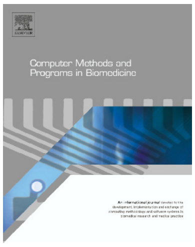

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

A data mining approach for diagnosis of coronary artery disease

Roohallah Alizadehsani[a], *Jafar Habibi*[a], *Mohammad Javad Hosseini*[a], *Hoda Mashayekhi*[a], *Reihane Boghrati*[a], *Asma Ghandeharioun*[a], *Behdad Bahadorian*[b], *Zahra Alizadeh Sani*[b,*]

[a] *Software Engineering, Department of Computer Engineering, Sharif University of Technology, Azadi Avenue, Tehran, Iran*
[b] *Rajaie Cardiovascular Medical and Research Center, Tehran University of Medical Science, Tehran, Iran*

ARTICLE INFO

Article history:
Received 20 September 2012
Received in revised form
20 February 2013
Accepted 5 March 2013

Keyword:
Classification
Data mining
Coronary artery disease
SMO
Bagging
Neural Networks

ABSTRACT

Cardiovascular diseases are very common and are one of the main reasons of death. Being among the major types of these diseases, correct and in-time diagnosis of coronary artery disease (CAD) is very important. Angiography is the most accurate CAD diagnosis method; however, it has many side effects and is costly. Existing studies have used several features in collecting data from patients, while applying different data mining algorithms to achieve methods with high accuracy and less side effects and costs. In this paper, a dataset called Z-Alizadeh Sani with 303 patients and 54 features, is introduced which utilizes several effective features. Also, a feature creation method is proposed to enrich the dataset. Then Information Gain and confidence were used to determine the effectiveness of features on CAD. Typical Chest Pain, Region RWMA2, and age were the most effective ones besides the created features by means of Information Gain. Moreover Q Wave and ST Elevation had the highest confidence. Using data mining methods and the feature creation algorithm, 94.08% accuracy is achieved, which is higher than the known approaches in the literature.

1. Introduction

Data mining is the process of extracting hidden knowledge from data. It can reveal the patterns and relationships among large amount of data in a single or several datasets. Data mining is used in various applications such as crime detection, risk evaluation and market analysis. Several industries like banking, insurance, and marketing use data mining to reduce costs, and increase profits [1].

Cardiovascular diseases are among the most common reasons of death all over the world. One major type of these diseases is coronary artery disease (CAD). Twenty five percent of people, who have CAD, die suddenly without any previous symptoms [2]. CAD is one of the most important types of diseases affecting the heart, and can cause severe heart attacks in patients. Being aware of disease symptoms, can aid in time treatment, and reduce the severity of disease's side effects.

Currently, angiography is used to determine the amount and place of heart vessels' stenosis. Being expensive and having several side effects, it has motivated many researchers to use data mining for diagnosing CAD. Several features and algorithms have been used in the literature.

Polat and Gunes [3] used fuzzy weighted pre-processing and AIRS and reached the accuracy of 92.59% for diagnosing CAD and as far as we know this is the best accuracy so far. Rajkumar and Reena [4] used decision tree and Naïve Bayes algorithms on the UCI dataset [5] and reached 52.33%

[*] *Corresponding author*. Tel.: +98 9153160452; fax: +98 2177952092.
E-mail addresses: dr_zahra_alizadeh@yahoo.com, alizadeh_roohallah@yahoo.com (Z.A. Sani).

0169-2607/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cmpb.2013.03.004

accuracy. Lavesson and Halling [6] applied Adaboost, Bagging, and Naïve Bayes algorithms on Chaps dataset and achieved 71% accuracy using Naïve Bayes. Shouman and Turner [7] used C4.5 decision tree for CAD diagnosis and used reduce error pruning, resulting in 84.1% accuracy. Itchhaporia et al. [8] applied Neural Network classification on 23 features and achieved 86% accuracy.

Existing studies do not include some important features. In this paper, several new features like EF, Region RWMA, Q Wave and TWaveInversion are considered in order to increase diagnosis accuracy, while discovering effect of features on CAD. A new feature creation method is used to add three new discriminative features to the patients' records which have a significant impact on prediction ability of the algorithms.

The Z-Alizadeh Sani dataset is constructed from the information provided by 303 random visitors to Shaheed Rajaei Cardiovascular, Medical and Research Center. 216 samples had CAD and the rest were healthy.

Sequential Minimal Optimization (SMO) [9], Naïve Bayes [10], Bagging [11] with SMO, and Neural Networks [12] classification algorithms are used to analyze the dataset. The results of the standard angiographic method are used as the base of comparison, to assess the prediction capability of classification algorithms.

The rest of this paper is organized as follows: The dataset is introduced in Section 2. Section 3 describes the technical aspects of the used data mining methods. The experimental results are discussed in Section 4, and finally Section 5 concludes the paper and discusses some future research directions.

2. The medical dataset

The Z-Alizadeh Sani dataset contains the records of 303 patients, each of which have 54 features. All features can be considered as indicators of CAD for a patient, according to medical literature [2]. However, some of them have never been used in data mining based approaches for CAD diagnosis. The features are arranged in four groups: demographic, symptom and examination, ECG, and laboratory and echo features. Table 1 presents the features of Z-Alizadeh Sani dataset along with their valid ranges, respectively. Each patient could be in two possible categories CAD or Normal. A patient is categorized as CAD, if his/her diameter narrowing is greater than or equal to 50%, and otherwise as Normal [2].

Some of the features in the presented tables should be further explained: 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.

The discretization ranges provided in Braunwald heart book [2] are used to enrich the dataset with discretized versions of some existing features. These new features are indicated by index 2 and are depicted in Table 2. Experiments show that these features which have been drawn from medical knowledge could help the classification algorithms to better classify a patient into CAD or Normal class.

3. Method

In this section, technical aspects of the data mining methods used to analyze the dataset are described. Sections 3.1–3.4, describe the classification algorithms. In Section 3.5, the feature selection approach is explained. A new feature creation algorithm is proposed in Section 3.6 to derive three new features from existing ones. Sections 3.7 and 3.8 describe two well-known scores to determine important features in classification and then, techniques for association rule mining from a dataset are discussed in Section 3.9. Finally the performance measures considered to evaluate the algorithm are described in Section 3.10.

3.1. SMO algorithm

Sequential Minimal Optimization (SMO) is an algorithm for efficiently solving the optimization problem which arises during the training of Support Vector Machines (SVMs). It was introduced by John Platt in 1998 at Microsoft Research. SMO is widely used for training SVM.

The publication of the SMO algorithm in 1998 has generated a lot of excitement in the SVM community, as previously available methods for SVM training were much more complex and required expensive third-party QP solvers [9].

3.2. Naïve Bayes algorithm

Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong independence assumption.

The probability of data record X having the class label C_j is:

$$P(C_j|X) = \frac{P(X|C_j) * P(C_j)}{P(X)} \qquad (1)$$

The class label C_j with largest conditional probability value determines the category of the data record.

3.3. Bagging algorithm

Bagging is an ensemble method, which trains some base classifiers using the input dataset. The base classifier used in this study is SMO. The results of base classifiers are combined in a poll, to achieve the final class label. For this study, the classifier type is SMO. The accuracy of the base classifiers must be a little bit higher than 50%. The learning algorithms used to construct these classifiers are called weak learners.

3.4. Neural Network algorithm

An Artificial Neural Network (ANN) or Simulated Neural Network (SNN), is an interconnected group of artificial neurons, that use a mathematical or computational model for information processing based on a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network. In more practical terms, Neural Networks are non-linear statistical data modeling or

Table 1 – Features of Z-Alizadeh Sani dataset.

Feature type	Feature name	Range
Demographic	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 (*Dyslipidemia*)	Yes, no
Symptom and examination	BP (blood pressure: mmHg)	90–190
	PR (pulse rate) (ppm)	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
	Dyspnea	Yes, no
	Function class	1, 2, 3, 4
	Atypical	Yes, no
	Nonanginal CP	Yes, no
	Exertional CP (Exertional Chest Pain)	Yes, no
	Low Th Ang (low Threshold angina)	Yes, no
ECG	Rhythm	Sin, AF
	Q Wave	Yes, no
	ST Elevation	Yes, no
	ST Depression	Yes, no
	T inversion	Yes, no
	LVH (left ventricular hypertrophy)	Yes, no
	Poor R progression (poor R wave progression)	Yes, no
Laboratory and echo	FBS (fasting blood sugar) (mg/dl)	62–400
	Cr (creatine) (mg/dl)	0.5–2.2
	TG (triglyceride) (mg/dl)	37–1050
	LDL (low density lipoprotein) (mg/dl)	18–232
	HDL (high density lipoprotein) (mg/dl)	15–111
	BUN (blood urea nitrogen) (mg/dl)	6–52
	ESR (erythrocyte sedimentation rate) (mm/h)	1–90
	HB (hemoglobin) (g/dl)	8.9–17.6
	K (potassium) (mEq/lit)	3.0–6.6
	Na (sodium) (mEq/lit)	128–156
	WBC (white blood cell) (cells/ml)	3700–18,000
	Lymph (Lymphocyte) (%)	7–60
	Neut (neutrophil) (%)	32–89
	PLT (platelet) (1000/ml)	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

decision making tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data [13].

3.5. *Feature selection*

In selecting features, “weights by SVM” [14] in RapidMiner [15] were considered.

It uses the coefficients of the normal vector of a linear SVM as feature weights. In contrast to most of the SVM based operators available in RapidMiner, this one works for multiple classes, too. The attribute values, however, still have to be numerical.

Among many features, 34 of them which had the weight higher than 0.6, were selected and the algorithms were applied on them.

Table 2 – Descritized features and their range of values.

Feature	Low	Normal	High
Cr2	$Cr < 0.7$	$0.7 \le Cr \le 1.5$	$Cr > 1.5$
FBS2	$FBS < 70$	$70 \le FBS \le 105$	$FBS > 105$
LDL2		$LDL \le 130$	$LDL > 130$
HDL2	$HDL < 35$	$HDL \ge 35$	–
BUN2	$BUN < 7$	$7 \le BUN \le 20$	$BUN > 20$
ESR2		If male and $ESR \le age/2$ or if female and $ESR \le age/2 + 5$	If male and $ESR > age/2$ or if female and $ESR > age/2 + 5$
HB2	If male and $HB < 14$ Or If female and $HB < 12.5$	If male and $14 \le HB \le 17$ or if female and $12.5 \le HB \le 15$	If male and $HB > 17$ or if female and $HB > 15$
K2	$K < 3.8$	$3.8 \le K \le 5.6$	$K > 5.6$
Na2	$Na < 136$	$136 \le Na \le 146$	$Na > 146$
WBC2	$WBC < 4000$	$4000 \le WBC \le 11{,}000$	$WBC > 11{,}000$
PLT2	$PLT < 150$	$150 \le PLT \le 450$	$PLT > 450$
EF2	$EF \le 50$	$EF > 50$	
Region with RWMA2	–	Region with $RWMA = 0$	Region with $RWMA \ne 0$
Age2[a]		If male and $age \le 45$ or if female and $age \le 55$	If male and $age > 45$ or if female and $age > 55$
BP2	$BP < 90$	$90 \le BP \le 140$	$BP > 140$
PulseRate2	$PulseRate < 60$	$60 \le PulseRate \le 100$	$PulseRate > 100$
TG2		$TG \le 200$	$TG > 200$
Function Class2		1	2, 3, 4

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

3.6. Feature creation

In this part, an algorithm is proposed for creating three new features named LAD recognizer, LCX recognizer, RCA recognizer. These features are specialized for recognizing whether three major coronary arteries, *Left Anterior Descending* (LAD), Left Circumflex (LCX), or Right Coronary Artery (RCA) is blocked, respectively. Higher values of any of these created features, indicates higher probability of having CAD. Each of these features is derived from set of available features in the dataset. Procedure 1 explains how to create LAD recognizer in detail. Available features of the dataset are first discretized into binary variables. The method is designed according to an assumption about the descritized features: value 1 for a feature indicates higher probabilities of the record being in the CAD class, while value zero indicates otherwise. LCX and RCA recognizers are created with similar methods. The record is classified as CAD, when at least one of the arteries, LAD, LCX, or RCA is blocked [2]. Therefore, these three created features will definitely have great importance in CAD diagnosis.

3.7. Information Gain

Information Gain measures the reduction in entropy of the data records because of a single split over a given attribute. The entropy before and after the split is computed as follows:

$$IG = -\sum_{c \in \text{classes}} P(c)\log(P(c)) \tag{2}$$

where *c* is the class value which can be CAD or Normal, and *P*(*c*) denotes the probability of a record being in class *c*. The higher values of Information Gain indicate preference of feature for discrimination of class values. For example, if a feature separates the two classes completely, it has the most Information Gain and is the best feature for classification [16].

Table 3 – Confusion matrix.

	Actual class C_1	Actual class C_2
Predicted class C_1	True positive (TP)	False positive (FP)
Predicted class C_2	False negative (FN)	True negative (TN)

3.8. Gini Index

Gini Index is a measure of how often a randomly chosen element from a set of elements would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. The probability of a correct labeling can be computed by summing the probability of choosing each item multiplied by the probability of correctly labeling it. In this setting, the probability of correctly labeling an item is equal to the probability of choosing that item. Therefore, Gini Index can be computed as:

$$GIdx = 1 - \sum_{c \in \text{classes}} P(c)^2 \tag{3}$$

Similar to Information Gain, the higher values of reduction in Gini Index implies that a feature is a better candidate in the classification task [16].

3.9. Association rule mining

A rule for a given dataset has the form A → B, where A and B are conditions on the values of the features. For each rule, two terms, namely support and confidence, are defined. Support of a rule means the proportion of data which satisfies both the left hand and right hand sides of that rule. Confidence means the probability of finding the right hand side of the rule in those item sets which satisfy the left hand side conditions as shown below:

$$\text{Confidence}(A \rightarrow B) = \frac{Pr(A \cup B)}{Pr(A)} \tag{4}$$

To obtain rules, firstly conditions on features that have the highest probabilities are derived from the dataset. Then these conditions are split in all possible ways to form smaller conditions A and B. Then rules of the form A → B that have the highest confidence are extracted. For our purpose, only rules that have the right hand side determining a condition on the label of the patients (CAD or Normal) are considered, since the relation between the features of the dataset and the label of patients are sought. In order to make rules, all features should be binomial [17]. Therefore, the same steps to make a feature binomial in the feature creation method should be done before the rules are extracted.

3.10. *Performance measure*

Accuracy, sensitivity, and specificity are the most important performance measures in the medical field [18], which are commonly used in the literature. So for measuring the performance of algorithms, these measures are used.

3.10.1. *Confusion matrix*

A confusion matrix is a table that allows visualization of the performance of an algorithm. In a two class problem (with classes C_1 and C_2), the matrix has two rows and two columns that specifies the number of false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN). These measures are defined as follows: TP is the number of samples of class C_1 which has been correctly classified. TN is the number of samples of class C_2 which has been correctly classified. FN is the number of samples of class C_1 which has been falsely classified as C_2. FP is the number of samples of class C_2 which has been falsely classified as C_1. Table 3 shows confusion matrix.

3.10.2. *Sensitivity and specificity*

According to confusion matrix, sensitivity and specificity are explained as following:

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} \tag{5}$$

$$\text{Specificity} = \frac{TN}{(TN + FP)} \tag{6}$$

3.10.3. *Accuracy*

Accuracy shows ratio of correctly classified samples to the total number of tested samples. It is defined as:

$$\text{Accuracy} = \frac{(TN + TP)}{(TN + TP + FN + FP)} \tag{7}$$

3.10.4. *ROC*

A Receiver Operating Characteristic (ROC) is a graphical plot which illustrates the performance of a binary classifier system. It is created by True Positive Rate (TPR) vs. False Positive Rate (FPR). The larger the area under ROC curve, the higher the performance of the algorithm is. FPR and TPR are explained as:

$$FPR = \frac{FP}{(FP + TN)} \tag{8}$$

$$TPR = \frac{TP}{(TP + FN)} \tag{9}$$

Table 4 – Information Gain.

Selected features?	Feature name	IG
Yes	LAD recognizer	1
	LCX recognizer	0.921
	RCA recognizer	0.750
	Typical Chest Pain	0.622
	Region RWMA2	0.270
	Age	0.217
	EF2	0.204
	HTN	0.158
	DM	0.138
	TWaveInversion	0.119
	ESR	0.097
	Q Wave	0.070
	ST Elevation	0.060
	PulseRate	0.059
	BMI	0.041
	Lymph	0.041
	BP2	0.037
	Dyspnea	0.028
	HDL	0.022
	CR2	0.021
	WBC2	0.018
	Weight	0.018
	VHD	0.017
	Function Class	0.017
	Airway disease	0.014
	HB	0.013
	TG2	0.013
	BBB	0.011
	Na2	0.007
	Sex	0.006
	LVH	0.003
	HB2	0.001
	FH	0.001
No	Atypical	0.323
	Nonanginal	0.130
	FBS2	0.106
	Diastolic murmur	0.037
	Current Smoker	0.011
	EX-Smoker	0.003

4. Experimental result

To apply the data mining algorithms, RapidMiner tool [15] was used. RapidMiner is an environment for machine learning, data mining, text mining and business analytics. It is used for research, education, training and industrial applications. In this study, version 5.2.003 of RapidMiner is used [15]. All algorithms were used in the default state.

In what follows the obtained results and discussions are presented.

Table 5 – High confidence features.

Feature	Number	Confidence
Q Wave	16	1
ST Elevation	14	1
Poor R progression	9	1
CRF	6	1
Week peripheral pulse	5	1
Region RWMA	86	0.953488
Typical Chest Pain	164	0.939024
Airway disease	11	0.909091
DM	90	0.888889
TWaveInversion	90	0.877778
FBS2	84	0.869048
CR2	22	0.863636
BP2	48	0.854167
WBC	27	0.851852
Edema	12	0.833333
ST Depression	71	0.830986
EF2	197	0.822335
HTN	179	0.821229
Lung rales	11	0.818182
LVH	20	0.8
Ex-Smoker	10	0.8
CVA	5	0.8
TG2	62	0.790323
ESR	46	0.782609
Age2	238	0.781513
Current smoker	63	0.777778
Function class	92	0.771739
Na2	34	0.764706
Neut2	89	0.764045
FH	48	0.75
PLT2	12	0.75
HB	157	0.732484
LDL2	62	0.725806
HDL2	87	0.724138
Systolic murmur	41	0.707317
Obesity	211	0.706161
DLP	112	0.705357
Dyspnea	134	0.649254
K2	37	0.648649
Thyroid Disease	7	0.571429
Atypical	93	0.430108
Diastolic murmur	9	0.333333

4.1. *Results*

The experiments are designed so that the different parts of the work could be evaluated. These include the evaluation of the features of the dataset, the feature selection and also the feature creation methods. To this aim, first the features which were selected by the feature selection method and their importance are discussed. Second, all the four possible combinations of the feature selection and creation methods are tested over the dataset. Finally, some new and useful rules with high confidences which are extracted by association rule mining techniques are presented.

4.1.1. *Results of feature selection*

In this section, the results of the applied feature selection method are discussed. List of the selected features, after applying the feature selection method, which was described in Section 3.5, can be seen in Table 4 as the selected features category.

In addition, for better recognizing effective features on CAD, the Information Gain and Gini Index indicators are computed for different features. However, the "weights by SVM" method is used for feature selection, as the experiments showed that the highest classification accuracy was achieved when this method was used.

Information Gains of the features selected by the used feature selection method, i.e. "weights by SVM", are shown in Table 4. It can be seen that the three created features have the most Information Gain. This indicates that the created features have the highest separation power for classifying records into CAD or Normal categories, among all features in the dataset.

Information Gains for the other features that have not been selected by the feature selection method are also shown in Table 4. As it can be seen, Atypical, Nonanginal and FBS2 also have great effect on CAD, according to Information Gain. The important features could also be extracted using Gini Index. However, the results are similar to Information Gain results and are presented in Appendix A.

In addition to the features which have high Information Gain or Gini Index, the features which have high confidence in diagnosing a patient as CAD should be mentioned. We define the confidence of a binomial feature *f*, with two possible values 0 and 1 as described in Procedure 1, in diagnosis a patient as CAD as the confidence of the rule $f=1 \rightarrow$ CAD. Table 5 shows the features in the descending order of their confidence along with the number of patients which have the value 1 for them.

Comparing the results of Tables 4 and 5, it can be concluded that the features with high Information Gain (or also Gini Index) does not have high confidence, necessarily. The reason is that a feature has high Information Gain (or Gini Index) if its values can effectively discriminate the class labels. For example, a binominal feature *f* with two values 0 and 1, as described in Procedure 1, has high Information Gain, if both of the following rules have high confidence:$f=1 \rightarrow$ CAD

$f=0 \rightarrow$ Normal

However, a feature has high confidence in diagnosis a patient as CAD, if the following rule has high confidence:

$f=1 \rightarrow$ CAD

Note that, unlike Information Gain, only high confidence of this rule is sufficient for a feature to be selected. In other words, if the value of feature *f* does not equal to 1, nothing needs to be concluded about the category of a patient.

Top features of Table 5 could be useful in decision making about a patient. Besides, the features with low confidence should not be considered as a sign of CAD in isolation. However, the combination of these features with others may help in classifying a patient.

4.1.2. *Performance comparison of different methods*

The performance measures for different algorithms, executed on the whole set of features without three created features, are represented in the first part of Table 6. As shown in this part of Table 6, the Bagging and SMO methods achieved nearly the same accuracy, which is above 89%. Neural Network also offers competitive accuracy of 85% but Naïve Bayes accuracy is

Table 6 – Comparing the performance of algorithms.

Used features	Algorithm used	Accuracy	Sensitivity	Specificity
All features without three created features	Bagging SMO	89.43 ± 6.78%	91.67%	83.91%
	Naïve Bayes	47.84 ± 6.35%	28.70%	95.40%
	SMO	89.76 ± 7.31%	92.13%	83.91%
	Neural Network	85.43 ± 7.02%	90.28%	73.56%
All features and three created features	Bagging SMO	90.10 ± 6.96%	91.67%	86.21%
	Naïve Bayes	63.31 ± 8.01%	50%	96.55%
	SMO	90.09 ± 6.49%	91.67%	86.21%
	Neural Network	87.11 ± 6.05%	91.67%	75.86%
Selected features without three created features	Bagging SMO	92.74 ± 6.43%	95.37%	86.21%
	Naïve Bayes	55.37 ± 9.62%	38.89%	96.55%
	SMO	93.39 ± 5.14%	95.37%	88.51%
	Neural Network	87.13 ± 5.84%	90.28%	79.31%
Selected features and three created features	Bagging SMO	93.40 ± 5.53%	95.83%	87.36%
	Naïve Bayes	75.51 ± 10.32%	67.59%	95.40%
	SMO	94.08 ± 5.48%	96.30%	88.51%
	Neural Network	88.11 ± 6.17%	91.20%	80.46%

Table 7 – The confusion matrix for different algorithms using three created features and selected features.

Algorithm		True Cad	True Normal
Bagging with SMO classifiers	Pred. Cad	206	12
	Pred. Normal	10	75
Naïve Bayes	Pred. Cad	146	4
	Pred. Normal	70	83
SMO	Pred. Cad	208	10
	Pred. Normal	8	77
Neural Network	Pred. Cad	197	17
	Pred. Normal	19	70

considerably lower. Noticing that sensitivity values are higher than specificity values for all algorithms other than Naïve Bayes, apparently Naïve Bayes is more inclined to identify Normal class. The other three methods are more capable of predicting the CAD samples in comparison to Normal samples.

The results for different algorithms, executed with the whole set of features are represented in the second part of Table 6. As observed in this part, the accuracy and specificity of all the algorithms have increased because of the created features. Also, the relative order of the used classification methods, according to different metrics (accuracy, sensitivity and specificity), is similar in the first and second part of Table 6. For the Naïve Bayes classifier, the accuracy and sensitivity are further enhanced in the second part in comparison to the first part. Finally, for SMO, sensitivity is reduced but specificity is increased.

The results of algorithms after feature selection and without the three created features are shown in the third part of Table 6. As shown in this part, feature selection has increased the accuracy of all the classification algorithms in comparison to the first and second parts, except for the Naïve Bayes algorithm. It shows that selecting a subset of effective features increases the accuracy, as irrelevant features could mislead the classifiers into false predictions. The highest accuracy in this case is obtained by SMO which is 93.39%.

The results of algorithms after feature selection with three created feature is shown in the last part of Table 6.

As shown in this part, the highest accuracy, sensitivity and specificity of algorithms are obtained when both feature selection and feature creation methods are used. Comparing the third and the forth parts shows that accuracy and sensitivity of all algorithms have increased after adding three new features. The highest increase in accuracy is for Naïve Bayes which is about 20%. Finally, SMO has achieved the highest accuracy which is 94.08%. To the best of our knowledge, this is the best achieved accuracy for diagnosis of CAD, compared to the literature [3].

Confusion matrices for the four classification algorithms along with feature selection and creation methods are shown in Table 7. The values presented in the first part of Table 7 reveal that upon applying the SMO classifier, TP, TN, FP, and FN are 206, 75, 12, and 10 respectively.

In Fig. 1, the ROC curves for Bagging, Naïve Bayes, Neural Network, and SMO models can be seen.

In ROC diagram, the more the area under the curve is, the higher the algorithm performance is. So as shown in the last part of Table 6, the highest performance is obtained by SMO, and afterwards by Bagging, Neural Network and Naïve Bayes classification algorithms, respectively.

4.1.3. *Results of association rule mining*

The rules produced by the association rule mining techniques are given below in decreasing support order. In these rules, C represents the Confidence and S represents the Support which were explained in Section 3.9. Because confidence is

Table 8 – Gini Index.

Selected features?	Feature name	GIdx
Yes	LAD ratio	1
	LCX ratio	0.939
	RCA ratio	0.643
	Typical Chest Pain	0.564
	Region RWMA2	0.213
	EF2	0.207
	Age	0.206
	HTN	0.157
	DM	0.120
	TWaveInversion	0.105
	ESR	0.088
	PulseRate	0.057
	BMI	0.045
	Lymph	0.043
	Qwave	0.041
	ST Elevation	0.035
	BP2	0.033
	Dyspnea	0.028
	HDL	0.025
	Weight	0.017
	VHD	0.017
	CR2	0.016
	Function class	0.016
	WBC2	0.015
	HB	0.013
	TG2	0.012
	BBB	0.012
	Airway disease	0.011
	Sex	0.006
	Na2	0.006
	LVH	0.003
	HB2	0.001
	FH	0.001
No	Atypical	0.331
	Nonanginal	0.144
	FBS2	0.101
	Diastolic murmur	0.041
	Current smoker	0.01
	EX-Smoker	0.004

more important than support, for extracting rules the minimum acceptable confidence and support values, C = 0.9 and S = 0.03, were considered for pruning the rules. Among created rules, 22 of them that had most confidence were selected.

1. [BMI > 25, Typical Chest Pain = true, TWaveInversion = true] ≥ [Cad], S = 0.12828, C = 1.
2. [Typical Chest Pain = true, PulseRate < 60 or PulseRate > 100, TWaveInversion = true] ≥ [Cad], S = 0.0874636, C = 1.
3. [BMI > 25, Typical Chest Pain = true, TG > 200] ≥ [Cad], S = 0.0845481, C = 1.
4. [EF≤50, Typical Chest Pain = true, HB = low or HB = High, PulseRate < 60 or PulseRate > 100] ≥ [Cad], S = 0.0787172, C = 1.
5. [HTN = true, Typical Chest Pain = true, HB = low or HB = High, PulseRate < 60 or PulseRate > 100] ≥ [Cad], S = 0.0728863, C = 1.
6. [Typical Chest Pain = true, TWaveInversion = true, ST Depression = true] ≥ [Cad], S = 0.0699708, C = 1.
7. [EF ≤ 50, HTN = true, TWaveInversion = true, ST Depression = true] ≥ [Cad], S = 0.0641399, C = 1.
8. [EF ≤ 50, HTN = true, HB = low or HB = High, ST Depression = true] ≥ [Cad], S = 0.0641399, C = 1.
9. [HB = low or HB = High, TWaveInversion = true, ST Depression = true] ≥ [Cad], S = 0.0641399, C = 1.
10. [Typical Chest Pain = true, Sex = female, TWaveInversion = true] ≥ [Cad], S = 0.0641399, C = 1.
11. [DM = true, ST Depression = true] ≥ [Cad], S = 0.0641399, C = 1.
12. [EF ≤ 50, Typical Chest Pain = true, HB = low or HB = High, ST Depression = true] ≥ [Cad], S = 0.0612245, C = 1.
13. [BMI > 25, EF ≤ 50, HB = low or HB = High, TWaveInversion = true] ≥ [Cad], S = 0.0612245, C = 1.
14. [HTN = true, FBS = low or FBS = high, ST Depression = true] ≥ [Cad], S = 0.0612245, C = 1.
15. [EF ≤ 50, HTN = true, PulseRate < 60 or PulseRate > 100, TWaveInversion = true] ≥ [Cad], S = 0.058309, C = 1.
16. [Typical Chest Pain = true, TWaveInversion = true, Current Smoker = true] ≥ [Cad], S = 0.058309, C = 1.
17. [BMI > 25, HTN = true, HDL < 35, TWaveInversion = true] ≥ [Cad], S = 0.0553936, C = 1.
18. [BMI > 25, TWaveInversion = true, Current Smoker = true] ≥ [Cad], S = 0.0524781, C = 1.
19. [HB = low or HB = High, PulseRate < 60 or PulseRate > 100, TWaveInversion = true] ≥ [Cad], S = 0.0524781, C = 1.
20. [HTN = true, Typical Chest Pain = true, Sex = female, ST Depression = true] ≥ [Cad], S = 0.0524781, C = 1.

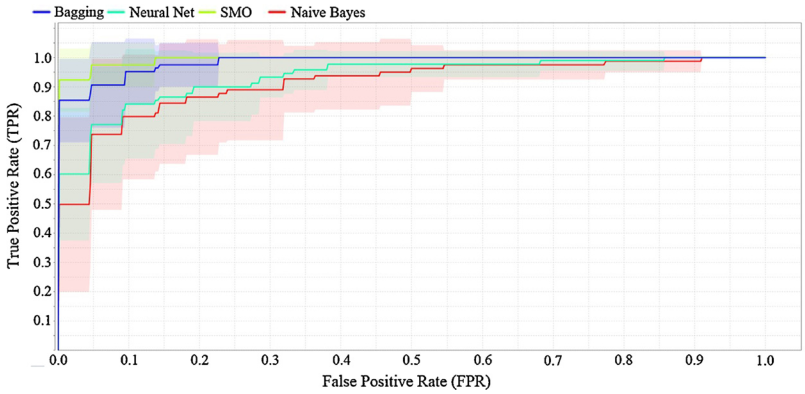

Fig. 1 – ROC diagram for four algorithms: the blue, red, green, and olive lines show the ROC curve for Bagging, Naïve Bayes, Neural Network, and SMO models, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

21. [Q Wave = true] ≥ [Cad], S = 0.0466472, C = 1.
22. [ST Elevation = true] ≥ [Cad], S = 0.0408163, C = 1.

Rules number 21 and 22 show that the patients with only ST Elevation or Q Wave in their ECG could be classified as CAD with confidence 1 in our dataset. These two rules that are extracted from the dataset have also been introduced in [19]. However, since the cases fit in these two rules are limited in the dataset, further examinations on more patients may confirm or reject these rules.

4.2. *Discussion*

In this study, several algorithms including Naïve Bayes, SMO, Bagging, and Neural Network were applied on Z-Alizadeh Sani dataset. Because of considering new important features like EF, Region RWMA, Q Wave and TWaveInversion, exploiting existing data mining approaches, and also proposing a new technique to create three new features from the dataset, remarkable accuracy values were achieved. The accuracy values were measured using ten-fold cross validation. As the last part of Table 6 shows, the highest accuracy is achieved by SMO algorithm along with the feature selection and feature creation methods. This accuracy is 94.08%. To the best of our knowledge, this is the highest accuracy value reported in the literature. For example, [4,20,21,22] had achieved accuracies 52.33%, 90%, 70% and 75% which are lower than the accuracy reported in this study.

The feature creation method which has been introduced in this study and was successful in increasing the classification accuracy is a general method which could be used in other applications of data mining.

Effective features on CAD are shown in Table 4. Most important features with respect to Information Gain include: Typical Chest Pain, Region RWMA2, age, EF2, HTN, DM, TWaveInversion, ESR, Q Wave, ST Elevation, PulseRate, and BMI, respectively. As a result, these features can have the greatest effect on the classification task. In contrast, other features could also be important because of their high confidence in detecting CAD. These features are shown in Table 5. However, these are not necessarily the best features for classification task and only could be used to make a high confidence decision for a patient to have CAD.

Bonow et al. [2] confirms that the selected features by both methods are important in CAD diagnosis.

Smoking and FH also significantly influence CAD disease [2]. These features have significant confidence values according to Table 5. However, Information Gain values of these features are not as high as many other features. This is because of the fact that the absence of these features for a patient does not necessarily indicate its health. On the other hand, the presence of these features would be a good reason for a patient to have CAD. However a feature should be determinant for all class labels, to be prominent in classification algorithms.

Table 9 – Gini Index for mutual features.

Feature	GIdx for Cleveland dataset	GIdx for our dataset
Age	0.292	0.206
BP	0.072	0.03
Sex	0.28	0.006

5. Conclusion and future work

In this study, several algorithms were applied on the Z-Alizadeh Sani dataset and the results were discussed. The features included in this dataset are possible indicators of CAD, according to our medical knowledge. In addition, data mining techniques including feature selection and creation were used to improve the accuracy. The accuracy value achieved in this study is, to the best of our knowledge, higher than currently reported values in the literature.

In addition, the features used in this study, can be measured with affordable costs and side effects. Hence, applying the proposed approach can identify the CAD state with high probability and low costs.

In future, we aim to consider predicting state of each artery independently. Moreover, it is obvious that true diagnosis of diseased people is more important than true identification of healthy ones. Therefore, another goal to meet is using cost sensitive algorithms to consider this factor. Finally, larger datasets, more features and also broader data mining approaches, could be used to achieve better and more interesting results.

Appendix A. Supplementary information on features

Gini Index for selected features by "weight by SVM" feature selection is shown in Table 8. As in Table 4, three features created have the most Gini Index. Also sequence of 12 first features is almost identical in both.

Gini Index for other features that have not been selected by the feature selection method is also shown in Table 8.

As in Table 4, Atypical, Nonanginal and FBS2 have a great effect on CAD.

In order to strengthen our assumptions about the used dataset, we compared Gini Index of its features with that of Cleveland's [5], the most well-known dataset which is used by previous studies. Table 9 shows the result for the mutual features of Cleveland dataset and our dataset, i.e. Age, BP and Sex. The results show that Gini Index of all these three features is higher for Cleveland dataset; however, the difference is not much for Age and BP features. These differences may be due to differences of the used samples of the datasets. Moreover, higher values of Gini Index for Cleveland dataset may lead to

Procedure 1 – Feature creation.

```
On train data:
  1. For any feature $f$ do
convert $f$ to a binomial feature using the following steps:
      a. If $f$ is numerical, discretize it by breaking its domain into
    intervals. For example, $K$ is descretized using these limits: Low if
    $K < 3.5$, Normal if $3.8 \leq K \leq 5.6$, and High otherwise.
```

Procedure 1 (*Continued*)

b. If f is binomial, feature values are considered as 1 and 0. The values that have positive effect on Cad are considered as 1 and the others considered as 0. For example: for Obesity, "Yes" is considered as 1 and "No" considered as 0.

c. If f is polynomial, change it to binomial by mapping the values having direct relationship to CAD, to 1 and others to 0: For example HB has three values: Low, Normal, and High. "Low" and "High" are mapped to 1 and group, and "Normal" is mapped to zero.

2. For all $f \in$ features calculate the following fraction on training data:

$w(f) = P(LAD = 1|f = 1)$,

where $LAD = 1$ means that LAD is clogged. For example, 179 records have HTN equal to 1, among which the LAD artery of 118 samples is clogged. Therefore

$W(HTN) = 118/179 = 0.66$

3. Choose K features that have the highest W value. Name as f_1, $f_2, \ldots, f_K$.

(K is set to 15 in the experiments.)

4. Compute

$$LAD \text{ recognizer} = \sum_{i=1}^{K} w(i) f(i).$$

higher accuracy if the features of Z-Alizadeh Sani dataset are extracted for the patients of Cleveland dataset.

REFERENCES

[1] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2005.

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

[3] K. Polat, S. Gunes, A hybrid approach to medical decision support systems: combining feature selection, fuzzy weighted pre-processing and AIRS, Computer Methods and Programs in Biomedicine 88 (2007) 164–174.

[4] A. Rajkumar, G.S. Reena, Diagnosis of heart disease using data mining algorithm, Global Journal of Computer Science and Technology 10 (2010) 38–43.

[5] UCI KDD Archive, available from: http://archive.ics.uci.edu/ml/ (accessed 07.07.12).

[6] N. Lavesson, A. Halling, Classifying the severity of an acute coronary syndrome by mining patient data, in: 25th Annual Workshop of the Swedish Artificial Intelligence Society, Linköping University Electronic Press, 2009, pp. 55–63.

[7] M. Shouman, T. Turner, Using decision tree for diagnosing heart disease patients, in: Proceedings of the 9-th Australasian Data Mining Conference (AusDM'11), Ballarat, Australia, 2011.

[8] D. Itchhaporia, R. Almassy, L. Kaufman, P. Snow, W. Oetgen, Artificial neural networks can predict significant coronary disease, Journal of the American College of Cardiology 25 (1995) 23–29.

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

[10] R. Caruana, A. Niculescu-Mizil, An empirical comparison of supervised learning algorithms, in: Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 161–168.

[11] L. Breiman, Bagging predictors, Machine Learning 24 (1996) 123–140.

[12] B.D. Ripley, Pattern Recognition and Neural Network, 1st ed., 1996.

[13] H. Lu, R. Setiono, H. Liu, Effective data mining using neural networks, IEEE Transactions on Knowledge and Data Engineering 8 (1996) 957–961.

[14] A. Ben-Hur, J. Weston, A user's guide to support vector machines, Methods in Molecular Biology 609 (2010) 223–239.

[15] http://sourceforge.net/projects/rapidminer/ (accessed 05.07.12).

[16] P.N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Pearson Addison Wesley, Boston, MA, 2006.

[17] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, in: ACM SIGMOD Conference on Management of Data, 1993, pp. 207–216.

[18] N. Lavrac, Selected techniques for data mining in medicine, Artificial Intelligence in Medicine 16 (1999) 3–23.

[19] P. Kligfield, L.S. Gettes, J.J. Bailey, et al., Recommendations for the standardization and interpretation of the electrocardiogram. Part I: the electrocardiogram and its technology: a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society, endorsed by the International Society for Computerized Electrocardiology, Journal of the American College of Cardiology 49 (2007) 1109–1127.

[20] H.G. Lee, K.Y. Noh, K.H. Ryu, A data mining approach for coronary heart disease prediction using HRV features and carotid arterial wall thickness, in: International Conference on Biomedical Engineering, 2008, pp. 200–206.

[21] C. Chu, W. Chien, A Bayesian expert system for clinical detecting coronary artery disease, Journal of Medical Science 29 (2009) 187–194.

[22] M.A. Karaolis, J.A. Moutiris, D. Hadjipanayi, Assessment of the risk factors of coronary heart events based on data mining with decision trees, IEEE Transactions on Information Technology in Biomedicine 14 (2010) 559–566.