Asian Journal of Information Technology 15 (8): 1314-1321, 2016
ISSN: 1682-3915

A Novel Feature Selection Method for Predicting Heart Diseases with Data Mining Techniques

[1]R. Suganya, [2]S. Rajaram, [1]A. Sheik Abdullah and [3]V. Rajendran
[1]Department of Information Technology, Thiagarajar College of Engineering, 625015 Madurai, India
[2]Department of ECE, Thiagarajar College of Engineering, 625015 Madurai, India
[3]Kovai Heart Foundation, Coimbatore, India

Abstract: This study introduces a new approach for prediction problem with the objective of attaining maximum classification accuracy with smallest number of features selected. Cardiac diseases are very common and one of the main reasons of death. According to a recent literature study by the Indian Council of Medical Research, about 25% of cardiac problems are between age group of 25-69 years. Hence, the main objective of this work is to predict the possibility of heart diseases at its early stages with less number of attributes. Our approach integrates anthropometric data and physiological data of cardiac diseases by proposing Novel Feature Selection method for prediction of heart diseases. The dataset used in this work is collected from Cleveland heart disease database. The results show the proposed approach leads to a superior feature selection process in terms of sinking the number of variable required and increased in classification accuracy for better prediction.

Key words: Data mining techniques, feature selection, prediction algorithm, neural network, heart diseases

INTRODUCTION

In the modern world, cardiovascular diseases are the highest flying diseases and in every year >12 million deaths occur worldwide due to heart problems. Heart diseases also cause maximum fatalities in India and its diagnosis is very difficult process. Due to limitation in potential of the medical specialists and their unavailability at certain places, puts the patients at a high threat. Normally, these diseases can be analyzed using perception of the medical specialist and it would be highly advantageous if the techniques used for diagnosis will be integrated with the medical information systems. A decision support or computer based information system which will assist physician for accurate diagnosis at reduced cost. The combination of WSNs data and data mining methods are used to identify heart diseases at the earliest stage itself.

There are several factors which increase the risk to occur heart problems are high blood pressure, cholesterol, genetic, lack of physical exercise, hypertension and smoking, etc. Data mining is the process of examining and summarizing the data from different view and converting these data into valuable information, it plays an important role in the intelligent medical system. The attributes/physiological data associated with the heart diseases and the patients' anthropometric data are quickly evaluated by the physicians via well developed software systems with different data mining approaches.

Medical data mining is a new area for exploring the hidden data patterns from h uge raw data sets. In medical organization like hospitals and medical centers, generates large amount of data which contains wealth of concealed information but these data is not used properly. Hence, that unused data can be converted into usefully information by using different data mining techniques.

Feature selection is more important topic in medical data mining. Feature selection is a procedure used in mining for extracting a subset of relevant features in order to create a less complex model of learning. The motivation behind this technique is the fact that when the initial pool consists of a huge number of features, some of them are redundant (not useful information). When we have a large number of features and a small number of examples, we need to discard some of the features in order to avoid over fitting. There are several approaches to feature selection: from ensemble learning methods to L1, L2 regularization, greedy selection (Couvreur and Bresler, 2000), genetic algorithms (Guo *et al*., 2011) and ant colony optimization (Kabir *et al*., 2012). Some advantages of feature selection are better understanding of the data, curse of dimensionality reduction and improvement of predictor performance.

Corresponding Author: R. Suganya, Department of Information Technology, Thiagarajar College of Engineering, 625015 Madurai, India

Wireless sensors are deployed for patients with heart diseases for continuous monitoring. For example, in the work proposed by Asada et al, a ring sensor worn on the base of the finger is implemented for the purpose of continuous monitoring for people with heart arrest problems (Asada *et al*., 2003). Patterson et al proposed an energy efficient ear worn PPG sensor for the function of heart monitoring. The comfortable placement of the sensor makes it suitable for long term monitoring (Patterson *et al*., 2009). Two major rewards of the aforesaid examples is the small size of the sensors and the low power consumption which are very important issues to be considered in any monitoring problem.

The main purpose of this study is to build up a system which predicts the possibility of risk of having heart diseases or heart arrest from sensor data in the earlier stages using Novel Feature Selection method and data mining techniques. For this work, we have collected data records of patients from Cleveland heart disease database and extract the essential attributes required for prediction. Heart diseases are among the most common reasons of fatality all over the country. Major types of heart diseases are Heart arrest & coronary artery diseases. Twenty five percent of people, any age, who have coronary artery disease, die suddenly without any previous symptoms. Coronary artery disease is one of the most important types of diseases affecting the heart, and can cause severe heart pains in patients. Being aware of disease symptoms, can aid in instant treatment, and reduce the rigorousness of disease's side effects.

Recently, angiography is used to find the quantity and place of heart vessels stenosis. Being costly and having several side effects, it has provoked many researchers to use data mining for diagnosing heart diseases. Some of the work done in heart disease monitoring using wireless sensor data by data mining techniques are discussed below: Prediction of heart diseases can be further boosted and extended by incorporating data mining techniques along with WSNs. WSNs continuously monitors and identifies the cardiovascular diseases experienced in patient at remote areas. Polat and Gunes (2007) used fuzzy weighted pre-processing and AIRS and reached the accuracy of 92.59% for diagnosing heart diseases.

Rajkumar and Reena (2010) used decision tree and Naïve Bayes algorithms on the ECG dataset and reached 53% accuracy. Lavesson *et al*. (2009) applied Adaboost, Bagging and Navie Bayes algorithms on ECG heart attack dataset and achieved 71% accuracy using navie bayes. Shouman and Turner used C4.5 decision tree for heart diagnosis and used reduce error pruning, resulting in 84.1% accuracy. Itchharporia *et al*.(1996) applied Neural Network classification methods on 24 features and achieved 86% accuracy. Sequential Minimal Optimization (SMO), Naive Bayes (Caruana and Niculescu-Mizil, 2006), Bagging with SMO and Neural networks classification algorithms are used to analyze the dataset. The results of the typical angiographic method are used as the base of comparison, to assess the prediction capability of classification algorithms.

Kappiarukudid and Ramesh (2010) present a real-time WSN system for prediction and monitoring of any upcoming heart diseases. The system has a capability of monitoring multiple patients at a time and delivers remote diagnosis and prescription to the patients. It also provides fast and effective alarms to doctors, relative and hospitals. Mai Shouman et al proposed various single and hybrid data mining techniques for heart diseases diagnosis. Using single data mining methods in heart disease diagnosis has been thoroughly investigated showing the considerable levels of accuracy. Recent investigation shows that for hybridizing more than one technique, will obtain enhanced result in diagnosis. Here author applied different data mining techniques like multilayer perceptron; naïve Bayes decision tree, neural network and kernel density on different heart disease datasets and measure the accuracy of each technique. Then applying hybrid data mining techniques on different heart diseases WSNs datasets shows the different accuracies.

Author presents a different classification approach for heart disease prediction. For result analysis, two essential functions namely training and testing will be performed. Accurate prediction will depend on algorithm and the selected minimal feature is applied on wireless sensor network dataset.

MATERIALS AND METHODS

Data mining techniques used for prediction of WSN dataset: In this section, technical aspects of data mining techniques are used to analyze WSN dataset are described (Jambhulkar and Aporikar, 2015). Wireless Sensor Networks are network that consist of sensor motes which are scattered in an ad hoc manner. These motes work with each other to sense some physical fact and then the information gathered is processed to get appropriate results. Wireless sensor network consists of protocols and algorithms with self organizing capabilities. Motes are physical unit in Wireless sensors. To explore novel data mining techniques, dealing with extracting knowledge

Table 1: Data mining techniques for WSN data

Data mining tech	Algorithms	Handling WSNs
Frequent and Sequential pattern mining	Apriori and Growth-based algorithms	To find association among large WSNs data
Cluster based technique	K-means, hierarchical and data correlation based clustering	Based upon the distance among the data point
Classification based technique	Decision tree, Random forest, NN, SVM and Logistic regression	Based on the application-Medical applications

from large communities arriving data from WSNs. Some of the data mining techniques for WSNs dataset are described in the Table 1.

Decision tree: Decision trees are recursive partitioning algorithms used to minimize the impurity present in the sensor dataset. The apex node is the root node specifying a testing provision of which the outcome corresponds to a branch leading up to an internal node. The fatal nodes of the tree allot the classifications and are also referred to as the leaf nods. Popular decision trees are C4.5 CART and Chi-Squared Analysis. This algorithm consists of splitting decision, stopping decision and assignment decision. Tree will start to fit the specificities or noise in the data, which is referred to as overfitting. In order to avoid this, the sensor data will be split into a training sample and validation sample. The training example will be used to construct the splitting assessment. The validation sample is an independent sample used to supervise the misclassification error. One way to determine impurity in a node is by calculating the mean squared error (MSE).

Decision tree is a classifier in the form of tree and classifies the occurrence by starting at the root of tree and moving through it until a leaf node where class label is assigned. The internal nodes are used to partition date into subset by applying test condition to separate instances that have different characteristics.

Neural Network: A Neural Network or Simulated Neural Network (SNN) is an interconnected group of artificial neurons that use a statistical or computational model for information processing based on a connection approach to computation. In several cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network. NN are non-linear statistical data modeling or decision making tools. They can be used to model multifarious relationships between inputs and outputs or to find similar patterns in data (Lu *et al*., 1996). The procedure adopted for NN is as follows:

- Train a neural network for WSNs data as much as possible in terms of connection, Split the data into training, validation and test set
- Assign the number of hidden neurons as 5 and categorize the hidden unit activation values using clustering
- Extract rules that explain the network output and inputs in terms of the categorized hidden unit activation rules and train a neural network on the training set with anthropometric data and measure the performance on the validation set
- Choose the number of hidden neurons with optimal validation set performance
- Measure the performance on the independent test set

Sequential Minimal Optimization: Sequential Minimal Optimization (SMO) is an algorithm for proficiently solving the optimization problem which arises during the training of SVM samples. It was introduced by John platt in 1998 at Microsoft research. SMO is widely used for training SVM. SVM techniques separate the data belonging to different classes by fitting a hyperplane between them which maximized the partition. The data is mapped into a higher dimensional feature space where it can easily partitioned by a hyperplane. SVM- Linear kernel is adopted for predicting heart diseases using WSNs dataset.

Random Forest: Random forest was first introduced by Breiman. It creates a forest of decision trees as follows:

- Given a data set with 13 observations and N inputs from 303 data instances
- Assume m = constant
- For t = 1….T
- Take a bootstrap sample with 13 observations
- Build a decision tree w hereby for each node of the tree, randomly choose m inputs on which to base the splitting decision based on the sensor information from motes
- Split on the best of this subset
- Fully grow each tree without pruning.
- Novel Feature Selection Method (NFS)
- Feature selection is more important topic in medical data mining. Feature selection is a procedure used in mining for extracting a subset of relevant features in order to create a less complex model of learning. The block diagram for proposed methodology using data mining techniques and Novel feature selection method is shown in Fig. 1.

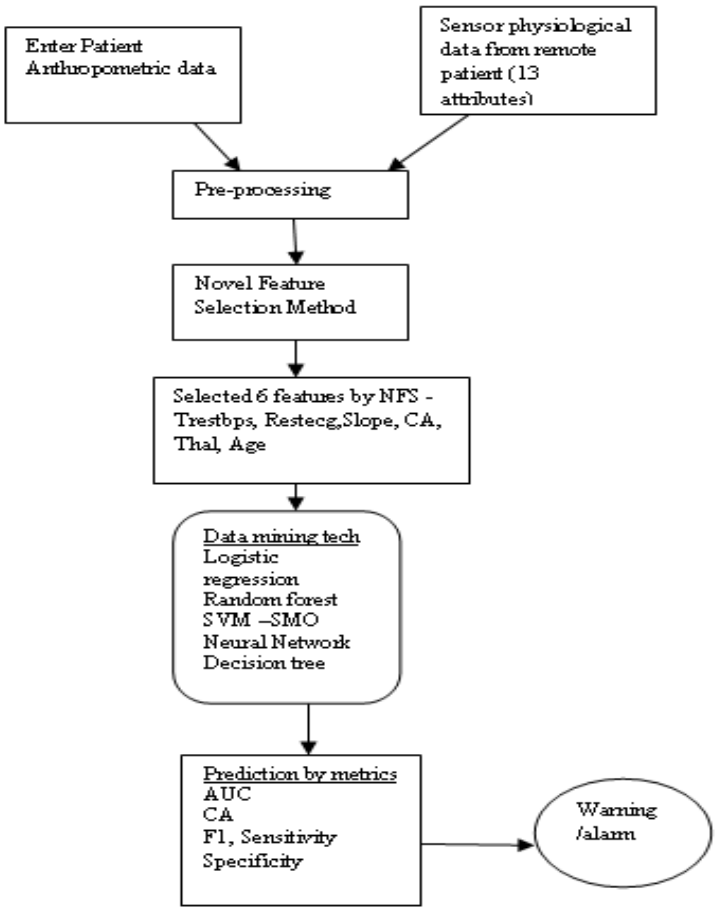

Fig. 1: Block diagram for novel feature selection method

Table 2: Selected attributes from cleveland database

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	1 = male; 0 = female
Cp	Discrete	Chest pain type; 1= typical angina;2 = atypical angina; 3 = non-anginal pa; 4 = asymptomatic
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol n mg dL^{-1}
Fbs	Discrete	Fasting blood sugar > 120 mg/d;1 = true; 0 = false
Restecg	Discrete	Resting electrocardiographic results;0 = normal; 1 = having ST-T wave abnormality; 2 = showing probable or define left ventricular hypertrophy by Estes criteria
Thalach	Continuous	Maximum heart rate achieveds
Exang	Discrete	Exercise induced angina; 1 = yes; 2 = no
Slope	Continuous	Temperature/humidity
Old peak	Continuous	ST depression induced by exercise relative to rest
CA	Discrete	Number of major vessels (0-3) colored by flourosopy
Thal	Discrete	3 = normal; 6 = fixed defect; 7 = reversible defect
Diagnosis	Discrete	Diagnosis classes; 0 = healthy; 1 = possible heart disease

Cleveland database: Medical databases have collected large quantities of anthropometric information about patient and their medical conditions. Record set with minimum significant medical attributes was obtained from the cleveland heart disease database. With the help of the dataset the patterns important to the heart prediction are extracted. The records were split the database into two equal parts namely training datasets and testing datasets. The records for each set were selected randomly. Table 2 shows that selected features from Cleveland clinic foundation for the prediction of heart diseases.

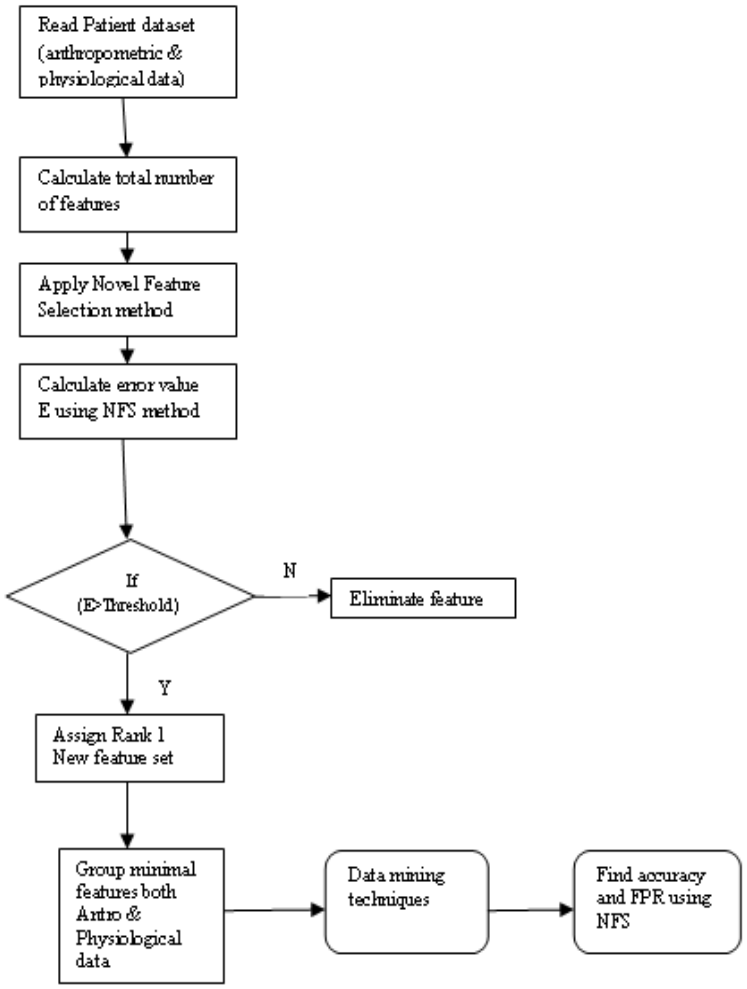

Fig. 2: Flow diagram for noval feature selection method

Cleveland Heart disease database consists of 303 instances with 13 features (anthropometric-age, gender, medical features - chest pain, rest SBP, cholestrol, blood sugar, rest ECG, max HR, exerc ind ang, ST by exercise, slope peak exc ST, major vessels colored, thal and diameter narrowing).

Novel feature selection method: Feature selection is the process of selecting a subset of relevant features to be used in constructing a classifier. It improves the prediction performance and provides a faster classifier. Because of these advantages, we execute a fourth trial to compare the classification accuracy using all the 13 features with that of the six best features selected by the Novel Feature Selection algorithm. The algorithm for NFS is shown below. The flow diagram for proposed Novel Feature Selection method is shown in Fig. 2.

- Step 1: Collect anthropometric data at time t of patient in remote using WSNs sensor
- Step 2: Collect Physiological data at t from the same patient
- Step 3: Predict sensor value at time t
- Step 4: Calculate the total number of features extracted from both anthropometric data and physiological data.
- Step 5: Feed the resultant dataset instances into Novel Feature Selection method.
- Step 6: NFS extract 5 physiological data and one anthropometric data that are more relevant to predict heart diseases in its early stages from 13 instances
- Step 7: Find error calculation
- Step 8: If error is greater than Threshold, then assign rank 1 and consider into new feature set

Table 3: Selected attributes from cleveland database using NFS method

Name	Type	Description
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Restecg	Discrete	Resting electrocardiographic results; 0 = normal; 1 = having ST-T wave abnormality; 2=showing feasible or define left ventricular hypertrophy by Estes criteria
Slope	Continuous	Temperature/humidity
CA	Discrete	Number of major vessels (0-3) colored by flourosopy
Thal	Discrete	3 = normal 6 = fixed defect; 7 = reversible defect
Age	Continuous	Ages in years

Table 4: Performance results before and after applying NFS

	Accuracy %		FPR	
Classification algorithm	Without FS	With FS	Without FS	With FS
Logistic Regression	73.1	79	1.9	1.5
Random Forest	78	80	1.5	1.3
SMO	77	89	1.6	1.3
Decision Tree	46.8	46.2	2.1	1.7
Neural Network	86.1	93.1	1.2	0.5

- Step 9: if error is not greater than threshold value, then eliminate this feature, (not much relevant to predict heart diseases in its early stage)
- Step 10: Calculate minimal set of features from whole dataset by using NFS method.
- Step 11: Feed this minimal features into classification pool which contains several data mining techniques.
- Step 12: Calculate the accuracy and FPR for both Without applying Feature selection and with feature selection.

After extracting the above said attributes from the processed Cleveland dataset, we can provide some of these attributes value in real-time way to prediction system using different sensors. Although the above attribute values are essential for prediction, but for getting precision in prediction we need the real-time values by using feature selection methods. Table 3 shows that the WSNs data: Attributes selected by Novel Feature Selection (NFS) method.

The above parameters value can be changes suddenly and immediately in a heart patient. Hence, by providing the above parameters real-time, we can achieve the level of accuracy in prediction in Heart disease detection system, first of all we need to use 3 types of sensors viz. temperature/ humidity sensor, blood pressure control sensor, heart rate sensor which are able to sense the real-time values for the above parameters from our body and send the data to system via Wireless motes, this will provide the mobility and flexibility. While the system receives the data, we have applied the classification technique on the combination of received real-time data and recorded remaining extracted attributes data with the processed Cleveland datasets, thus identify the class code according to the code labels system

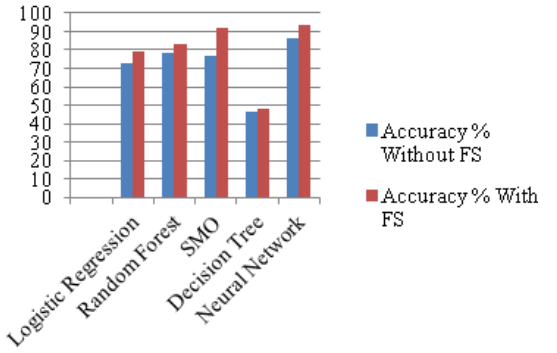

Fig. 3: Accuracy of data mining techniques before and after applying NFS method

provides the results, i.e., probability of risk of having heart disease or not. We choose 8-folds cross validation for results validation. Table 4 shows a comparison of classification performance results-Accuracy and FPR results before and after applying NFS method.

It is clear that from the above table, the feature selection process for WSNs dataset has improved the classification accuracy results across the different classifiers with corresponding FPR also recorded in Table 4. In Fig. 3, accuracy of data mining techniques before and after NFS method is clearly shown. From this, it is concluded that NFS method selects more relevant features that improves the good classification accuracy by SMO and Neural network methods.

RESULTS AND DISCUSSION

Confusion matrix: A confusion matrix is a table that allows visualization of the performance of an algorithm. In a two class problem (with classes C1 and C2), the matrix has two rows and two columns that specifies the number of False Positives (FP), False Negatives (FN), True Positives (TP) and True Negatives (TN). These measures are defined as follows: TP is the number of samples of class C1 which has been correctly classified. TN is the number of samples of class C2 which has been correctly classified. FN is the number of samples of class C1 which has been falsely classified as C2. FP is the number of

Table 5: Comparison of performance results with different data mining techniques

Data miningTech	AUC	CA	F1	Sensitivity	Specificity
Logistic Regression	0.844	0.792	0.821	0.853	0.791
Random Forest	0.802	0.805	0.784	0.799	0.770
SMO	0.887	0.890	0.869	0.849	0.791
Decision Tree	0.454	0.467	0.543	0.511	0.435
Neural Network	0.812	0.931	0.819	0.889	0.799

samples of class C2 which has been falsely classified as C1. Based upon the confusion matrix, the following performance metrics can be measured.

Sensitivity and Specificity: Sensitivity and specificity are explained as following:

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

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

Classification Accuracy: Accuracy shows ratio of correctly classified samples to the total number of tested samples. It is defined as:

$$\text{Classification accuracy} = (TP+TN)/(TP+FP+FN+TN)$$

$$\text{Classification error} = (FP+FN)/(TP+FP+FN+TN)$$

ROC: A Receiver Operating Characteristics (ROC) is a graphical plot which illustrates the performance of a binary classifier system. It is created by True Positive Rate (TPR) vs False Positive Rate (FPR). The larger the area under ROC curve, the higher the performance of the algorithm is:

$$FPR = FP/ (FP+TN)$$

$$TPR = TP / (TP+FN)$$

All the above classification measures depend on the cut-off of 0 (1), classification accuracy becomes 40 percent (60%), the error 60%(40%), the sensitivity 100%(0) and the specificity 0 (100%).

We have applied several data mining techniques for WSNs heart dataset which has been extracted from Novel Feature Selection (NFS) method and analyzed AUC, Classification Accuracy (CA), F1, Sensitivity and specificity. The comparative performance results for different data mining methods are shown in the Table 5.

CONCLUSION

In this study, we have shown that the Novel Feature Selection (NFS) method extracts more relevant features that support for attaining maximum classification accuracy by neural network and SMO-SVM with minimum number of features selected. NFS method integrates both anthropometric data and physiological data for prediction of heart diseases. We have extracted five major physiological data (Trestbps, Restecg, Slope, CA, thal) with the support of heart physician and one anthropometric data (Age) for classification from Wireless sensor network dataset Cleveland dataset. This selection of minimum features is used to predict the possibility of heart diseases at its early stages and generate a alarm for physicians and patient as well. We ran experiments in all data mining algorithms and proved that the neural network predicts 93% of accuracy and SMO predicts 89% of accuracy along with NFS. The results show the proposed approach leads to an superior feature selection process in terms of sinking the number of variable required and an increased in classification accuracy for better prediction. This decision support system can use for providing better healthcare services to heart patient. Thus, the early diagnosis of heart disease detection may decrease the chances of death in cardiac.

REFERENCES

Asada, H.H., P. Shaltis, A. Reisner, S. Rhee and R.C. Hutchinson, 2003. Mobile monitoring with wearable photoplethysmographic biosensors. IEEE Eng. Med. Biol. Mag., 22: 28-40.

Caruana, R. and A. Niculescu-Mizil, 2006. An empirical comparison of supervised learning algorithms. Proceedings of the 23rd International Conference on Machine Learning, June 25-29, 2006, Pittsburgh, PA., USA., pp: 161-168.

Couvreur, C. and Y. Bresler, 2000. On the optimality of the backward greedy algorithm for the subset selection problem. SIAM. J. Matrix Anal. Appl., 21: 797-808.

Guo, J., J. White, G. Wang, J. Li and Y. Wang, 2011. A genetic algorithm for optimized feature selection with resource constraints in software product lines. J. Syst. Software, 84: 2208-2221.

Itchhaporia, D., P.B. Snow, R.J. Almassy and W.J. Oetgen, 1996. Artificial neural networks: Current status in cardiovascular medicine. J. Am. Coll. Cardiol., 28: 515-521.

Kabir, M.M., M. Shahjahan and K. Murase, 2012. A new hybrid ant colony optimization algorithm for feature selection. Expert Syst. Appl., 39: 3747-3763.

Kappiarukudil, K.J. and M.V. Ramesh, 2010. Real-time monitoring and detection of" heart attack" using wireless sensor networks. Proceedings of the 2010 Fourth International Conference on Sensor Technologies and Applications (SENSORCOMM), July 18-25, 2010, IEEE, Venice, Italy, ISBN: 978-1-4244-7538-4, pp: 632-636.

Lavesson, N., A. Halling, M. Freitag, J. Odeberg and H. Odeberg *et al*., 2009. Classifying the severity of an acute coronary syndrome by mining patient data. Proceedings of the 25th Annual Workshop of the Swedish Artificial Intelligence Society, May 27-28, 2009, Linkoping University Electronic Press, Linkoping, Sweden, pp: 55-63.

Lu, H., R. Setiono and H. Liu, 1996. Effective data mining using neural networks. Knowl. Data Eng. IEEE. Trans., 8: 957-961.

Patterson, J.A., D.G. McIlwraith and G.Z. Yang, 2009. A flexible, low noise reflective PPG sensor platform for ear-worn heart rate monitoring. Proceedings of the Conference on BSN 2009 Sixth International Workshop Wearable and Implantable Body Sensor Networks, June 3-5, 2009, IEEE, Berkeley, California, USA., ISBN: 978-0-7695-3644-6, pp: 286-291.

Polat, K. and S. Gunes, 2007. A hybrid approach to medical decision support systems: Combining feature selection, fuzzy weighted pre-processing and AIRS. Comput. Methods Programs Biomed., 88: 164-174.

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