A System for Diagnosis of Coronary Artery Disease based on Neural Networks and Machine Learning Algorithms

Najmeh Samadiani
Kosar University of Bojnord
Bojnord, Iran

Zeinab Hassani
Kosar University of Bojnord
Bojnord, Iran

ABSTRACT

Today, computer aided systems play an important role in various fields of medical science such as diagnosis and treatment of diseases; therefore, selected tools should minimize error and maximize the confidence. In this study, considering the importance of cardiovascular disease in the world, the coronary artery disease is diagnosed by neural networks and machine learning algorithms. The proposed system employs three types of artificial neural networks, decision tree and Adaboost algorithm to distinguish people who suffer from heart disease and the healthy individuals using Cleveland's dataset. Among these algorithms, the multilayer perceptron neural network has the best performance and is able to predict coronary artery disease with the accuracy, sensitivity and specificity of 94.53%, 86.77%, and 99.39%, respectively. The superiority of the proposed system is obvious comparing to other existing studies because it diagnoses the disease with higher accuracy, sensitivity and more reliability.

General Terms

Neural Networks, Decision Trees, Heart Disease.

Keywords

Multi-layer Perceptron Neural Network (MLP), Self-Organizing Map Neural Network (SOM), Decision Tree, Adaboost, Coronary Artery Disease, Heart Disease.

1. INTRODUCTION

Coronary artery disease (CAD) as the most common type of heart disease is the cause of a growing incidence of death in both women and men [1]. This disease affects 17 million people worldwide and is the leading cause of death among other cardiovascular diseases [2] as 11.1 million deaths in 2020 are estimated by the World Health Organization (WHO). In coronary artery disease the coronary arteries of the heart are clogged by fat deposits (atherosclerosis). This condition limits the blood and oxygen needed in the heart especially when doing physical activity. For many people the first sign of heart dysfunction is a heart attack that occurs when the blood clot in the coronary arteries blocks the flow of blood into a part of the heart muscle [3].

There are several diagnostic methods for coronary artery disease such as Exercise Tolerance Test (ETT), electrocardiography (ECG), angiography, or cardiac catheterization. But patient pains and inadequate accuracy in diagnosis limit using all these methods; therefore, doctors are encouraged to use computer aided systems [4]. Computer aided methods which extract effective features and use them in the classifications for the early detection of the diseases, overcome these problems.

In general, some risk factors are responsible for coronary artery disease. Reviewing various sources marks the risk factors for coronary artery disease are: smoking, high blood pressure, high fat (high cholesterol, high triglycerides, high LDL and low HDL), diabetes, physical inactivity, obesity, age, gender and family history [5]. Based on these risk factors, various algorithms such as decision tree [6-7], linear support vector machine [4-8] and various neural networks are developed and proposed for detecting and preventing of coronary artery disease.

The artificial neural network is a useful and effective classification and prediction methods in various fields of science. In medicine, it is used to categorize people into healthy and patient classes or predict the patient's state based on risk factors. In [9], a model is proposed for the diagnosis of coronary artery disease based on the neural network using the genetic factors of the disease. The neural network included two hidden layers has been able to recognize the genetic data of 487 individuals into the healthy and patient classes with an accuracy of 64 to 94% depending on the input factors. Authors of [10] have presented a multi-layered neural network included a hidden layer, a back propagation algorithm and LM training technique to predict the incidence of coronary artery angiography. This method can diagnose 88 patients with 95.5% accuracy. A coronary artery disease diagnostic system is proposed by averaging of the performance of several different neural networks [11]. It has obtained accuracy of 89.01%. Sajja et al. have proposed a MLP network model with an accuracy of 91.75% [12]. Kurt et al. have used 1245 patient records using logistic regression, decision tree and neural network techniques and an accuracy of 78.7% is obtained by MLP network as the best performance [13].

Various hybrid methods have been developed to address the problems of detecting coronary artery disease. A hybrid approach for diagnosis of coronary artery disease is presented in [14]. This method can increase the learning power and the neural network performance by 10% through optimizing weights by genetic algorithm. The proposed network has got 88.25% accuracy in the diagnosis of coronary artery disease. Akila and Chandramath have proposed a hybrid method for classifying healthy and coronary artery disease patients using the physical and biological factors [15]. They combined decision tree and neural network to improve the accuracy. In [16], a system has been proposed using the artificial intelligence recognition system and k nearest neighbor (KNN). It can diagnose coronary artery disease 87% accuracy. The authors of [17] have used exercise stress test and SVM to achieve an accuracy of 79.22% in diagnosing coronary artery disease. The authors of [18] classified the Cleveland dataset with accuracy of 77.55% and 78.54% into two healthy and patient groups. They used fast and C4.5 decision tree algorithms.

Performing invasive diagnostic methods, such as angiography cause many risks while data mining techniques gain great success and can effectively predict the disease without any danger. Therefore, we decided to propose a model for the diagnosis of coronary artery disease by intelligent algorithms. First, a diagnostic model is proposed for coronary artery disease using MLP neural network, probabilistic neural network (PNN) and self-organization map (SOM). Then least important features are removed to improve performance by decision tree and the coronary artery disease is predicted. Finally, after removing the outliers, Adaboost algorithm classifies data and the patient and healthy groups are recognized with high accuracy.

The Data set is introduced in the second section. In Section 3, the methodology is described in details and the results of applying the proposed method to the data set are reported in Section 4. Section 5 concludes and summarizes the paper.

2. DATA AND PREPROCESSING

In this paper, University of California heart database is applied. These data are collected by the Hungarian Cardiac Disease Center and Cleveland Clinical Data [19]. They include 76 features which only 13 of them are used as input features in published experiments as mentioned in [19]. Also, the output is shown by one feature. The dataset features are presented in Table 1. Since a binary problem is discussed here, the output has two values "0" for healthy people and "1" for a heart patient.

The non-binary features are normalized to the interval [0, 1] to improve the results and control the range. The normalization formula is as follows:

$$x = (x - x_{MIN}) / (x_{MAX} - x_{MIN}) \qquad (1)$$

Where x_{MIN} and x_{MAX} are the minimum and maximum values of each feature.

Table 1. Description of the selected features

Feature	Value
Age	A number, age in years
Maximum heart rate	A number
Resting blood pressure in mm Hg	A number
Fasting blood sugar	1: true, 0: false
Exercise induced angina	1: yes, 0:no
The slope of the peak exercise ST segment	1: up sloping, 2: flat, 3: down sloping
Scan thallium	3: normal, 6: fixed defect, 7: reversible defect
Sex	1: female, 0: male
Chest pain type	1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic
Serum cholesterol in mg/dl	A number
Resting electrocardiographic results	A number: Abnormal St-T wave
ST depression induced by exercise relative to rest	A number
Number of major vessels colored by fluoroscopy	A number
Output	0: healthy person, 1: patient person

3. METHODOLOGY

We aim to diagnosis the heart disease in this paper. The proposed method consists of two separated steps: detecting unhealthy people by neural networks in the first stage and distinguishing them by decision tree and Adaboost as the second phase. Firstly, we explain the used neural networks and the corresponding proposed structure of them in this Section. Secondly, the process of removing the unnecessary features and outliers is described, and we detect the heart disease by applying proposed decision tree and Adaboost.

3.1 Neural Networks

In this study three types of artificial neural networks are used to distinguish the patient and healthy person using the introduced data. The artificial neural network is a method for data processing inspired by biological nervous systems. It is formed of a set of neurons as the processing units and able to model complex systems, relationships and nonlinear functions [20]. The connection between neurons with adjustable weights is dependent on the conditions controlling the problem. The neurons of each layer are connected to the next layers' neurons with different weights in which store the information. Neural network implementation consists of three stages: providing training samples, training and testing phase. The learning algorithms used in the training phase are categorized to supervised and unsupervised learning [21]. In this study, we applied two types of supervised learners and an unsupervised one. They are as follows: multi-layer perceptron network (MLP), probabilistic neural network, and self-organization map.

3.1.1 Multi-layer neural network (MLP)

The most common neural networks are recursive networks. A model of recursive networks is the multilayer perceptron network which maps input to output data by adjusting the weights of layers. A multi-layer perceptron network is trained by back propagation algorithm [22]. It also has at least three layers of neurons (input, hidden, and output layers) [20].

In this paper, the designed MLP network has three layers included 13 neurons in input, 16 neurons in hidden and one neuron in output layer. The output neuron corresponds to the existence of coronary artery disease. The number of neurons in hidden layer is obtained by pruning approaches [23]. Figure 1a shows the structure of the designed network.

3.1.2 Probabilistic neural network (PNN)

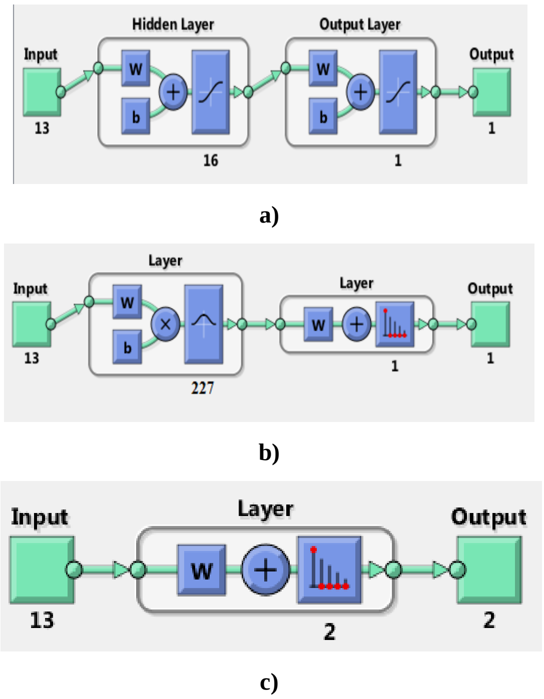
The probabilistic neural network, one of the radial basis function networks, minimizes the expected risk of classification based on optimal decision theory in nonlinear calculations. The probabilistic neural network consists of three neurons layers: an input, a hidden of the radial basis layer, and a competitive output layer. The hidden layer uses the Gaussian function and the number of neurons in this layer is equal to the number of training data [24-25]. In classification problems, it has some advantages included more computing speed than a recursive network, lack of sensitivity to the outliers, high accuracy and simple training [26].

The designed PNN network contains 13 and 227 neurons in input and hidden layer, respectively since we used 75% of data as training set and 25% of them as test set. There is a neuron in output layer corresponding to the existence of disease. The structure of this network is shown in Figure 1b.

3.1.3 Self-Organizing Map network (SOM)

In a self-organizing map network that is sometimes referred to as Self-Organization Feature Map (SOFM), the processor units are settled into neurons of multi-dimensional network. Units are organized in a competitive learning process adaptable to input patterns. One unit just wins at when a competition stage is finished. Winner weights are changed differently relative to the weights of other units [27].

The records are clustered by SOM network. The network is designed with a two-dimensional structure, consisted of 13 features as its input neurons. Considering the goal of separating healthy people from patients, the number of output neurons is 2 clusters, one for people suffering from heart disease, and another for healthy people. Figure 1c shows the structure of SOM network. When training of network is finished, the weight of each neuron in output layer represents one of the healthy or patient clusters. By applying test samples to the network, the nearest weight to that sample is selected and data clustering is performed accordingly.



a)

b)

c)

Fig 1: a) MLP network included 13 neurons in input, 16 neurons in hidden and one neuron in output layer; b) PNN which contains 13 neurons in input, 227 neurons in hidden and one neuron in output layer; c) SOM network.

3.2 Decision tree

The decision tree is one of the most widely used data mining algorithms. The decision tree is a learning algorithm based on training data and the advantages are easiness, clarity and the ability to extract rules. A decision tree is a classifier that is described as a recursive part of the sample space [28]. The learning system of a common decision tree obeys a top-down strategy where a simple tree is built but not definitely the simplest one. Each inner or non-leaf node is characterized by a feature. There are several branches from each internal node which are equal to number of possible answers. Each branch shows one value of answer. Also, the leaves represent a class or a set of possible solutions [29].

In this paper, we used decision tree to classify data included 13 features. Moreover, the decision tree was applied for determining importance of features in order to improve the accuracy of classification.

3.3 Adaboost

Group algorithms, or "meta-algorithms", combine several learning algorithms to build a stronger model. The accuracy of hybrid model is higher than initial models in these algorithms.

Boosting is a hybrid meta-algorithm in machine learning used to reduce imbalances as well as variances. This algorithm repeatedly trains weak learners and adds them to the previous set to reach a strong classifier eventually. Weak learners are weighed when added to the set which is usually based on the accuracy of classification. AdaBoost, short for “Adaptive Boosting”, is an algorithm for constructing a strong classifier as a linear combination of weak algorithms. It is a popular Boosting algorithm for binary classification. The algorithm trains learners sequentially. For each learner with an index t, Adaboost calculates the weighted classifier error:

$$\epsilon_t = \sum_{n=1}^{N} d_n^{(t)} I(y_n \neq h_t(x_n)) \qquad (2)$$

Where x_n is a vector of predicator values for n observation. Y_n is the correct class label and h_t is the predicted value of the learner with index t. I is the predictive function and $d_n^{(t)}$ is the weight of observation n in step t.

Adaboost then increases weights of observations misclassified by learner t and lessens weights of samples correctly classified to further consider and carefully categorize in next steps (by new learners). The next learner t + 1 train’s data with the updated weights. $d_n^{(t+1)}$When the training is completed, Adaboost computes the prediction for new data using equation (3):

$$f(x) = \sum_{t=1}^{T} a_t h_t(x_n) \qquad (3)$$

Where $a_t = \frac{1}{2}\log\frac{1-\epsilon_t}{\epsilon_t}$ is the weak learner's weight in the ensemble. Training by Adaboost is considered as an exponential loss minimization:

$$\sum_{n=1}^{N} w_n \exp(-y_n f(x_n)) \qquad (4)$$

Where, $y_n \in \{-1,+1\}$ is the correct class label. W_n are normalized weights of the observations to add up to 1 and $f(x_n) \in (-\infty,+\infty)$ is the predicted classification score [30-31].

In this study, Adaboost algorithm was used by a cross-validation method. This method specifies the extent of being generalized the results on dataset and they are independent of training set. In this validation, data is divided into K subsets. One of K subsets is used for validation and K-1 is applied for training in each step. This procedure is repeated K times and all data is used exactly once for training and once for validation purposes. Finally, the average result of K validation is selected as a final estimate. We applied 10-fold as it is common in other researches. Also, 200 trees were trained as weak learners in this algorithm.

4. RESULTS

As mentioned, 303 records included 14 features are used in this study. The output is the last feature (having heart disease). To demonstrate better performance of the proposed method, we have compared it to other diagnosis studies of coronary artery disease. Table 3 shows the results of comparison. As it can be observed, classification accuracy is higher than the other available methods. Simplicity and speed of the proposed method are other advantages. As shown in Table 3, different datasets have been used to test and evaluate the methods; however, the performance of proposed system is better than those with similar data.

Among 303 records, we have 139 patients and 164 records of healthy people. MATLAB software is applied for implementation of the proposed system.

In the first stage, we applied the data to neural networks. By using 10-fold cross validation for MLP, we evaluated performance of the designed networks; therefore, the accuracy of MLP and PNN is 94.539% and 88.78% according to equation 5. Also, two sensitivity and specificity criteria (equations 6, 7) are calculated to more precise assessment of proposed method. Specificity is ratio of the detected patients (TN) to all patients (TN+FP) while sensitivity is ratio of the diagnosed healthy people (TP) to all healthy individuals (TP+FN). The sensitivity and specificity of MLP network are 86.77% and 99.39%, and they are 72% and 100% for PNN network, respectively. Clustering accuracy, sensitivity and specificity of SOM network are 79.54%, 79.14% and 79.88% respectively.

$$\text{Accuracy} = \frac{(\text{TN}+\text{TP})}{\text{n}} \qquad (5)$$

(TN+TP) are the number of detected records and n is number of all records.

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \qquad (6)$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN}+\text{FP})} \qquad (7)$$

In the second step, we used decision tree to classify data included 13 features. The accuracy was obtained 83.87%. Then we attempted to extract the important features and remove the less important ones. According to Figure 2, there were only eight features that had a positive impact on classification. Therefore, in second experiment, we eliminated five features and data with 8 important features were applied to the tree. Finally, the classification accuracy was obtained 93.07%. In fact, 282 samples were correctly classified while 21 records were labeled wrong classes. Then, the outliers were removed by using K-means algorithm [32-33] and the remained data were applied to Adaboost algorithm by 10-fold cross-validation method. Figure 3 illustrates the lowest classification error for weak learners. Therefore, the classification accuracy is 90.37%. Table 2 shows the results of the algorithms.

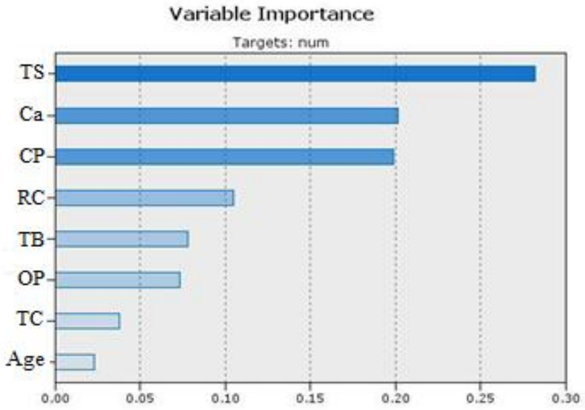


Fig 2. The importance of features. As it seen, Scan thallium is the most important feature and age is the least important one. (TS = Thallium Scan, CP = Chest Pain type, OP = Old Peak, Ca= number of major vessels colored by flourosopy, RC= resting electrocardiographic results, TB= resting blood pressure, TC= maximum heart rate achieved)

5. CONCLUSIONS

Coronary artery disease is one of the most common cardiac diseases and a major cause of deaths in the world every year. There are difficulties of real methods such as angiography and they may often yield insufficient accuracy. Therefore, using computer systems has become customary to diagnose the disease. In this paper, we proposed a new system to improve the performance of coronary artery detection. The disease is diagnosed by multi-layer perceptron (MLP), probabilistic (PNN) and self-organization map (SOM) neural networks, decision tree and Adaboost algorithm. The best accuracy, sensitivity and specificity have been obtained 94.539%, 86.33% and 99.39% respectively. The results of comparing the experiment with other existing studies prove the superiority of the proposed method.

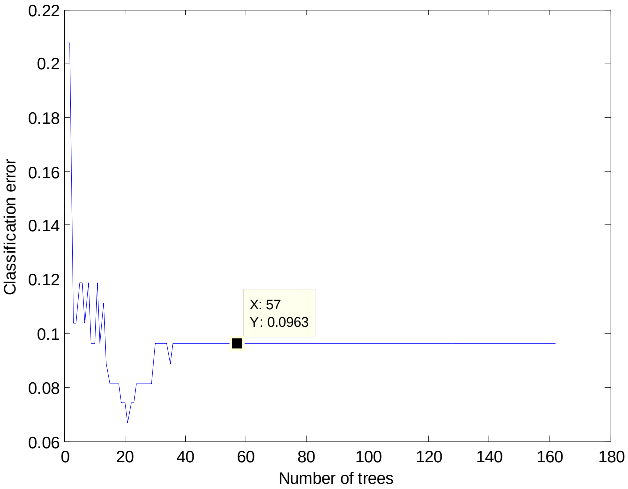


Fig3. The lowest classification error obtained for weak learners

Table 2. Results of applying different algorithms on heart disease dataset

Algorithm	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**
Perceptron Neural network	94.539	86.77	99.39
Probability Neural network	88.78	72	100
SOM neural network	79.54	79.14	79.88
Decision tree	93.07	91.33	95.48
Adaboost	90.37	89.26	91.67

Table 3. Comparison of the proposed method with other existing studies

Work	**Dataset**	**Methodology**	**The best accuracy %**	**Sensitivity %**	**Specificity %**
[16]	303 records	Hybrid method: KNN+ artificial intelligence algorithms	87	92.30	78.57
[17]	480 records	Exercise stress test and SVM	79.22	-	-
[11]	303 records	The average of several neural networks	89.01	80.95	95.91
[14]	303 records	Neural network optimized genetic algorithm	88.25	88	91
The proposed method	303 records	MLP PNN SOM	**94.539**	86.33	99.39
The proposed method	303 records	Decision tree via reduced features	**93.07**	95.48	91.33
The proposed method	303 records	Adaboost	**90.37**	91.67	89.26

6. REFERENCES

[1] [Internet]. 2017 [cited 21 December 2017]. Available from: https://www.nlm.nih.gov/medlineplus/coronaryarterydisease.htm

[2] Wong N. 2014. Epidemiological studies of CHD and the evolution of preventive cardiology. Nature Reviews Cardiology. 11(5): 276-289.

[3] Buchan K, Filannino M, Uzuner Ö. 2017. Automatic prediction of coronary artery disease from clinical narratives. Journal of Biomedical Informatics. 72:23-32.

[4] Davari Dolatabadi A, Khadem S, Asl B. 2017. Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM. Computer Methods and Programs in Biomedicine. 138:117-126.

[5] Zengin E, Bickel C, Schnabel R, Zeller T, Lackner K, Rupprecht H et al. 2015. Risk Factors of Coronary Artery Disease in Secondary Prevention—Results from the AtheroGene—Study. PLOS ONE. 10(7):e0131434.

[6] Shouman M, Turner T, Stocker R. 2011. Using Decision Tree for Diagnosing Heart Disease Patients. AusDM '11 Proceedings of the Ninth Australasian Data Mining Conference.121, 23-30.

[7] Kochurani O.G, Aji S, Kaimal M.R. 2007. A Neuro Fuzzy Decision Tree Model for Predicting the Risk in Coronary Artery Disease. IEEE 22nd International Symposium onIntelligent Control (ISIC).

[8] El-Bialy R, Salamay M, Karam O, Khalifa M. 2015. Feature Analysis of Coronary Artery Heart Disease Data Sets. Procedia Computer Science. 65:459-468.

[9] Atkov O, Gorokhova S, Sboev A, Generozov E, Muraseyeva E, Moroshkina S et al. 2012. Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters. Journal of Cardiology. 59(2):190-194.

[10] H.S N. 2013. ANN Model to Predict Coronary Heart Disease Based on Risk Factors. Bonfring International Journal of Man Machine Interface. 3(2):13-18.

[11] Das R, Turkoglu I, Sengur A. 2009. Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications. 36(4):7675-7680.

[12] Kurt I, Ture M, Kurum A. 2008. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Systems with Applications. 34(1):366-374.

[13] Schapire R.E. 2003. The boosting approach to machine learning: An overview. Nonlinear Estimation and Classification. Springer. 149-171.

[14] Arabasadi Z, Alizadehsani R, Roshanzamir M, Moosaei H, Yarifard A. 2017. Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. Computer Methods and Programs in Biomedicine. 141:19-26.

[15] Akila S, Chandramathi S. 2015. A Hybrid Method for Coronary Heart Disease Risk Prediction using Decision Tree and Multi-Layer Perceptron. Indian Journal of Science and Technology. 8(34).

[16] Sunitha S. 2010. Data Mining of Medical Datasets with Missing Attributes from Different Sources [Master of Science in Mathematics]. Youngstown State University, Department of Mathematics and Statistics.

[17] Polat K, Şahan S, Güneş S. 2007. 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. 32(2):625-631.

[18] Babaoğlu I, Fındık O, Bayrak M. 2010. Effects of principle component analysis on assessment of coronary artery diseases using support vector machine. Expert Systems with Applications. 37(3):2182-2185.

[19] Newman, D. J., Hettich, S., Blake, C. L. S., & Merz, C. J. 1998. UCI repository of machine learning database. Irvine, CA: University of California.

[20] Offor U, Alabi S. 2016. Artificial Neural Network Model for Friction Factor Prediction. Journal of Materials Science and Chemical Engineering. 04(07):77-83.

[21] Ebrahimabadi A, Azimipour M, Bahreini A. 2015. Prediction of roadheaders' performance using artificial neural network approaches (MLP and KOSFM). Journal of Rock Mechanics and Geotechnical Engineering. 7(5):573-583.

[22] Durão R, Mendes M, João Pereira M. 2016. Forecasting O 3 levels in industrial area surroundings up to 24 h in advance, combining classification trees and MLP models. Atmospheric Pollution Research. 7(6):961-970.

[23] Thoma M. 2017. Analysis and optimization of convolutional neural network architectures, master Thesis.

[24] Sa'di S, Hashemi R, Abdollapour A, Chalabi K, Salamat M. 2015. A Novel Probabilistic Artificial Neural Networks Approach for Diagnosing Heart Disease. International Journal in Foundations of Computer Science & Technology. 5(6):47-53.

[25] Specht. 1988. Probabilistic neural networks for classification, mapping, or associative memory. IEEE International Conference on Neural Networks.

[26] Ghaderzadeh M. 2014. An Intelligent System Based on Back Propagation Neural Network and Particle Swarm Optimization for Detection of Prostate Cancer from Benign Hyperplasia of Prostate. Journal of Health & Medical Informatics. 05(03).

[27] Mingoti S, Lima J. 2006. Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. European Journal of Operational Research. 174(3):1742-1759.

[28] Rokach L, Maimon O. 2010. Data mining with decision trees: theory and applications: Series in machine perception and artificial intelligence. World Scientific.

[29] Weihong W. 2006. A preliminary study on constructing decision tree with gene expression programming. In Proc. 1st international conference on innovative computing, information and control; 222-225.

[30] Freund Y, Schapire R. 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences. 55(1):119-139.

[31] Schapire R.E. 2003. The boosting approach to machine learning: An overview. Nonlinear Estimation and Classification. Springer; 149-171.

[32] Rodger J. 2015. Discovery of medical Big Data analytics: Improving the prediction of traumatic brain injury survival rates by data mining Patient Informatics Processing Software Hybrid Hadoop Hive. Informatics in Medicine Unlocked. 1:17-26.

[33] Sun Y, Clark O. 2009. Implementing an Intuitive Reasoner for Predicting Continuous Weather Variables. 2009 International Conference on Computer Modeling and Simulation.