]. The internal node denotes a test on attribute, each branch represents an outcome of the test, and the leaf node represents classes. It is a graphical representation of possible solutions based on the condition of the solutions, the optimum course of action was carried out [12]. In this work, Zero R, MLP and J48 classifiers were used to classify the heart disease data set.

PROPOSED SYSTEM's ARCHITECTURE

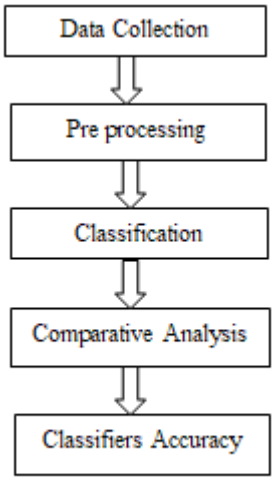

Figure 1. The proposed architecture model

The data was collected from the patients' heart disease data sets. The preprocessing technique was applied to the dataset to reduce the noise from data, transform the format of the data, and remove the irrelevant attributes using the ranking method. Classification algorithms such as J48, Zero R and Multilayer Perceptron were applied on the preprocessed data to classify the data set. The performance of the classifiers was analyzed by the confusion matrix.

EXPERIMENTS WITH WEKA

Waikato Environment for knowledge Analysis 3.7(WEKA) software is an open source software which can be downloaded from the website [13].

Performance Measure of Classifiers

In this experiment, the data was supplied to the classifiers of J48, MLP and Zero R Algorithm to classify the data. The classifiers' performance was analyzed by the confusion Matrix.

a. Confusion Matrix

In the confusion matrix, correctly identified instances were calculated by adding the value of the diagonal elements TP (True Positive) and TN (True Negative). FP (false positive) and FN (False Negative) were incorrectly called classified instances.

b. Accuracy

It was defined as the ratio of correctly classified instances to the total number of instances in the dataset.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

RESULT ANALYSIS

There were totally 270 records in the heart disease dataset. Among these 270 instances, 150 instances belonged to the present class, and 120 belonged to the absent class. The following tables represent the confusion matrix with 14 attributes.

The following Table 3 shows the confusion matrix of ZERO R Algorithm.

Table 3: Confusion matrix for Zero R Algorithm

Target class	Present(heart disease)	Absent(heart disease)
Present(heart disease)	150	0
Absent(heart disease)	120	0

In ZERO R classifier, the correctly classified instances were 150, and incorrectly identified instances were 120.

The following Table 4 shows the confusion matrix of MLP Algorithm.

Table 4: Confusion matrix for MLP Algorithm

Target class	Present(heart disease)	Absent(heart disease)
Present(heart disease)	117	33
Absent(heart disease)	26	94

In MLP classifier, the correctly identified instances were 211, and incorrectly identified instances were 59.

The following Table 5 shows the confusion matrix of J48Algorithm.

Table 5: Confusion matrix for J48 Algorithm

Target class	Present(heart disease)	Absent(heart disease)
Present(heart disease)	119	31
Absent(heart disease)	32	88

In J48 classifier, the correctly identified instances were 207, and incorrectly identified instances were 63.

The following Table 6 depicts the detailed accuracy, time, error rate for J48, MLP and Zero R algorithm for 14 attributes.

Table 6: Accuracy, error rate and time taken to build the model of Algorithms for 14 attributes (Before adding the other attributes (only 14 attributes))

Classifier	Accuracy	Time taken to build model (Seconds)	Error rate
Zero R	55.55%	0	0.493
Multilayer perceptron	**78.14%**	**0.38**	**0.224**
J48	76.66%	0.04	0.274

Table 6 shows that Multilayer perceptron gave the highest accuracy and the minimum error rate. But compared with the time, it was high in MLP classifier.

Table 7: Accuracy, error rate and time taken to build model of Algorithms for 18 attributes (After adding the four attributes (18 attributes))

Classifier	Accuracy	Time taken to build model(Seconds)	Error rate
Zero R	55.55%	0.2	0.493
Multilayer perceptron	99.52%	0.57	0.48
J48	**100%**	**0**	**0**

Table 7 shows that the accuracy of Zero R was (55.55%), the accuracy of MLP was (99.52%), and the accuracy of J48 was (100%). Zero R took 0.2 seconds to build the model, MLP took 0.57 seconds to build the model, and J48 took 0 seconds to build the model. The error rate of Zero R was 0.493, the error rate of MLP was 0.48 and the error rate of J48 was 0. While comparing J48 with other classifier, J48 gave the highest accuracy (100%), the minimum error rate (0%) and the minimum time (0 seconds) comparing with the other algorithms.

The following chart depicts the accuracy, error rate and time taken to build the model of various classifiers.

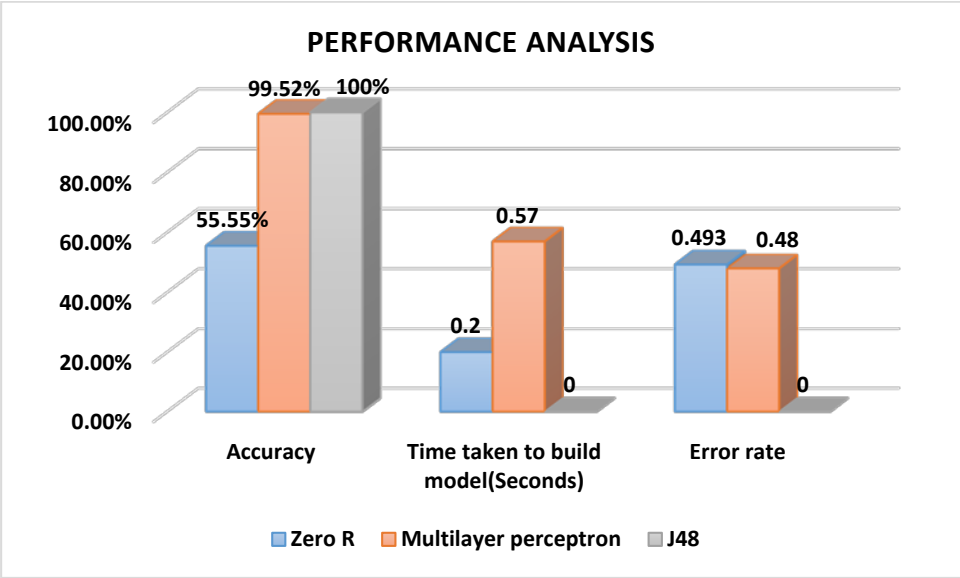

Chart 1: Performance Analysis of classifiers

In this chart, X axis represents the classifier and Y axis represents the accuracy, the error rate and the time. It shows that the accuracy of J48 classifier was 100%, the time to build the model was 0 seconds, and the error rate was 0% which was the best among the other algorithms.

DISCUSSION

There have been many data mining methods used to predict diseases. In 2005, Dr. Yashpal Singh proved that artificial neural network provides best results while compared with other methods [14, 15]. In 2012, Nidhi et al used neural network with 15 attributes to predict heart diseases. They achieved the accuracy of 96.5 % using neural network [16]. In 2017, Assari used KNN, Naïve Bayes, SVM and decision tree to diagnose heart diseases with 13 attributes, and the performance of the classifiers was also evaluated. It was proved that SVM gave the best accuracy (84.33%) among the other classifiers [17]. To improve the accuracy of classifiers, 18 attributes were selected from the database using the ranking method in the preprocessing stage. We have used Multi-Layer Perceptron algorithms, Zero R and J48 algorithms to diagnose heart disease. While J48 is compared with MLP algorithm and Zero R using weka tool, J48 is giving 100% accuracy, 0 seconds to build the model and 0% error rate.

CONCLUSION AND FUTURE SCOPE

Diagnosis of disease is a very challenging task in the field of health care. Many data mining techniques are used in decision making process. In this study, the dimensionality reduction was used to select the subset of attributes from the original data, and J48, Zero R and Multi-Layer perceptron data mining classification techniques were applied to classify the presence and absence of heart diseases. The performance of classifiers was evaluated

through the confusion matrix in terms of accuracy. The J48 algorithm gave 100% accuracy which provided better accuracy than the other algorithms, it also took less time and the minimum error rate. To analyze the patients' behavior and safety, other data mining techniques such as clustering and association rule mining were used. One of the important trends in data mining is the text mining to mine the vast amount of data. In health care organizations, mostly the available database is in the unstructured form. As a future work, the same technique can be used to apply in the other disease datasets such as eye disease, Diabetes, Lung cancer and so on.

ACKNOWLEDGEMENT

The authors would like to thank their family members, colleagues and friends for their great support to complete this research. They would also like to thank the University for providing an environment to carry out this research.

Conflict of Interest

There was no conflict of interests between the authors.

REFERENCES

1. Frawley and G. Piatetsky -Shapiro, "Knowledge Discovery in Databases: An Overview", published by the AAAI Press/ The MIT Press, Menlo Park, C.A 1996.
2. Geneva: 2011. World Health Organization, global status report on non-communicable diseases 2010. Available from: www.who.int/nmh/publications/ncd_report_full_e n.pdf.
3. Lim SS, Vos T, Flaxman AD, Danaei G, Shibuya K, Adair-Rohani H, "A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010", Lancet. 2012; 380:2224–2260.doi:10.1016/S0140-6736 (12)61766-8.
4. Mathers CD, Loncar D, "Projections of global mortality and burden of disease from 2002 to 2030", PLoS Med.2006; 3:e442; Available from: www.ncbi.nlm.nih.gov/17132052.
5. Al-Nozha MM, Arafah MR, Al-Mazrou YY, et al. "Coronary artery disease in Saudi Arabia", Saudi Med J.2004;25(9):1165–1171.
6. Rumelhart, David E., Geoffrey E. Hinton, and R. J. Williams. "Learning Internal Representations by Error Propagation". David E. Rumel hart, James L. McClelland, and the PDP research group. (Editors), Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations. MIT Press, 1986.
7. Dr.N.Sasikala, Dr.G.Rasitha Banu, Dr.Murtaza Ali," Utilization OF Health Informatics in Cardiovascular Disease", ISSN 0975-6299, Int J Pharm Bio Sci 2016 July; 7(3): 303 – 307.
8. S. B. Patil and Y. S. Kumaraswamy, "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction," International Journal of Computer Science and Network Security (IJCSNS), vol. 9, no. 2, pp. 228–235, 2009.
9. M. Jabbar, P. Chandra, and B. Deekshatulu, "Cluster Based Association Rule Mining For Predicting Heart Disease," Journal of Theoretical & Applied Information Technology, vol. 32, no. 2, pp. 196–201, 2011.
10. A. K. Sen, S. B. Patel, and D. P. Shukla, "A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level," International Journal of Engineering and Computer Science, vol. 2, no. 9, pp. 1663–1671.
11. "UCI Machine Learning Repository of machine learning database", University of California, school of Information and Computer Science, Irvine. C.A. Available from: http://www.ics.uci.edu.
12. Available at: http://en.wikipedia.org.
13. Available at: http://www.cs.waikato.ac.nz/ml/weka.
14. Dr. Yashpal Singh, Alok Singh Chauhan "Neural Networks in data mining" Journal of Theoretical and Applied Information Technology , 2005 - 2009 JATIT. 5(1). Pp.37-42.
15. Rosenblatt, Frank. x. Principles of Neuro dynamics: Perceptron and the Theory of Brain Mechanisms. Spartan Books, Washington DC, 1961.

16. Nidhi Bhatla and Kiran Jyoti, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, 2012 ; Vol. 1 Issue 8; 1-4.
17. Assari R, Azimi P, Taghva MR, "Heart Disease Diagnosis Using Data Mining Techniques". ISSN: 2162-6359. Int J Econ Manag Sci; 2017; volume 6 Issue 3 . doi: 10.4172/2162- 6359.1000415.