Journal of Engineering and Applied Sciences 14 (10): 3289-3294, 2019
ISSN: 1816-949X

Design of a New Hybrid Model using k-means Clustering and Two Class Neural Network for Heart Disease Prediction

[1]Animesh Hazra, [1]Subrata Kumar Mandal, [2]Asmita Mukherjee and [2]Arkomita Mukherjee
[1]Department of Computer Science and Engineering, Jalpaiguri Government Engineering College, Jalpaiguri, 735102 West Bengal, India
[2]Department of Information Technology, Jalpaiguri Government Engineering College, Jalpaiguri, 735102 West Bengal, India

Abstract: Heart diseases have become one of the most common and severe ailments to affect the well being of mankind in recent times. Approximately, 17.5 million deaths occur worldwide due to cardiovascular diseases. The healthcare industry generates millions and trillions of data which unfortunately is not efficiently reserved for mining and disease prediction purposes. The process of exploring relevant and intelligible information from extensive amounts of data is known as data mining. Clustering, Naive Bayes, decision trees, regression, artificial neural networks are some popular data mining techniques. It has been observed from the previous research that instead of applying a single mining algorithm, better results are obtained if a combination of mining algorithms can be applied. These combinational models, referred to as hybrid models, have set new trends in mining techniques. This study focuses on the predictive analysis of such a hybrid mining model that is a combination of k-means clustering and two class neural network on a heart disease dataset. The implementation of the project is done using Microsoft Azure Machine Learning Studio. Accuracy of the proposed model is 95.83%.

Key words: Heart disease, hybrid model, k-means clustering, two class neural network, analysis, mining

INTRODUCTION

Heart plays a pivotal role in the circulatory system. It transports blood, oxygen and other required materials to different parts of the body. Cardiovascular Diseases (CVD) are a class of diseases that involve the heart and blood vessels. These cardiovascular diseases also include Coronary Artery Diseases (CAD) such as heart attack and Coronary Heart Disease (CHD) like atherosclerosis which are usually fatal. It occurs when plaque, a waxy substance builds up in the coronary arteries and eventually narrows the artery and reduces blood flow to the heart. It may also happen that the plaque ruptures and forms a large blood clot which completely blocks blood flow. Failure to restore the stopped blood flow quickly will render the heart muscles impaired. Some of the characteristic symptoms of heart attack are chest pain, nausea, heartburn, stomachache, painful arms, fatigue, dizziness, sweating, etc. Heart failure, hypertensive heart disease, cardiomyopathy, heart arrhythmia, congenital heart disease, valvular heart disease, aortic aneurysms are some other prevalent CVDs. Cardiovascular diseases are also linked to sedentary lifestyle choices, smoking and certain eating habits. The severity of this disease, thus, calls for appropriate and timely diagnosis followed by medication and a healthy lifestyle. The proposed model can be used in medical facilities to diagnose heart disease with better accuracy. Moreover, deploying a software for analyzing such diseases will also reduce the drawbacks that occur due to erroneous diagnosis made manually by health specialists.

Literature review: Chen *et al.* (2011) predicted heart disease using artificial neural network model and Learning Vector Quantization (LVQ) classification algorithm in C# environment. Accuracy of the system was about 80%. Hannan *et al.* (2010) used a RBFNN (Radial Basis Function-Neural Network) Model in MATLAB to figure out instances of heart diseases. Ordonez (2006) studied association rule mining deploying the train and test concept on a dataset for heart disease prediction and evaluated the results through support, confidence and lift. Patil and Kumaraswamy (2009) mined crucial patterns from heart disease warehouse using MAFIA algorithm (Maximal Frequent Itemset Algorithm) which deployed clustering and significant weightage methods. Shetty and Naik (2016) developed a hybrid system of genetic algorithm and neural network on WEKA and MATLAB which showed impressive results. Aydin *et al.* (2016) used techniques such as bagging, AdaBoostM1, random forest, Naive Bayes, RBF network, IBK and NN in WEKA and concluded that RBF network has the highest accuracy of 88.2%. Chhabbi *et al.* (2016) applied Naive Bayes and modified k-means on a heart disease dataset

Corresponding Author: Animesh Hazra, Department of Computer Science and Engineering, Jalpaiguri Government Engineering College, Jalpaiguri, 735102 West Bengal, India

and concluded that modified k-means gave better accuracy. Pravabathi and Chitra (2015) studied research being performed using DNFS (Decision tree based Neural Fuzzy System) and concluded that decision trees and Naive Bayes classifiers are prominent for cardiovascular disease diagnosis with an accuracy reaching more than 95%. Chaurasia and Pal (2013) used CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and C4.5 in their research. They concluded that CART gave the best result. Sangwan and Tazeem (2015) developed a hybrid algorithm using k-means clustering and A-priori through a "bottom up" approach for better accuracy. Ghadge *et al*. (2015) have developed a prediction model that works on heart disease dataset. They have used three techniques namely neural network, Naive Bayes and decision tree. Rajkumar and Reena (2010) have studied various supervised classification techniques on a heart disease dataset and explored the statistical results. They have concluded that Naive Bayes has lower error rates. Dessai (2013) has implemented a model that uses Probabilistic Neural Network (PBN). PBN is a part of the class of Radial Basis Function (RBF) network. The model is 94.6% accurate.

Dataset description: The heart disease dataset originally contains 76 attributes but for data mining purpose only 14 are required. Here is a list of the 14 attributes used in this experiment along with their descriptions in Table 1.

Table 1: Heart disease dataset used in the experiment

Attribute	Attribute description
Age	Age in years
Sex	1 = Male, 0 = Female
CP	Chest pain type
	Value 1: Typical angina
	Value 2: A typical angina
	Value 3: Non-anginal pain
	Value 4: Asymptomatic
Trestbps	Resting blood pressure (in mmHg on admission to the hospital)
Chol	Serum cholestoral (mg/dL)
FBS	Fasting blood sugar >120 mg/dL
	1 = True, 0 = False
Restecg	Restecg: resting
	Electrocardiographic results
	Value 0: Normal
	Value 1: Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05 mV)
	Value 2: Showing probable or definite left ventricular hypertrophy by Estes criteria
Thalach	Maximum heart rate achieved
Exang	Exercise induced angina
	1 = Yes, 0 = No
Oldpeak	ST depression induced by exercise relative to rest
Slope	The slope of the peak exercise ST segment
	Value 1: Upsloping
	Value 2: Flat
	Value 3: Downsloping
CA	Number of major vessels (0-3) colored by flourosopy
Thal	3 = Normal, 6 = Fixed defect, 7 = Reversible defect
Num	Diagnosis of heart disease (angiographic disease status)
	Value 0: <50% diameter narrowing
	Values 1-4: >50% diameter narrowing

MATERIALS AND METHODS

The dataset used here is obtained from UCI machine learning repository. It contains 14 features that are crucial in the context of mining important information to predict heart diseases. For data pre-processing, the missing values are replaced with zero. The label (classification) of this dataset contains the values 0-4. Here, 0 indicates the absence of heart disease but the numbers 1-4 reflect the presence of heart disease with different intensities. The proposed prototype is a hybrid model that combines two machine learning algorithms, one is k-means clustering and the other one is two class neural network. With the help of k-means clustering, the data is clustered into two classes out of the five classes provided in the dataset. Cluster 0 indicates the absence of heart disease and cluster 1 signifies the presence of heart disease. Next, two class neural network algorithm is applied on the clustered dataset to achieve a better prediction accuracy of the heart disease. Figure 1 shows the workflow diagram of the proposed model.

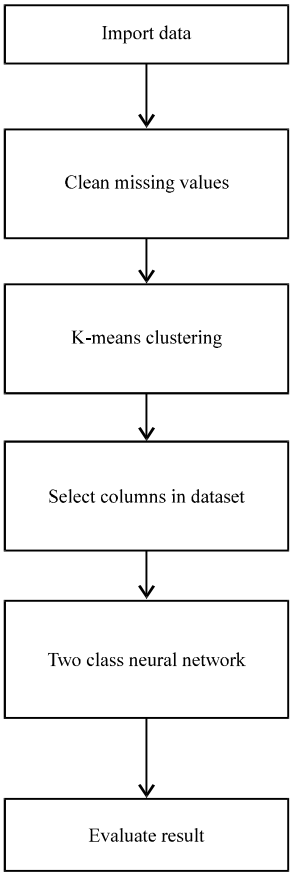

Fig. 1: Workflow diagram of the proposed model

Initially, the dataset is split into two sections, each containing 60 and 40% of the total data. The first part is used to train the clustering model. The second partition is used to assign data to the clusters, namely 0 and 1. Once the dataset containing only two class labels is obtained, it is again segregated into two sections. Now, the first part which comprises 60% of the data is used to train the second machine learning algorithm, i.e., two class neural network. Testing is done with the remaining 40% of data.

Proposed algorithm for the hybrid model: The following steps have been carried out for implementation of the algorithm.

Step 1: Initially, the dataset is taken from UCI machine learning repository. It contains 14 attributes. The class label 0 represents the absence of heart disease and class labels 1-4 represent the presence of heart disease with various intensities.

Step 2: The dataset has some missing values. Such inconsistent data leads to reduced performance of machine learning models. These missing values are substituted with zero to get more accurate result.

Step 3: Next, the dataset is divided into two sections for clustering. First 60% of the data is used for training the k-means clustering model and then the remaining 40% is used to assign the data into their respective clusters. Thus, a dataset containing binary values for class label is obtained. In this binary class label, 0 denotes the absence of heart disease and 1 indicates the presence of heart disease.

Step 4: The dataset thus obtained is segregated further into two parts of 60 and 40%, respectively. The implementation of the hybrid model is achieved by incorporating a two class neural network algorithm on this dataset. The first partition (60%) is used for training purpose and the second one (40%) for testing phase.

Step 5: The two class neural network classifier used here has 200 hidden nodes and undergoes 100 iterations with a learning rate of 0.1. The training is carried out using the new class label which is obtained after clustering.

Step 6: The results are evaluated by finding out the accuracy, precision, recall and F1 score. It gives a proper overview about how well the proposed model is performing.

Data mining techniques used in the proposed hybrid model

k-means clustering: Clustering is an unsupervised machine learning technique that can be defined as the process of combining objects into clusters, so that, objects inside a particular cluster exhibit similarities but is dissimilar to the objects in the other clusters. This partitioning is based on the assessment of the attribute values describing the data. In the context of data mining, cluster analysis gives insight into data distribution and characterization.

k-means is a centroid based clustering technique. Centroid of a cluster conceptually refers to its center. Let the centroid of cluster i be denoted by C_i. The dataset containing n objects is partitioned into k clusters C_1, ..., C_k that is $C_i \cap D$ (D being the dataset) and $C \subseteq C \neq \Phi$. An objective function is used to determine the partitioning which focuses towards providing a higher intracluster similarity and decreased intercluster similarity.

The difference between an object $p \in C_i$ and c_i(c_i represents the cluster) is obtained by dist(p, c_i) where dist (x, y) is defined as the Euclidean distance between two points x and y.

Within-cluster variation is used to measure the quality of the clusters. For every object in a cluster, the distance between the object and its cluster center is squared and all the distances are summed. It can be mathematically expressed as follows:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \text{dist}(p, c_i)^2 \quad (1)$$

where, E is the summation of the squared error for all objects in the given data. The working mechanism of k-means algorithm includes defining the count of clusters (k) at first. The centroid of a cluster is obtained as the mean value of points within the cluster. Initially, the mechanism randomly selects k of the objects in the dataset, each of which represents a cluster mean. As the algorithm proceeds to its further iterations, the objects are assigned into their respective clusters based on their Euclidian distance with the aim of enhanced inter-cluster variations. Figure 2 is an example of k-means clustering algorithm with three clusters.

Two class neural network: This is a binary classification model that uses neural network as its underlying algorithm. The network consists of a large number of nodes which are arranged in layers namely input, output and hidden layer. The network is trained by adjusting the weights that appear on the edges of the network. The functioning of the network starts from the input layer,

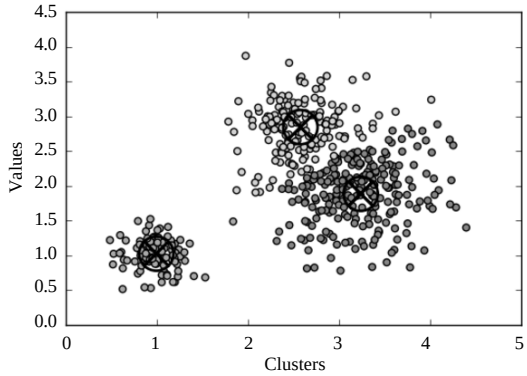

Fig. 2: An example of k-means clustering algorithm with three clusters

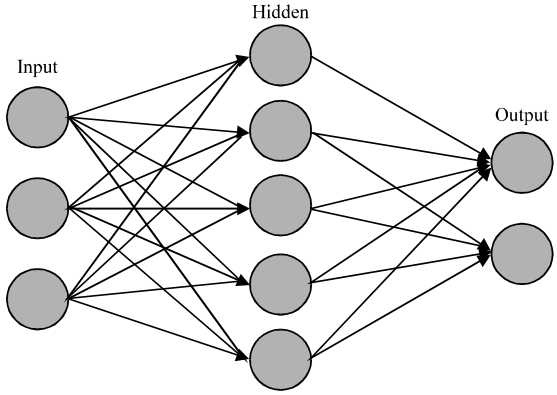

Fig. 3: Neural network with one hidden layer

proceeds through the hidden layers and finally to the output layer. During this training of the network, a specific value is calculated at each node. This value is obtained by using an activation function on the weighted summation of the previous nodes.

Neural networks are stimulated from the biological structure of the human axioms. Neural networks currently facilitate remarkable solutions to problems of machine learning. Some well known algorithms that are used to train the network are gradient descent and back propagation. The intrinsic instability of various layers in deep neural network makes it difficult for training. Sometimes the early layers learn faster compared to the later ones and sometimes the later layers show significant learning ability. As such, training the neural network in accordance with the problem provides the key solution. The vast number of hidden layers in deep neural networks gives them extreme computational powers and is a remarkable thing indeed. Figure 3 shows a neural network with one hidden layer.

The proposed model has 200 hidden nodes and undergoes 100 iterations with a learning rate of 0.1. Neural

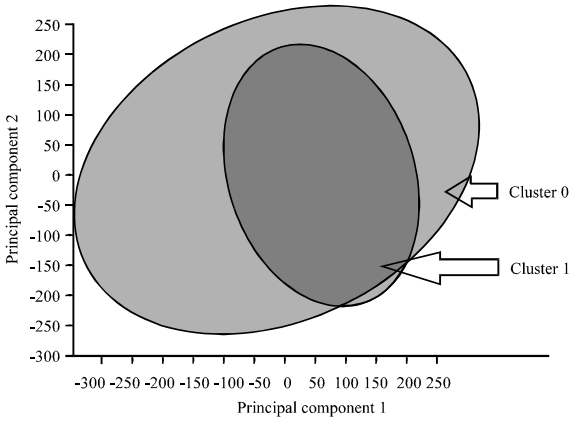

Fig. 4: The output of k-means clustering algorithm in Azure machine learning studio

network being a supervised machine learning algorithm, the training is done on the binary class label obtained after k-means clustering.

RESULTS AND DISCUSSION

This experiment builds a hybrid model using two machine learning algorithms namely k-means clustering and two class neural network to predict whether a patient has a heart disease or not. The implementation is done using Microsoft Azure Machine Learning Studio.

The k-means clustering algorithm divides the dataset into two clusters. It further prepares the dataset for training the two class neural network. The output of the clustering is shown in Fig. 4.

The accuracy of the proposed model is 95.83% and precision is 1.0. The ROC curve, precision/recall curve, lift curve and confusion matrix for this model are discussed. The ROC curve is a graphical plot that demonstrates the diagnostic ability of a binary classifier as its discrimination threshold is varied. The ROC curve of the proposed model is illustrated in Fig. 5.

Precision is the ratio of correct positive observations. Recall, also known as sensitivity or true positive rate, is the ratio of correctly predicted positive events. It is calculated by dividing true positives with the summation of true positives and false negatives. So, basically in a precision/recall curve, precision represents the fraction of retrieved instances that are relevant whereas recall signifies the fraction of admissible instances that are retrieved (Fig. 6). Lift curve is a variation of the ROC curve which measures the fraction of true positives in relation to the target response probability (Fig. 7). A confusion matrix, also known as an error matrix is a distinct table

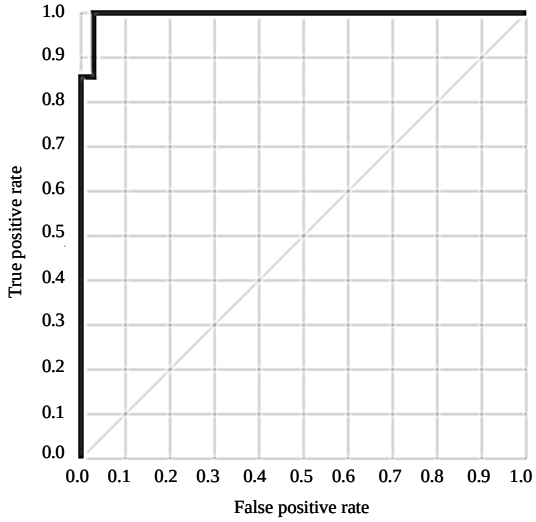

Fig. 5: ROC curve of the proposed model

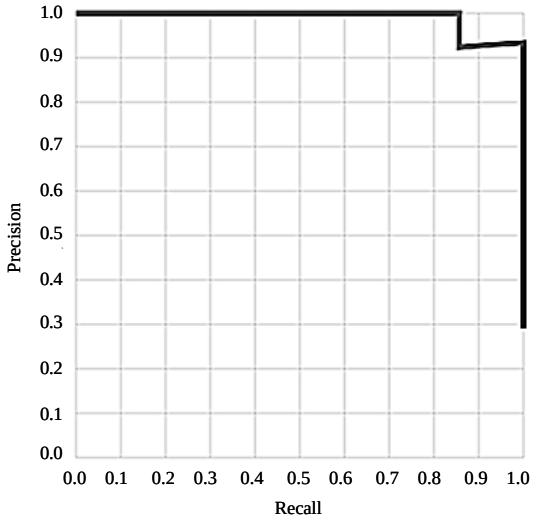

Fig. 6: Precision/recall curve of the proposed hybrid model

layout that aids in visualization of the performance of an algorithm. Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class (or vice versa). The dataset which is used in this final phase of testing contains 14 cases of heart diseases out of which 12 have been predicted correctly and 2 have been misclassified. Of the 34 cases where heart disease was absent, all have been predicted correctly. It is clearly shown in Table 2. The confusion matrix gives four measures namely true positive, true negative, false positive and false negative.

True positive: Those instances where the value of actual class is yes and the value of predicted class is also yes.

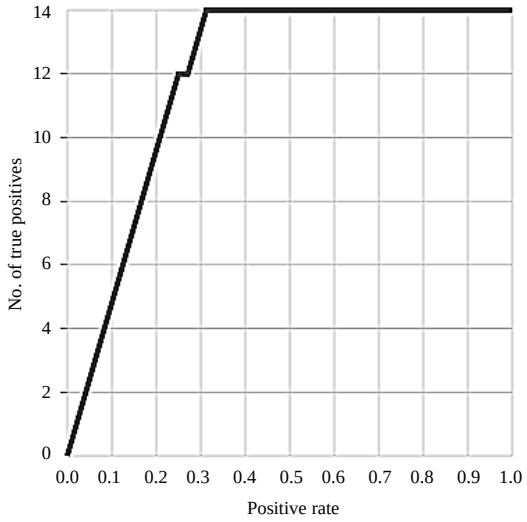

Fig. 7: Lift curve of the proposed hybrid model

Table 2: Confusion matrix of the proposed model

	Predicted	
Actual	Yes	No
Yes	12 (True positive)	2 (False negative)
No	0 (False positive)	34 (True negative)

True negative: Those instances where the class label and the predicted label both are false.

False positive: Those instances where class label is false but predicted label is true.

False negative: Those instances where class label is true but predicted label is false.

The confusion matrix can be used to calculate the four measures namely accuracy, precision, recall and F1 score.

Accuracy: It is the degree of correctness of a system. It is given by the following formula:

$$\text{Accuracy} = (\text{True positive}+\text{True negative})/(\text{True negative}+\text{True positive}+\text{False negative}+\text{False positive}) = (12+34)/(12+2+34+0) \equiv 95.83\%$$

Precision: It refers to the closeness of the measured values. Mathematically it can be defined as follows:

$$\text{Precision} = (\text{True positive})/(\text{True positive}+\text{False positive}) = 12/(12+0) = 1.0$$

Recall: Recall is the ratio of correctly predicted positive observations to all the observations in which the actual class was true:

$$\text{Recall} = (\text{True positive})/(\text{True positive}+\text{False negative}) = 12/(12+2) = 0.857$$

F1 score: It is the weighted average of precision and recall:

$$\text{F1 score} = 2*(\text{Recall}*\text{Precision})/(\text{Recall}+\text{Precision}) = 2*(0.857*1)/(0.857+1) = 0.922$$

CONCLUSION

Heart diseases being complicated and often fatal, call for accurate and timely diagnosis. The number of heart disease patients is rising rapidly. As such, people need to be aware of the causes and effects of heart diseases. Regular exercise, ditching unhealthy eating habits at the earliest, succulent helpings of fruits and vegetables daily are some of the key contributors to a healthy heart. Detecting the disease at the earliest is a crucial factor in ensuring a better and long survivability of the patient. Recent research suggests that in place of applying a single classification algorithm, better accuracy can be achieved by applying a combination of several machine learning algorithms. The proposed approach implements a hybrid model that runs on two data mining techniques. This research analyses the accuracy of a heart disease prediction model using a combination of k-means clustering and two class neural network. K-means clustering algorithm groups the data into two clusters. Thus, data having a binary class label (either 0 or 1) is obtained. Next, the dataset is trained by a two class neural network. This training is carried out using the new class label which is obtained after clustering. K-means followed by two class neural network gives an elevated accuracy of 95.83%. Also, instead of splitting the clustered data in 60-40 ratio, if it is partitioned in a ratio of 80-20, the accuracy becomes 100%. But this second 80-20 partitioning scheme works with considerably less number of tuples for testing. The hybrid model analyzed in this paper works on 121 tuples for the final training and testing phase. Therefore, running this hybrid model on bulk volumes of data will help in achieving more balanced results.

RECOMMENDATION

In future, this hybrid approach can also be implemented into a software that can be used in clinics to predict the chances of development of heart disease in a patient thereby elevating awareness well in advance.

REFERENCES

Aydin, S., M. Ahanpanjeh and M. Sogol, 2016. Comparison and evaluation of data mining techniques in the diagnosis of heart disease. Intl. J. Comput. Sci. Appl., 6: 1-15.

Chauraisa, V. and S. Pal, 2013. Early prediction of heart diseases using data mining techniques. Carib. J. Sci.Tech., 1: 208-217.

Chen, A.H., S.Y. Huang, P.S. Hong, C.H. Cheng and E.J. Lin, 2011. HDPS: Heart disease prediction system. Proceedings of the Conferences on Computing in Cardiology, September 18-21, 2011, IEEE, Hangzhou, China, ISBN: 978-1-4577-0612-7, pp: 557-560.

Chhabbi, A., A. Lakhan, A. Sahil and Y.K. Sharma, 2016. Heart disease prediction using data mining techniques. Intl. J. Res. Advent Technol., 2016: 104-106.

Dessai, I.S.F., 2013. Intelligent heart disease prediction system using probabilistic neural network. Intl. J. Adv. Comput. Theor. Eng., 2: 38-44.

Ghadge, P., V. Girme, K. Kokane and P. Deshmukh, 2015. Intelligent heart attack prediction system using big data. Intl. J. Recent Res. Math. Comput. Sci. Inf. Technol., 2: 73-77.

Hannan, S.A., A.V. Mane, R.R. Manza and R.J. Ramteke, 2010. Prediction of heart disease medical prescription using radial basis function. Proceedings of the 2010 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), December 28-29, 2010, IEEE, Coimbatore, India, ISBN: 978-1-4244-5965-0, pp: 1-6.

Ordonez, C., 2006. Association rule discovery with the train and test approach for heart disease prediction. IEEE. Trans. Inf. Technol. Biomed., 10: 334-343.

Patil, S.B. and Y.S. Kumaraswamy, 2009. Extraction of significant patterns from heart disease warehouses for heart attack prediction. Intl. J. Comput. Sci. Network Secur., 9: 228-235.

Prabhavathi, S. and D.M. Chitra, 2015. Analysis and prediction of various heart diseases using DNFS techniques. Intl. J. Innovations Sci. Eng. Res., 2: 1-7.

Rajkumar, A. and G.S. Reena, 2010. Diagnosis of heart disease using datamining algorithm. Global J. Comput. Sci. Technol., 10: 38-43.

Sangwan, S. and A.K. Tazeem, 2015. Review paper automatic console for disease prediction using integrated module of A-priori and K-mean through ECG signal. Intl. J. Technol. Res. Eng., 2: 1368-1372.

Shetty, A. and C. Naik, 2016. Different data mining approaches for predicting heart disease. Intl. J. Innovative Res. Sci. Eng. Technol., 5: 277-281.