Original Article

www.jmss.mui.ac.ir

Diagnosis of Coronary Arteries Stenosis Using Data Mining

Roohallah Alizadehsani, Jafar Habibi, Behdad Bahadorian[1], Hoda Mashayekhi, Asma Ghandeharioun, Reihane Boghrati, Zahra Alizadeh Sani[1]

Department of Computer Engineering, Sharif University of Technology, [1]Rajaie Cardiovascular Medical and Research Center, Tehran University of Medical Science, Tehran, Iran

Submission: 21-06-2012 Accepted: 30-07-2012

ABSTRACT

Cardiovascular diseases are one of the most common diseases that cause a large number of deaths each year. Coronary Artery Disease (CAD) is the most common type of these diseases worldwide and is the main reason of heart attacks. Thus early diagnosis of CAD is very essential and is an important field of medical studies. Many methods are used to diagnose CAD so far. These methods reduce cost and deaths. But a few studies examined stenosis of each vessel separately. Determination of stenosed coronary artery when significant ECG abnormality exists is not a difficult task. Moreover, ECG abnormality is not common among CAD patients. The aim of this study is to find a way for specifying the lesioned vessel when there is not enough ECG changes and only based on risk factors, physical examination and Para clinic data. Therefore, a new data set was used which has no missing value and includes new and effective features like Function Class, Dyspnoea, Q Wave, ST Elevation, ST Depression and Tinversion. These data was collected from 303 random visitor of Tehran's Shaheed Rajaei Cardiovascular, Medical and Research Centre, in 2011 fall and 2012 winter. They processed with C4.5, Naïve Bayes, and k-nearest neighbour (KNN) algorithms and their accuracy were measured by tenfold cross validation. In the best method the accuracy of diagnosis of stenosis of each vessel reached to 74.20 ± 5.51% for Left Anterior Descending (LAD), 63.76 ± 9.73% for Left Circumflex and 68.33 ± 6.90% for Right Coronary Artery. The effective features of stenosis of each vessel were found too.

Key words: *C4.5 Algorithm, coronary artery disease, data mining, feature, KNN algorithm, Naïve Bayes algorithm*

INTRODUCTION

Large data sets are idle in database of companies, universities, etc. Using the hidden information in these databases is based on efficient management. Data mining is looking for hidden relationships in databases. This process is more than just a simple retrieving data and lets researchers to find new information from data. One of the most important algorithms of data mining is classification which is used when the tag of samples is obvious.

Cardiovascular diseases are one of the most common diseases which cause a large number of deaths each year. The most common type of these diseases is Coronary Arteries Disease (CAD).[1] This disease makes Coronary arteries hard and tight. Due to its danger and causing of about 1/3 deaths in the world,[2] early diagnosis of CAD is very vital.

The best way to diagnose heart vessels stenosis is angiography. Because of its complications, researchers are looking for alternative methods. A lot of studies have been done so far. They have tried to diagnose CAD by using data mining methods and collecting features based on noninvasive methods.

In the studies that have been done so far, stenosis of the LAD, Left Circumflex (LCX) and Right Coronary Artery (RCA) vessels have been rarely examined separately. The majority of papers consider just being CAD or normal. The ones whom LAD, LCX or RCA vessel is clogged are considered as CAD patients, others as healthy. As is said, most of the papers have not examined and analyzed stenosis of vessels separately and the effective features on them.

Babaoglu *et al.*[3] used exercise test data and neural network algorithm to diagnose stenosis of each vessel separately and reached 73%, 64.85% and 69.39% accuracy for LAD, LCX and RCA vessels respectively.

Srinivas *et al.*[4] examined stenosis of each vessel separately and find some rules for them.

Sony *et al.*[5] used genetic algorithm and decision tree and find some rules for stenosis of each vessel.

Address for correspondence:
Dr. Zahra Alizadeh Sani, Unit 2, Second Floor, IsatisTower, Shahid Ansari ST, Valieasr Blvd,Tehran, Iran. E-mail: dr_zahra_alizadeh@yahoo.com

Ordonez *et al*.[6] put some constraints on association rule and find some rules for stenosis of each vessel.

Some papers have worked on diagnosing CAD and achieved 52.33%, about 90%, 70% and 75% accuracy respectively.[7-10] But they have not considered stenosis of each vessel separately.

In this study, the effect of some new features which don't consider in previous studies is investigated on stenosis of LAD, LCX and RCA vessels. For this purpose, C4.5,[11] Naïve Bayes,[12] and k-nearest neighbour (KNN) algorithms were used.

In next sections data set, Algorithms used in this study are described, and results are given. In last section the study is concluded and is mentioned about some future work.

Used Medical Data Set

The data set which is used in this study is collected from visitors with chest pain to the Shaheed Rajaei hospital. They were suspicious of having CAD and were volunteers for angiography.

ECG abnormality is not common among CAD patients. Also it is easy to diagnose stenosed coronary artery when significant ECG abnormality exists. The aim of this study is to find a way for specifying the lesioned vessel when there is not enough ECG changes and only based on Demographic, Symptom, Examination information and ECG Features. The class feature is Cath which is determined by angiography. The class value is Cad when diameter narrowing is greater or equal to 50 and Normal otherwise. The features along with their valid ranges in data set are given in Tables 1-3.

In the above features, HTN identifies history of Hypertension, DM is history of Diabetes Mellitus, Current Smoker is current consumption of cigarettes, Ex-Smoker is history of previous consumption of cigarettes and FH is history of heart disease in first-degree relatives.

MATERIALS AND METHODS

Three classification algorithms were used to analyse the data set. In all algorithms, default setting of RapidMiner was used. In the subsequent sections, the data mining algorithms used to analyze the data set is described.

C 4.5

One of the best decision tree algorithms is C4.5. This algorithm can manage continuous data. It uses pruning algorithms to increase accuracy and Gain Ratio to select features. One of the pruning algorithms which is used by C4.5 is reduce error pruning and it increases accuracy of the algorithm. One of its factors is M. it shows the minimum instances that a leaf should have. C means the confidence threshold which is considered for pruning. By changing these two factors the accuracy of algorithm is changed.

Naïve Bayes

One of the Bayesian methods is Naïve Bayes. All of these algorithms use Bayes formula:

Table 1: Demographic features

Demographic features	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 (Hyper tension)	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 (Dyslipidaemia)	Yes, No

Table 2: Symptom and examination features

Symptom and examination features	Range
BP (Blood pressure)	90-190
PR (Pulse rate)	50-110
Edema	Yes, No
Weak peripheral pulse	Yes, No
Lung rales	Yes, No
Systolic murmur	Yes, No
Diastolic murmur	Yes, No
Typical chest pain	Yes, No
Dyspnoea	Yes, No
Function class	1, 2, 3, 4
Atypical	Yes, No
Non anginal CP	Yes, No
Exertional CP (Exertional chest pain)	Yes, No
LowThAng (low threshold angina)	Yes, No

Table 3: ECG features

ECG features	Range
Rhythm	Sin, AF
Q wave	Yes, No
ST elevation	Yes, No
ST depression	Yes, No
Tinversion	Yes, No
LVH (Left ventricular hypertrophy)	Yes, No
Poor R Progression (Poor R wave progression)	Yes, No

$$P(A|B) = \frac{P(B|A)*(PA)}{P(B)} \quad (1)$$

The hypothesis of this algorithm is independency of features. Thus when some of features depend on each other, this algorithm may work badly.

KNN

KNN is a method for classifying instances based on closest training examples so that an instance is classified base on majority vote of its neighbours. It is a lazy learning because all computation is deferred until classification. K is one of the KNN factors which shows considered neighbours for an instance for determining its label. By changing the value of K, the accuracy is changed.

EXPERIMENTAL RESULTS

To apply the data mining algorithms, the RapidMiner tool is used.[13] RapidMiner is a tool for experimenting with machine learning and data mining algorithms. In this study the default setting of RapidMiner is used, the difference is in C4.5 algorithm where C and M was set to 0.4 and 11 respectively.

Performance Measure

For measuring the performance of algorithms, Accuracy, Sensitivity, and Specificity were used because these three criteria are used more in the medical field.

Confusion matrix

For calculation of Sensitivity, Specificity and accuracy confusion matrix is required.

In Table 4 Confusion Matrix is shown.

In confusion matrix:

- Actual class is the class which determined by angiography and it is existed in dataset. Predicted class is the one which is predicted by algorithms
- TP is number of samples of class C1 which has been correctly classified
- TN is number of samples of class C2 which has been correctly classified
- FN is number of samples of class C1 which has been falsely classified as C2
- FP is number of samples of class C2 which has been falsely classified as C1

Sensitivity, specificity and accuracy

According to Confusion Matrix, Sensitivity, Specificity and Accuracy are calculated as follows:

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

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

$$Accuracy = \frac{(TN + TP)}{(TN + TP + FN + FP)}$$

ROC

It is a diagram for comparing performance of algorithms.

It is created by True Positive Rate (TPR) vs. False Positive Rate (FPR).

The more the area under ROC curve is, the higher the performance of the algorithm is. FPR and TPR are explained as follows:

$$FPR = \frac{FP}{(FP + TN)}$$

$$TPR = \frac{TP}{(TP + FN)}$$

Evaluation Results

The evaluation results are presented in the next subsections.

Output

For getting results, K is set to 3 for KNN algorithm and M and C are set to 11 and 0.4 respectively for C4.5 algorithm.

The highest accuracy in this study is related to these values of K, M, and C for KNN and C4.5 algorithms.

The performance measures for different algorithms to diagnose stenosis of LAD are represented in Table 5.

As shown in Table 5, the highest accuracy is related to

Table 4: Confusion matrix

Predicted class	Actual class	
	C1	C2
C1	True positive (TP)	False positive (FP)
C2	False negative (FN)	True negative (TN)

Table 5: Performance of classification algorithms for diagnosing stenosis of left anterior descending vessel

Algorithm used	Accuracy %	Sensitivity %	Specificity %
Naïve Bayes	51.81±5.02	19.77	96.83
C 4.5	74.20±5.51	83.62	61.11
k-nearest neighbour	59.65±9.53	65.54	51.59

C4.5 which is 74.20%. Notice that the Sensitivity values are higher than Specificity values in C4.5 and KNN. These two methods are more capable of predicting the CAD samples in comparison to normal ones unlike Naïve Bayes.

As far as we know, the best accuracy for diagnosing stenosis of the vessels separately was reached by[3] which is 73% for the LAD vessel stenosis and it is lower than the accuracy of this study.

The performance measures for different algorithms to diagnose LCX vessel stenosis are represented in Table 6.

As shown in Table 6, the C4.5, Naïve Bayes, and KNN methods achieved nearly the same accuracy values, which is above 61%. Also in this Table, C4.5 in comparison to Naïve Bayes and KNN is more capable of predicting the normal samples.

In[3] the accuracy for diagnosing LCX vessel stenosis was reached 64.85% which is nearly the same as this article.

The performance measures for different algorithms to diagnose RCA vessel stenosis are represented in Table 7.

As shown in Table 7 the C4.5 and Naïve Bayes methods achieved nearly the same accuracy values, which is above 68%. However, KNN algorithm has lower accuracy.

In[3] the accuracy for diagnosing stenosis of RCA vessel was reached 69.39% which is nearly the same as this article.

As shown in Tables 5-7 diagnosing stenosis of LAD vessel is easier than others and diagnosing stenosis of RCA vessel is easier than LCX vessel.

The information gain was used to determine the effect of features on these three vessels. A feature will get a higher weight if one of its values causes CAD and other one causes Normal. The results are shown in Tables 8-10.

As shown in Table 8, Typical Chest Pain, Atypical, Age, ST Elevation, Nonanginal, Tinversion, Q Wave, BP, PR, HTN, ST Depression and Diastolic Murmur have the highest influence on stenosis of LAD vessel respectively.

As shown in Table 9, Age, Typical Chest Pain, BP, HTN, Atypical, DM, Weight, length, BMI, EX-Smoker, Function class, PR, Sex, and Non anginal have the highest influence on stenosis of LCX vessel respectively.

As shown in Table 10, Typical Chest Pain, DM, Age, Atypical, Poor R Progression, Function Class, Non anginal, HTN, length, PR, BP, Weight, Sex, Q wave, Diastolic Murmur and Dyspnea have the highest influence on stenosis of RCA vessel respectively.

Table 6: Performance of classification algorithms for diagnosing left circumflex vessel stenosis

Algorithm used	Accuracy %	Sensitivity %	Specificity %
Naïve Bayes	62.73±6.18	75.54	42.86
C 4.5	63.76±9.73	67.39	57.98
k-nearest neighbour	61.39±9.82	70.11	47.90

Table 7: Performance of classification algorithms for diagnosing right coronary artery vessel stenosis

Algorithm used	Accuracy %	Sensitivity %	Specificity %
Naïve Bayes	67.29±4.50	47.37	79.37
C 4.5	68.33±6.90	51.75	78.31
k-nearest neighbour	59.11±8.57	34.21	74.07

Table 8: Information gain of features for left anterior descending

Feature	Weight
Typical chest pain	1
Atypical	0.573
Age	0.389
ST elevation	0.235
Non anginal	0.233
Tinversion	0.213
Q Wave	0.164
BP	0.162
PR	0.155
HTN	0.153
ST depression	0.107
Diastolic murmur	0.076
DM	0.071
BMI	0.064
Poor R Progression	0.063
Weight	0.060
Edema	0.054
Length	0.053
Function class	0.046
Sex	0.044
Lung rales	0.043
Thyroid disease	0.039
Current smoker	0.034
Dyspnea	0.033
Ex-smoker	0.033
LowTHAng	0.033
CRF	0.027
Obesity	0.018
CHF	0.016
CVA	0.011

For comparing performance of algorithms, ROC diagrams have been shown in Figures 1-3.

In Figures 1-3 the blue, red, and green lines show the KNN, C4.5, and Naïve Bayes models respectively.

As Figure 1 shows the best performance for diagnosing LAD vessel stenosis is related to C4.5 because of the wider area under its curve.

As Figure 2 shows the best performance is related to C4.5 for diagnosing LCX vessel stenosis.

As Figure 3 shows the performance of three algorithms for diagnosing RCA vessel stenosis is approximately the same.

RESULTS AND DISCUSSION

In this study Naïve Bayes, C4.5, and KNN algorithms were applied on new features that some of them had not been considered in pervious papers for diagnosing stenosis of each vessel separately.

Table 9: Information gain of features for left circumflex

Feature	Weight
Age	1
Typical chest pain	0.930
BP	0.454
HTN	0.327
Atypical	0.311
DM	0.259
Weight	0.176
Length	0.172
BMI	0.169
EX-smoker	0.139
Function class	0.130
PR	0.129
Sex	0.124
Nonanginal	0.115
Airway disease	0.096
Thyroid disease	0.075
CRF	0.065
Tinversion	0.050
Diastolic murmur	0.043
Systolic murmur	0.041
Poor R progression	0.035
CHF	0.035
Current smoker	0.031
CVA	0.030
Q Wave	0.028
Dyspnea	0.026
Edema	0.021
Obesity	0.019
ST depression	0.012
Lung Rales	0.006

Table 10: Information gain of features for right coronary artery

Feature	Weight
Typical chest pain	1
DM	0.824
Age	0.723
Atypical	0.677
Poor R progression	0.402
Function class	0.358
Nonanginal	0.341
HTN	0.254
Length	0.252
PR	0.244
BP	0.231
Weight	0.217
Sex	0.170
Q wave	0.162
Diastolic murmur	0.125
Dyspnea	0.119
BMI	0.114
CHF	0.074
LowTHAng	0.072
Tinversion	0.043
Weak peripheral pulse	0.039
CVA	0.039
ST elevation	0.035
EX-smoker	0.025
DLP	0.019
Current smoker	0.017
ST depression	0.015
Airway disease	0.011
Thyroid disease	0.010
Systolic murmur	0.009

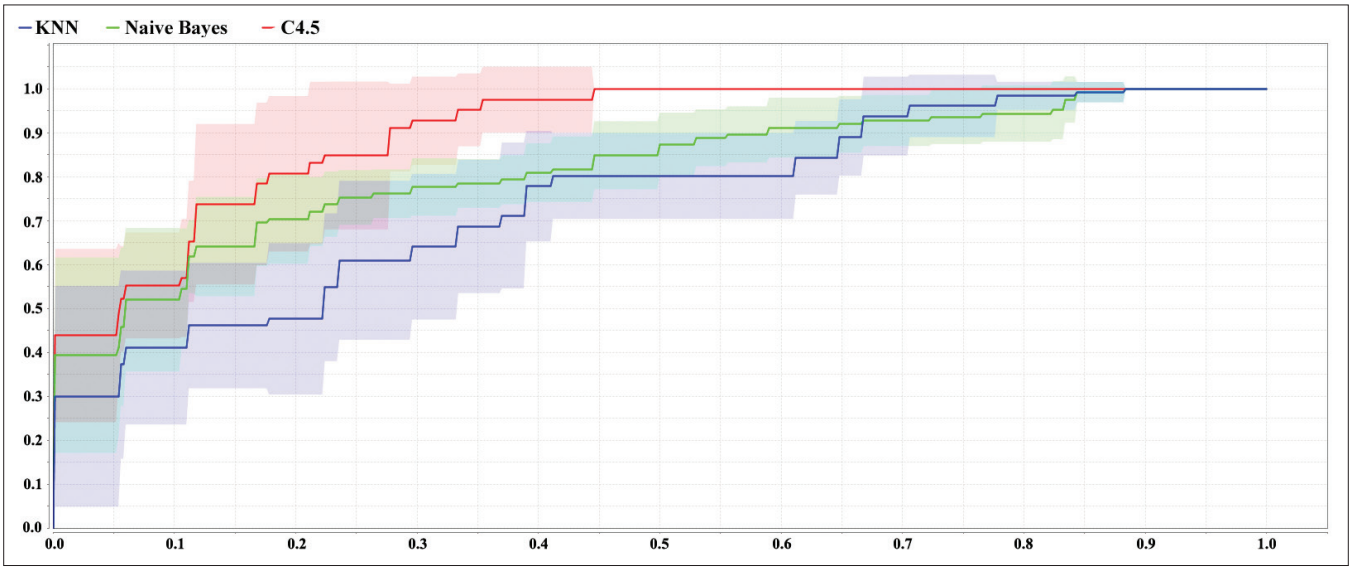


Figure 1: ROC diagram for diagnosing LAD vessel stenosis using C4.5, Naïve Bayes, and KNN algorithms

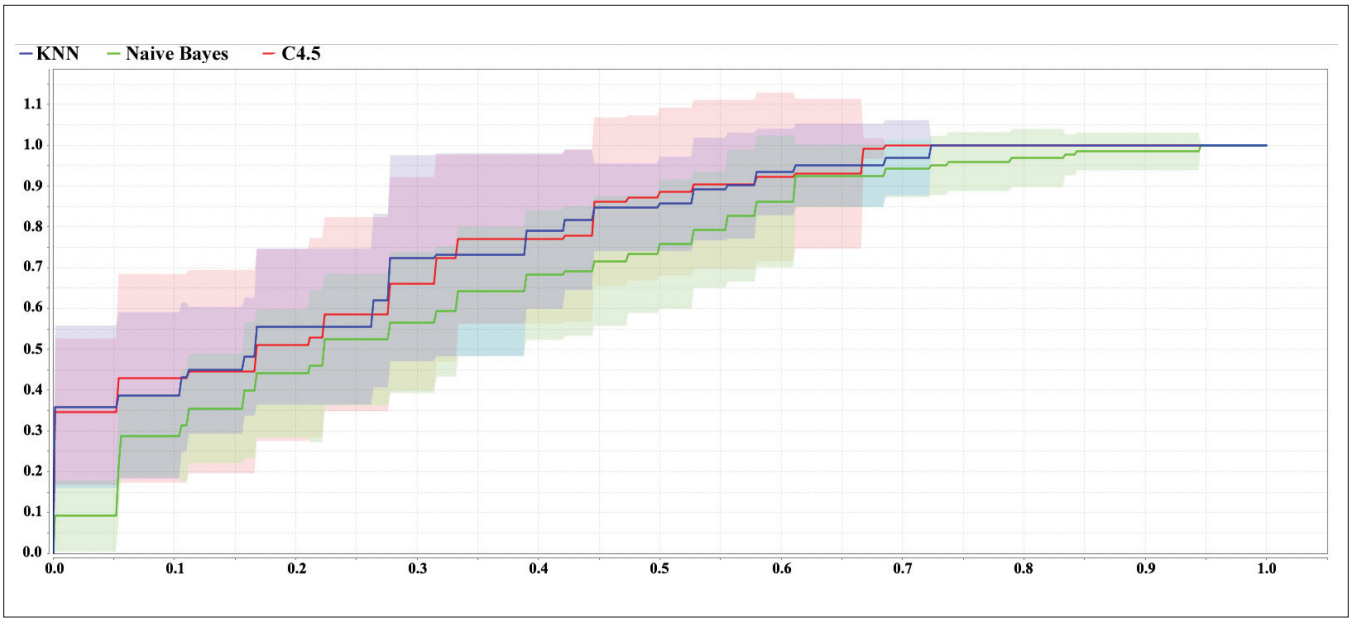


Figure 2: ROC diagram for diagnosing LCX vessel stenosis using C4.5, Naïve Bayes, and KNN algorithms

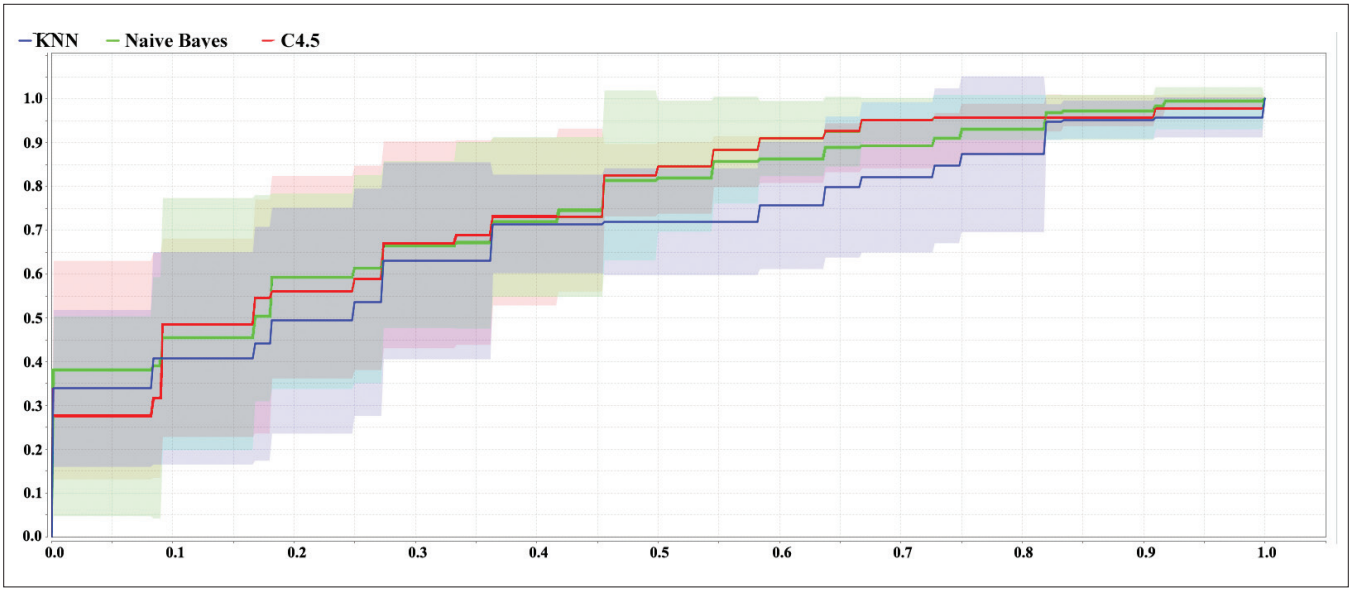


Figure 3: ROC diagram for diagnosing RCA vessel stenosis using C4.5, Naïve Bayes, and KNN algorithm

This study showed that, C4.5 method is better than Naïve Bayes and KNN on this data set and Typical Chest Pain, Age and Atypical features have a significant impact on vessels stenosis.

The accuracy for diagnosing stenosis of LAD vessel which is reached by this study is higher than the others and the accuracy for diagnosing stenosis of LCX and RCA vessels is approximately the same as[3] which had the best accuracy among the other papers.

CONCLUSION AND FUTURE WORK

In this study, three data mining algorithms were used and the highest accuracy was related to C4.5 which achieved 74.20%, 63.76%, and 68.33% accuracy for LAD, LCX, and RCA vessels respectively. As far as we know this accuracy is the best one for diagnosing LAD stenosis and was not achieved by previous studies.

The effects of 37 features were examined on stenosis of vessels. Age and Typical Chest Pain had a large effect on these vessels stenosis.

In the future, the goal is to add other features like Lab data and Echo data to find the impact of these features on stenosis of vessels and to achieve higher accuracy for diagnosing them.

REFERENCES

1. Bonow RO, Mann DL, Zipes DP, Libby P. Braunwald'sheart disease: A textbook of cardiovascular medicine. 9th ed. New York: Saunders; 2012.
2. Das R, Turkoglu I, Sengur A. Effective diagnosis of heart disease through neural networks ensembles. Expert Syst Appl 2009;36:7675-80.
3. Babaoglu I, Baykan OK, Aygul N, Ozdemir K, Bayrak M. Assessment of exercise stress testing with artificial neural network in determining coronary artery disease and predicting lesion localization. Expert Syst Appl 2009;36:2562-6.
4. Srinivas K, Rani BK, Govrdhan A. Applicationsof data mining techniques in healthcare and prediction of heart attacks. Int J Comput Sci Eng 2010;2:250-5.
5. Soni J, Ansari U, Sharma D, Soni S. Predictive datamining for medical diagnosis: An overview of heart disease prediction. Int J Comput Appl 2011;17:43-8.
6. Ordonez C, Omiecinski E, Santana C, Ezquerra N. Mining constrainedassociation rules to predict heart disease. In Proceedings of IEEE ICDM Conference; California 2001. p. 433-40.
7. Rajkumar A, Reena GS. Diagnosis of heart diseaseusing datamining algorithm. Global J Comput Sci Technol 2010;10:38-43.
8. Lee HG, Noh KY, Ryu KH. A data mining approach for coronary heart disease prediction usinghrv features and carotid arterial wallthickness. Singapore: International Conference on Biomedical Engineering; 2008. p. 200-6.
9. Chu C, Chien W. A bayesian expert system for clinical detecting coronary artery disease. J Med Sci 2009;29:187-94.
10. Karaolis MA, Moutiris JA, Hadjipanayi D. Assessment of the risk factors of coronary heart events based on data mining with decision trees. IEEE Trans Inf Technol Biomed 2010;14:559-66.
11. Quinlan JR. Improved use of continuous attributes in c4.5. J Artif Intell Res 1996;4;77-90.
12. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. Proceedings of the 23rd international conference on Machine learning; Pennsylvania: 2006. p. 161-8.
13. Available from: http://www. sourceforge.net/projects/rapidminer/ [Last accessed on 2012 July 01].

How to cite this article: How to cite this article: Alizadehsani R, Habibi J, Bahadorian B, Mashayekhi H, Ghandeharioun A, Boghrati R, Sani ZA. Diagnosis of Coronary Arteries Stenosis Using Data Mining. J Med Sign Sens 2012;2:153-60.

Source of Support: Sharif University of Technology and Tehran University of Medical Science, **Conflict of Interest:** None declared

BIOGRAPHIES

Roohallah Alizadehsani obtained a Bachelor of Science degree in computer engineering-software from Sharif University of Technology, and then received a Master in Science in computer engineering-software from Sharif University of Technology in 2012. He has been worked in Software engineering Lab under supervision of an Associate Professor, Jafar Habibi since 2011. Roohallah's research interests include mainly in the areas of data mining, machine learning, bioinformatics, and evaluation of computer systems performance. He has published three journal papers and one proceeding paper so far based on his thesis, "Diagnosis of heart disease using data mining". His approach for these topics includes new algorithms in data mining and mathematically prove them. He is a member of the IEEE.

E-mail: alizadeh_roohallah@yahoo.com

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. He has published more than 20 journal papers including Expert Systems with Applications, Applied Mathematics and Computation, and Pattern Recognition.

E-mail: jhabibi@sharif.ir

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.

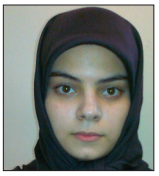
E-mail: 3ehdad@gmail.com

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.

E-mail: h.mashayekhi@gmail.com

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.

E-mail: asma.ghandeharioun@gmail.com



Reihane Boghrati is currently a senior Bachelor of Science student in computer engineering (software engineering) at Sharif University of Technology. She works in Software engineering laboratory under supervision of associate prof. Jafar Habibi since one of her interests is data-mining. In addition she is interested in computer science, machine learning, evaluation of computer systems performance, human science, and art. By attending online and university courses about Human-Computer Interaction field, she found this area as her main interest as it beautifully fulfills her interests. She has published three journal papers and one proceeding paper so far.

E-mail: r.boghrati@gmail.com

Zahra Alizadeh Sani received her MD degree (Oct. 1991-Sep. 1998) and cardiology specialty (Oct. 2002-Sep. 2006) from the Mashhad University of Medical science, Echocardiographic fellowship degree (Sep.2007-Mar.2009) in Tehran University of Medical science and cardiac MRI fellowship from Society of Cardiac MRI (Feb. 2011). At present, she is an assistant professor and the head of cardiac MRI department at Tehran University of Medical science, Rajaei research and medical cardiovascular center. Her research interests are mainly in the areas of diagnosis of cardiovascular disease and cardiovascular imaging. She has published more than ten journal papers including: computed Tomography journal, Physiol. Meas Journal, and JCMR Journal. She is Reviewer of Cardiovascular Journal of Tehran University.

E-mail: dr_zahra_alizadeh@yahoo.com