Using Decision Trees in Data Mining for Predicting Factors Influencing Heart Disease

Moloud Abdar
Undergraduate of Computer Engineering, Department of Engineering
University of Damghan
Damghan, Iran
m.abdar1987@gmail.com

Abstract—**Statistics from the World Health Organization (WHO) shows that heart disease is one of the leading causes of mortality all over the world. Because of the importance of heart disease, in recent years, many studies have been conducted on this disease using data mining. The main objective of this study is to find a better decision tree algorithm and then use the algorithm for extracting rules in predicting heart disease. Cleveland data, including 303 records are used for this study. These data include 13 features and we have categorized them into five classes. In this paper, C5.0 algorithm with a accuracy value of 85.33% has a better performance compared to the rest of the algorithms used in this study. Considering the rules created by this algorithm, the attributes of Trestbps, Restecg, Thalach, Slope, Oldpeak, and CP were extracted as the most influential causes in predicting heart disease.**

Keywords— Data Mining; Heart Disease; Classification; Decision Tree; C5.0 Algorithm.

I. INTRODUCTION

In recent years, the volume of accumulated data has rapidly been increased. Regarding this project, using a method, which can extract beneficial information from these data, has highly been considered. Data mining is used in most of the scientific fields, including medical sciences. So far, data mining techniques have been used for diagnosing diseases such as heart disease, diabetes, neurology, depression, breast cancer, liver disease, etc. There are different methods and algorithms in data mining, and according to different provided data, the power, and performance of each of these algorithms is different. For example, algorithms of Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Trees, Artificial Neural Network (ANN), etc. can be named in this area. The WHO has reported that heart disease as a major cause of death around the world is highly important [1]. Because of the importance of this disease in the world, this paper examines the performance of decision algorithms of C5.0, CHi-squared Automatic Interaction Detection (CHAID), Quick Unbiased and Efficient Statistical Tree (QUEST), and The Classification and Regression Tree (C&R Tree) on heart patients' data.

A. Data Mining and Classification

With the advancement of science, the volume of stored data in various fields has been increased. Analyzing the accumulated data can extract useful information existing in them. Applying data mining as a new science on data can extract the science lying within the data. Data mining reveals the beneficial relationship between data and the right decisions can be made based on these relationships [2], [3]. Utilizing the related tools to show the results, data mining uses analytical modeling, classification, and prediction of information. To be able to extract information easily, data mining algorithms need a set of pre-processing on the data and post-processing on the extracted patterns. The techniques used for data mining can be classified as follows:

1. Classification (Predictive Technique): In this method, a sample is classified into one of several predefined categories.
2. Regression (Preventive Technique): Prediction of a variable amount based on other variables
3. Clustering (Descriptive Technique): A data set mapped into one of the several clusters. Clusters are defined as *groupings of data categories*, which are shaped based on the similarity of some criteria.
4. Discovery of association rules (Descriptive Technique): It states the relationship of dependence among the various features.
5. Sequence Analysis: It models sequence patterns, such as time series.

One of the divisions in data mining is classification, which acts using *If-Then* rule. Its purpose is to predict a feature (characteristic) based on other features (characteristics) that are known as predictors. In classification, data are divided into two classes of *Training and Testing*, and having the training data mining algorithms extract the rules. The purpose feature and the value of prediction features should be placed into data mining algorithms. Algorithms of KNN, SVM, Decision Tree, and ANN are from among classification algorithms.

B. Heart Disease

The human body has a complex mechanism so that any dysfunction of any body parts influences the other. The human heart is about the size of a fist, while it is one of the strongest muscles in the body. It begins to throb 21-28 days after forming the fetus in the womb and beats 100,000 times daily on average. Average heart rate is about 70 beats per minute, which doubles or multiplies because of physical activity. The human heart is a body part playing an important role in his/her life. Any dysfunction of human heart leads to dysfunction in the body's total system, especially blood supply and respiratory system.

According to the WHO, the World Heart Federation (WHF) and the USA's Centers for Disease Control and Prevention (CDC) in 2020, the number of deaths due to “heart disease and stroke” reaches up to 20 million, whereas the mentioned number will be increased up to 24 million deaths by the year 2030 [4]. The rising number of deaths due to heart disease is the reason for the high importance of research on heart disease. The above-mentioned problem causes a lot of spending by patients and governments to manage and cure heart disease. There are many types of heart disease including "coronary heart disease, stroke, hypertensive heart disease, inflammatory heart disease, and rheumatic heart disease [5]. Like other diseases, heart diseases have also certain symptoms, which we can refer to chest pain, discomforts in chest area, cough, palpitations, and fluid retention from among [6]. There are many data concerning heart patients, and one of the most popular sources of data is for Cleveland. The source includes 303 records with 13 features in five classes [7].

C. Risk factors of Heart Disease

For each disease there are some factors causing the illness or intensifying its effects. The effect of these factors varies with each patient. Each of the following factors also has a variety of effects on different types of heart disease. The most common risk factors for heart disease include: smoking, gender (Sex), age, ethnicity, family history of the disease, high blood pressure, high blood cholesterol, diabetes, poor diet, Lack of exercise, obesity, stress and blood vessel inflammation.

II. RELATED WORKS

One of these studies conducted by Jasmin Nahar, Tasadduq Imam, Kevin S. Tickle, and Yi-Ping Phoebe Chen [6] deals with identifying the role of relationship among risk factors for heart disease in women and men. It refers to the fact that men are more likely to develop Coronary heart disease than women are. Men and women can overcome chest pain doing exercise. One of the extracted points in this article is that factor of Rest ECG are introduced in either Normal or Hyper forms, and Slope being flat is defined as risk factors. However, for men *Rest ECG* only as a *Hyper* is a risk. Therefore, the result is that Rest ECG as a factor to predict heart disease in women has to be considered as well. In this research, the techniques of Apriori, Predictive Apriori and Tertius were compared with each other, resulting in the report as follows: With the level of confidence of 90%, the high accuracy rate of 99% and the confirming level of 79%, respectively for the models of Apriori, Predictive Apriori and Tertius.

Another valuable research by Kemal Polat and Salih Güneş [8] utilizing fuzzy weighted pre-processing and artificial immune recognition system (AIRS) reported 92.59% accuracy. The article by Roohallah and et al. [9] have used the algorithms of SMO, Bagging, and ANN. They have used Z-Alizadeh Sani data involving 303 patients with 54 features. This study has then noted that "Chest Pain and Age" have had the greatest impact on productivity than the rest of the features, and their reported accuracy has been equal to 94.08%. Another article written by Abushariah, Mohammad, Assal AM Alqudah, Omar Y. Adwan, and Rana MM Yousef [10] deals with the comparison of ANN and ANFIS using MATLAB software. The accuracy of training data for ANFIS 100% and ANN equal to 90.74% has been calculated. However, the accuracy of experimental data for the ANN 87.04% and for ANFIS equal to 75.93% has been obtained. Negar Ziasabounchi and Iman Askerzade [11] in their paper have discussed the accuracy of ANFIS technique. In their study, UCI data with seven features have been considered as an input. The results of their study report the accuracy of 92.30%. An investigation conducted by Bhatia, Sumit, Praveen Prakash, and G. N. Pillai [12] has examined the SVM technique, and using five classes, the accuracy obtained has been 72.55%, but when it has been examined on two classes including patients and healthy people, the obtained accuracy is 90.57%.

Mai Shouman, Tim Turner, and Rob Stocker [13] in their paper discusses some kinds of decision trees defined as J4.8, Gain Ratio and binary discretization, then introduces two types of less used decision trees namely Gini Index and Information Gain. After comparing to J4.8 and Bagging, this paper has finally proved the effectiveness of the chosen method with the accuracy of 84.10%. The paper by Thanh Nguyen, Abbas Khosravi, Douglas Creighton, and Saeid Nahavandi [14], introduces GSAM obtained from the integration of "Fuzzy Standard Additive Model" (SAM) with "Genetic algorithm". Then, the paper has compared it to Probabilistic Neural Network (PNN), SOM, Fuzzy ARTMAP (FARTMAP), Adaptive Neuro-Fuzzy Inference System or Adaptive Network-based Fuzzy Inference System (ANFIS), and Original Standard Additive Model (SAM). Finally, the highest level of accuracy has obtained as follows: in the Original SAM the ANFIS method with 73.10%, in the Principal Component Analysis (PCA) the GSAM method with 64.25% and in the Wavelets the GSAM method with 78.78%.

To predict heart disease and breast cancer using classification algorithms, Hlaudi Daniel Masethe and Mosima Anna Masethe [15] have compared the algorithms of J4.8, Bayes Net, Naïve Bayes, Simple Cart, and REPTREE with each other in their paper, and they have obtained the accuracy of 99.0741% for each of the algorithms of J4.8, REPTREE, and Simple Cart for heart patients.

Articles presented by Kay Chen Tan, Eu Jin Teoh, Q. Yu, and K. C. Goh [3], which examines the diseases Iris, Diabetes, Breast-Cancer, Heart-c, Hepatitis using GA-SVM obtained

from the integration of Genetic algorithm and SVM, gives the best accuracy of 85.81% in relation to heart patients.

Another article about heart disease presented by Jyoti Soni , Ujma Ansari, Dipesh Sharma, and Sunita Soni [16] has compared the techniques of Naïve Bayes, Decision Tree, and Classification via clustering. It has found the accuracy of Decision Tree equal to 99.2%. Another article done by Chaitrali S. Dangare, and Sulabha S. Apte [17], deals with the application of data mining to predict heart disease using ANN. In the article, it is important to note that adding the two factors of obesity and smoking to 13 previous factors causes the accuracy level to rise up to 100%. In the same vein, Acharya, U. Rajendra et al. [18] have studied the coronary artery disease. For this research, they have provided 400 healthy controls and 400 cases of patients. The researchers in this study have chosen the Gaussian Mixture Model (GMM), finally the research accuracy has been reported equal to 100%. Nura Esfandiari, Mohammad Reza Babavalian, Amir-Masoud Eftekhari Moghadam, and Vahid Kashani Tabar [19] in a study based on data collected between the years 1999-2013; deal with knowledge discovery in medicine. Each section of the paper has been devoted to one of the six medical tasks, which includes the following: screening, diagnosis, treatment, prognosis, monitoring, and management. For each of the six tasks, five data mining approaches have been considered: classification, regression, clustering, association and hybrid. The main purpose of this paper is the investigation of performed tasks between the years 1999-2013, as well as the integration of them to extract medical information and data mining from 291 articles published in the mentioned periods. Classifying the frequency cycle of cardiovascular disease in different regions of Texas, that is done by Kyle E Walker and Sean M. Crotty [20] show that the main cause of mortality in this state is cardiovascular disease. To achieve this goal and to help the development of health care policies in combating the disease, the present paper has examined the area most vulnerable to the disease. After investigations, it was concluded that although factors such as poor health, social and economic deprivation can cause this disease, in some areas, also this disease affected on people with high living standards.

III. METHOD AND DATASET

To prepare the present paper, decision trees, which is one of the most important algorithms used in data mining was employed; they included the algorithms of C5.0, C&R Tree, CHAID, and QUEST. Decision tree structure in machine learning is a predictive model, which turns the observed facts about a phenomenon into some inferences about the purpose value of that phenomenon. Machine learning techniques to infer a decision tree from data is called *Decision Tree Learning*, which is one of the most common methods of data mining. Decision trees are able to produce an understandable description for humans, from the relationships in a data set that can be used for performing classification and prediction tasks. This technique is widely used in various fields such as diagnosis, classification of plants and customer marketing strategies. Then each of the algorithms is briefly described:

A. C5.0 Algorithm

C5.0 algorithms developed from ID3 and C4.5 algorithms is one of the most important and widely used algorithms in data mining. C5.0 tree is a classification tree, which finds an attribute (feature) based on the analysis of the input data, aiming to use it for making decisions on each Node. Since each Node is likely to have different features, all of them will be examined to choose one feature from among, so as to selecting the feature would lead to entropy (disorder) reduction. This process goes on to reach the last Node (Leaf). The algorithm has the capacity to be applied to classify into a decision tree or a set of rules. In many applications, it is preferred to the other rules because the set of rules are easier to understand [21], [22], [23], [24].

B. C&R Tree Algorithm

This algorithm was introduced in 1984 by Leo Breiman, Jerome Friedman, Charles J. Stone, and Richard A. Olshen [25]. Using this algorithm, it is possible to create a decision tree with single-variable binary division. In fact, this algorithm has been developed for quantitative variables but it can also be used for other variables. In this algorithm, the standard Gini coefficient (Gini Index) is used to divide the data into different groups, and it is also possible to use index such as entropy at higher speed. C&R Tree algorithm generates a univariate binary tree. This algorithm can also develop regression tree. From among the weaknesses of this algorithm, we can refer to biased selection of variables and misleading results in qualitative variables with more than two levels.

C. CHAID Algorithm

This algorithm is a type of decision tree developed and introduced by Kass in 1980 [26]. It stands for CHi-squared Automatic Interaction Detection algorithm that can be used for prediction, classification, and also establishing relationship between the various factors. Decision trees usually provide simple and understandable results. One of the advantages of this algorithm is also simplicity of results to understand and interpret. CHAID algorithm can be used for grouped qualitative and quantitative variables. Using three steps of merging, splitting, and stopping which is done iteratively, CHAID Algorithm moves from root Node toward the bottom of tree. At each step, CHAID chooses the best choice to predict and the best choice continues to reach the end of the tree. The algorithm uses p-values to find the best attributes (features) on each Node, so each variable with lower p-values will be considered in the first stage to split the node.

D. QUEST Algorithm

The algorithm was designed and introduced for nominal variables by Loh and Shih in 1997 [27]. The Tree formed by this algorithm has binary division just like C & R Algorithm. This algorithm creates a single variable tree using the linear separation standard. This tree is an upgraded version of FACT tree. The decision criterion for selecting variables in this algorithm concerning F statistics using the P-Value is ANOVA test for quantitative variables and P-Value of chi-square statistic concerning correspondence tables for

qualitative variables. Because of the P-Value for deciding QUEST algorithm does not cause an unbiased tree to be created. The algorithm accuracy is the same as C & R Tree's, but its speed is higher.

E. Dataset

To perform the study, Cleveland heart disease data were used [7], which included 303 records with 13 features as shown in TABLE I. The data have divided into 5 classes from among *class 0* shows lack of heart disease and *class 1 to 4* indicates respectively increasing severity of heart disease. Features of this data are as follows:

TABLE I. CLEVELAND HEART DISEASE DATASET

1. age: age in years [29.0, 77.0]
2. sex: gender (1 = male; 0 = female) [0.0, 1.0]
3. cp: chest pain type[1.0, 4.0]
4. trestbps: resting blood pressure (in mm Hg on admission to the hospital) [94.0, 200.0]
5. chol: serum cholestoral in mg/dl[126.0, 564.0]
6. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false) [0.0, 1.0]
7. restecg: resting electrocardiographic results[0.0, 2.0]
8. thalach: maximum heart rate achieved[71.0, 202.0]
9. exang: exercise induced angina (1 = yes; 0 = no) [0.0, 1.0]
10. oldpeak = ST depression induced by exercise relative to rest[0.0, 6.2]
11. slope: the slope of the peak exercise ST segment[1.0, 3.0]
12. ca: number of major vessels (0-3) colored by flourosopy [0.0, 3.0]
13. thal: 3 = normal; 6 = fixed defect; 7 = reversible defect [3.0, 7.0]
14. num: the predicted attribute: 0=healthy, 1-4: increasingly sick) [0,1,2,3,4]

IV. RESULT

In this paper, the Clementine 12 was used. The specifications for the system to implement as follows: Intel core i7, 3610 QM, 2.3 GHz with 8 GB Installed memory. C5.0, C&R Tree, CHAID QUEST algorithms implemented on heart patients' data with the aim of extracting knowledge underlying in the data under investigation. In this regard, field of diagnosis containing category model label (TABLE II) was considered as the output. The sample data was divided into two groups (70% for training and 30% for testing).

TABLE II: CATEGORY MODEL LABLE

Category Label	Description
0	Patients with no heart failure
1-4	Heart Disease

A. Evaluation

Evaluating the results obtained from C5.0 model causes the model to be improved and usable. There are various parameters such as specificity, sensitivity, precision, and accuracy to evaluate the classification methods, which are calculated in accordance with the following 1 to 4 formula. We can use the confusion matrix to compute the indices. This matrix is a useful tool to analyze the performance of classification method in data detection or observation of various categories. Ideally, a large amount of data relevant to observation must be located on the main diagonal of the matrix, and the remaining values of matrix are zero or near zero [28], [29]. The accuracy of the generated models was calculated to better model for knowledge extraction as shown in TABLE III. Then the accuracy of the generated models to choose better model for knowledge extraction was calculated as shown in TABLE III. Since the model generated by the C5.0 algorithm had the highest accuracy, this model was selected to extract knowledge.

TP = number of positive data labels, which have been properly classified,
FP = number of negative data labels, which have been falsely classified as positive,
FN = number of positive data labels, which have been falsely classified as negative,
TN = number of negative data labels, which have been properly classified,

$$\text{Specificity} = TN / TN+FP \quad (1)$$

$$\text{Sensitivity} = TP / TP+FN \quad (2)$$

$$\text{Precision} = TP / TP+FP \quad (3)$$

$$\text{Accuracy} = TP+TN / TP+TN+FP+FN \quad (4)$$

TABLE IV. shows the values of four indices for each class labels. The values have been calculated for C5.0 algorithm using the confusion matrix. The error rate or misclassification rate can also be calculated based on accuracy (formula 5) index [28], [29].

TABLE III: ACCURCY FOR ALGORITHMS IN THIS STUDY

Algorithm Name	Accuracy (%)
C5.0	85.33
C&R Tree	60.82
CHAID	59
QUEST	59.36

TABLE IV. THE INDICES FOR C5.0 ALGORITHM

Category(Class) Label	Specificity (%)	Sensitivity (%)	Precision (%)	Accuracy (%)
0	54.67	95.12	71.23	76.56
1	95.96	16.36	47.36	81.51
2	89.51	50	39.13	84.81
3	98.88	11.42	57.14	88.77
4	97.58	38.46	41.66	95.04

$$\text{Error Rate} = 1 - \text{accuracy} \quad (5)$$

The average of model accuracy calculated by confusion matrix is 85.338%. Hence, the error rate of the model is 14.662% and this rate indicates a rather good precision and accuracy for the model.

B. Findings

The value of 85.33% for accuracy, 51.30% for precision, 87.32% for specificity, and 42.27% for sensitivity calculated by C5.0 algorithm shows that the existing tree could present comprehensive rules for predicting the future mood of patients. The extracted knowledge or rules have been shown in TABLE V. Since the class 0 is associated with heart disease, if the condition is Cp<= 3 and also if Cp> 3 and Oldpeak <= 0.7, Cp> 3, Oldpeak> 0.7 and Thalach> 168, the result is lack of heart disease. Classes 1 to 4 indicating an increase in symptoms and severity of heart disease in ascending form, is more important. There are symptoms of heart disease in class 1, but they are less severe than in classes 2, 3 and 4. In this class, if Cp> 3, Oldpeak> 0.7, Thalach <= 168, Restecg> 1, Slope <= 2, Oldpeak <= 2.6 and Trestbps <= 142, there are weak signs of heart disease. In class 2, the severity of heart disease is more than that of class 1 If Cp>3, Oldpeak >0.7, Thalach <=168, Restecg<=1 and If Cp>3 and Oldpeak >0.7, Thalach <=168, Restecg>1, Slope<=2 and Oldpeak>2.6, the precision in the evaluation of heart disease is increased, therefore it should be considered more sensitively. Classes 4 and 3, which have a great number of heart diseases, must be seriously taken into consideration.

TABLE V. A SAMPLE OF RULES MADE BY C5.0 ALGORITHM

Category (Class) Label	Rules
0	If Cp<= 3 then ClassLable=0 If Cp>3 and Oldpeak <=0.7 then Classlable=0 If Cp>3 and Oldpeak >0.7 and Thalach >168 then ClassLable=0
1	If Cp>3 and Oldpeak >0.7 and Thalach <=168 and Restecg>1 and Slope<=2 and Oldpeak<=2.6 and Trestbps<=142 then ClassLable=1
2	If Cp>3 and Oldpeak >0.7 and Thalach <=168 and Restecg<=1 then ClassLable=2 If Cp>3 and Oldpeak >0.7 and Thalach <=168 and Restecg>1 and Slope<=2 and Oldpeak>2.6 then ClassLable=2
3	If Cp>3 and Oldpeak >0.7 and Thalach <=168 and Restecg>1 and Slope>2 then ClassLable=3
4	If Cp>3 and Oldpeak >0.7 and Thalach <=168 and Restecg>1 and Slope<=2 and Oldpeak<=2.6 and Trestbps>142 then ClassLable=4

In class 3, if Cp>3, Oldpeak >0.7, Thalach <=168, Restecg>1 and Slope>2, the severity of heart disease is high, and the patients need more care than the previous classes. Class 4, which is the most important and principally the most risky for heart patients, is of extraordinary importance. Patients in this class are likely to be subjected to death. Therefore, If Cp> 3, Oldpeak> 0.7, Thalach <= 168 and Restecg> 1, Slope <= 2, Oldpeak <= 2.6, and Trestbps> 142, the associated patients include in the class with patients having very bad conditions, so it can be concluded that these patients must be under the perfect care.

V. CONCLUSION

This paper examined the factors influencing heart disease using data mining techniques. The data consisted of 303 patients associated with Cleveland heart disease dataset. After implementation of the C5.0, C&R Tree, CHAID, and QUEST algorithms on the data, the C5.0 algorithm with the accuracy of 85.33 percent had the best performance in detection of coronary heart disease causes. Attributes of Cp, Old Peak, Slope, Thalach, Rest Ecg, Trestbps were obtained as important and influential factors in coronary heart disease. One of the main distinguished issues in the study of rules generated by the C5.0 algorithm is paying attention to Thalach attribute value, so that if the value of this factor equals to Thalach> 168, it will belong to the class representing lack of heart disease. However, sifting the rules we made it clear that if Thalach <=168, the attribute value of [71-168] belongs to all classes of patients afflicted with heart disease. This relationship suggests that the attribute can be a great help in detecting patients suffering from heart disease. Another important point was paying attention to the value of Oldpeak attribute in class 4 patients. The value of Oldpeak attribute along with consideration of other attributes was in intervals of (0.7-2.6]. Therefore, after careful examination we can say that this model can serve as an appropriate model to help identify the factors influencing heart patients.

REFERENCES

[1] WHO Report, The Top 10 Causes of Death, Last Accessed 12/9/2013 From Http:// Who.Int/Mediacentre/Factsheets/Fs310/En/ [Accessed 02/03/2015].

[2] V. Paramasivam, T. S. Yee, S. K. Dhillon and A. S. Sidhu, "A methodological review of data mining techniques in predictive medicine: An application in hemodynamic prediction for abdominal aortic aneurysm disease" , Biocybernetics and Biomedical Engineering, Vol. 34, Issue. 3, pp. 139-145, 2014.

[3] K. C. Tan, E. J. Teoh, Q. Yu and K. C. Goh , "A hybrid evolutionary algorithm for attribute selection in data mining" , Expert Systems with Applications, Vol. 36, No. 4, pp. 8616-8630, 2009.

[4] S. Mendis, David Porter, Judith Mackay, Lauren O'Brien." Http://Www.Who.Int/Mediacentre/News/Releases/2004/Pr68/En/" [Accessed 03/04/2015].

[5] J. W. V. Goethe. "Types Of Cardiovascular Disease", ,"Www.Who.Int/Cardiovascular_Diseases./Cvd,[Accessed04.04.2015].

[6] J. Nahar, T. Imam, K. S. Tickle and Y. P. P. Chen, "Association rule mining to detect factors which contribute to heart disease in males and females" , Expert Systems with Applications, Vol. 40, No. 4, pp. 1086-1093, 2013.

[7] KEEL. Cleveland Heart Disease Datset, "A Software Tool To Assess Evolutionary Algorithms For Data Mining Problem",Http://Sci2s.Ugr.Es/Keel/Dataset.Php?Cod=57, [Accessed 02/12/2015].

[8] K. Polat and Salih Güneş, "A hybrid approach to medical decision support systems: Combining feature selection, fuzzy weighted pre-processing and AIRS" , computer methods and programs in biomedicine, Vol. 88, No. 2, pp.164-174, 2007.

[9] R. Alizadehsani, J. Habibi, M. J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian and Z. A. Sani., "A data mining approach for diagnosis of coronary artery disease" , Computer methods and programs in biomedicine, Vol. 111, No. 1, pp. 52-61, 2013.

[10] M. A. M. Abushariah, A. A. M. Alqudah, O. Y. Adwan and R. M. M. Yousef, "Automatic Heart Disease Diagnosis System Based on Artificial Neural Network (ANN) and Adaptive Neuro-Fuzzy Inference Systems

(ANFIS) Approaches." Journal of Software Engineering and Applications, 7, No. 12, pp. 1055-, 2014.

[11] N. Ziasabounchi and I. Askerzade, "ANFIS Based Classification Model for Heart Disease Prediction" ,International Journal Of Electrical & Computer Sciences IJECS-IJENS, Vol. 14, No. 02, pp. 7-12, 2014.

[12] S. Bhatia, P. Prakash and G.N. Pillai, "SVM based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features" , In Proceedings of the World Congress on Engineering and Computer Science, WCECS, San Francisco, USA, pp. 22-24. 2008.

[13] M. Shouman, T. Turner and R. Stocker, "Using decision tree for diagnosing heart disease patients" , In Proceedings of the Ninth Australasian Data Mining Conference-Volume 121, Australian Computer Society, Inc., pp. 23-30, 2011.

[14] T. Nguyen, A. Khosravi, D. Creighton and S. Nahavandi, "Classification of healthcare data using genetic fuzzy logic system and wavelets" , Expert Systems with Applications, Vol. 42, No. 4, pp. 2184-2197, 2015.

[15] H. D. Masethe and M. A. Masethe, "Prediction of Heart Disease using Classification Algorithms." In Proceedings of the World Congress on Engineering and Computer Science, WCECS, San Francisco, USA, Vol. 2. 2014.

[16] J. Soni, U. Ansari, D. Sharma and S. Soni ,"Predictive data mining for medical diagnosis: An overview of heart disease prediction", International Journal of Computer Applications, Vol. 17, No. 8, pp. 43-48, 2011.

[17] C. S. Dangare and S. S. Apte. ,"A data mining approach for prediction of heart disease using neural networks." International Journal of Computer Engineering and Technology (IJCET), Vol. 3, No. 3, 2012.

[18] U. R. Acharya, S. VSree, M. M. R. Krishnan, N. Krishnananda, S. Ranjan, P. Umesh, and J. S. Suri, "Automated classification of patients with coronary artery disease using grayscale features from left ventricle echocardiographic images" , Computer methods and programs in biomedicine, Vol. 112, No. 3, pp. 624-632, 2013.

[19] N. Esfandiari, M. R. Babavalian, A-M E. Moghadam and V. Kashani Tabar, "Knowledge discovery in medicine: Current issue and future trend" , Expert Systems with Applications, Vol. 41, No. 9, pp. 4434-4463, 2014.

[20] K. E. Walker and S. M. Crotty ,"Classifying high-prevalence neighborhoods for cardiovascular disease in Texas" , Applied Geography, Vol. 57, pp. 22-31, 2015.

[21] J. R. Quinlan, "Induction of decision trees" , Machine learning, Vol. 1, No. 1, pp. 81-106, 1986.

[22] J. R. Quinlan , "C4. 5*:* programs for machine learning" , Elsevier, 2014.

[23] J. R. Quinlan , "Bagging, boosting, and C4. 5" , In *AAAI/IAAI, Vol. 1*, pp. 725-730. 1996.

[24] J. R. Quinlan , "C5", Http://Rulequest.Com, 2007.

[25] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, "Classification and regression trees, CRC press, 1984.

[26] G. V. Kass ,"An exploratory technique for investigating large quantities of categorical data" , Applied statistics, pp. 119-127, 1980.

[27] W-Y. Loh and Y-S. Shih. "Split selection methods for classification trees" , Statistica sinica, Vol. 7, No. 4, pp. 815-840, 1997.

[28] S. Alizadeh, M. Ghazanfari, and B. Teimorpour, "Data Mining and Knowledge Discovery" , Publication of Iran University of Science and Technology, 2011.

[29] J. Han, M. Kamber, and J. Pei. Data mining: concepts and techniques: concepts and techniques. Elsevier, 2011.

Moloud Abdar. He received his Undergraduate degree in Computer Engineering from the University of Damghan, Iran in 2015. He has more than 8 conference and journal papers about the Data Mining. Currently, His research interests include data mining, web and text mining, artificial intelligence and image processing.