Testing the Probability of Heart Disease Using Classification and Regression Tree Model

Mohammad Subhi Al-batah1*

1Department of Computer Science and Software Engineering, Faculty of Science and Information Technology, Jadara University, B.O. Box: 733, Irbid 21110, Jordan.

Author's contribution

This whole work was carried out by the author MSA.

ABSTRACT

The objective of this study is to predict the presence of heart disease with reduced number of attributes using data mining techniques. The term heart disease is related to all diverse diseases affecting the heart. The exposure of heart disease from various factors is an issue which is not free from false presumptions often accompanied by unpredictable effects. Researchers have been using several data mining techniques to help health care professionals in the diagnosis of heart disease. In our work, Classification and Regression Tree (CRT) is proposed to determine the attributes which contribute more towards the diagnosis of heart ailments, which indirectly may reduce the number of tests which are needed to be taken by a patient. The dataset used consists of 270 cases. Originally, thirteen attributes were involved (age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate, exercise induced angina, depression induced by exercise relative to rest, slope, number of vessels colored by fluoroscopy, and exercise thallium scintigraphic defects). To evaluate the performance of the model the sensitivity, specificity, and accuracy are calculated. Comparison with other data mining techniques is presented. The simulation result obtained from the model enables to establish significant patterns and relationships between the medical factors and heart disease.

Keywords: Modeling; Testing; Data mining; Decision tree; Heart disease; Features reduction; Classification and regression tree.

*Corresponding author: Email: albatah@jadara.edu.jo;
1. INTRODUCTION

Today, medical services have come a long way to treat patients with various diseases. Today, diagnosing patients correctly and administering effective treatments have become quite a challenge. Poor clinical decisions may end inpatient's death which could not be tolerated by the hospital as it loses its reputation. The cost to treat a patient with a heart problem is quite high and not affordable by every patient. To achieve a correct and cost-effective treatment, computer-based information and/or decision support systems can be developed to do the task. Most hospitals today use some sort of hospital information systems to manage their healthcare or patient data. These systems typically generate huge amounts of data which take the form of numbers, text, charts and images [1].

There have been several efforts in developing indices to estimate cardiovascular risk, since cardiovascular diseases are the main cause of morbidity and mortality in the world [2]. There is a wealth of hidden information in these data that is largely untapped. The diagnosis of diseases is a vital and intricate job in medicine [3]. The recognition of heart disease from diverse features or signs is a multi-layered problem that is not free from false assumptions and is frequently accompanied by impulsive effects. Thus the attempt to exploit knowledge and experience of several specialists and clinical screening data of patients in the form of databases to assist the diagnosis procedure is regarded as a valuable option [4].

Knowledge of the risk factors associated with heart disease helps health care professionals to identify patients at high risk of having heart disease [5]. Statistical analysis has identified the risk factors associated with heart disease to be age, blood pressure, smoking habit, total cholesterol, diabetes, hypertension, family history of heart disease, obesity, and lack of physical activity [6-8].

Data mining techniques play an important role in finding patterns and extracting knowledge from a large volume of data [9]. It is very helpful to provide better patient care and effective diagnostic capabilities [10-12]. Several data mining techniques are used in the diagnosis of heart disease such as: Genetic algorithm, classification via clustering, direct kernel self-organizing map, naïve bayes, decision tree, neural network, kernel density, automatically defined groups, bagging algorithm, and support vector machine showing different levels of accuracies [13-14]. Among these classification algorithms decision tree algorithms is the most commonly used because it is easy to understand and cheap to implement. It provides a modeling technique that is easy for humans to comprehend and simplifies the classification process [15-17].

In particular, researchers have been investigating the application of the decision tree technique in the diagnosis of heart disease with considerable success. Andreeva [18] used C4.5 decision tree in the diagnosis of heart disease showing accuracy of 75.73%. Sitar-taut et al. [19] used the weka tool to investigate naïve bayes and J4.8 decision trees for the detection of coronary heart disease. The results showed that there is no significant difference between naïve bayes and decision trees in the ability to realize a correct prediction of coronary heart disease. Tu et al. [20] used the bagging algorithm in the weka tool and compared it with J4.8 decision tree in the diagnosis of heart disease. The bagging algorithm showed a better accuracy of 81.41% while the decision tree showed an accuracy of 78.91%. Rajkumar and Reena [21] developed an Intelligent Heart Disease Prediction System to predict the heart disease using three classifiers; decision tree, naïve bayes and neural networks. Kaur and Wasan [22] examined the potential use of classification data mining technique like decision tree, rule induction and artificial neural network for the
diagnosis of diabetic patients. Ordonez [23] implemented an efficient search for the
diagnosis of heart disease comparing association rules with decision trees. Xing et al. [24]
conducted a survey of 1000 patients, the results of which showed SVM to have 92.1%
accuracy, artificial neural networks to have 91.0% and decision trees with 89.6% using TNF,
IL6, IL8, HICRP, MPO1, TNI2, sex, age, smoke, hypertension, diabetes, and survival as the
parameters. Similarly, Chen et al. [25] compared the accuracy of SVM, neural networks,
Bayesian classification, decision tree and logistic regression. Considering 102 cases, SVM
had the highest accuracy of 90.5%, neural networks 88.9%, Bayesian 82.2%, decision tree
77.9%, and logistic regression 73.9%. Karaolis et al. [26] has developed a data mining
system for the assessment of heart event related risk factors. It is found that the data mining
technique could help in the identification of high and low risk subgroups of patients. Decision
tree was used for extracting rules based on the risk factors. Palaniappan and Awang [27]
have developed an Intelligent Heart Disease Prediction System (IHDPS) using decision
trees, neural networks and naive bayes. A scalable, reliable, expandable, and user friendly
web-based system was developed on the .NET platform.

As stated, decision tree is one of the successful data mining techniques used in the
diagnosis of heart disease. Our work aims to predict efficiently diagnosis with reduced
number of factors that contribute more towards the heart disease using Classification and
Regression Tree (CRT) model, and to use Gini index to measure the impurity of a partition or
set of training tuples [28]. It can handle high dimensional categorical data.

2. CLASSIFICATION AND REGRESSION TREE MODEL

Decision tree induction is one of the classification techniques used in decision support
systems and machine learning process [29]. With this technique the training data set is
recursively partitioned using depth-first (Hunt’s method) or breadth-first greedy technique
until each partition is pure or belongs to the same class/leaf node [30]. The model is
preferred among other classification algorithms because it is an eager learning algorithm and
easy to implement. Decision tree algorithms can be implemented serially or in parallel.
Despite the implementation method adopted, most decision tree algorithms in literature are
constructed in two phases: tree growth and tree pruning phase. Tree pruning is an important
part of decision tree construction as it is used in improving the classification/prediction
accuracy by ensuring that the constructed tree model does not over fit the data set [31].

The decision tree is based on a multistage or hierarchical decision scheme (tree structure).
The tree is composed of a root node, a set of internal nodes, and a set of terminal nodes
(leaves). Each node of the decision tree structure makes a binary decision that separates
either one class or some of the classes from the remaining classes. The processing is
carried out by moving down the tree until the terminal node is reached. In a decision tree,
features that carry maximum information are selected for classification, while remaining
features are rejected, thereby increasing computational efficiency [32]. The top-down
induction of the decision tree indicates that variables in the higher order of the tree structure
are more important [33]. In this study we focused on one type of decision tree called CRT
model which are memory resident, fast and easy to implement.

CRT is a recursive partitioning method to be used both for regression and classification; is
constructed by splitting subsets of the data set using all predictor variables to create two
child nodes repeatedly, beginning with the entire data set. The best predictor is chosen using
a variety of impurity or diversity measures (Gini, twoing, ordered twoing and least-squared
deviation). The goal is to produce subsets of the data which are as homogeneous as
possible with respect to the target variable [34]. In this study, we used the measure of Gini impurity that is used for categorical target variables.

Gini Impurity Measure:

The Gini index at node \( t \), \( g(t) \), is defined as

\[
g(t) = \sum_{j \neq i} p(j|t) (i|t)
\]

where \( i \) and \( j \) are categories of the target variable. The equation for the Gini index can also be written as

\[
g(t) = 1 - \sum_{j} p^2(j|t)
\]

Thus, when the cases in a node are evenly distributed across the categories, the Gini index takes its maximum value of \( 1 - \left( \frac{1}{k} \right) \), where \( k \) is the number of categories for the target variable [35]. When all cases in the node belong to the same category, the Gini index equals 0. If costs of misclassification are specified, the Gini index is computed as

\[
g(t) = \sum_{j \neq i} C(i|j)p(j|t) p(i|t)
\]

where \( C(i|j) \) is the probability of misclassifying a category \( j \) case as category \( i \). The Gini criterion function for split \( s \) at node \( t \) is defined as

\[
\emptyset(s,t) = g(t) - p_L g(t_L) - p_R g(t_R)
\]

where \( p_L \) is the proportion of cases in \( t \) sent to the left child node, and \( p_R \) is the proportion sent to the right child node. The split \( s \) is chosen to maximize the value \( \emptyset(s,t) \). This value is reported as the improvement in the tree [36].

3. DATA COLLECTION

A total of 270 cases with 13 medical attributes were obtained from the Statlog Heart Disease database [37]. Papers that cite this data set [38-41]. The dataset contained 150 patients without heart disease and 120 patients with heart disease. The diagnosis classes is identified as value "1" for patients with no heart disease and value "2" for patients with heart disease. The attributes are acronym as: age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, and thal. The data don't have any missing values for all attributes [42-44]. Table 1 lists the name of the attributes with its type and description.
Table 1. Heart disease data set attributes

| 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 pain, 4 = asymptomatic |
| trestbps  | Continuous | Resting blood pressure (in mm Hg)                                           |
| chol      | Continuous | Serum cholesterol (in mg/dl)                                                |
| fbs       | Discrete   | Fasting blood sugar > 120 mg/dl: 1 = true, 0 = false                        |
| restecg   | Discrete   | Resting electrocardiographic results: 0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria |
| thalach   | Continuous | Maximum heart rate achieved                                                 |
| exang     | Discrete   | Exercise induced angina: 1 = yes, 0 = no                                    |
| oldpeak   | Continuous | ST depression induced by exercise relative to rest                           |
| slope     | Discrete   | The slope of the peak exercise ST segment: 1 = up sloping, 2 = flat, 3 = down sloping |
| ca        | Discrete   | Number of major vessels colored by fluoroscopy that ranged between 0 and 3. |
| thal      | Discrete   | Exercise thallium scintigraphic defects: 3 = normal, 6 = fixed defect, 7 = reversible defect |

The data is divided into 90% (239 cases) for training, and 10% (31 cases) for a test set. To evaluate the performance of the CRT model the sensitivity, specificity, and accuracy (the proportion of cases that are correctly classified [36]) were calculated.

Sensitivity = True Positive/Positive  
Specificity = True Negative/ Negative  
Accuracy = (True Positive + True Negative) / (Positive +Negative)

4. RESULTS AND DISCUSSION

CRT algorithm is used to predict absence or presence of heart disease based on values of all independent variables. The tree diagram (Fig. 1) shows tree construction based on the test sample of 31 cases, 0.05 adjustments of the probabilities, a minimum parent node size of 30, a minimum child nodes size of 10 and equal misclassification costs. The Gini index was selected as a splitting criterion. There are totally 10 nodes that consist of 6 terminal nodes; the first node placed in the tree is the root node. The depth of the tree is equal to 3. Parent node has 17 absence (54.8%) and 14 presence (45.2%) of heart disease. Prior probabilities used are the observed probability 0.548 and 0.452 for absence to presence of heart disease, respectively. The first discriminator "exercise thallium scintigraphic defects" splits the root node into two child nodes: normal (node 1, n=17), and fixed or reversible defect (node 2, n=14). The improvement for this classification is 0.136. The classifier "chest pain type" is for normal defect (0.028 improvement), and the "number of vessels colored by
fluoroscopy" is for fixed or reversible defects (0.031 improvement). The next discriminator is "chest pain type" split into atypical angina or non-anginal pain (terminal node 3), and typical angina, or asymptomatic (Node 4). The next discriminator is "number of vessels colored by fluoroscopy" which produces two nodes; node 5 for that number of major vessels colored by fluoroscopy is 0, and terminal node 6 for those numbers of major vessels colored by fluoroscopy that ranged between 1 and 3. When "chest pain type" is typical angina, or asymptomatic, then the next discriminator is "number of vessels colored by fluoroscopy" (0.027 improvement), which is split into terminal node 7 (number of vessels colored by fluoroscopy is equal to 0) and terminal node 8 (number of vessels colored by fluoroscopy is ranged between 1 and 3). Also, when the "number of vessels colored by fluoroscopy" is 0, then the last discriminator is "chest pain type" (0.019 improvement), which produces two terminal nodes; terminal node 9 for asymptomatic and terminal node 10 for typical angina, atypical angina, or non-anginal pain. Percentages in each category and in each joint category are shown in Fig. 1, and in Table 2.

Fig. 1. CRT classification tree for test samples
Table 2. Tree table for test samples

| Node | absence | presence | Total | Predicted Category | Parent Node | Primary Independent Variable | Improvement | Split Values |
|------|---------|----------|-------|-------------------|-------------|-------------------------------|-------------|--------------|
| 0    | 17      | 14       | 31    | 100.0%            | absence     |                               |             |              |
| 1    | 13      | 4        | 17    | 54.8%             | absence     | exercise thallium scintigraphic defects | 0.136       | 3.00         |
| 2    | 4       | 10       | 14    | 45.2%             | presence    | exercise thallium scintigraphic defects | 0.136       | 7.00; 6.00   |
| 3    | 8       | 0        | 8     | 100.0%            | absence     | chest pain type                | 0.028       | 3.00; 2.00   |
| 4    | 5       | 4        | 9     | 29.0%             | absence     | chest pain type                | 0.028       | 4.00; 1.00   |
| 5    | 3       | 3        | 6     | 19.4%             | presence    | number of vessels colored by fluoroscopy | 0.031       | 0.00         |
| 6    | 1       | 7        | 8     | 25.8%             | presence    | number of vessels colored by fluoroscopy | 0.031       | 3.00; 1.00; 2.00 |
| 7    | 4       | 0        | 4     | 12.9%             | absence     | number of vessels colored by fluoroscopy | 0.027       | 0.00         |
| 8    | 1       | 4        | 5     | 16.1%             | presence    | number of vessels colored by fluoroscopy | 0.027       | 3.00; 1.00; 2.00 |
| 9    | 1       | 3        | 4     | 12.9%             | presence    | chest pain type                | 0.019       | 4.00         |
| 10   | 2       | 0        | 2     | 6.5%              | absence     | chest pain type                | 0.019       | 3.00; 2.00; 1.00 |
The improvement measure shown in Table 2, measures the increase of the effect of child node on the dependent variable, it is determined by the largest difference in the proportions of the dependent variable in the child nodes [45]. Thus, an improvement of 0.136 means that “exercise thallium scintigraphic defects” contributes 13.6% in the discrimination between absence and presence of heart disease; "chest pain type" makes an additional 4.7% improvement, and "number of vessels colored by fluoroscopy" makes another 5.8% improvement.

The results of CRT are summarized in Table 2, it is clear that not all predictor's categories contribute in the classification process. The predictors: age, sex, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate, exercise induced angina, depression induced by exercise relative to rest, and slope do not contribute to the classification tree used.

The Dependency viewer in CRT model shows the results from the most significant to the least significant (weakest) medical predictors. The normalized importance of predictors in the classification is given in Table 3. The CRT model shows that the most significant factors influencing heart disease are chest pain type, number of vessels colored by fluoroscopy, and exercise thallium scintigraphic defects. Other significant factors include depression induced by exercise relative to rest, maximum heart rate, exercise induced angina, slope, and age. The model shows sex, resting blood pressure, serum cholesterol and fasting blood sugar as the weakest factors. The attribute "resting electrocardiographic results" is excluded from the recognition. Physicians can use this information to further analyze the strength and weakness of the medical attributes associated with heart disease.

| Table 3. Independent variable importance |
|-----------------------------------------|
| Independent Variable | Importance | Normalized Importance |
|-----------------------|------------|-----------------------|
| chest pain type        | 0.179      | 100.0%                |
| number of vessels colored by fluoroscopy | 0.174       | 97.5%                 |
| exercise thallium scintigraphic defects | 0.149       | 83.4%                 |
| depression induced by exercise relative to rest | 0.137       | 76.9%                 |
| maximum heart rate     | 0.120      | 67.1%                 |
| exercise induced angina | 0.101        | 56.8%                 |
| slope                  | 0.090      | 50.6%                 |
| age                    | 0.060      | 33.8%                 |
| sex                    | 0.050      | 27.8%                 |
| resting blood pressure | 0.024       | 13.6%                 |
| serum cholesterol       | 0.023      | 13.1%                 |
| fasting blood sugar     | 0.002      | 1.1%                  |

The classification accuracy is the percent predicted correct (absence and presence) over the total sample size used. It was found that the prediction accuracy is 90.3%, with sensitivity 82.4% and specificity 100.0%. The misclassification rate (risk resubstitution estimate) is 0.097 with standard error 0.053.

Comparison of a study with other in literature is an indispensable part of a study to evaluate the obtained results from its applications. Whereas the comparison results cannot be generalized to all problems related with developed systems, at least this comparison gives an alternative about the place of the proposed system among the compared classifiers in literature.
Work on the heart disease datasets can be dated since 1989, when Detrano et al. [46] used logistical regression to obtain 77% accuracy of prediction. The accuracies of different models on the same datasets have been tabulated in Table 4 [47-51]. As can be seen, the highest classification accuracy was obtained via GA-AWAIS with a classification accuracy of 87.43%. While, the lowest classification accuracy among the studies conducted is reached by InductH with an accuracy of 58.5%. As stated, the proposed CRT model achieved an accuracy of 90.3% which is higher than other methods. This result is also important when it is taken into account that the average classification accuracy of classifiers applied for this problem is in 75.8%. Thus, this problem can be seen as a hard medical classification problem and CRT model has reached a considerably good result for this problem.

Table 4. Classification accuracies obtained by other classifiers

| Author (Ref #)          | Technique              | Accuracy (%) |
|-------------------------|------------------------|--------------|
| Ozsen and Gunes [47]    | GA-AWAIS               | 87.43        |
| Polat et al. [48]       | k-nn based Fuzzy-AIRS  | 87.00        |
| Polat et al. [49]       | AIRS                   | 84.50        |
| WEKA, RA                | Naïve-Bayes            | 83.60        |
| Sahan et al. [50]       | AWAIS                  | 82.59        |
| Cheung [51]             | Naïve Bayes            | 81.48        |
| Cheung [51]             | BNND                   | 81.11        |
| Cheung [51]             | C4.5                   | 81.11        |
| Cheung [51]             | BNNF                   | 80.96        |
| Detrano et al. [46]     | Logistic regression    | 77.00        |
| WEKA, RA                | K*                     | 76.70        |
| WEKA, RA                | IB1c                   | 74.00        |
| WEKA, RA                | 1R                     | 71.40        |
| WEKA, RA                | T2                     | 68.10        |
| Tool Diag, RA           | MLP + BP               | 65.60        |
| WEKA, RA                | FOIL                   | 64.00        |
| Tool Diag, RA           | RBF                    | 60.00        |
| WEKA, RA                | Induct H               | 58.50        |

5. CONCLUSION

Diagnosis of heart disease is a significant task in medicine. This paper investigates applying the CRT model to help healthcare professionals in this task. A data set of 270 cases and thirteen attributes were selected as the input variables of the model. Based on the proposed model, the ascending order of important variables are chest pain type, number of vessels colored by fluoroscopy, exercise thallium scintigraphic defects, depression induced by exercise relative to rest, maximum heart rate, exercise induced angina, slope, age, sex, resting blood pressure, serum cholesterol, and fasting blood sugar. The attribute resting electrocardiographic results is excluded from the recognition. The simulation results show the capability of this model for the prognosis of heart disease with good accuracy and appropriate convergence. Also, compared to other methods in literature, CRT has obtained reasonable results. This type of research can play an important role in improving patient outcomes, cost reduction of medicine, and further advanced clinical studies. The proposed work can be further enhanced and expanded for the automation of heart disease prediction. Real data from health care organizations and agencies needs to be collected and all the
available techniques will be compared for the optimum accuracy. It is hoped that deeper investigations using this method in the future will use more recent datasets.

ACKNOWLEDGEMENT

The author would like to thank Malak Fuad Subhi Al-Battah, a student in Department of Medicine and Surgery, Faculty of Medicine in Jordan University of Science and Technology, for her ideas about heart disease that were insightful and most helpful for enhancing the article.

COMPETING INTERESTS

Author has declared that no competing interests exist.

REFERENCES

1. Ishtake SH, Sanap SA. Intelligent Heart Disease Prediction System Using Data Mining Techniques. International J. of Healthcare & Biomedical Research. 2013;1(3):94-101.
2. Heller RF, Chinn S, Pedoe HD, Rose G. How well can we predict coronary heart disease? Findings in the United Kingdom Heart Disease Prevention Project. British Medical Journal. 1984;288(6428):1409-1411.
3. Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of Coronary Heart Disease Using Risk Factor Categories. American Heart Association Journal. 1998;97:1837-1847.
4. Simons LA, Simons J, Palaniappan L, Friedlander Y, McCallum J. Risk functions for prediction of cardiovascular disease in elderly Australians: the Dubbo Study. Medical Journal of Australia. 2003;178(3):113-116.
5. Salahuddin, Rabbi F, Qadir F, Khattak M. Statistical Analysis of Risk Factors for Cardiovascular disease in Malakand Division. Pakistan journal of statistics and operation research. 2006;3(2):49-56.
6. Shahwan-Akl L. Cardiovascular Disease Risk Factors among Adult Australian-Lebanese in Melbourne. International Journal of Research in Nursing. 2010;1(1):1-7.
7. Srinivas K, Rani BK, Govrghan A. Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. International Journal on Computer Science and Engineering (JCSE). 2010;2(2):250-255.
8. Kangwanariyakul Y, Nantasenamat C, Tantimongcolwat T, Naenna T. Data mining of magnetocardiograms for prediction of ischemic heart disease. EXCLI Journal. 2010;9:82-95.
9. Rajkumar A, Reena GS. Diagnosis of Heart Disease Using Datamining Algorithm. Global Journal of Computer Science and Technology. 2010;10(10):38-43.
10. Beule MD, Maes E, Winter ODE, Vanlaere W, Impe RV. Artificial neural networks and risk stratification: A promising combination. Mathematical and Computer Modelling. 2007;46(1-2):88-94.
11. Tantimongcolwat T, Naenna T, Ayudhya Cin, Embrechts MJ, Prachayasittikul V. Identification of ischemic heart disease via machine learning analysis on magnetocardiograms. Computers in Biology and Medicine. 2008;38(7):817-825.
12. Anbarasi M, Anupriya E, Iyengar NCHSN. Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. International Journal of Engineering Science and Technology. 2010;2(10):5370-5376.
13. Kavitha KS, Ramakrishnan KV, Singh MK. Modeling and design of evolutionary neural network for heart disease detection. International Journal of Computer Science Issues (IJCSI). 2010;7(5):272-283.
14. Kaur H, Wasan SK. Empirical Study on Applications of Data Mining Techniques in Healthcare. Journal of Computer Science. 2006;2(2):194-200.
15. Mythili T, Mukherji D, Padalia N, Naidu A. A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL). International Journal of Computer Applications. 2013;68(16):11-15.
16. Pandey AK, Pandey P, Jaiswal KL, Sen AK. A Heart Disease Prediction Model using Decision Tree. IOSR Journal of Computer Engineering (IOSR-JCE). 2013;12(6):83-86.
17. Anyanwu MN, Shiva SG. Comparative Analysis of Serial Decision Tree Classification Algorithms. International Journal of Computer Science and Security. 2009;3(3):230-240.
18. Andreeva P. Data Modelling and Specific Rule Generation via Data Mining Techniques. International Conference on Computer Systems and Technologies--Comp Sys Tech. 2006;17:1-6.
19. Sitar-tăut AV, Zdrenghea D, Pop D, Sitar-tăut DA. Using Machine Learning Algorithms in Cardiovascular Disease Risk Evaluation. Journal of Applied Computer Science & Mathematics. 2009;5(3):29-32.
20. Tu MC, Shin D, Shin D. Effective Diagnosis of Heart Disease through Bagging Approach. Biomedical Engineering and Informatics. BMEI '09. 2nd International Conference. 2009;1-4.
21. Rajkumar A, Reena GS. Diagnosis of Heart Disease Using Data mining Algorithm, GJCST. 2010;10(10):38-43.
22. Kaur H, Wasan SK. Empirical Study on Applications of Data Mining Techniques in Healthcare. Journal of Computer Science. 2006;2(2):194-200.
23. Ordonez C. Comparing Association Rules and Decision Trees for Disease Prediction. ACM, HIKM'06, Arlington, Virginia, USA. 2006;17-24.
24. Xing Y, Wang J, Zhao Z, Gao Y. Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease. Convergence Information Technology. International Conference. 2007;868-872.
25. Chen J, Xi G, Xing Y, Chen J, Wang J. Predicting Syndrome by NEI Specifications: A Comparison of Five Data Mining Algorithms in Coronary Heart Disease. Life System Modeling and Simulation. Lecture Notes in Computer Science. 2007;4689:129-135.
26. Karaolis MA, Moutiris JA, Hadjipanayi D, Pattiichis CS. Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees. IEEE Transactions on Information Technology in Biomedicine. 2010;14(3):559-566.
27. Palaniappan S, Awang R. Intelligent Heart Disease Prediction System Using Data Mining Techniques. IJCSNS International Journal of Computer Science and Network Security. 2008;8(8):343-350.
28. Han J, Kamber M. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers, 2nd Edition; 2006.
29. Shafer J, Agrawal R, Mehta M. Sprint: A scalable parallel classifier for data mining. Proceedings of the 22th International Conference on Very Large Data Bases, CA, USA, 1996;544-555.
30. Hunt EB, Marin J, Stone PJ. Experiments in induction, Academic Press, New York. 1966;1-247.
31. Mehta M, Agrawal R, Rissanen J. SLIQ: A fast scalable classifier for data mining. Advances in Database Technology-EDBT’96. Lecture Notes in Computer Science. 1996;1057:18-32.
32. Xu M, Watanachaturaporn P, Varshney PK, Arora MK. Decision tree regression for soft classification of remote sensing data. Remote Sensing of Environment. 2005;97:322-336.
33. Witten IH, Frank E. Data Mining - Practical Machine Learning Tools and Techniques, Second Edition. Elsevier, Amsterdam; 2005.
34. Lewis RJ. An Introduction to Classification and Regression Tree (CART) Analysis. Presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine. San Francisco, California. 2000;1-14.
35. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Monterey: Wadsworth and Brooks/Cole; 1984.
36. Bramer M. Principles of data mining. Second Edition, Springer; 2007.
37. Blake CL, Mertz CJ. UCI Repository of Machine Learning Databases. 1998. Accessed 20 March 2013. Available: http://mlearn.ics.uci.edu/databases/heart-disease/.
38. Brown G. Diversity in Neural Network Ensembles. PhD Thesis, The University of Birmingham, United Kingdom; 2004.
39. Kononenko I, Šimc E, Robnik-Šikonja M. Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF. Applied Intelligence. 1997;7(1):39-55.
40. Smirnova EN, Sprinkhuizen-Kuyper IG, Nalbantis IG, Ikat A. Unanimous Voting using Support Vector Machines. 2004;1-8.
41. Shouman M, Turner T, Stocker R. Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients. International Journal of Information and Education Technology. 2012;2(3):220-223.
42. Negar A, Ahad Z. A New Approach to Heart Failure Prognosis based on Combination of Machine Learning and Fuzzy Clustering. International Journal of Computer Technology and Applications. 2010;2(1):199-206.
43. Patel SB, Yadav PK, Shukla DP. Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques. IOSR Journal of Agriculture and Veterinary Science. 2013;4(2):61-64.
44. Soni J, Ansari U, Sharma D, Soni S. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. International Journal of Computer Applications. 2011;17(8):43-48.
45. Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. Annals of Behavioral Medicine. 2003;26(3):172-181.
46. Detrano R, Janosi A, Steinbrunn W, Pfisterer M, SchmidJJ, Sandhu S, Guppy KH, Lee S, Froelicher V. International application of a new probability algorithm for the diagnosis of coronary artery disease. The American Journal of Cardiology. 1989;64(5):304-310.
47. Ozsen S, Gunes S. Attribute weighting via genetic algorithms for attribute weighted artificial immune system (AWAIS) and its application to heart disease and liver disorders problems. Expert Systems with Applications. 2009;36(1):386-392.
48. Polat K, Sahan S, Gunes S. Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing. Expert Systems with Applications. 2007;32(2):625-631.
49. Polat K, Sahan S, Kodaz H, Günes S. A new classification method to diagnose heart disease: Supervised artificial immune system. In Proceedings of the Turkish Symposium on Artificial Intelligence and Neural Networks (TAiNN). 2005;2186-2193.

50. Şahan S, Polat K, Kodaz H, Güneş S. The Medical Applications of Attribute Weighted Artificial Immune System (AWAIS): Diagnosis of Heart and Diabetes Diseases. Artificial Immune Systems. Lecture Notes in Computer Science. 2005;3627:456-468.

51. Cheung N. Machine learning techniques for medical analysis. School of Information Technology and Electrical Engineering, B.Sc. Thesis, University of Queensland; 2001.

© 2014 Al-batah; This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history.php?iid=437&id=32&aid=3600