Academic Platform Journal of Engineering and Science

journal homepage: http://apjes.com/

A Study on Performance Improvement of Heart Disease Prediction by Attribute Selection Methods

*[1]Kemal Akyol, [2]Ümit Atila

[1] Kastamonu University, Computer Engineering Department, Turkey, kakyol@kastamonu.edu.tr,

[2] Computer Engineering Department, Karabuk University, Turkey, umitatila@karabuk.edu.tr,

Research Paper Arrival Date: 20.12.2018 Accepted Date: 17.03.2019

Abstract

Heart pumps blood for all tissues of the body. The deteriorate of this organ causes a severe illness, disability and death since cardiovascular diseases involve the diseases that related to heart and circulation system. Determination of the significance of factors affecting this disease is of great importance for early prevention and treatment of this disease. In this study, firstly, the best attributes set for Single Proton Emission Computed Tomography (SPECT) and Statlog Heart Disease (STATLOG) datasets were detected by using feature selection methods named RFECV (Recursive Feature Elimination with cross-validation) and SS (Stability Selection). Secondly, GBM (Gradient Boosted Machines), NB (Naive Bayes) and RF (Random Forest) algorithms were implemented with original datasets and with datasets having selected attributes by RFECV and SS methods and their performances were compared for each dataset. The experimental results showed that maximum performance increases were obtained on SPECT dataset by 14.81% when GBM algorithm was applied using attributes provided by RFECV method and on STATLOG dataset by 6.18% when GBM algorithm was applied using attributes provided by RFECV method. On the other hand, best accuracies were obtained by NB algorithm when applied using attributes of SPECT dataset provided by RFECV method and using attributes of STATLOG dataset provided by SS method. The results showed that medical decision support systems which can make more accurate predictions could be developed using enhanced machine learning methods by RFECV and SS methods and this can be helpful in selecting the treatment method for the experts in the field.

Keywords: Cardiovascular disease, attribute importance, attribute selection, stability selection, recursive feature selection.

1. INTRODUCTION

The heart is the organ which pumps blood to all tissues of the body. If the heart fails, vital organs degenerate. Moreover, death is inevitable if the heart stops working at all [1]. Cardiovascular diseases (CDs) cover related to heart and circulation system diseases including coronary heart disease, angina, heart attack, congenital heart disease and stroke [2]. These bring on a severe disease, disability and death [3]. Expert medical decision support systems are developed to improve the ability of the field-specialists about the disease [1]. Determination of the significance of factors affecting the disease is of value importance for early preclusion and treatment of this disease.

Diverse studies have highlighted this subject in literature. These studies have been carried out using various datasets. Some of these are as follows: Das et al. proposed an integrated system of software solutions SAS 9.1.3 for heart disease diagnose. By combining the predicted values obtained from multiple predecessor models, Sensitivity of 80.95%, Specificity of 95.91% and Accuracy of 89.01% were obtained in experiments performed on the data taken from Cleveland heart disease dataset [1]. Ciecholewski discussed the performances of Support Vector Machine (SVM) and CLIP3 which is a combination of the decision tree and rule induction algorithm on the SPECT images [4]. Ebenezer et al. modeled an intelligent system by using feed forward multilayer perceptron and SVM, and they obtained accuracies of 85% and 87.5% respectively by using these algorithms [5]. Yang and Garibaldi introduced an information extraction system for automatically identifying risk factors for heart disease. They achieved promising performance on the test data with an overall micro-averaged 0.915 of F-measure [6]. Kurgan et al. described a computerized process of myocardial perfusion diagnosis from cardiac SPECT images using six-step knowledge discovery process. A set of features were extracted from these images, and then rules were implemented by utilizing the machine learning and heuristic approaches in their studies [7]. Padmavathi et al. analyzed the performance of predictive model on different medical datasets. The datasets which include heart datasets, cancer and diabetes datasets are of binary class and each dataset has a different number of attributes. SVM classifier produced better percentage of accuracy in classification according to their studies [8]. Rafaie et al. analysed the SPECT dataset containing the records of 267 patients with a variety of heart diseases using a combined the Rough Sets and neural network approach. The feature space was reduced from 22 to 10 essential

* Corresponding Author: Faculty of Engineering, Department of Computer Engineering, Kastamonu University, Kastamonu, Turkey, kakyol@kastamonu.edu.tr +90 366 280 2978

Doi: 10.21541/apjes.500131

features by using the Rough Sets analysis. The reduced feature set was tested in order to measure of classification accuracy by using a neural network approach [9]. Prasad and Biswas proposed two models, Binary Particle Swarm Optimization-SVM and Novel Particle Swarm Optimization-SVM, for classification of several datasets. According to their studies, the accuracy on test data is 84.64% for SPECT dataset by utilizing radial basis function-SVM classifier [10]. Vanisree and Singaraju presented a decision support system for Congenital Heart Disease diagnosis occurring in the baby's heart during pregnancy. Classification accuracy of 90% was achieved by using the multi-layer feed forward neural network (MLP) [11]. Nalluri et al. proposed hybrid intelligent systems in order to diagnose ailments on benchmark datasets. SVM and multilayer perceptron algorithms were optimized using individual classifier parameters in order to evaluate the efficiency of the models [12]. Durairaj and Sivagowry implemented feature reduction using Particle Swarm Optimization (PSO) algorithm and Ant Colony Optimization (ACO) algorithm. PSO was better than ACO in terms of accuracy [13]. Setiawan et al. developed a rule selection method for filtering large number of extracted rules from CAD dataset. The method has better quality compared to previous rule selection methods for this disease [14]. In another study, Setiawan developed a decision support system including three stages: rule generation, rule selection and rule fuzzification. Furthermore, the reduction of attributes by using Rough Set theory was proposed and applied to select the most important rules [15]. Raghu et al. developed a decision support system for heart disease prediction using medical situations such blood pressure and blood sugar. Also, the author implemented web-based questionnaire application [16]. Vijayashree and Narayanalyengar examined the decision support systems supported by data mining and hybrid intelligent techniques for the prediction and diagnosis of heart disease [17].

The main aim of this study is two-fold. First is to detect the importance of attributes for the disease on two datasets. Second is to demonstrate the performance improvement of probability based and tree based machine learning algorithms by utilizing feature selection methods. In this context, machine learning algorithms are implemented on best attributes datasets and their performances are discussed in the paper.

The rest of the paper is organized as follows. Section 2 presents the materials and methods. Section 3 gives experimental study and results. Finally, the paper ends with conclusions in section 4.

2. MATERIALS AND METHODS

2.1. Datasets

The performances of machine learning algorithms are evaluated on SPECT dataset [18] which consists of cardiac disease data including 22 partial diagnosis features (F1-F22) and 267 instances. SPECT dataset consist of binary features describing the original SPECT images. Each instance is classified as ‘normal’ and ‘abnormal’. STATLOG dataset [19], which consists of heart disease data including 13 features and 270 instances. Absence or presence of heart disease for each instance is categorized as 1 or 2 respectively.

2.2. Feature Selection and Machine Learning

Learning is the knowledge acquisition process. The knowledge obtained from the real world is improved by utilizing machine learning algorithms [20]. There are several machine learning algorithms proposed in the literature. These algorithms use datasets as input data for learning. High-dimensional data analysis is a difficult process in machine learning and data mining. Feature selection presents an effective solution for this problem by removing irrelevant and redundant data. This approach reduces computation time, improves learning accuracy and facilities better understanding for the model developed [21]. It aims to select a subset of features from all features [22]. Many studies have highlighted on this subject. For example, Liu et al. described the importance of feature selection, and reviewed its developments [23]. Zhou et al. implemented an online feature selection system [24].

In this study, Recursive Feature Elimination with cross-validation (RFECV) and Stability Selection (SS) methods were utilized for improving performance of tree-based and probability-based machine learning algorithms.

The selected learning algorithms for this purpose were Gradient Boosted Machines (GBM), Random Forest (RF) and Naive Bayes (NB).

RFE method selects best attributes using an iterative procedure as follows [25]:

a) Classifier training
b) Calculating the ranking criterion for all features
c) Eliminating the feature with lowest ranking

RFECV fit the RFE and automatically tune the number of selected features.

The SS method provides information about the attributes of the output variable. The method perturbs the dataset many times. A small subset of features in dataset is selected with the combination of ‘The Least Absolute Shrinkage and Selection Operator (Lasso)’ and its successive regressions to explain the output variable [26]. Randomized Lasso method [27] can consistently select variables even if the required constraints for consistency of the original Lasso method are violated [26].

GBM fits new models consecutively during learning to estimate the response variables more accurately. The main idea behind this algorithm is to construct new base-learners to have maximum correlation with the negative gradient of the loss function associated with the entire ensemble [28].

RF introduced by Breiman [29] is an ensemble learning algorithm that is created by random decision trees. Main difference from decision tree is that RF searches for the best feature among the random subsets of features while decision tree searches for the most important feature when splitting a node. Therefore, this provides a wide diversity that results with a better model.

NB classifier is based on Bayes' theorem in which every pair of features been classified is independent of each other [30]. It uses the probability theory in order to find the most possible classifications [31]. This algorithm is suited when the dimensionality of the input is high.

3. EXPERIMENTS AND RESULTS

Firstly, the datasets were divided as training and test set with 70% and 30% rates respectively and then RFECV (cross validation value=5) and SS methods were implemented to original datasets in order to determine best attributes. The experiments are carried out with Python 3.6 programming language by utilizing the *sklearn* library.

RFE method returns 'True' or 'False' values to indicate whether a feature is important or non-important respectively based on mathematical calculations. Besides, SS method gives result considering a threshold value which is taken 0.25 as default. Attributes with the importance value greater than this threshold value are accepted as important. In this context, the numbers of attributes chosen by RFECV and SS methods were presented in Figure 1.

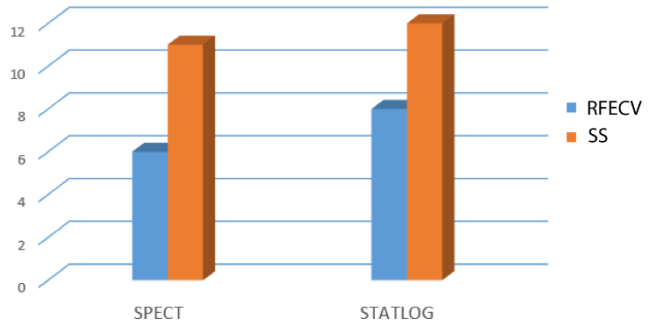

Figure 1. Statistical information about the numbers of selected attributes.

Also, the best attributes derived by SS and RFECV methods which compute the importance of attributes for the SPECT and STATLOG datasets demonstrated in Table 1 and 2. According to these results;

a) *Attributes;* F1, F7, F8, F11, F14 and F22 are found important by RFECV method for SPECT dataset. So, the optimal number of important attributes is 6.
b) *Attributes;* F1, F2, F6, F7, F11, F13, F14, F16, F17, F19 and F22 are found important by the SS method for SPECT dataset. So, the optimal number of important attributes is 11.
c) Attributes; *sex, chest paint type, fasting blood sugar, resting electrocardiographic results, exercise induced angina, old peak, the number of major vessels and thal* are found important by RFECV method for STATLOG dataset. So, the optimal number of important attributes is 8.
d) Attributes; *sex, chest paint type, resting blood pressure, serum cholestoral, fasting blood sugar, resting electrocardiographic results, maximum heart rate, exercise induced angina, old peak, the slope of the peak exercise ST segment, the number of major vessels and thal* are found important by SS method for STATLOG dataset. So, the optimal number of important attributes is 12.

Table 1. Information of attribute importance for SPECT dataset obtained by feature selection methods.

	RFECV*	**SS****
Attribute	**Importance**	**Importance value**
F1	**True**	**0.64**
F2	False	**0.26**
F3	False	0.19
F4	False	0.22
F5	False	0.04
F6	False	**0.51**
F7	**True**	**0.76**
F8	**True**	0.05
F9	False	0.09
F10	False	0.06
F11	**True**	**0.67**
F12	False	0.09
F13	False	**0.64**
F14	**True**	**0.99**
F15	False	0.03
F16	False	**0.3**
F17	False	**0.58**
F18	False	0.24
F19	False	**0.38**
F20	False	0.02
F21	False	0.18
F22	**True**	**0.93**

Table 2. Information of attribute importance for STATLOG dataset obtained by feature selection methods.

	RFECV*	**SS****
Attribute	**Importance**	**Importance value**
age	False	0.125
sex	**True**	**1.0**
chest paint type	**True**	**0.995**
resting blood pressure	False	**0.755**
serum cholestoral	False	**0.765**
fasting blood sugar	**True**	**0.57**
resting electrocardiographic results	**True**	**0.42**
maximum heart rate	False	**0.905**
exercise induced angina	**True**	**1.0**
old peak	**True**	**0.97**
the slope of the peak exercise ST segment	False	**1.0**
the number of major vessels	**True**	**1.0**
thal; fixed defect; reversible defect	**True**	**1.0**

* 'True' indicates important attribute
** Value greater than 0.25 indicates important attribute

In this study, accuracy (Acc), sensitivity (Sen) and specificity (Spe) metrics were used for the evaluation of the performances of these algorithms. The following equations (1-3) define these metrics respectively [32]:

$$Acc = (TP + TN)/(TP + FP + TN + FN) \quad (1)$$

$$Sen = TP/(TP+FN) \quad (2)$$

$$Spe = TN/(TN + FP) \quad (3)$$

where TP is the number of patients correctly classified as having heart disease, TN is the number of patients correctly classified as not having heart disease, FP is the number of patients incorrectly classified as having heart disease and FN is the number of patients incorrectly classified as not having heart disease.

Experiments were performed on both original dataset and best attributes sets to discuss effects of the attribute selection methods on learning algorithms. Results presented in Table 3 indicate that;

For SPECT dataset, RFECV method increased the accuracy of GBM from 56.79% to 71.6%, accuracy of RF was increased from 72.84% to 76.54% and accuracy of NB was increased from 71.6% to 77.78%. On the other hand, while SS method was able to increase the accuracy of GBM to 59.26%, this method could not improve the accuracy of RF and caused a decrease on accuracy of NB from 71.6% to 70.37%. Therefore, it can be clearly seen that the RFECV outperforms the SS for SPECT dataset.

For STATLOG dataset, while RFECV method increased the accuracy of GBM from 76.54% to 82.72%, accuracy of RF from 80.25% to 82.72% and it decreased the accuracy of NB from 85.19% to 81.48%. Besides, SS method increased the accuracy of GBM to 79.01%, accuracy of RF to 83.95% and accuracy of NB to 86.42%. As understood from results, while RFECV method was more successful than SS method when applied with GBM algorithm, success of the method is less than SS when applied with RF. On the other hand, while accuracy of NB was decreased when used with RFECV method, it was increased when applied with SS method.

For SPECT dataset, it can be said that RFECV method was more successful than SS method and best couple was RFECV method with NB algorithm. On the other hand, maximum increase on accuracy was achieved as 14.81% when GBM algorithm was applied with RFECV method. Besides, for STATLOG dataset, maximum increase on accuracy was obtained as 6.18% when RFECV method was applied with GBM algorithm and most successful couple was SS method with NB algorithm.

Our proposed method achieved accuracy values 77.78% and 86.42% on the publicly available datasets SPECT and STATLOG respectively. The performance of this study is compared with existing methods as shown in Table 4. Results of previous studies summarized in this table show that hybrid use of metaheuristic optimization methods for feature selection with machine learning algorithms such as SVM and MLP give better performance enhancement than using SS or RFECV methods.

Table 3. The performance results of learning algorithms.

	Machine Learning Algorithms								
	GBM			**RF**			**NB**		
SPECT Dataset	Acc	Sen	Spe	Acc	Sen	Spe	Acc	Sen	Spe
Original Dataset	56.79%	51.43%	60.87%	72.84%	54.29%	86.96%	71.6%	48.57%	89.13%
RFECV-SPECT Dataset	71.6%	48.57%	89.13%	76.54%	62.86%	86.96%	**77.78%**	62.86%	89.13%
SS-SPECT Dataset	59.26%	62.86%	56.52%	72.78%	57.14%	84.78%	70.37%	37.14%	95.65%
STATLOG Dataset									
Original Dataset	76.54%	56.25%	89.80%	80.25%	65.63%	89.8%	85.19%	71.88%	93.88%
RFECV-STATLOG	82.72%	65.63%	93.88%	82.72%	59.38%	97.96%	81.48%	68.75%	89.80%
SS-STATLOG	79.01%	62.5%	89.8%	83.95%	65.63%	95.92%	**86.42%**	71.88%	95.92%

Table 4. The comparison of the studies.

	Acc %	Sen %	Spe %
SPECT Dataset			
Rafaie et al. [9]	93.0	95.0	85.0
Prasad and Biswas [10]	84.64	-	-
Nalluri et al. [12] - parameter optimized MLP base	89.51	91.93	77.27
Proposed Study - SS and NB base	**77.78**	**62.86**	**89.13**
STATLOG Dataset			
Nalluri et al. [12] - parameter optimized MLP base	90.74	92.16	89.88
Ebenezer et al. [5]	87.5	84.44	89.8
Proposed Study - RFECV and NB base	**86.42**	**71.88**	**95.92**

4. CONCLUSION AND DISCUSSION

The determination of importance of attributes for any disease play an important role in the detection and treatment of the disease. And also, it helps to field specialists' examinations. CDs cause to death of many people. Any risk factor affecting the CDs is of great importance for early hindrance and treatment of this disease. In this context, the importance of attributes for this disease are investigated by utilizing RFECV and SS methods, and the best attributes sets are obtained. Then, the machine learning models are carried out. Experimental results showed that to achieve best results by our proposed methods on heart disease prediction, it was better to use NB algorithm with RFECV method on SPECT dataset and with SS method on STATLOG dataset. It was also observed that NB algorithm was affected badly with SS method on SPECT dataset and with RFECV method on STATLOG dataset. On the other hand, maximum accuracy increases were obtained with GBM algorithm when used with RFECV method on both SPECT and STATLOG datasets. It was observed in the results that using SS or RFE method for feature selection on SPECT and STATLOG datasets showed lower performance than other methods given in Table 4. Although the performances of SS and RFE are lower than other methods applied to the problem, the identification of important attributes can at least be a guide in the field specialists' examinations. Small biomedical datasets such as SPECT and STATLOG involve too much noise and many local minima. This situation makes those datasets resistant to classical machine learning algorithms and this causes a decrease on performance of these algorithms. Better solution to the problem could be achieved using Rough Set theory which can be considered successful on noise reduction and increasing the performance of machine learning algorithms. On the other hand, hybrid use of metaheuristic methods for optimizing the parameters of machine learning algorithms could be also preferred for having better results.

ACKNOWLEDGMENT

The authors thank the UCI Machine Learning Repository for providing publically available SPECT and STATLOG datasets.

REFERENCES

[1]. R. Das, I. Turkoglu, A. Sengur, "Effective Diagnosis of Heart Disease through Neural Network Ensemble", Expert Syst Appl, vol. 36, no 4, pp. 7675- 7680, May 2009.

[2]. Coronary heart disease, URL: https://www.bhf.org.uk/heart-health/conditions/coronary-heart-disease (accessed time; July, 1, 2018).

[3]. J. S. Sonawane, D. R. Patil, V. S. Thakare, "Survey on Decision Support System for Heart Disease",

International Journal of Advancements in Technology, vol. 4, no 1, pp. 89-96, March, 2013.
[4]. M. Ciecholewski "Support Vector Machine Approach to Cardiac SPECT Diagnosis", In: J.K. Aggarwal, R.P. Barneva, V.E. Brimkov, K.N. Koroutchev, E.R. Korutcheva (eds) Combinatorial Image Analysis. IWCIA 2011. Lecture Notes in Computer Science, vol. 6636, pp. 432-443, 2011.
[5]. O. Ebenezer, K. O. Oyebade, A. Khashman, "Heart Diseases Diagnosis Using Neural Networks Arbitration", J. Intelligent Systems and Applications, vol. 7, no 12, pp. 72-79, 2015.
[6]. H. Yang, J. M. Garibaldi, "A hybrid model for automatic identification of risk factors for heart disease", J Biomed Inform, vol. 58, pp. 171-82, 2015.
[7]. L. A. Kurgan, K. J. Cios, R. Tadeusiewicz, M. Ogiela, L.S. Goodenday, "Knowledge discovery approach to automated cardiac SPECT diagnosis", Artif Intell Med, vol. 23, no 2, pp.149-169, 2001.
[8]. J. Padmavathi, L. Heena, S. Fathima "Effectiveness of Support Vector Machines in Medical Data mining", Journal of Communications Software and Systems, vol. 11, no 1, pp.25-30, 2015.
[9]. S. El Rafaie, M. S. Abdel-Badeeh, K. Revett, "On the use of SPECT imaging datasets for automated classification of ventricular heart disease", Informatics and Systems, 8th International Conference on Cairo, Egypt, pp. 14-16 May 2012.
[10]. Y. Prasad, K. K. Biswas, "PSO - SVM Based Classifiers: A Comparative Approach", In: Ranka S. et al. (eds) Contemporary Computing. IC3 2010. Communications in Computer and Information Science, vol. 94, pp. 241-252, 2010.
[11]. K. Vanisree, J. Singaraju, "Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks", International Journal of Computer Applications, vol. 19, no 6, pp.6-12, 2011.
[12]. M. R. Nalluri, K. Kannan, M. Manisha, D.S. Roy, "Hybrid Disease Diagnosis Using Multiobjective Optimization with Evolutionary Parameter Optimization", J Healthc Eng, vol. 5907264, pp. 1-27, 2017.
[13]. M. Durairaj, S. Sivagowry, "Feature Diminution by Using Particle Swarm Optimization for Envisaging the Heart Syndrome", International Journal of Information Technology and Computer Science, vol. 2, pp. 35-43, 2015.
[14]. N. A. Setiawan, P. A. Venkatachalam, M. H. Ahmad Fadzil, "Rule Selection for Coronary Artery Disease Diagnosis Based on Rough Set", International Journal of Recent Trends in Engineering, vol. 2, no 5, pp. 198-202, 2009.
[15]. N. A. Setiawan, "Fuzzy Decision Support System for Coronary Artery Disease Diagnosis Based on Rough Set Theory", International Journal of Rough Sets and Data Analysis, vol. 1, no 1, pp. 65-80, 2014.
[16]. D. Raghu, T. Srikanth, R. Jacub, "Probability Based Heart Disease Prediction using Data Mining Technique", International Journal of Computer Science and Technology, vol. 2, no 4, pp. 66-68, 2011.
[17]. J. Vijayashree, N. C. S. Narayanalyengar, "Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques: A Review", International Journal of Bio-Science and Bio-Technology, vol. 8, no 4, pp.139-148, 2016.
[18]. Spect dataset, URL: https://archive.ics.uci.edu/ml/datasets/spect+heart (accessed time; June, 01, 2018)
[19]. Statlog dataset, URL: http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29 (accessed time; June, 01, 2018)
[20]. P. Ivens, A. Paulo, C. Donald, "The use of machine learning algorithms in recommender systems: A systematic review", Expert Syst Appl, vol. 97, pp. 205-227, 2018.
[21]. J. Cai, J. Luo, S. Wang, S. Yang, "Feature selection in machine learning: a new perspective. Neurocomputing", vol. 300, pp. 70-79, 2018.
[22]. L. Huan, H. Motoda, "Computational Methods of Feature Selection", Chapman & Hall/Crc Data Mining and Knowledge Discovery Series, 2007.
[23]. H. Liu, H. Motoda, R. Setiono, Z. Zhao, "Feature Selection: An Everlasting Frontier in Data Mining", Proceedings of the Fourth International Workshop on Feature Selection in Data Mining, pp. 4-13, 2010.
[24]. P. Zhou, X. Hu, P. Li, X. Wu, "Online feature selection for high-dimensional class-imbalanced data", Knowledge-Based Systems, vol. 136, pp. 187–199, 2017.
[25]. A. Filali, C. Jlassi, N. Arous, "Recursive Feature Elimination with Ensemble Learning Using SOM Variants", International Journal of Computational Intelligence and Applications, vol. 16, no 1, pp.1-25, 2017.
[26]. F. Mordelet, J. Horton, A. J. Hartemink, B. E. Engelhardt, R. Gordân, "Stability selection for regression-based models of transcription factor–DNA binding specificity", Bioinformatics, 29:i117–i125, 2013.
[27]. N. Meinshausen, P. Bühlmann, "Stability selection", J. R. Statist Soc. B, vol. 72, no. 4, pp.417–473, 2010.
[28]. J. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine", The Annals of Statistics, vol. 29, no. 5, 2001.
[29]. L. Breiman, "Random forests", Mach Learn, vol. 45, no 1, pp.5-32, 2011.
[30]. J. Han, M. Kamber, J. Pei, "Data Mining Concepts and Techniques", 3rd ed, Waltham, USA: Elsevier, 2012.
[31]. M. Bramer, "Principles of Data Mining, Undergraduate Topics in Computer Science", 2nd ed. London: Springer, 2013.
[32]. S. A. Shaikh, "Measures derived from a 2x2 table for an accuracy of a diagnostic test", J Biom Biostat, vol.2, no 128, pp.1-4, 2011.