ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, Vol 8(34), DOI: 10.17485/ijst/2015/v8i34/85947, December 2015 ISSN (Online) : 0974-5645

A Hybrid Method for Coronary Heart Disease Risk Prediction using Decision Tree and Multi Layer Perceptron

S. Akila[1*] and S. Chandramathi[2]

[1]Department of Computer Science, Vellalar College for Women, Thindal, Erode - 638012, Tamil Nadu, India; akilraz@yahoo.co.in

[2]Electrical Sciences, Sri Krishna College of Technology, Kovaipudur, Coimbatore - 641042, Tamil Nadu, India; chandrasrajan@gmail.com

Abstract

Background/Objectives: The diagnosis of Coronary Heart Disease (CHD) risk prediction is a vital and complicated job in medicine which is closely linked with lifestyle related behaviors. The main intention of this work is to build up a rapid and automatic prediction of CHD risk by integrating Decision Tree (DT) and Multi Layer Perceptron (MLP). **Methods/Statistical Analysis:** The proposed hybrid method consists of two stages in which risk identification is carried out in the first stage and level prediction is done in the second stage. In the first stage the physical and biochemical factors are classified using C4.5 algorithm in DT. In the second stage the CHD risk identified instances from DT are analyzed using MLP with habitation and medical history attributes. **Findings:** The obtained classification accuracies of this system are 98.66% for DT and 96.66% for MLP. The performance analysis of the proposed method is evaluated using sensitivity and specificity which helps to reduce the healthcare costs, further invasive CHD risk examination and waiting time of the individuals. **Application/Improvements:** The proposed work structured an optimal predictive tool for diagnosing CHD risk, which can further serve as an implication sketch for physicians in clinical diagnosis.

Keywords: Coronary Heart Disease, Data Mining, Decision Tree, 10-Fold Cross Validation, Multi Layer Perceptron

1. Introduction

CHD are a group of disorders that influence the heart's capability to function normally. It is due to the growth of atherosclerotic plaques within the walls of coronary arteries resulting in narrowing of lumen of coronary arteries and subsequently occlusion that causes to Myocardial Infarction (MI) or unexpected death[1–3]. Major reasons for CHD are tobacco use, sedentary life style, unhealthy and irregular diet and harmful use of alcohol[4]. Several research works has been carried out with the aid of statistical and data mining tools to help healthcare experts in the findings of CHD[5]. The major risk factors of CHD are due to the variations from the normal values of blood sugar, blood pressure and serum cholesterol levels[6].

The diagnosis of CHD risk requires much experience and knowledge. Nowadays huge amount of medical data is available which contains hidden information. However, the effective analysis tool to discover the hidden patterns in data is a challenging process for the predication of CHD. The hidden knowledge inside the enormous amount of data can be discovered using data mining[7]. It can expose the patterns and relationships using statistical analysis, pattern recognition and machine learning techniques. Data mining is used in various applications such as agriculture, medicine, marketing, banking, insurance, crime detection, privacy preservation, etc[8]. The fast innovation in the information and computation technology derives various data mining methods which have been implemented to medical data set for diagnosis. The data mining methods are used to locate some concealed knowledge that provides a novel medical tool

* Author for correspondence

for diagnosis which is especially useful when certain diagnostic instruments are costly and invasive.

The objective of the paper is to build up a data mining technique using C4.5 decision tree classification algorithm for the identification of CHD risk and to analyze the risk level using multi layer perceptron. Different parameters are used to build a diagnostic model that presents the association between input and output variables to find vital factors and rules that affect the heart and to create an optimal predictive tool for diagnostic enhancement that can further serve as referential strategy in clinical diagnosis for physicians. The main proposal of this study is early identification of the CHD risk which helps to reduce the CHD events and to reduce the diagnostic cost.

2. Methods

2.1 Data Description

To implement this hybrid system the data was collected from occupational drive's master health checkup at Institute of Road and Transport Perundurai Medical College and Hospital. The professional drivers in the transportation industry are at higher risk due to irregular diet and sedentary behavior. They undergo high psychological demands and low physical activity at work which has been connected with Ischemic Heart Disease (IHD)[9,10]. Furthermore, drivers have long working hours, shift work, exposure to nitrous oxide, carbon monoxide and traffic noises which have also been related with IHD[11].

Each individual is described by a set of nineteen attributes that includes screening of both clinical and biochemical data. The medical data records are transformed into operational format and a database is created. Database cleaning is done by removing the irrelevant attributes related to this study and duplicate records from the dataset. The missing values for the attributes are filled based on the data estimation and 375 instances are taken for this study. The numerical and categorical variables are converted into binary data based on the cutoff points. A unique number is given to each patient for recognition in the database for further reference.

After rigorous assessment, only six predominant attributes out of nineteen attributes has been taken for first level analysis using DT. The predominant attributes includes three biophysical parameters Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Body Mass Index (BMI) and three blood chemical parameters Fasting Blood Sugar (FBS), Post Prandial Blood Sugar (PPBS) and Triglycerides (TG). The threshold for all these attributes are BMI> = 25kg/m^2, BP>140/90 mmHg, FBS> = 126mg/dl, PPBS> = 200mg/dl, TG> = 170mg/dl, fixed based on the standards provided by World Health Organization (WHO)[12]. In the second level of analysis, outputs of the CHD risk patients from first level are analyzed with the other predominant habitation and medical history attributes using MLP. It includes two habitation attributes smoking, alcohol and three historical attributes treatment history, past history, family history along with Electrocardiogram (ECG).

2.2 Decision Tree Induction

It is an induction learning algorithm which has the benefits of simplicity, transparency and ability to extract decision rules which are frequently used for classification to predict the set in which a individual belongs to. There are many decision tree algorithms like ID3, Classification and Regression Trees (CART), Chi-squared Automatic Interaction Detection (CHAID), C4.5 and C5.0. For this study C4.5 algorithm is applied since it is a top down divide and conquer strategy that splits a given set of instances into smaller and smaller subsets in step with the growth of the tree.

C4.5 algorithm that has been applied in the proposed task for the inquiry of CHD risk is described as follows. The CHD dataset is selected as an input to the algorithm for analysis. The attribute selection is carried out using information gain and gain ratio that provide a ranking for each attribute. Based on the ranking the algorithm identifies the most significant independent variable and positions it as a root, which is tracked by the next finest variables[13]. The C4.5 algorithm finds the variable-threshold for each leaf node that maximizes the homogeneity and splits the input observation into two or more subgroups[14]. Based on the splitting criterion, branches are grown for every leaf node until the tree is completed.

2.2.1 Information Gain

Information gain is based on Claude Shannon's information theory[15]. The C4.5 algorithm uses information gain as its attribute selection measure. In the information gain, the content of the dataset is the predictable information necessary to classify an instance. Attribute's information

of the dataset is the new quantity of information required to classify an instance of the dataset after partitioning by that attribute. Thus to discover the splitting attribute of the tree, the information gain must be calculated for each attribute and the attribute that have maximum information gain must be selected. Each attribute's information gain is calculated using the formula.

$$\text{Information Gain (A)} = \text{Info } (D) - \text{Info}_A(D) \quad (1)$$

Where attribute investigated is denoted as A:

$$Info(D) = -\sum_{i=1}^{k} p_i \log_2(p_i) \quad (2)$$

Where,
P_i - Probability (class I in dataset D)
k – Number of class values

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Info(D_j) \quad (3)$$

Where,
$|D_j|$ - Number of observations with attribute values j in dataset D.
$|D|$ - Total number of observations in dataset D.
$|D_j|$ - Sub dataset of D that contains attribute values j.
v - All attribute values.

2.2.2 Gain Ratio

The information gain chooses to pick attributes with huge number of distinct values which results in large number of partitions which may lead to a useless classification. To overcome the effect of the bias in information gain, a variant known as gain ratio was introduced by the C4.5 algorithm. The gain ratio adjusts the information gain for each attribute by applying a kind of normalization to each attribute by using the formula:

$$GainRatio(A) = \frac{InformationGain(A)}{SplitInfo_A(D)} \quad (4)$$

$$SplitInfo\ (D) = -\sum \frac{|D_j|}{|D|} \times \log \left(\frac{|D_j|}{|D|} \right) \quad (5)$$

2.3 Multilayer Perceptron

The most important model in Artificial Neural Network (ANN) is MLP, the multilevel neural feed forward network trained by the back propagation method for finding errors. The sets of variable parameters are chosen in order to adjust the ANN by pair wise correlation between the database parameters and CHD diagnosis. The MLP consists of one input layer, one output layer and one or more hidden layers. Each layer consists of one or more nodes with lines between them indicate the flow of information from one node to another node. The input data is provided to input layer for processing to get a predicted output. The error is calculated by subtracting the predicted output from the actual output. The network then uses a back propagation algorithm which adjusts the weights between output layer nodes and hidden layer nodes and works backwards through network. When back propagation is completed, the forward procedure begins again. This process is repeated to minimize the error between the predicted value and actual output.

3. Results and Discussion

In the first stage of this study using DT the classes are taken as positive instances when CHD = No, while it is negative instances when CHD = Yes. P is the amount of positive instances i.e. 287 and N is the amount of negative instances i.e. 88. For each instance, a comparison is made with the classifier's predicted class label and the actual class label. The correctly classified instances are CHD = No for 285 instances and CHD = Yes for 85 instances and the misclassification occurs for 5 instances. For the second stage of this study correctly classified 85 instances with CHD = Yes along with 5 misclassified instances are taken for analysis using MLP. In this stage the habitation, medical history and ECG attributes are analyzed for the 90 instances. All the misclassified instances from the first stage taken for the second stage of analysis produce the output as low risk patients. The following four additional terms are used to compute the evaluation measures for both DT and MLP.

3.1 True Positive (TP)

Refers to the positive instances (CHD = No) that were correctly labeled by the classifier as CHD = No for DT and positive instances (CHD = Low Risk) that is correctly labeled by the classifier as CHD = Low Risk for MLP.

3.2 True Negative (TN)

Refers to the negative instances (CHD = Yes) that were properly marked by the classifier as CHD = Yes for DT and negative instances (CHD = High Risk) that were properly marked by the classifier as CHD = High Risk for MLP.

3.3 False Positive (FP)

These are negative instances (CHD = Yes) that were inaccurately marked by the classifier as positive i.e. CHD = No for DT and negative instances (CHD = High Risk) that were improperly marked by the classifier as positive i.e. CHD = Low Risk for MLP.

3.4 False Negative (FN)

These are positive instances (CHD = No) that were wrongly marked by the classifier as negative i.e. CHD = Yes for DT and positive instances (CHD = Low Risk) that were incorrectly marked by the classifier as negative i.e. CHD = High Risk for MLP.

The confusion matrix is a valuable tool for evaluating the algorithms by recognizing the instances of different classes. TP and TN are used to know when the classifier is getting the correct data while FP and FN are used to know when the classifier is getting wrong data. The confusion matrix for first level of prediction is shown in Table 1. It contains predicted classification information done by the C4.5 classification system as TP = 285, TN = 85, FP = 3, FN = 2.

The confusion matrix for second level of prediction is shown in Table 2. It contains predicted classification information done by MLP system as TP = 63, TN = 24, FP = 2, FN = 1. From the confusion matrix the accuracy of the classifier is calculated using the formula:

$$Accuracy = \frac{TP + TN}{P + N} \qquad (6)$$

The error rate or misclassification rate of the classifier is calculated using the formula:

$$ErrorRate = \frac{FP + FN}{P + N} \qquad (7)$$

The sensitivity is estimated by the formula:

$$Sensitivity = \frac{TP}{P} \qquad (8)$$

The specificity is estimated by the formula:

$$Specificity = \frac{TN}{N} \qquad (9)$$

The above said performance measures are tabulated in Table 3.

Table 1. Confusion matrix for DT

Class predicted	No CHD risk	CHD risk
No CHD risk	285	2
CHD risk	3	85

Table 2. Confusion matrix for MLP

Class predicted	Low Risk	High Risk
Low Risk	63	1
High Risk	2	24

Table 3. Performance summary of CHD

S.No.	Performance Estimator	Values (%)	
		DT	MLP
1	Sensitivity	98.95	98.43
2	Specificity	97.70	92.3
3	Accuracy	98.66	96.66
4	Positive predictive value	99.30	98.43
5	Negative predictive value	96.59	92.30

Table 4. Information gain and gain ratio for selecting the root of the tree

S.No.	Attributes	Information Gain	Gain Ratio
1.	TG	0.184	0.236
2.	BMI	0.138	0.175
3.	DBP	0.104	0.133
4.	PPBS	0.100	0.127
5.	SBP	0.099	0.126
6.	FBS	0.079	0.102

The accuracy rate of DT is 98.66% and MLP is 96.66% which cannot be acceptable as the classifier could be correctly labeling simply the CHD = Yes instances and misclassifying all the CHD = No tuples. In order to find how well a classifier finds the positive instances and negative instances, sensitivity and specificity measures are used. From the measures it is observed that the both classifiers have high accuracy and it has accurately classified the positive instances i.e. sensitivity, but its ability to find the negative instances i.e. specificity is comparatively low.

The positive predictive value in Table 3 shows the probability that the patients with CHD = No truly have no CHD risk and the negative predictive value shows the probability that the patients with CHD = Yes truly have CHD risk. The comparison between sensitivity and specificity is plotted using Receiver Operating Characteristics (ROC) curve in Figure 1for DT and Figure 2 for MLP.

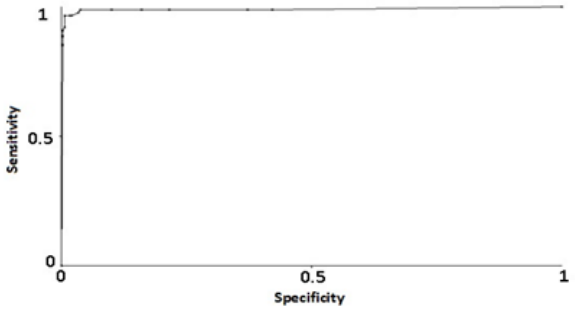

Figure 1. ROC graph for specificity versus sensitivity has a larger area under the curve which shows higher performance of the algorithm.

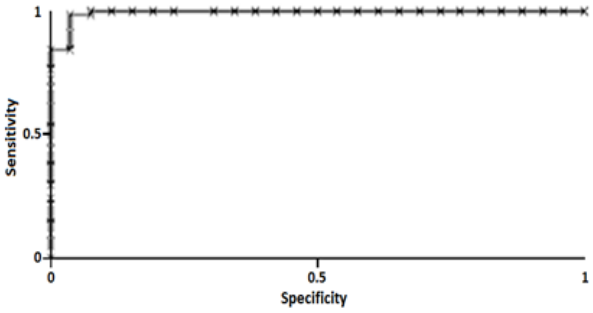

Figure 2. ROC graph for specificity versus sensitivity of MLP shows higher performance of the algorithm.

The ROC graph has been extensively used in signal detection theory to show trade-offs between true positive rate (sensitivity) and false positive rate (specificity)[16]. In this work an ROC curve permits one to imagine the trade-offs between the rates at which the classifier can exactly recognize positive instance versus the rate at which it incorrectly identifies negative instances as positive. Similarly the dataset were evaluated by both classifiers on the above said performance measures on percentage split and supplied train and test set method but the accuracy is low. But the most significant improvement was obtained only by using 10-fold cross validation method. This method has been used in C4.5 algorithm to build the decision tree and MLP. The fold versus accuracy graph is shown in Figure 3. This shows that the accuracy increases and becomes stable for higher folds. In order to fix the predominant attribute the information gain and gain ratio for all the attributes of DT are computed by using the equations 1 to 4 and tabulated as revealed in Table 4.

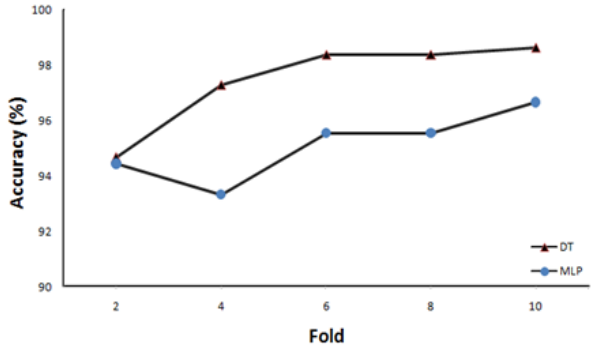

Figure 3. Fold versus accuracy graph for both DT and MLP.

The information gain and gain ratio of the attribute TG is higher than other attributes so the decision tree selects TG as its root. The process of selecting the next node to the root is done by selecting TG = High as the target and information gain is calculated for the remaining attributes. The information gain of BMI is found to be higher than other attributes and is selected as the next node to the root in the tree. This process continues for each leaf node until every attribute has been included along this path through the tree. The final decision tree learned by C4.5 algorithm for 375 instances is shown in Figure 4.

The MLP structure is shown in Figure 5. This structure has three types of neuron layers: input, hidden and output layers with non linear activation function. The prime task of the neuron in the input layer is to divide the input signal along with neurons in the hidden layer. Each neuron in the hidden layer adds its input signals once it influence them with the strength of the respective connection from the input layer and decides its output. The back propagation algorithm is able to regulate the network weights and biasing values to reduce the square sum of the difference between the given CHD class and predicted CHD values figured by the network.

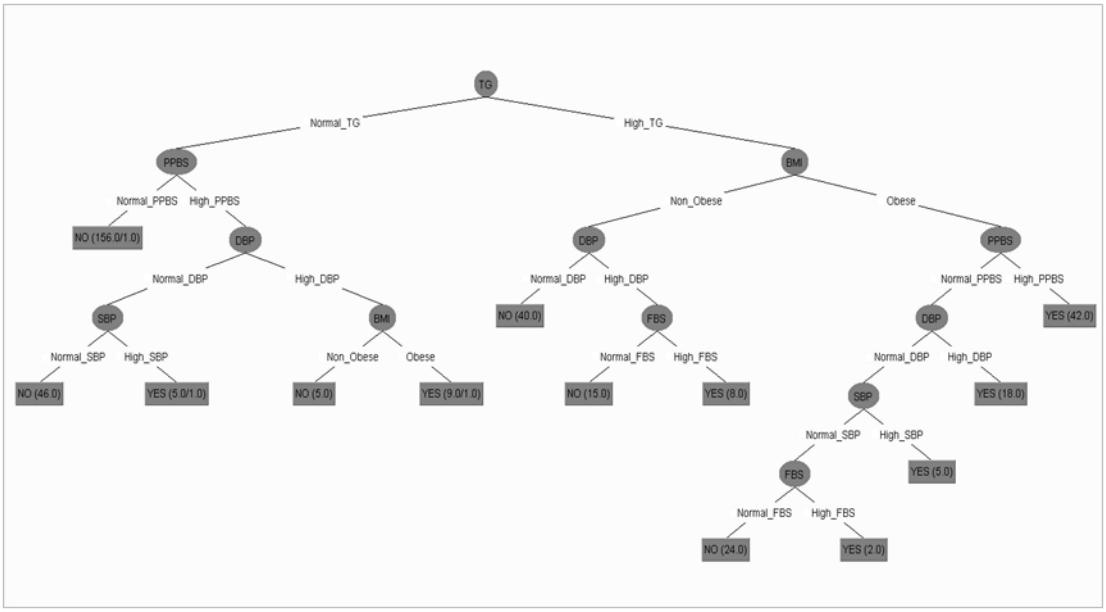

Figure 4. Decision tree for CHD risk prediction shows the leaf node with YES for CHD risk and NO for No CHD risk.

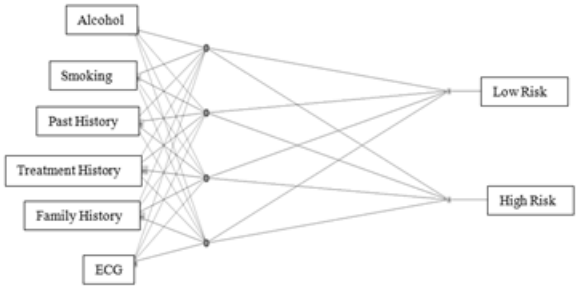

Figure 5. The structure of MLP for CHD risk level prediction.

4. Conclusion

This study demonstrated a hybrid method for CHD risk prediction of occupational drivers with regard to clinical, chemical, habitation parameters using DT and MLP. In the first stage of prediction using DT the accuracy of 98.66% has been achieved. In the second level of prediction using MLP the CHD risk instances along with misclassified instances from the first level are analyzed and achieved the accuracy of 96.66%. The predictive power of medical data mainly depends on the specificity and sensitivity than accuracy. The clinical experts in particularly wanted to know the possibilities of increasing the specificity of the predictive process without affecting the sensitivity too much which may lead to the danger of number of patients who actually has the CHD risk, but omitted without any examination. In the second stage due to higher sensitivity (98.43%) the percentage of misclassification is very less which would minimize the no of patients going for further invasive investigations and also cut down the waiting time of the really ill patients. The above said one misclassification is very less in this work which shows the high effective classification of MLP along with 10-fold cross validation method. Out of 90 CHD risk patients 63 patients are identified as low risk which reduces the further invasive CHD risk examination cost for the individuals. Considering the correctly classified instances, 70% of the instances need not go for further analysis as they have low CHD risk and only 26% has to undergo for further analysis. This shows the improved diagnostics performance of the proposed work with very few predominant attributes. The misclassified 3.33% of instances can be corrected by further analyzing the instances by applying different soft computing technique for further analysis.

5. References

1. Polat K, Gunes S. A hybrid approach to medical decision support systems: Combining feture selection, fuzzy weighted pre-processing and AIRS. Computer Methods and Programming in Biomedicine. 2007 Nov; 88(2):164–74.
2. Tsipouras MG, Exarchos TP, Fotiadas, Kotsia AP, Vakalis KV, Naka KK, Michalis LK. Automated diagnosis of Coronary Artery Disease based on data mining and fuzzy mod-

eling. IEEE Transactions on Information Technology in Biomedicine. 2008 Jul; 12(4):447–58.
3. Muthukaruppan S, Er MJ. A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease. Expert Systems with Applications. 2012 Oct; 39(14):11657–65.
4. Srinivas K. Analysis of Coronary Heart Disease and prediction of heart attack in coal mining regions using data mining techniques. Proceedings of 5th ICCSE: Hefei China; 2010 Aug 24-27. p. 1344–9.
5. Yanwei X, Jie W, Zhihong Z, Yonghong G. Combination data mining methods with new medical data to predicting outcome of coronary heart disease. Proceedings of ICCIT: Gyeongju Republic of Korea; 2007 Nov 21-23. p. 868–72.
6. Matsumori R, Miyazaki T, Shimada K, Kume A, Kitamura Y, Oshida K, Yanagisawa N, Kiyanagi T, Hiki M, Fukao K, Hirose K, Ohsaka H, Mokuno H, Daida H. High levels of very long-chain saturated fatty acid in erythrocytes correlates with atherogenic lipoprotein profiles in subjects with metabolic syndrome. Diabetes research and clinical practice. 2013 Jan; 99(1):12–8.
7. Alizadehsani, R, Habibi J, Hosseini MJ, Mashayekhi H, Boghrati R, Ghandeharioun A, Bahadorian B, Sani ZA. A data mining approach for diagnosis of coronary artery disease. Computer Methods and Programs in Biomedicine. 2013 Jul; 111(1):52–61.
8. Podgorelec V, Kokol P, Stiglic B, Rozman I. Decision trees: an overview and their use in medicine. Journal of Medical Systems. 2002 Oct; 26(5):445–63.
9. Bigert C, Gustavsson P, Hallqvist J, Hogstedt C, Lewne M, Plato N, Reuterwall C, Scheele P. Myocardial Infarction among professional driver. Epidemiology. 2003 May; 14(3):333–9.
10. Ragland DR, Krause N, Greiner BA, Fisher JM. Studies of health outcomes in transit operators: policy implications of the current scientific database. Journal of Occupational Health Psychology. 1998 Apr; 3(2):172–87.
11. Claire HQ, Jen MN, Troy HP. Does occupational driving increase the risk of cardiovascular disease in people with diabetes? Diabetes Research and Clinical Practice. 2013 Jan; 99(1):e9–e11.
12. Prevention of Cardio Vascular Disease pocket guidelines for assessment and management of cardio vascular risk. 2015. Available from: http://www.who.int/ cardiovascular_diseases/guidelines/PocketGL.ENGLISH.AFR-D-E.rev1.pdf
13. Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, Pidetcha P, Prachayasittikul V. Identification of metabolic syndrome using decision tree analysis. Diabetes Research and Clinical Practice. 2010 Oct; 90(1):e15–e18.
14. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine. 2005 Jun; 34(2):113–27.
15. Shannon CE. A mathematical theory of communication. Bell System Technical Journal. 1948 Jul; 27(3):379–423.
16. Kukar M, Kononenko I, Groselj C, Krali K, Fettich J. Analyzing and improving the diagnosis of ishaemic heart disease with machine learning. Artificial Intelligence in Medicine. 1999 May; 16(1):25–50.