Original Article

A New Hybrid Method for Improving the Performance of Myocardial Infarction Prediction

Hojat Hamidi [1*], Atefeh Daraei[1]

[1.] Information Technology Engineering Department, K. N. Toosi University of Technology, Tehran, Iran.

Received: 2016/04/05 **Accepted:** 2016/05/14

Abstract

Introduction: Myocardial Infarction, also known as heart attack, normally occurs due to such causes as smoking, family history, diabetes, and so on. It is recognized as one of the leading causes of death in the world. Therefore, the present study aimed to evaluate the performance of classification models in order to predict Myocardial Infarction, using a feature selection method that includes Forward Selection and Genetic Algorithm.

Materials & Methods: The Myocardial Infarction data set used in this study contains the information related to 519 visitors to Shahid Madani Specialized Hospital of Khorramabad, Iran. This data set includes 33 features. The proposed method includes a hybrid feature selection method in order to enhance the performance of classification algorithms. The first step of this method selects the features using Forward Selection. At the second step, the selected features were given to a genetic algorithm, in order to select the best features. Classification algorithms entail Ada Boost, Naïve Bayes, J48 decision tree and simpleCART are applied to the data set with selected features, for predicting Myocardial Infarction.

Results: The best results have been achieved after applying the proposed feature selection method, which were obtained via simpleCART and J48 algorithms with the accuracies of 96.53% and 96.34%, respectively.

Conclusion: Based on the results, the performances of classification algorithms are improved. So, applying the proposed feature selection method, along with classification algorithms seem to be considered as a confident method with respect to predicting the Myocardial Infarction.

Keywords: Ada Boost; Forward Selection; J48 decision tree; Myocardial Infarction; SimpleCART

[*] **Corresponding author: Tel:** 02188464143 **email:** H_Hamidi@kntu.ac.ir

Introduction

Cardiovascular diseases are known as the most important causes of death all over the world. Based on Ministry of Health and Medical Education report in 2012, cardiac diseases involve 39% of the mortality rates, that the mortality rate due to Myocardial Infarction has been reported 85 per 100,000 [1]. Since a rapid increase has been demonstrated in heart diseases in the world, this disease is likely to become the most common cause of death by 2020 [2]. Myocardial Infarction, known as a heart attack, means the death of heart muscle due to a sudden blockage of a coronary artery by a blood clot [3].

On the other hand, using data mining allows extraction of useful knowledge from the data [4, 5]. Hence, data mining can be considered as a tool in order to turn the raw data to knowledge in the field of Myocardial Infarction [6]. Data mining can be categorized into three steps: preprocessing, modeling and post-processing, among which modeling step contains two category tasks: first, the predictive category which includes classification algorithms as well as regression methods; second, the descriptive category entails clustering algorithms and association rule algorithms. Classification algorithms specify the class label for the test data using the training data with specific labels [7]. In classification, data are divided into two sections. First section uses the training data for learning, and the second section utilizes the test data as validation [4, 7]. Tsien et al. [8] used the decision tree and logistic regression to diagnose Myocardial Infarction. They evaluated information of 9856 patients, of which 9656 had referred to the Royal Hospital in Edinburgh and 500 to a hospital in Sheffield, England. Conforti et al. [9] applied Support Vector Machine (SVM) whit five core functions, including Linear, Gaussian, Laplacian, Polynomial and Sigmoid for diagnosis of acute Myocardial Infarction (AMI). The best accuracy in average in different feature selection (FS) methods was equal to 86.43% for Laplace core function. Baxt et al. [10] proposed a prediction model for acute Myocardial Infarction which used artificial neural network. The data set includes information of 2204 patients and 40 features. Their results showed the high performance of neural network for prediction of acute Myocardial Infarction. Qazi et al. [11] are proposed an abnormality detection system for detection the abnormal heart rate based on real data sets. After feature selection, SVM classification was applied to the data. The results showed that selecting the three important features achieved the highest efficiency. Masethe and Masethe [12] proposed five classification algorithms including, Bayesian networks, J48 decision tree, CART, Naïve Bayes (NB), and REPTREE in order to predict the heart attack. The results revealed that NB, CART and J48 have achieved higher accuracy, equal 99.07%, compared to REPTREE and Bayesian network. Patil and Kumaraswamiin [13] proposed a neural network method applied to a data set with 13 features, to detect the heart attacks. After preprocessing, the classes were determined using clustering. Finally, a three layer neural network is used

for classification. In another study, Bhaskar [14] used the ECG signals to detect Myocardial Infarction. Regarding the two methods of neural network and support vector machine on the data set, the results reached higher accuracy for the SVM than the artificial neural network. Karaolis et al. [15] proposed a prediction model with regard to Myocardial Infarctions and coronary artery bypass graft. Ultimately, using the C4.5 tree, relevant rules were extracted and the most important factors, affecting Myocardial Infarction were determined. In addition, Krishnaraj and Vinothkumar [16] proposed a feed-forward neural network model for predicting the heart attacks, combined with genetic algorithm (GA). Applying GA optimal weights, the neural network was determined. UCI Data Repository was used in this study containing 270 patients and 13 features. The proposed model reduced the features from 13 to 6 main features. Their proposed model achieved accuracy of 88%. Hachesu et al. [17] used data mining approaches to predict the Long of Stay (LOS) of the cardiac patients in the hospitals. The data set they used entailed 4948 cases of patients suffering from CAD, which contained 36 features. Three algorithms namely Decision Tree, SVM and Artificial Neural Network were used with respect to LOS prediction. As a matter of fact, they reached accuracy 96.4% for SVM, and concluded that the LOS for single patients with marital status “single” was equal or less than 5 days, whereas it was reported to be more than 10 days for the married patients. Furthermore, the findings of the present study revealed that renal diseases and high blood pressure seem to cause longer LOS.

Due to the high mortality rate resulted in Myocardial Infarction in the world, the side effects of the treatment methods and the drugs used for Myocardial Infarction, the current study aimed to apply classification methods for predicting Myocardial Infarction, utilizing a hybrid feature selection method. Regarding the proposed method in this study, new feature selection approach was used along with the classification algorithms to predict the Myocardial Infarction. Feature selection method includes Forward Selection at the first phase and then a genetic algorithm at the second stage. Ultimately, this method extracts the best features for classification algorithms, which can lead to a higher performance in comparison with absence of feature selection. The algorithms of AdaBoost, Naïve Bayes, J48 decision tree and simpleCART were used to achieve this goal.

Materials and Methods

RapidMiner is a software used to implement the proposed model. In fact, this is a powerful and easy-to- use Graphical User Interface for designing an analytic process [18]. In this study, version 7.0.1 of RapidMiner was applied.

The data used in this study was obtained out of the information collected from records of 519 visitors to Shahid Madani Hospital of Khorramabad, of which 297 patients suffered from Myocardial Infarction and 222 cases revealed no symptoms in this regard. This data

set contains 33 features. The features are selected based on the risk factors which are specified in books [2] and [19] for Myocardial Infarction. The features of the Myocardial Infarction data set are illustrated in Table 1.

Table 1. Summary of the Features in Myocardial Infarction Data Set

Features	Range	Features	Range
Sex	M, F	Systolic Blood Pressure	80 – 210
Age	28-93	Diastolic Blood Pressure	40 – 190
Weight	43-120	Pulse Rate	50 – 190
BMI (Body Mass Index)	16-42	Edema	Yes , No
FH (Family History)	Yes , No	Fatigue and weakness	Yes , No
DM(Diabetes Mellitus)	Yes, No	Lung Rales	Yes , No
Smoker	Yes , No	Typical Chest Pain	Yes , No
Obesity	Yes , No	Distribution of pain to arms and neck	Yes , No
HTN (History of Hypertension)	Yes , No	Dyspnea	Yes , No
CRF (Chronic Renal Failure)	Yes , No	Atypical Chest Pain	Yes , No
CVA (Cerebrovascular Accident)	Yes , No	Non-anginal Chest Pain	Yes , No
Airway Disease	Yes , No	Exertional Chest Pain	Yes , No
Thyroid Disease	Yes , No	ST Elevation	Yes , No
HLP (Hyperlipidemia)	Yes , No	ST Depression	Yes , No
Blood Pressure	Yes , No	T inversion	Yes , No
CHF(Congestive Heart Failure)	Yes, No	Poor R Progression	Yes , No
Class	MI= 0 Healthy=1		

- **Preprocessing**

The nature of raw data is usually incomplete and noisy. These noisy data, missing values and unrelated values could be due to human errors [20]. For handling these incomplete and noisy data, preprocessing methods are used. Preprocessing is considered as an important stage in data mining [21]. Therefore, the data is prepared for the data mining [22]. Preprocessing involves cleansing, integration, reduction and transformation of data, among which cleansing is related to handling the missing values with the appropriate values [20]. Min-max normalization is a process in which the feature values are scaled to a smaller range.

- **Hybrid Feature Selection Method**

In this study a hybrid feature selection method was applied to the data set. In general, this method is divided into two steps:

1) **Forward Selection:** This procedure starts with an empty set. In each step, the best feature is determined and added to the set. Then in each iteration of the method, the best feature is selected of the remained features [20]. The maximum number of attribute is set to 32. K-NN (K=5) is

considered as the algorithm nested in operator, for selecting the features.

2) **Genetic Algorithm:** Genetic algorithm achieves to the best individuals by searching the population space. Initial population is randomly generated. In order to evaluate each population, a fitness function is used, in which the greater fitness has the higher probability of being chosen for the next generation. Moreover, the other operations, such as crossover and mutation are applied with respect to producing the new population in the next generation [23]. Regarding genetic algorithm, the population size is set to 7and the number of generations is considered 10. For the other parameters, mutation and crossover, the defaults of the software are considered.

- **Classification Algorithms**

Classification algorithms can be stated as the next step after feature selection, utilized in order to predict MI. The algorithms, used in this study, include AdaBoost, Naïve Bayes, J48 decision tree and simpleCART.

1) **AdaBoost:** This algorithm trains different classifiers on similar training set in an iterative process. This is obtained by different distributions of data and can determine the data weights based on the classification results [24].
2) **Naïve Bayes (NB):** The Naïve Bayes algorithm is a probabilistic classifier based on the Bayes rule of conditional probability. Naive Bayes classifier uses probability to classify the new instance [6, 25]. In this algorithm, the features are considered independent of each other, which means the importance of them are equal [26].
3) **J48 Decision Tree:** J48 algorithm is a binary tree, which is a simple form of C4.5 tree. In tree approaches, classification process is modeled by a tree, which is applied on all the data in the data set and determine the class label for each data in data set [27].
4) **SimpleCART:** This algorithm presents the results in the form of a decesion tree or a regression tree. This algorithm uses a minimal cost complixity method in regard with the classification [28].

- **Evaluation Method**

In order to evaluate the performance of the proposed classification models, accuracy, sensitivity, and specificity were used.

1) **Accuracy:** Accuracy of a classification method is the ratio of the number of truly classified instances, predicted by a classifier, on the total of instances [29].
2) **Sensitivity** and **specificity:** Sensitivity or true positive rate is the ratio of positive instances which were truly classified into positive class. Specificity is the ratio of negative instances which truly classified in negative class [30].

Results

In this section, the results of implementing the proposed model on the data set, are presented to predict the Myocardial Infarction. After applying the Forward Selection, as the first part of the feature selection, 11 features were selected, which included CVA, ST Elevation,

T inversion, HLP, CHF, DM, Edema, ST depression, Atypical CP, BMI, and Smoker. These selected features are given to the genetic algorithm, considered as the second step of feature selection. Finally, genetic algorithm selects 7 best features, including CHF, DM, Edema, Atypical CP, ST depression, T inversion and ST Elevation.

The classification algorithms are executed via considering these features as the best ones. The results of the classification method are shown in Tables 2, 3, 4 and 5 in the three states.

Table 2. Results of AdaBoost in different states of feature selection

AdaBoost	Accuracy	Sensitivity	Specificity
Not using FS	95.18	95.29	95.05
FS using Forward Selection	95.57	96.63	94.14
FS using Proposed Method	95.95	96.30	95.50

First state shows the results in the lack of using feature selection, in which all the features were used. Second state demonstrates the results of using the first step of feature selection, Forward Selection. Third state presents the results of using the proposed hybrid feature selection. Table 2 reports the results of the AdaBoost algorithm in the mentioned 3 states. Table 3 indicates the results of the Naïve Bayes algorithm in 3 states. Table 4 shows the results of the J48 algorithm, and Table 5 indicates the results of implementing simpleCART.

Table 3. Results of Naïve Bayes in different states of feature selection

Naïve Bayes	Accuracy	Sensitivity	Specificity
Not using FS	93.26	90.57	96.85
FS using Forward Selection	94.61	92.26	97.75
FS using Proposed Method	95.57	94.28	97.30

Table 4. Results of J48 in different states of feature selection

J48	Accuracy	Sensitivity	Specificity
Not using FS	94.99	95.29	94.59
FS using Forward Selection	95.95	95.29	96.85
FS using Proposed Method	96.34	95.62	97.30

Table 5. Results of simpleCART in different states of feature selection

simpleCART	Accuracy	Sensitivity	Specificity
Not using FS	94.99	95.96	93.69
FS using Forward Selection	95.76	95.29	96.40
FS using Proposed Method	96.53	95.62	97.75

Discussion

Based on the Tables 2, 3, 4 and 5, simpleCART and J48 algorithms have achieved the highest accuracy. In spite of highest accuracy of these algorithms, the other algorithms demonstrated small differences in accuracy values. The results of Tables 2, 3, 4 and 5 can be presented as charts in the figures below.

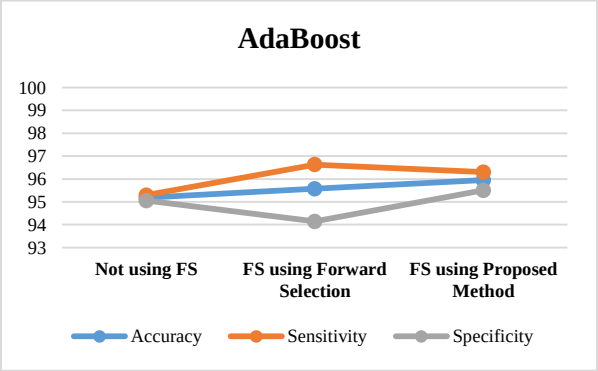

Figure 1. Performance of AdaBoost based on different states of feature selection

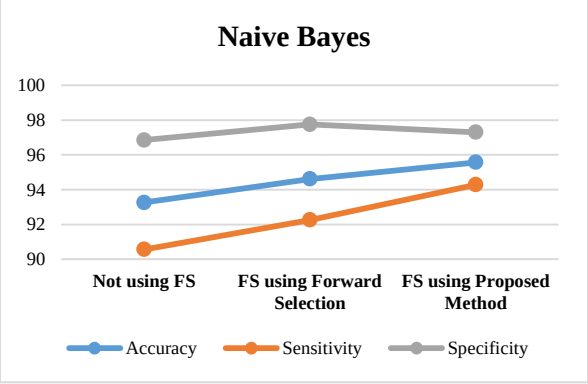

Figure 2. Performance of Naïve Bayes based on different states of feature selection

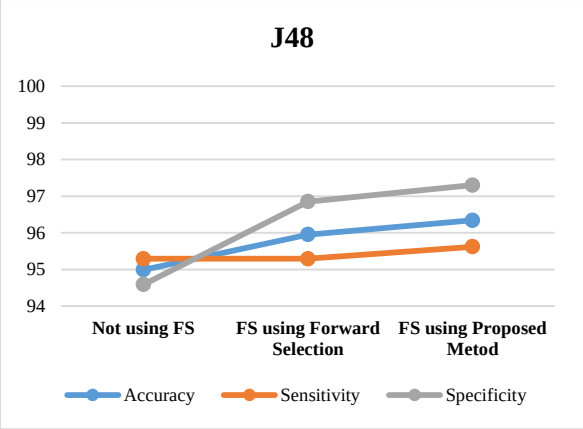

Figure 3. Performance of J48 based on different states of feature selection

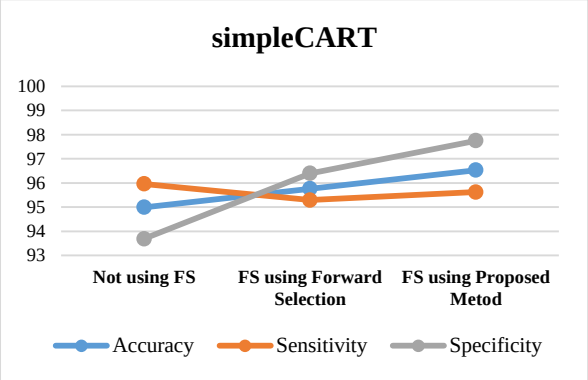

Figure 4. Performance of simpleCART based on different states of feature selection

Figure 1 shows the results of AdaBoost algorithm in the mentioned 3 states. Figure 2, 3 and 4 demonstrates the results of the Naïve Bayes, J48 and simpleCART algorithms, respectively. According to the results demonstrated in figures 1, 2, 3 and 4, applying proposed hybrid feature selection method on data set was determined to lead to improvement of the accuracy of all algorithms. It is worth mentioning that although Naïve Bayes had the lowest accuracy, using proposed feature selection caused a 2% increase of the accuracy.

In general, as the Figures 1, 2, 3 and 4, revealed, applying the feature selection method has improved the sensitivity of algorithms. Higher sensitivity means the algorithm tended to predict Myocardial Infarction cases. Regarding specificity, except for Naïve Bayes, the feature selection method caused higher specificity; that is to say the tendency of algorithms to predict healthy cases is enhanced. It is notable that using the proposed feature selection method enhanced the specificity for J48 and simpleCART algorithms about 3% and 4%, respectively as well as the sensitivity for Naïve Bayes about 4%. In spite of the good performance of the models in present study, these models have not always led to the best results. In comparison to the studies reviewed, the findings of the

present study revealed a better performance of the proposed method; but despite these notable accuracy values, the reported accuracy values in [12] are better than three of algorithms in our study. The accuracies of Naïve Bayes, J48 and simpleCART algorithms, were lower than the accuracy reported in [12]. This can be due to the different types of features or even different numbers of the data, which cause different results and performances. These methods and the results can be used in predicting Myocardial Infarction. Since sensitivity means the ratio of truly labeling the positive cases, it determines how much the model in tended to predict disease cases. Also, based in the definition of specificity, mentioned earlier, it determines how much the model is tended to predict healthy instances. So, the high sensitivity and specificity of the model shows the good performance of the model in predicting healthy or Myocardial Infarction. Hence, the high accuracy resulted in these prediction methods can be considered reliable in regard with predicting Myocardial Infarction. These approaches can be used as an application, a software or a decision support system in hospitals. It can help the cardiologists in their decisions. Moreover, in the absence of cardiologists in emergency cases, it can be used by general practitioners or even nurses. Using this method, in form an application, general practitioners and nurses can predict Myocardial Infarction; so, to prevent Myocardial Infarction, they can perform the primary cares, until achieving a cardiologist.

Conclusion

The current study proposed a hybrid feature selection method including Forward Selection and Genetic Algorithm. This feature selection method along with the classification method can improve the performance of this algorithm. In the present study, AdaBoost, Naïve Bayes, J48 and simpleCART were applied to the data set. High performance of models showed that using the combination of the hybrid feature selection and classification algorithms can be considered as reliable approaches in prediction of Myocardial Infarction before occurrence. SimpleCART and J48 algorithms have achieved the highest accuracies, 96.53% and 96.34%, respectively. AdaBoost with highest sensitivity is regarded as an algorithm with the highest tendency for predicting Myocardial Infarction cases, though Naïve Bayes, J48 and simpleCART with higher specificity tended to predict healthy cases. In general, it is concluded the proposed method, is almost considered as a reliable approach for predicting Myocardial Infarction. In other words, high performance of the algorithms can be mentioned as the benefit of our method. Therefore, the results of this study can be applied for early prediction of Myocardial Infarction which can lead to reduction of Myocardial Infarction caused mortality as well as the costs of treatments.

Authors would employ the other classification algorithms in their future study, and assess the proposed feature selection method. Moreover, this method is recommended to be extended using the other feature selection methods and t

classification methods, which can be used for the other fields like fraud detection.

Acknowledgements

We would like to thank personnel of Shahid Madani Hospital of Khorramabad for assisting in collecting the data. We are also thankful to Dr. Maryam Kooshki, for her guidance in the medical fields.

References

1. Ahmadi A, Soori H, Mehrabi Y, et al. Incidence of acute Myocardial Infarction in Islamic Republic of Iran: a study using national registry data in 2012. Eastern Mediterranean Health Journal. 2014; 21(1):5-12.
2. Wiener C, Brown C, Hemnes A, et al. Harrison's principles of internal medicine. 19th ed. New York: McGraw-Hill Medical; 2015. 406-72.
3. Dhingra R,Shaw J and Kirshenbaum A.molecular regulation of apoptosis signaling pathway in heart. in:Preedy R. Apoptosis: Modern Insights into Disease from Molecules to Man. CRC Press; 2010;382-5.
4. Esfandiari N, Babavalian MR, Moghadam AM, et al. Knowledge discovery in medicine: Current issue and future trend. Expert Systems with Applications. 2014; 41(9):4434-63.
5. Deekshatulu BL, Chandra P. Classification of heart disease using K-nearest neighbor and genetic algorithm. Procedia Technology. 2013; 10: 85-94.
6. Kumar S, Sahoo G. Classification of Heart Disease Using Naïve Bayes and Genetic Algorithm. In: Computational Intelligence in Data Mining-Volume 2. Springer India; 2015; 269-82.
7. Tan P, Steinbach M and Kumar V. Introduction to data mining. 1st ed. Boston: Pearson Addison Wesley. 2005; 1-18.
8. Tsien CL, Fraser HS, Long WJ, Kennedy RL. Using classification tree and logistic regression methods to diagnose Myocardial Infarction. Studies in health technology and informatics. 1998; (1):493-7.
9. Conforti D, Constanzo D, Guido R. Medical decision making: A case study within the cardiology domain. Journal on Information Technology in Healthcare. 2007; 5(6):343-56.
10. Baxt WG, Shofer FS, Sites FD, et al. A neural computational aid to the diagnosis of acute Myocardial Infarction. Annals of emergency medicine. 2002; 39(4):366-73.
11. Qazi M, Fung G, Krishnan S, Bi J, Bharat Rao R, Katz AS. Automated heart abnormality detection using sparse linear classifiers. Engineering in Medicine and Biology Magazine, IEEE. 2007; 26(2):56-63.
12. Masethe HD, Masethe MA. Prediction of heart disease using classification algorithms. In: Proceedings of the World Congress on Engineering and Computer Science. 2014; 22-4.
13. Patil SB, Kumaraswamy YS. Intelligent and effective heart attack prediction system using data mining and artificial neural network. European Journal of Scientific Research. 2009; 31(4):642-56.
14. Bhaskar NA. Performance Analysis of Support Vector Machine and Neural Networks in Detection of Myocardial Infarction. Procedia Computer Science. 2015; 46:20-30.

15. Karaolis M, Moutiris JA, Papaconstantinou L, Pattichis CS. Association rule analysis for the assessment of the risk of coronary heart events. In: Engineering in Medicine and Biology Society. EMBC 2009. Annual International Conference of the IEEE. 2009 (3); 6238-41
16. Krishnaraj N, Vinothkumar MR. Heart Disease Prediction using GA and MLBPN. Heart Disease. 2014; 2(4):17-24.
17. Hachesu PR, Ahmadi M, Alizadeh S, Sadoughi F. Use of data mining techniques to determine and predict length of stay of cardiac patients. Healthcare informatics research. 2013; 19(2):121-9.
18. Rangra K, Bansal KL. Comparative study of data mining tools. International Journal of Advanced Research in Computer Science and Software Engineering. 2014; 4(6):216-23.
19. Benjamin I, Griggs RC, Wing EJ, et al. Andreoli and Carpenter's Cecil Essentials of Medicine. 8th ed. Saunders; 2015. 117-45.
20. Han J, Kamber M, Pei J. Data mining: concepts and techniques. 3rd ed. Elsevier. 2012; 3-120.
21. Amma NB. Cardiovascular disease prediction system using genetic algorithm and neural network. In: Computing, Communication and Applications (ICCCA), 2012 International Conference on 2012(22); 1-5.
22. Fayyad R. Data mining and knowledge discovery in databases. Communications of the ACM. 1996; 39(11):24-6.
23. Pacharne M, Nayak VS. Feature Selection Using Various Hybrid Algorithms for Speech Recognition. Computational Intelligence and Information Technology. 2011:652-6.
24. 9. Li P, Wang Y, Tian Y, et al. An Automatic User-adapted Physical Activity Classification Method Using Smartphones. IEEE Transactions on Biomedical Engineering. 2016; 1-1.
25. Alizadehsani R, Habibi J, Bahadorian B, et al. Diagnosis of Coronary Arteries Stenosis Using Data Mining. Journal of Medical Signals and Sensors. 2012; 2(3):153-9.
26. Hand DJ, Yu K. Idiot's Bayes—not so stupid after all?. International statistical review. 2001; 69(3):385-98.
27. Dunham MH. Data mining: Introductory and advanced topics. Pearson Education India; 2006.
28. Rogulj D, Konjevoda P, Milić M, et al. Fatty liver index as an indicator of metabolic syndrome. Clinical biochemistry. 2012; 45(1):68-71.
29. Onan A. A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. Expert Systems with Applications. 2015; 42(20):6844-52.
30. Heydari M, Teimouri M, Heshmati Z, et al. Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran. International Journal of Diabetes in Developing Countries. 2015:1-7.