Diagnosis of Coronary Artery Disease Using Data Mining Based on Lab Data and Echo Features

Roohallah Alizadehsani, Jafar Habibi, Zahra Alizadeh Sani, Hoda Mashayekhi, Reihane Boghrati, Asma Ghandeharioun, and Behdad Bahadorian

Abstract—**According to American Heart Association report, cardiovascular diseases are one of the five leading causes of death in the world. Coronary Artery Disease (CAD) is the most common fatal heart disease, and is the subject of large body of studies. According to prevalence of CAD, early diagnosis of this disease is very important. The most reliable method for CAD diagnosis is angiography, but it is costly, time-consuming, and hazardous. Therefore in order to predict such diseases, study of non-invasive methods such as analysis and mining of patients' medical information is becoming popular, and has proved to be effective. Unfortunately, majority of approaches in the literature rely on a limited and small set of medical features for disease detection. This paper aims to examine effects of set of features; including lab data and echo information on CAD diagnosis which some of them were not considered in previous studies. The data set consists of the information gathered from 303 random visitors to Tehran's Shaheed Rajaei Cardiovascular, Medical and Research Center which is one of the largest heart hospitals in Asia. The method used in this research was data mining. Several classification algorithms were adopted to analyze the data set, including SMO, Naïve Bayes, C4.5 and AdaBoost. According to the comprehensive set of features used, the obtained classification accuracy exceeded 82 percent. Results showed that new added features including Region with RWMA and Ejection Fraction (EF) have a large effect on CAD.**

Index Terms— **Coronary artery disease, data mining, naïve bayes, sequential minimal optimization (SMO), adaboost.**

I. Introduction

Cardiovascular diseases entail a large number of deaths in the world annually. The most common type of them is CAD which is the reason of about 1/3 of deaths [1]. Angiography determines number, location and rate of blocked coronary arteries. This information helps to find the proper treatment. However, angiography is an invasive method which is costly, time-consuming, and somehow dangerous.

While angiography can accurately decide on the disease status, many other medical features can be used to approximately predicate existence of CAD. Determining and analyzing these features, however, is less costly and often previously available by different medical examinations taken from the patient. Among the non-invasive methods, data mining has attracted a lot of attention in the literature. Data mining methods can extract hidden knowledge from existing data.

Data mining is one of the ten developing fields in knowledge which is likely to make a revolution in technology in the next decades. Nowadays this technology is applied in various fields.

In the context of predicting heart diseases, data mining methods have proved to be safe and effective in building analytical models based on the patient's medical information, which can later be used to analyze new cases. Based on the constructed models, the disease can be diagnosed with relatively acceptable values of accuracy.

A lot of researchers have attempted to apply different data mining methods to diagnose CAD. The most important data sets used in different proposals are based on Doppler signals [2-4], and heartbeat [5].

Many available proposals, however, use the existing data sets which consist of limited and few features. Setiawan *et al.* [6] used KNN and decision tree on UCI dataset [7] which has 13 features per sample. Srinivas *et al.* [8] used Bayesian methods and C4.5 on the same data set.

Tian-hua *et al.* [5] investigated the impact of extracted features from heartbeat on the disease, and Latifoglu et al. [9] used Doppler signals to diagnose of the disease.

To the best of our knowledge, the impact of some lab data and echo information on CAD has not been examined yet. In this paper, effects of a broad set of features on CAD diagnosis were explored. Some of effective features considered are made up of Hb, WBC, Proteinuria, HDL, LDL, VHD, Region with RWMA, and TG which have been not studied in the previous works.

Also, the data used in this research has gathered from 303 random visitors to Tehran's Shaheed Rajaei Cardiovascular, Medical and Research Center from which 87 were healthy. Several classification algorithms were adopted to analyze the data set including SMO [10], Naïve Bayes [11], C4.5 [12] and AdaBoost [13], and evaluated using 10-fold cross validation.

II. Material and Method

In this section, first the feature set collected per each examined visitor is described. Afterwards, the employed classification methods are elaborated.

A. Features

The features used in this paper, include demographical and

Manuscript received July 5, 2012; revised July 30, 2012.

Roohallah Alizadehsani, Jafar Habibi, Hoda Mashayekhi, Reihane Boghrati, and Asma Ghandeharioun are with the Software Engineering, Department of Computer Engineering, Sharif University of Technology, Tehran, Iran.

Zahra Alizadeh Sani and Behdad Bahadorian are with the Tehran University of Medical Science, Tehran, Iran (e-mail: d_zahra_alizadeh@yahoo.com; Tel.: +989153160452; fax: +982177952092).

doi: 10.12720/jomb.1.1.26-29

laboratory data. The details of the features along with valid ranges are presented in Tables I and II.

TABLE I: DEMOGRAPHICAL FEATURES

Biographical Feature	Range
Age	30-86
Weight	48-120
Sex	Male, Female
BMI (Body Mass Index Kg/m^2)	18-41
DM (Diabetes Mellitus)	Yes, No
HTN (History of hypertension)	Yes, No
Current Smoker	Yes, No
Ex-Smoker	Yes, No
FH (Family History)	Yes, No
Obesity	Yes if MBI>25, No otherwise
CRF (Chronic Renal Failure)	Yes, No
CVA (Cerebrovascular Accident)	Yes, No
Airway Disease	Yes, No
Thyroid Disease	Yes, No
CHF (Congestive Heart Failure):	Yes, No
DLP (Dyslipidemia):	Yes, No

TABLE II: LABORATORY AND ECHO FEATURES

Laboratory Features	Range
FBS (Fasting Blood Sugar)	62- 400
Cr (creatine)	0.5- 2.2
TG (Triglyceride):	37- 1050
LDL (Low density lipoprotein)	18- 232
HDL (High density lipoprotein)	15- 111
BUN (Blood Urea Nitrogen)	6- 52
ESR (Erythrocyte Sedimentation rate)	1- 90
Hb (Hemoglobin)	8.9- 17.6
K (Potassium)	3.0- 6.6
Na (Sodium)	128- 156
WBC (White Blood Cell)	3700- 18000
Lymph (Lymphocyte)	7- 60
Neut (Neutrophil)	32- 89
PLT (Platelet)	25- 742
EF (Ejection Fraction)	15- 60
Region with RWMA (Regional Wall Motion Abnormality)	0,1,2,3,4
VHD (Valvular Heart Disease)	Normal, Mild, Moderate, Severe
Cath(Class Attribute)	CAD if diameter narrowing ≥50%, Normal otherwise

In features of Table I, HTN is history of hypertension, and DM is history of Diabetes Mellitus. Current Smoker denotes the current consumption of cigarettes, while Ex-Smoker is history of the previous consumption of cigarettes. FH is history of heart disease in first-degree relatives.

Discrete of continuous range of features may enhance accuracy of data mining algorithms. For this purpose range of some features in Tables I and II break into three intervals: low, normal and high, extracted from Braunwald Heart Disease [1], shown in Table III. New features were also considered and distinguished by index 2.

TABLE III: LEVEL OF FEATURES

Feature	Low	Normal	High
Cr2	Cr<0.7	0.7≤Cr≤1.5	Cr>1.5
FBS2	FBS<70	70≤FBS≤105	FBS>105
LDL2		LDL≤130	LDL>130
HDL2	HDL<35	HDL>=35	-
BUN2	BUN<7	7≤BUN≤20	BUN>20
ESR2		if male & ESR≤age/2 or if female & ESR≤age/2+5	if male & ESR>age/2 or if female & ESR>age/2+5
Hb2	if male & Hb<14 Or If female & Hb<12.5	if male & 14≤Hb≤17 or if female & 12.5≤Hb<=15	if male & Hb>17 or if female & Hb>15
K2	K<3.8	3.8≤K≤5.6	K>5.6
Na2	Na<136	136≤Na≤146	Na>146
WBC2	WBC<4000	4000≤WBC≤11000	WBC>11000
PLT2	PLT<150	150≤PLT≤450	PLT>450
EF2	EF≤50	EF>50	
Region with RWMA2	-	Region with RWMA=0	Region with RWMA≠0
Age2[1]		if male & age≤45 or if female & age≤55	if male & age>45 or if female & age>55

B. Algorithms

Four classification algorithms were used to analyze the data set and built decision models for future prediction. In the subsequent sections, the feature selection method is described followed by the data mining algorithms used to analyze the data set.

1) C4.5 algorithm

C4.5 is one of the decision tree algorithms. It uses pruning, gain ratio criterion, and also is able to manage continuous data. C4.5 was presented to improve the ID3 algorithm.

2) Naïve bayes algorithm

Naïve Bayes is based on probabilities. In applications such as text classification and medical diagnosis, this method has high efficiency. This method assumes a simple premise that features values are independent. It essentially uses the Bayes formula.

3) Adaboost

The AdaBoost algorithm is one of the ensemble algorithms whose result is expressed based on multiple classifier voting. The base classifier used in this study is Naïve Bayes. In each round of the algorithm, the weight of each incorrectly classified sample is increased, while the weight of the correctly classified samples is reduced. This causes the classifier to focus on samples which are harder to determine, in order to learn examples.

4) SMO

The SMO algorithm implements John Platt's sequential minimal optimization algorithm for training a support vector classifier. This implementation globally replaces all missing values and transforms nominal features into binary ones. It also normalizes all features by default. Multi-class problems are solved using pair wise classification Feature Selection.

5) Feature selection

The information gain criterion used to select the features, was that, the higher this value for a feature, the better it separates the classes. Consequently 16 features with the highest information gain were selected.

6) Association rule

To produce association rules, first the frequent item set were extracted. From these patterns, rules with a predefined confidence value were extracted. Confidence is the probability of finding the right hand side of the rule in transactions under the condition that these transactions also

[1] Given that women under 55 years and men under 45 years are less affected by CAD, the range of age is partition at these values

contain the left hand side. To make laws, all features should be bi-value. Therefore, features that had low, normal and high, low and high as a group were considered. And only some of the features that had two value and index 2 were used.

III. Result and Discussion

RapidMiner is one of the open-source systems for data mining. It is available as a stand-alone application for data analysis and as a data mining engine for the integration into own products. In this study, available algorithms in RapidMiner have been used.

A. Performance Measure

Accuracy, sensitivity and specificity were used to compare the algorithms. Accuracy is the ratio of correctly diagnosed samples to all. Sensitivity and specificity are the ratio of correctly diagnosed CAD and normal samples respectively. Receiver Operating Characteristics (ROC) compares the algorithms in a way that the more the area under the curve is, the better the algorithm works. In this section the evaluation results are presented and discussed. Four algorithms are compared according to their ROC in Fig. 1.

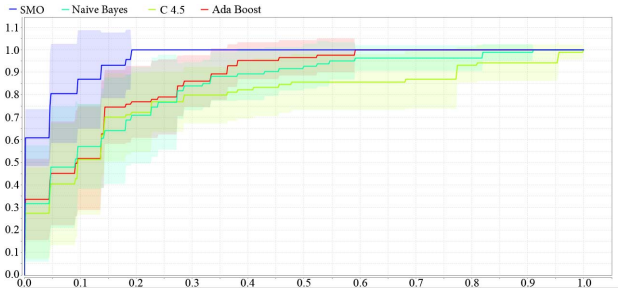


Fig. 1. ROC diagram for four algorithms.

In Fig.1 the blue, red, green, and olive lines show the SMO, AdaBoost, Naïve Bayes and C4.5 models. As Fig. 1 shows the best performance is related to SMO because of the wider area it involves under its curve.

TABLE IV: Information Gain for Selected Features

Feature	Weight
Region RWMA	1
EF2	0.832
Age	0.806
HTN	0.589
DM	0.516
FBS2	0.389
VHD	0.356
BMI	0.158
WBC	0.093
HDL	0.088
Hb	0.057
TG2	0.055
Current Smoker	0.039
Sex	0.031
LDL2	0.009
BUN2	0.002

The information gain of features is shown in Table IV. Effects of features on CAD were determined by information gain. As shown in Table IV the greatest impact respectively is related to the Region RWMA, EF2, Age, HTN, DM, and FBS2 with respect to the information gain. In Table V the accuracy of the algorithms after feature selection have been compared. In Table VI the accuracy of the algorithms with all features has been compared.

TABLE V: Comparing the Accuracy of Algorithms with Selected Features

Algorithm Used	Accuracy	Sensitivity	Specificity
Naïve Bayes	74.89%±9.34%	72.22%	81.61%
C4.5	78.23%±4.09%	87.50%	55.17%
AdaBoost	76.86%±5.88%	78.70%	72.41%
SMO	82.16%±5.45%	90.74%	60.92%

TABLE VI: Comparing the Accuracy of Algorithms with All Features

Algorithm Used	Accuracy	Sensitivity	Specificity
Naïve Bayes	72.59% ± 9.28%	71.30%	75.86%
C4.5	74.24% ± 4.90%	84.26%	49.43%
AdaBoost	77.89% ± 8.54%	86.11%	57.47%
SMO	79.86% ± 5.88%	87.96%	59.77%

As shown in Table V, the highest accuracy is related to the SMO. C4.5, AdaBoost and Naïve Bayes have the highest accuracy respectively. Also Srinivas *et al.* [8] reached 82% accuracy. The accuracy achieved by all algorithms increased about 2 percent by using a subset of the total data (Tables V and VI), but AdaBoost was not increased.

The rules were produced by the association rule and are given below in decreasing support order. The confidence in the all rules is 1. In these rules, C represents the Confidence and S represents the Support per each rule which were explained in section II.B.6.

1) [EF <55, HTN=True, FBS= High or Low, BUN= High or Low]=>[CAD], *S*=0.0495627, *C*=1
2) [EF<55, FBS= High or Low, DM=True, BUN= High or Low]=>[CAD], *S*=0.0466472, *C*=1
3) [Age = Old, EF <55, DM=True, BUN= High or Low]=>[CAD], *S*=0.0408163, *C*=1
4) [Age = Old, Obesity = True, Sex = Female, LDL = High]=>[CAD], *S*=0.0408163, *C*=1
5) [EF <55, HTN=True, Sex = Female, HDL= Low]=>[CAD], *S*=0.0379009, *C*=1
6) [EF <55, Hb= High or Low, FBS= High or Low, BUN= High or Low]=>[CAD], *S*=0.0349854, *C*=1
7) [Age = Old, EF <55, Sex = Female, LDL = High]=>[CAD], *S*=0.0349854, *C*=1
8) [HTN=True, Hb= High or Low, Current Smoker = True]=>[CAD], *S*=0.0349854, *C*=1
9) [Sex = Female, FBS= High or Low, DM=True, BUN= High or Low]=>[CAD], *S*=0.03207, *C*=1
10) [HTN=True, Sex = Female, DM=True, BUN= High or Low]=>[CAD], *S*=0.0291545, *C*=1
11) [Age = Old, HTN=True, BUN= High or Low, LDL = High]=>[CAD], *S*=0.0291545, *C*=1
12) [Hb= High or Low, FBS= High or Low, DM=True, BUN= High or Low]=>[CAD], *S*=0.0262391, *C*=1
13) [HTN=True, Hb= High or Low, DM=True, BUN= High or Low]=>[CAD], *S*=0.0262391, *C*=1
14) [EF <55, FBS= High or Low, HDL= Low, BUN= High or Low]=>[CAD], *S*=0.0262391, *C*=1
15) [EF <55, Sex= Female, DM=True, BUN= High or Low]=>[CAD], *S*=0.0262391, *C*=1

16) [DM=True, BUN= High or Low, TG= High]=>[CAD], *S*=0.0262391, *C*=1
17) [EF <55, TG = High, Current Smoker = True]=>[CAD], *S*=0.0262391, *C*=1
18) [Hb= High or Low, FBS= High or Low, HDL= Low, BUN= High or Low]=>[CAD], *S*=0.0233236, *C*=1
19) [HTN=True, FBS= High or Low, BUN= High or Low, TG= High]=>[CAD], *S*=0.0233236, *C*=1
20) [EF <55, Hb= High or Low, DM=True, BUN= High or Low]=>[CAD], *S*=0.0233236, *C*=1

IV. Conclusion

In this study the impact of new features like Hb, WBC, Proteinuria, HDL, LDL, VHD, Region with RWMA, and TG on CAD, which were not examined in previous studies, was considered. Table IV is concluded the high impact of the features Region RWMA, and EF2 on CAD.

Findings showed that use of classification can predict the existence of disease with high accuracy and low cost overheads. Results of this study can be used for early detection of disease and reduction of mortality and treatment costs. Several data mining algorithms were used. Initially, a subset of the data was selected and this increases the accuracy about 2 percent. Many other potentially influential features exist, that the evaluation of their impact on CAD remains as future work.

References

[1] R. O. Bonow, D. L. Mann, D. P. Zipes, and P. Libby, "Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine," *9th Edition*: New York, Saunders, 2012.

[2] E. Çomaka, A. Arslana, and I. Türkoˇ glub, "A decision support system based on support vector machines for diagnosis of the heart valve diseases," *Computers in Biology and Medicine*, pp. 21 – 27, 2007.

[3] H. Uguz, A. Arslan, anad I. Turkoglu, "A biomedical system based on hidden Markov model for diagnosis of the heart valve diseases," *Pattern Recognition Letters*, pp. 395–404, 2007.

[4] E. Avci and I. Turkoglu, "An intelligent diagnosis system based on principle component analysis and ANFIS for the heart valve diseases," *Expert Systems with Applications*, pp. 2873–2878, 2009.

[5] C. Tian-hua, X. Su-xia, G. Pei-yuan, and Y. Zhen, "The Research of Non-invasive Method of Coronary Heart Disease Based on Neural Network and Heart Sound Signals," *Information Engineering and Computer Science*, *IEEE*, pp.1-4, 2009.

[6] N. A. Setiawan, P. A. Venkatachalam, and A. Fadzil M. H., "Rule Selection for Coronary Artery Disease Diagnosis Based on Rough Set," *International Journal of Recent Trends in Engineering*, vol. 2, PP. 198-202, 2009.

[7] "UCI KDD Archive," [online]. Available from http://mlearn.ics.uci.edu/MLSummary.html

[8] K. Srinivas, G. Raghavendra Rao, and A. Govardhan, "Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques," *The 5th International Conference on Computer Science and Education*, China, pp. 1344-1349, 2010.

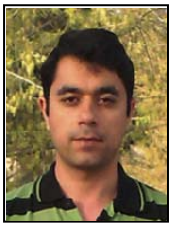
[9] F. Latifoglu, H. Kodaz, S. Kara, and S. Güne¸, "Medical application of Artificial Immune Recognition System (AIRS): Diagnosis of atherosclerosis fromcarotid arteryDoppler signals," *Computers in Biology and Medicine*, pp. 1092 – 1099, 2007.

[10] J. C. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," *Technical Report MSR-TR-98-14*, Microsoft Research, 1998.

[11] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," *Proceedings of the 23rd international conference on Machine learning*, pp. 161 – 168, 2006.

[12] J. R. Quinlan, "Improved use of continuous attributes in c4.5," *Journal of Artificial Intelligence Research*, vol. 4, pp. 77-90, 1996.

[13] T. Zhang, "Statistical behavior and consistency of classification methods based on convex risk minimization," *Annals of Statistics*, vol. 32, pp. 56-85, 2004.



Roohallah Alizadehsani received his B.S. degree in computer engineering from Sharif University of Technology. In present, M. S. student in computer engineering at the Sharif University of Technology. His research interests are mainly in the areas of data mining, machine learning, and evaluation of computer systems performance.

Dr. Jafar Habibi received his B.S. degree in computer engineering from the Supreme School of Computer, his M. S. degree in industrial engineering from Tarbiat Modares University and his Ph.D. degree in Computer engineering from Manchester University. At present, he is an assistant professor and the head of computer engineering department at Sharif University of Technology. He is supervisor of Sharif's RoboCup Simulation Group. His research interests are mainly in the areas of computer engineering, simulation systems, MIS, DSS and evaluation of computer systems performance.

Zahra Alizadeh Sani received her MD degree and cardiology specialty from the Mashhad University of Medical science, Echocardiographic fellowship degree in Tehran University of Medical science and cardiac MRI fellowship from SCMR(society of cardiac MRI). At present, she is an assistant professor and the head of cardiac MRI department at Tehran University ,Rajaei research and medical cardiovascular center .Her research interests are mainly in the areas of diagnosis of cardiovascular disease and cardiovascular imaging.

Hoda Mashayekhi is currently a Ph.D. candidate in department of computer engineering at Sharif University of Technology. She received her B.Sc. and M.Sc. degrees from the same university in the field of computer engineering. Her research interests include parallel and distributed computing, data mining and decision making, peer-to-peer networks and semantic structures.

Reihane Boghrati is B.S. student in computer engineering at the Sharif University of Technology. Her research interests are mainly in the areas of data mining and human-computer interaction.

Asma Ghandeharioun is currently a B.S. student in computer engineering at Sharif University of Technology. Her interest in algorithmic reasoning led her to take part in several computer vision, machine learning, and data mining seminars and semi-projects. On the other hand, human-computer interaction motivated her to participate in design projects, e.g: software development, game, and web design.
She has won a silver national medal in Informatics Olympiad, 2008, Tehran, Iran. Also she is recognized as talented student by the Iranian Elites Foundation, receiving grant for undergraduate studies.

Behdad Bahadorian was studying medicine in Tehran University of Medical Sciences from 1997 to 2005 and graduation with MD degree in 2005. He is currently a cardiology resident in Rajaee Heart Center. He is the sole author or co-author in five books on teaching English for high-school students between 1999 and 2003 as a part-time job while studying medicine. He published one article in "*Heart Asia*" journal and also has another one under publication.