Open access • Posted Content • DOI:10.20944/PREPRINTS202001.0220.V1

Coronary Artery Disease Diagnosis: Ranking the Significant Features Using Random Trees Model — Source link

Javad Hassannataj Joloudari, Edris Hassannataj Joloudari, Hamid Saadatfar, Mohammad GhasemiGol ...+5 more authors

Published on: 20 Jan 2020

Topics: Coronary artery disease, Ranking (information retrieval) and Heart disease

International Journal of
Environmental Research and Public Health

MDPI

Article

Coronary Artery Disease Diagnosis; Ranking the Significant Features Using a Random Trees Model

Javad Hassannataj Joloudari [1], Edris Hassannataj Joloudari [2], Hamid Saadatfar [1], Mohammad Ghasemigol [1], Seyyed Mohammad Razavi [3], Amir Mosavi [4,5,6,7], Narjes Nabipour [8,*], Shahaboddin Shamshirband [9,10,*] and Laszlo Nadai [4]

1 Department of Computer Engineering, Faculty of Engineering, University of Birjand, Birjand 97175/615, Iran; Javad.hassannataj@birjand.ac.ir (J.H.J.); saadatfar@birjand.ac.ir (H.S.); ghasemigol@birjand.ac.ir (M.G.)
2 Department of Nursing, School of Nursing and Allied Medical Sciences, Maragheh Faculty of Medical Sciences, Maragheh, Iran; edris.hn2000@gmail.com
3 Department of Electronics, Faculty of Electrical and Computer Engineering, University of Birjand, Birjand 9717434765, Iran; smrazavi@birjand.ac.ir
4 Kalman Kando Faculty of Electrical Engineering, Obuda University, 1034 Budapest, Hungary; amir.mosavi@kvk.uni-obuda.hu (A.M.); nadai@uni-obuda.hu (L.N.)
5 Institute of Structural Mechanics, Bauhaus Universität-Weimar, D-99423 Weimar, Germany
6 Department of Mathematics and Informatics, J. Selye University, 94501 Komarno, Slovakia
7 Faculty of Health, Queensland University of Technology, 130 Victoria Park Road, Queensland 4059, Australia
8 Department Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam
9 Department for Management of Science and Technology Development, Ton Duc Thang University, Ho Chi Minh City, Vietnam
10 Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam
* Correspondence: narjesnabipour@duytan.edu.vn (N.N.); shahaboddin.shamshirband@tdtu.edu.vn (S.S.)

Received: 20 December 2019; Accepted: 20 January 2020; Published: 23 January 2020

Abstract: Heart disease is one of the most common diseases in middle-aged citizens. Among the vast number of heart diseases, coronary artery disease (CAD) is considered as a common cardiovascular disease with a high death rate. The most popular tool for diagnosing CAD is the use of medical imaging, e.g., angiography. However, angiography is known for being costly and also associated with a number of side effects. Hence, the purpose of this study is to increase the accuracy of coronary heart disease diagnosis through selecting significant predictive features in order of their ranking. In this study, we propose an integrated method using machine learning. The machine learning methods of random trees (RTs), decision tree of C5.0, support vector machine (SVM), and decision tree of Chi-squared automatic interaction detection (CHAID) are used in this study. The proposed method shows promising results and the study confirms that the RTs model outperforms other models.

Keywords: heart disease diagnosis; coronary artery disease; machine learning; health informatics; data science; big data; predictive model; ensemble model; random forest; industry 4.0

1. Introduction

Today, the healthcare industry has accumulated big data [1]. Data science has been greatly empowering the advancement of novel technologies for bringing insight into big data for smart diagnose, disease prevention, and policy-making in healthcare industry [2,3]. Among the data science technologies, machine learning for big data has been reported as the most effective strategies with an ever-increasing popularity in a broad range of applications in preventive healthcare [4]. The relatively low- cost computation, high performance, robustness, generalization ability, and high accuracy have been often associated with machine learning methods [5].

Common methods such as angiography [6] for diagnosing is costly and has adverse side effects. Consequently, the healthcare industry has been investing on computer-aided disease diagnosis methods such as machine learning. Whereas, the data mining process by utilizing machine learning science and database management knowledge [1] has become a robust tool for data analysis and management of big data which ultimately leads to knowledge extraction. It should be noted that with the progress of health informatics, including industry 4.0, the healthcare systems have been more intelligent, automated, and equipped with early diagnosis, warning systems, and predictive strategies [7]. In this new generation, with the development of new medical devices, equipment, and tools, new knowledge can be gained in the field of disease diagnosis. One of the best ways to quickly diagnose diseases is to use computer-assisted decision making, i.e., machine learning to extract knowledge from data. In general, knowledge extraction from data is an approach that can be very crucial for the medical industry in diagnosing and predicting diseases. In other words, the purpose of knowledge extraction is the discovery of knowledge from databases in the data mining process. Data mining is used as a suitable approach to reduce costs and for quick diagnosis of the disease.

Therefore, the purpose of the data mining process, known as database knowledge discovery (KDD), is to find a suitable pattern or model of data that was previously unknown so that these models can be used for specific disease diagnosis decisions in the healthcare environment [1]. Steps of the KDD process [1] include data cleaning (to remove disturbed data and conflicting data) and data integration (which may combine multiple data sources), data selection (where appropriate data is retrieved from the database for analysis), and data transformation (where data are synchronized by performing summary or aggregation operations and transformed appropriately for exploring), data mining (the essential process in which intelligent methods are used to extract data patterns), pattern evaluation (to identify suitable patterns that represent knowledge based on fit measurements), and knowledge presentation (where visualization and presentation techniques are used to provide users with explored knowledge) are shown in Figure 1.

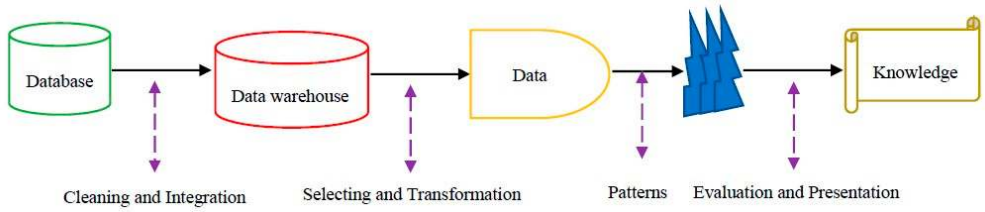

Figure 1. Database knowledge discovery (KDD) process steps [1].

The subject of this study is in the field of heart disease. Heart disease encompasses a variety of conditions, including congenital diseases, coronary artery disease, and heart rheumatism. Among these conditions, coronary artery disease is the most common, therefore, comprehensive reports of heart disease have been conducted in recent years on heart disease. The World Health Organization (WHO) has declared coronary artery disease (CAD) as the most common type of cardiovascular disease [8]. More than 30% of deaths worldwide were due to CAD, which resulted in more than 17 million deaths in 2015 [9]. Additionally, more than 360,000 Americans have died from heart attacks [10]. As a result, heart disease costs alone total more than $200 billion in the United States annually [10]. In addition, health care costs for heart disease will double by 2030, according to the American Heart Association [11].

Hence, in this paper, attention has been paid to a heart disease case study in order to apply a prediction method to coronary artery disease [6,12]. One way to accurately diagnose this disease is to use data mining methods to build an appropriate and robust model that is more reliable than medical imaging tools, including angiography in the field of diagnosis of coronary heart disease [4–6]. A main challenge in model learning is the feature selection problem, as the feature selection step is so important in data mining and its purpose is to eliminate unnecessary and unimportant features [1,13–15]. The

method used in this study is the feature ranking-selection method to choose the best subset of features in dataset. In this method, we utilize various data mining methods including random trees (RTs), decision tree of C5.0, Chi-squared automatic interaction detection (CHAID), and support vector machine (SVM). Through these methods, the selection of the subset of features according to their order of priority takes place. For this purpose, the subset of features is ranked from the least important to the most important due to the different weightings to the features associated with the classification models that these features were assigned to in the output simulator.

Finally, among the classification models used in this study, obtaining the most appropriate subset feature by random trees model with the best classification set and the most accurate classification of coronary-heart disease diagnosis is the main purpose of this study. As a result, in terms of accuracy, area under the curve (AUC) and Gini value criteria for CAD diagnosis, random trees model is the best model compared to other prediction models.

The rest of the paper is organized as follows: data mining classification methods are presented in Section 2 and related works are described in Section 3. The proposed methodology is explained in Section 4. Section 5 represents the evaluated results of the experiment. Section 6 presents findings of the research and the conclusions, namely "Results and Discussion" and "Conclusions" in Section 7.

2. Data Mining Classification Methods

In this section, we describe the classification methods used in this study. These methods include CHAID decision tree, C5.0 decision tree, random trees, and support vector machine (SVM). Among the mentioned methods, except for the support vector machine, CHAID, C5.0, because random trees (RTs) are based on the decision tree, rules are extracted that are useful for the diagnosis of CAD, especially rule extraction using RTs.

2.1. Decision Tree of CHAID

The Chi-squared automatic interaction detection (CHAID) is one of the oldest tree classifications, and it is a supervised learning method by building the decision tree, which is evidence of the rules extraction, which is proposed by Kass [16]. This classification model is a statistical method based on the diagnosis of Chi-squared automatic interaction, and it is a recessive partitioning method that can be obtained by input features as predictors and the predictive class, a Chi-squared statistic test between target class and the predictive features are computed [17–19] so that the predictive features are ranked in order of their priority. As such, the most significant predictors of subset feature with the highest probability of their weight to diagnose CAD can be gained. It should be noted that the process of selecting a significant predictor feature is based on data sample segmentation so that until we reach an external node i.e., the leaf, the samples partition continues into smaller subdivisions [17,20].

In general, the CHAID model includes the following steps [17–19]:

1. Reading predictors: the first step is to make classified predictors or features out of any consecutive predictors by partitioning the concerned consecutive disseminations into a number of classifiers with almost equal numbers of observations. For classified predictors, the classifiers or target classes are determined.
2. Consolidating classifiers: the second step is to round through the features to estimate for each feature the pair of feature classifiers that is least significantly different with regard to the dependent variable. In this process, the CHIAD model includes two types of statistical tests. One, for the classification dataset, it will gain a Chi-square test or Pearson Chi-square. The assumptions for Chi-square test are as follows:

N_{ij} = The value of observations concerned with feature fields or sample size,

G_{ij} = The gained expected feature fields for datasets, for example, the training dataset $(xn = i, yn = j)$,

V_n = The value weight (W_n) concerned with per sample of dataset,

D_f = The most number of logically independent values, which are values that have the freedom to vary, in the dataset, namely, Degrees of Freedom. D_f is equal to ($Nij-1$).

C = The corresponsive data sample, afterward:

$$X^2 = \sum_{j=1}^{j} \sum_{i=1}^{D} \frac{(N_{ij} - G_{ij})^2}{G_{ij}}, \tag{1}$$

$$N_{ij} = \sum_{N \epsilon C} F_n D_f (X_n = i \cap Y_n = j). \tag{2}$$

Two, for regression datasets where the dependent variable is consecutive, for measure-dependent variables, F-tests are used. If the concerned test for a given pair of feature classifiers is not statistically significant as defined by an alpha-to-consolidate value, then it will consolidate the concerned feature classifiers and iterate this step, i.e., obtain the next pair of classifiers, which now may include previously consolidated classifiers. If the statistical significance for the concerned pair of feature classifiers is significant, i.e., less than the concerned alpha-to-consolidate value), then it will gain optionally a Bonferroni adopted p-value for the set of classifiers for the concerned feature.

$$F = \frac{\sum_{i=1}^{D} \sum_{N \epsilon C} W_n V_n D_f (X_n = i) \left(Y_i' - Y\right)^2 / (D_f - 1)}{\sum_{i=1}^{D} \sum_{N \epsilon C} W_n V_n D_f (X_n = i) (Y_n - Y')^2 / (N_f - D_f)}, \tag{3}$$

given that the functions Y_n, Y, and N_f are formulated as follows:

$$Y_n = \frac{\sum_{N \epsilon C} W_n V_n Y_n D_f (X_n = i)}{\sum_{N \epsilon C} W_n V_n D_f (X_n = i)}, \tag{4}$$

$$Y = \frac{\sum_{N \epsilon C} W_n V_n Y_n D_f}{\sum_{N \epsilon C} W_n V_n D_f}, \tag{5}$$

$$N_f = \sum_{N \epsilon C} V_n. \tag{6}$$

3. Selecting the partition variable: the third step is to select the partition of the predictor variable with the smallest adapted p-value, i.e., the predictor variable that will gain the most significant partition. The p-value is formulated in a $\left(P = pr\left(X_c^e > X^2\right)\right)$. If the smallest (Bonferroni) adopted p-value for any predictor feature is greater than some alpha-to-partition value, then no further partitions will be done, and the concerned node is a final node. This process is continued until no further partitions can be done, i.e., given the alpha-to-consolidate and alpha-to-partition values). Eventually, according to step 2, the p-value is obtained as follows:

$$P = P\left(F\left(C - 1,\ N_f - 1\right) > F\right). \tag{7}$$

2.2. Decision Tree of C5.0

Following is the process of improving decision tree models including ID3 [21,22], C4.5 [23–25], the C5.0 tree model [26–29] as the latest version of decision tree models developed by Ross. The improved C5.0 decision tree is manifold faster than its ally models in terms of speed. In terms of memory usage, the memory gain in this model is much higher than the other models mentioned. The model also improves trees by supporting boosting and bagging [25] so that using it increases accuracy of diagnosis. As one of the common characteristics among decision trees is weighting of disease features, but the C5.0 model allows different features and types of incorrect classifies to be weighted.

One of the crucial advantages of the C5.0 model to test the features is the gain ratio, with increased information gain, i.e., the information entropy, and the bias is reduced [1,17,29]. For example,

the assumptions for the information entropy, information gain, and gain ratio problems are as follows [1,17,25]:

We assume the S as a set of training dataset and split S into n subsets, and, N_i = the sample dataset of K features.

Therefore, we obtain the features to diagnosis CAD selected with the least information entropy, and the most information gain and gain ratio. The information entropy, information gain, and gain ratio are formulated as follows.

$$Info_{\,Gain}(S,K) = \sum_{i=1}^{N\epsilon C_i} P_i \times Info_{\,Entropy}(S_i), \tag{8}$$

$$Info_{\,Entropy}(S) = -\sum_{j=1}^{N\epsilon C_i} P_i log_2 P_i. \tag{9}$$

Based on Equations (8) and (9), the number of K features, a partition S according to values of K, and where P is the probability distribution of division ($C1$, $C2$, ... , Ci):

$$\mathrm{P} = (|\mathrm{C1}|/|\mathrm{S}|, |\mathrm{C2}|/|\mathrm{S}|, \ldots, |\mathrm{Ci}|/|\mathrm{S}|). \tag{10}$$

Based on Equation (10), Ci is the number of disjoint classes and $|S|$ is the number of samples in set of S. The value of Gain is computed as follows.

$$Gain\ (S,K) = Info_{\,Entropy}(S) - Info_{\,Gain}(S,K). \tag{11}$$

Ratio instead of gain was suggested by Quinlan so that *Split Info (K,C)* is the information due to the division of C on the basis of value of categorical feature K, using the following:

$$Split\ Info(K,C) = Info_{\,entropy}\left(\frac{|C1|}{|C|}, \frac{|C2|}{|C|}, \ldots, \frac{|Ci|}{|C|}\right), \tag{12}$$

$$GainRatio(K,C) = Gain(K,C)\ /\ Split\ Info(K,C). \tag{13}$$

For Equations (12) and (13), where $(\mathrm{C1},\ \mathrm{C2},\ \ldots\ \mathrm{Ci})$ is the partition of C induced by value of K.

2.3. Support Vector Machine

Support Vector Machine (SVM) is a supervised learning model based on statistical learning theory and structural risk minimization [29,30] presented by Vapnic that only the data assigned in the support vectors are based on machine learning and model building. The SVM model is not sensitive to other data points and its aim is to find the best separation line, i.e., the optimal hyperplane between the two classes of samples so that it has the maximum distance possible to all the two classes of support vectors [29–32]. The predictor feature is determined by the separator line for each predictive class. Figure 2 shows the scheme of the support vector machine in two-dimensional space.

Regarding Figure 2, a description of SVM model is as follows:

Allow training data sample $\{(xi,\ yi)\}i = 1 \ldots n,\ xi \in R^d$ and the data of the two classes labeled $yi\ \in\ \{-1,\ 1\}$ to be separated by a optimal hyperplane in a $\{x\langle w,x\rangle + b = 0\}$ so it is assigned in the middle of the other two lines, i.e., $\{x\langle w,x\rangle\ +\ b\ =\ +1\}$ *and* $\{x\langle w,x\rangle\ +\ b = -1\}$ with margin M that the margin $\left(M = \frac{2}{\|W\|}\right)$ of the separator is the distance between support vectors, data samples closest to the hyperplane are support vectors, and also, b represents the offset between the optimal hyperplane and the origin plane. Then for each training sample $(xi,\ yi)$:

$$\left\{ \begin{array}{l} W^T X_i\ +\ b\ \leq\ -\frac{M}{2}\ if\ yi\ =\ -1 \\ W^T X_i\ +\ b\ \geq\ \frac{M}{2}\ if\ yi\ =\ 1 \end{array} \right\} \leftrightarrow Y_i\left(W^T X_i + b\right) \geq \frac{\mathrm{M}}{2}. \tag{14}$$

According to the hyperplane optimization that the SVM model was meant to solve, the optimization problem is as follows [29]:

$$Minimize: \frac{1}{2} \parallel W \parallel^2 \ subjected\ to: \ y_i(w.x + b) - 1 \geq 0 \ \forall i. \tag{15}$$

To solve the problem of Equation (15), one has to obtain the dual of the problem using the Lagrange Method, namely, (*Lp*). To obtain the dual form of the problem, the nonnegative Lagrangian coefficients are multiplied by $\alpha i \geq 0$. L_p is defined as follows:

$$L_p = \frac{1}{2} |w|^2 - \sum_i \alpha_i \ (y_i(w.x + b) - 1. \tag{16}$$

Finally, Equation (16) is transformed into the following equation [29]:

$$\text{Maximize} \ : L_D = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j Y_i Y_j \left(X_i X_j\right). \tag{17}$$

Equation (17) is called the dual problem, namely, L_D. However, for nonlinear SVM due to the absence of trade-off between maximizing the margin and the misclassification. Therefore, it could not obtain the linear separate hyperplane in the data sample. In the nonlinear space, the best solution, the basic data of higher dimension, i.e., feature space, of the linear separate is transformed. At the end, kernel functions are used, such as linear, polynomial, radial basis function (RBF), and sigmoid [29]. Based on Equation (18), L_D for the nonlinear data sample is obtained. In Equation (18), parameter *C* is the penalty agent and determines the measure of penalty placed to a fault, so that the "*C*" value is randomly selected by the user.

$$\begin{gathered} \text{Maximize } \Phi(W, b, \xi, \alpha, \beta) : L_D = \sum_{i=1}^{N} \alpha_i - \tfrac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j Y_i Y_j K\left(X_i, X_j\right) \\ \text{subjected to } \sum_j \alpha_j Y_j = 0, \ 0 \leq \alpha_i \leq C, \ i = 1, \ldots, N. \end{gathered} \tag{18}$$

N is the number of data samples in Equation (18). In this study, the radial basis function (RBF) [29] is selected as the kernel function as shown in Equation (19):

$$K\left(X_i, X_j\right) = \ \exp(-\Upsilon \parallel X_i - X_j \parallel^2 . \tag{19}$$

In Equation (19), the kernel parameter of Υ with respect to $(\Upsilon \ \geq 0)$ represents the width of the RBF.

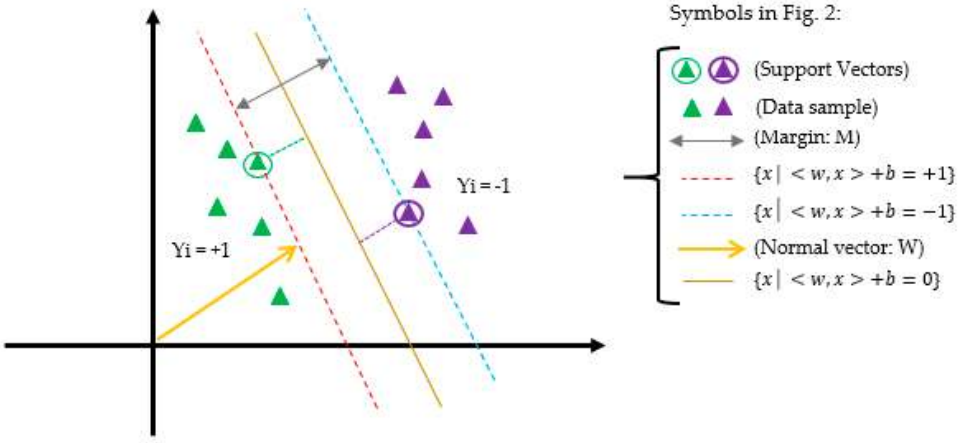

Figure 2. Support vector machine in two-dimensional space.

2.4. Random Trees

The model of random trees (RTs) is one of the robust predictive models, better than other classification models in terms of accuracy computing, data management, more information gain with eliminating fewer features, extracting better rules, working with more data, and more complex networks. Therefore, the model for disease diagnosis is suitable. This model consists of multiple trees randomly with high depth so that the most significant votes are chosen from a set of possible trees having K random features at each node. In other words, in the set of trees, each tree has an equal probability of being assigned. Due to the experiments performed in the classification of the dataset, the accuracy of the RTs model is more accurate than the other models because it uses the evaluation of several features and composes functions. Therefore, RTs can be constructed efficiently and the combination of large datasets of random trees generally leads to proper models. There has been vast research in recent years of RTs in the field of machine learning [33]. Generally, random trees is confirmed as a crucial performance as compared to the classifiers presented as a single tree in this study.

If we consider random trees at very high dimensions with a complex network, then it can include the following steps [33,34]:

1. Using the N data sample randomly, in the training dataset to develop the tree.
2. Each node as a predictive feature grasps a random data sample selected so that $m < M$ (m represents the selected feature and M represents the full of features in the corresponding dataset. Given that during the growth of trees, m is kept constant.)
3. Using the m features selected for generating the partition in the previous step, the P node is computed using the best partition path from points. P represents the next node.
4. For aggregating, the prediction dataset uses the tree classification voting from the trained trees with n trees.
5. For generating the terminal RTs, the model uses the biggest voted features.
6. The RTs process continues until the tree is complete and reaches only one leaf node.

3. Related Works

In recent years, several studies have been conducted on the diagnosis of CAD on different datasets using data mining methods. The most up-to-date dataset that researchers have used recently is the Z-Alizadeh Sani dataset in the field of heart disease. To this end, we review recent research on the Z-Alizadeh Sani dataset [35,36].

Alizadeh Sani et al. have proposed the use of data mining methods based on ECG symptoms and characteristics in relation to the diagnosis of CAD [37]. In their research, they used sequential minimal optimization (SMO) and naïve Bayes algorithms separately and in combination to diagnose the disease. Finally, using the 10-fold cross-validation method for the SMO-naïve Bayes hybrid algorithm, they achieved greater accuracy of 88.52% than the SMO (86.95%) and naïve Bayes (87.22%) algorithms.

In another study, Alizadeh Sani et al. developed classification algorithms such as SMO, naïve Bayes, bagging with SMO, and neural networks for the diagnosis of CAD [12]. Confidence and information gain on CAD have also been used to determine effective features. As a result, among these algorithms, SMO algorithm with information gain has the best performance, with accuracy of 94.08% using the 10-fold cross-validation method.

Alizadeh Sani et al. have used computational intelligence methods to diagnose CAD, and they have separately diagnosed three major coronary stenosis using demographics, symptoms and examination, ECG characteristics, laboratory analysis, and echo [38]. They have used analytical methods to investigate the importance of vascular stenosis characteristics. Finally, using the SVM classification model with 10-fold cross-validation method, along with feature selection of combined information gain and average information gain, they obtained accuracies of 86.14%, 83.17%, and 83.50% for left anterior descending (LAD), left circumflex (LCX), and right coronary arteries coronaries (RCA), respectively.

Arabasdi et al. have presented a neural network-genetic hybrid algorithm for the diagnosis of CAD [39]. For this purpose, in their research, genetic and neural network algorithms have been used separately and a hybrid to analyze the dataset, and the accuracy of the neural network algorithm and neural network-genetic algorithm using the 10-fold cross-validation method was 84.62% and 93.85%, respectively.

Alizadeh Sani et al. have performed a feature engineering algorithm that used the naïve Bayes, C4.5, and SVM classifiers for non-invasive diagnosis of CAD [36]. They increased their dataset from 303 records to 500 samples. The accuracy obtained using the 10-fold cross-validation method for naïve Bayes, C4.5, and SVM algorithms were 86%, 89.8%, and 96.40%, respectively.

In a study conducted by Abdar et al. [40], the authors used a two-level hybrid genetic algorithm and NuSVM called N2Genetic-NuSVM. The two-level genetic algorithm was used to optimize the SVM parameters and to select the features in parallel. Using their proposed method, the accuracy of CAD diagnosis was 93.08% through a 10-fold cross-validation method.

4. Proposed Methodology

In this section, we follow the proposed methodology in Figure 3 by using IBM Spss Modeler version 18.0 software for implementation of classification models.

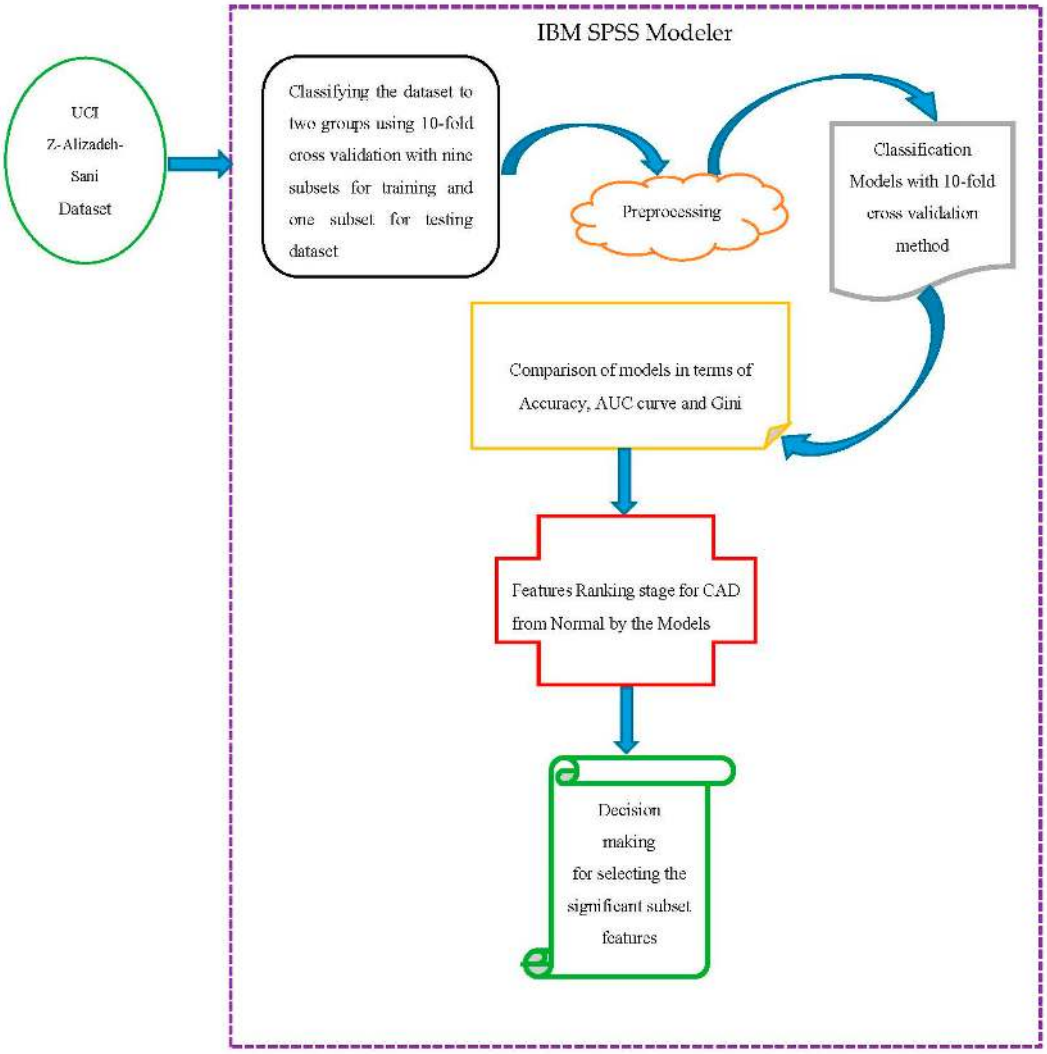

Figure 3. Proposed methodology.

4.1. Description of the Dataset

Initially based on Figure 3, to diagnose the CAD, the Z-Alizadeh Sani dataset was used in this study [35]. This dataset contains information on 303 patients with 55 features, 216 patients with CAD, and 88 patients with normal status. The features used in this dataset were divided into four groups that were features of CAD for patients including demographics, symptoms, and examination, electrocardiogram (ECG), and laboratory and echo features, described in Table 1. For categorizing the CAD from Normal, the diameter narrowing above 50% represents a patient as CAD, and its absence is stated as Normal [12].

Table 1. Description of the features used in the Z-Alizadeh-Sani dataset with their valid ranges.

Feature Type	Feature Name	Range	Measurement			
			Mean	Std. Error of Mean	Std. Deviation	Variance
Demographic	Age	(30–80)	58.90	0.6	10.39	108
Demographic	Weight	(48–120)	73.83	0.69	11.99	143.7
Demographic	Length	(140–188)	164.72	0.54	9.33	87.01
Demographic	Sex	Male, Female	—	—	—	—
Demographic	BMI (body mass index Kb/m^2)	(18–41)	27.25	0.24	4.1	16.8
Demographic	DM (diabetes mellitus)	(0, 1)	0.3	0.03	0.46	0.21
Demographic	HTN (hypertension)	(0, 1)	0.6	0.03	0.49	0.24
Demographic	Current smoker	(0, 1)	0.21	0.02	0.41	0.17
Demographic	Ex-smoker	(0, 1)	0.03	0.01	0.18	0.03
Demographic	FH (family history)	(0, 1)	0.16	0.02	0.37	0.13
Demographic	Obesity	Yes if MBI > 25, No otherwise	—	—	—	—
Demographic	CRF (chronic renal failure)	Yes, No	—	—	—	—
Demographic	CVA (cerebrovascular accident)	Yes, No	—	—	—	—
Demographic	Airway disease	Yes, No	—	—	—	—
Demographic	Thyroid disease	Yes, No	—	—	—	—
Demographic	CHF (congestive heart failure)	Yes, No	—	—	—	—
Demographic	DPL (dyslipidemia)	Yes, No	—	—	—	—
Symptom and examination	BP (blood pressure mm Hg)	(90–190)	129.55	1.09	18.94	358.65
Symptom and examination	PR (pulse rate ppm)	(50–110)	75.14	0.51	8.91	79.42
Symptom and examination	Edema	(0, 1)	0.04	0.01	0.2	0.04
Symptom and examination	Weak peripheral pulse	Yes, No	—	—	—	—
Symptom and examination	Lung rates	Yes, No	—	—	—	—
Symptom and examination	Systolic murmur	Yes, No	—	—	—	—
Symptom and examination	Diastolic murmur	Yes, No	—	—	—	—
Symptom and examination	Typical chest pain	(0, 1)	0.54	0.03	0.5	0.25
Symptom and examination	Dyspnea	Yes, No	—	—	—	—
Symptom and examination	Function class	1, 2, 3, 4	0.66	0.06	1.03	1.07

Table 1. *Cont.*

Feature Type	Feature Name	Range	Measurement			
			Mean	Std. Error of Mean	Std. Deviation	Variance
Symptom and examination	Atypical	Yes, No	—	—	—	—
Symptom and examination	Nonanginal chest pain	Yes, No	—	—	—	—
Symptom and examination	Exertional chest pain	Yes, No	—	—	—	—
Symptom and examination	Low TH Ang (low-threshold angina)	Yes, No	—	—	—	—
ECG	Rhythm	Sin, AF	—	—	—	—
ECG	Q wave	(0, 1)	0.05	0.01	0.22	0.05
ECG	ST elevation	(0, 1)	0.05	0.01	0.21	0.04
ECG	ST depression	(0, 1)	0.23	0.02	0.42	0.18
ECG	T inversion	(0, 1)	0.3	0.03	0.46	0.21
ECG	LVH (left ventricular hypertrophy)	Yes, No	—	—	—	—
ECG	Poor R-wave progression	Yes, No	—	—	—	—
Laboratory and echo	FBS (fasting blood sugar mg/dL)	(62–400)	119.18	2.99	52.08	2712.29
Laboratory and echo	Cr (creatine mg/dL)	(0.5–2.2)	1.06	0.02	0.26	0.07
Laboratory and echo	TG (triglyceride mg/dL)	(37–1050)	150.34	5.63	97.96	9596.05
Laboratory and echo	LDL (low-density lipoprotein mg/dL)	(18–232)	104.64	2.03	35.4	1252.93
Laboratory and echo	HDL (high-density lipoprotein mg/dL)	(15–111)	40.23	0.61	10.56	111.49
Laboratory and echo	BUN (blood urea nitrogen mg/dL)	(6–52)	17.5	0.4	6.96	48.4
Laboratory and echo	ESR (erythrocyte sedimentation rate mm/h)	(1–90)	19.46	0.92	15.94	253.97
Laboratory and echo	HB (hemoglobin g/dL)	(8.9–17.6)	13.15	0.09	1.61	2.59
Laboratory and echo	K (potassium mEq/lit)	(3.0–6.6)	4.23	0.03	0.46	0.21
Laboratory and echo	Na (sodium mEq/lit)	(128–156)	141	0.22	3.81	14.5
Laboratory and echo	WBC (white blood cell cells/mL)	(3700–18.000)	7562.05	138.67	2413.74	5,826,137.52
Laboratory and echo	Lymph (lymphocyte %)	(7–60)	32.4	0.57	9.97	99.45
Laboratory and echo	Neut (neutrophil %)	(32–89)	60.15	0.59	10.18	103.68
Laboratory and echo	PLT (platelet 1000/mL)	(25–742)	221.49	3.49	60.8	3696.18
Laboratory and echo	EF (ejection fraction %)	(15–60)	47.23	0.51	8.93	79.7
Laboratory and echo	Region with RWMA	(0–4)	0.62	0.07	1.13	1.28
Laboratory and echo	VHD (valvular heart disease)	Normal, Mild, Moderate, Severe	—	—	—	—
Categorical	Target class: Cath	CAD, Normal	—	—	—	—

4.2. Classifying the Dataset

Data was classified into nine subsets i.e., 90% for training the classifiers and one subset i.e., 10% for testing dataset using 10-fold cross-validation.

4.3. Preprocessing the Dataset

The preprocessing step was performed after the data was classified. In general, a set of operations lead to the creation of a set of cleaned data that can be used on the dataset, investigation operation, so-called data preprocessing. The samples values in the Z-Alizadeh-Sani dataset [35] were numeric and string. The purpose of preprocessing the data in this study was to homogenize them so that all data was in the domain of (0, 1), which is called the normalization operation, so that the standard normalization operation was employed using the Min-Max function. After normalizing numbers, the string data was transformed to numeric. In this regard, given the nature of the string data, the value was assigned to them in the interval (0, 1). For example, the sex feature had male and female values that transformed to zero and one, respectively.

4.4. Classifying the Models Using the 10-Fold Cross-Validation Method

For classifying the models the 10-fold cross-validation method was used [41], where the dataset was randomly divided into the same K-scale for the division so and the k-1 subset being used to train the classifiers. The rest of the division was also used to investigate the output performance at each step, and repeated 10 times. For this purpose, classification of the prediction models was performed based on the 10-fold cross-validation method so that the average of the criteria was obtained 10-fold [1,42], as 90% of the data was used for training and 10% was used for testing the data. Finally, this cross-validation process was executed 10 times so that the results were demonstrated by averaging each 10 times.

5. Evaluating the Results

In this section, we examine the evaluation in two subdivisions. First, evaluation based on the classification criteria, including ROC curve, Gini, gain, confidence, return on investment (ROI), profit, and response. Second, the evaluation was based on significant predictive features.

5.1. Evaluation Based on Classification Criteria

We used a confusion matrix [1,39,43,44] to evaluate classification models such as SVM, CHAID, C5.0, and RTs in the diagnosis of CAD on the Z-Alizadeh Sani dataset described in Table 2.

Table 2. Confusion matrix for detection of coronary artery disease (CAD).

The Actual Class	The Predicted Class	
	Disease (CAD)	**Healthy (Normal)**
Positive	True Positive	False Positive
Negative	False Negative	True Negative

In the following, through the confusion matrix method, the AUC [1,45] and the Gini index [46] criteria were obtained, and the comparison between the models mentioned for this AUC criterion are shown in Figure 4a,b.

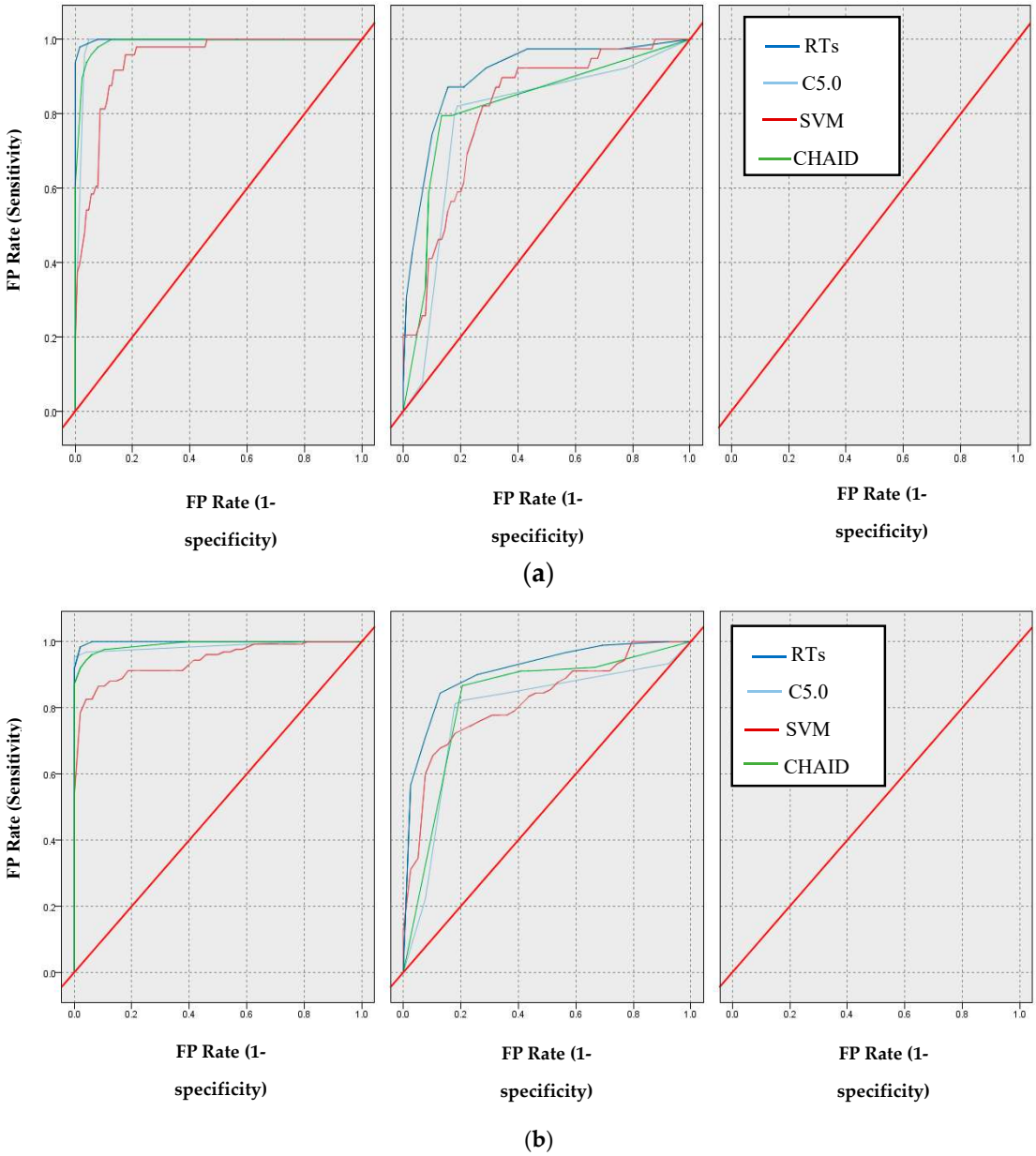

Figure 4. Comparison based on ROC of models: (**a**) Normal class (**b**) CAD class.

According to Figure 4b, the AUC values for the SVM, CHAID, C5.0, and RTs models are 80.90%, 82.30%, 83.00%, and 90.50%, respectively. Additionally, the Gini values for SVM, CHAID, and RTs models are 61.80%, 64.60%, 66.00%, and 93.40%, respectively.

In addition, the gain, confidence, profit, ROI, and response criteria for evaluating the models were examined, and comparisons between models through these criteria are shown in Figures 5–9.

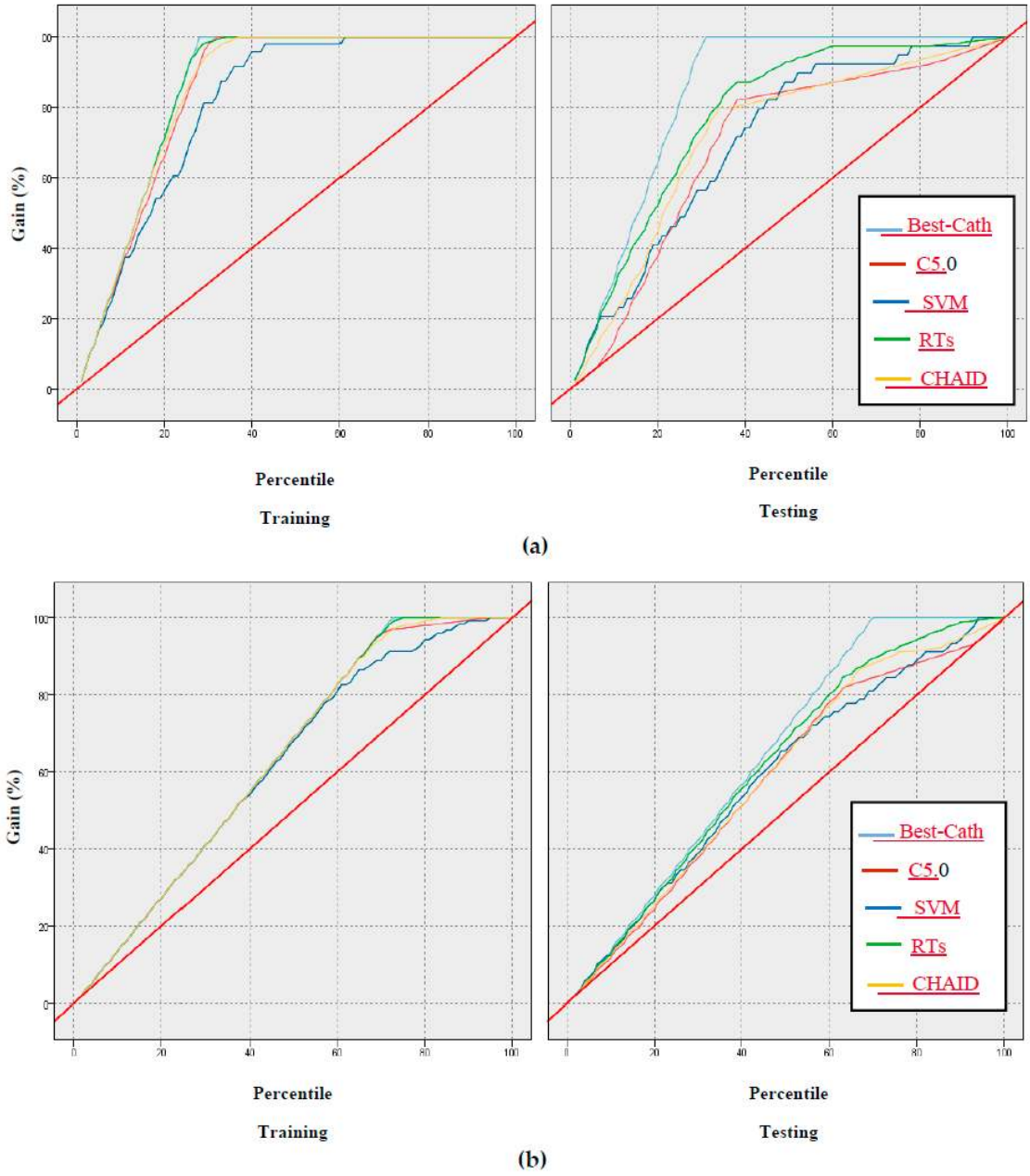

Figure 5. Results based on gain of models: (**a**) Normal class, (**b**) CAD class.

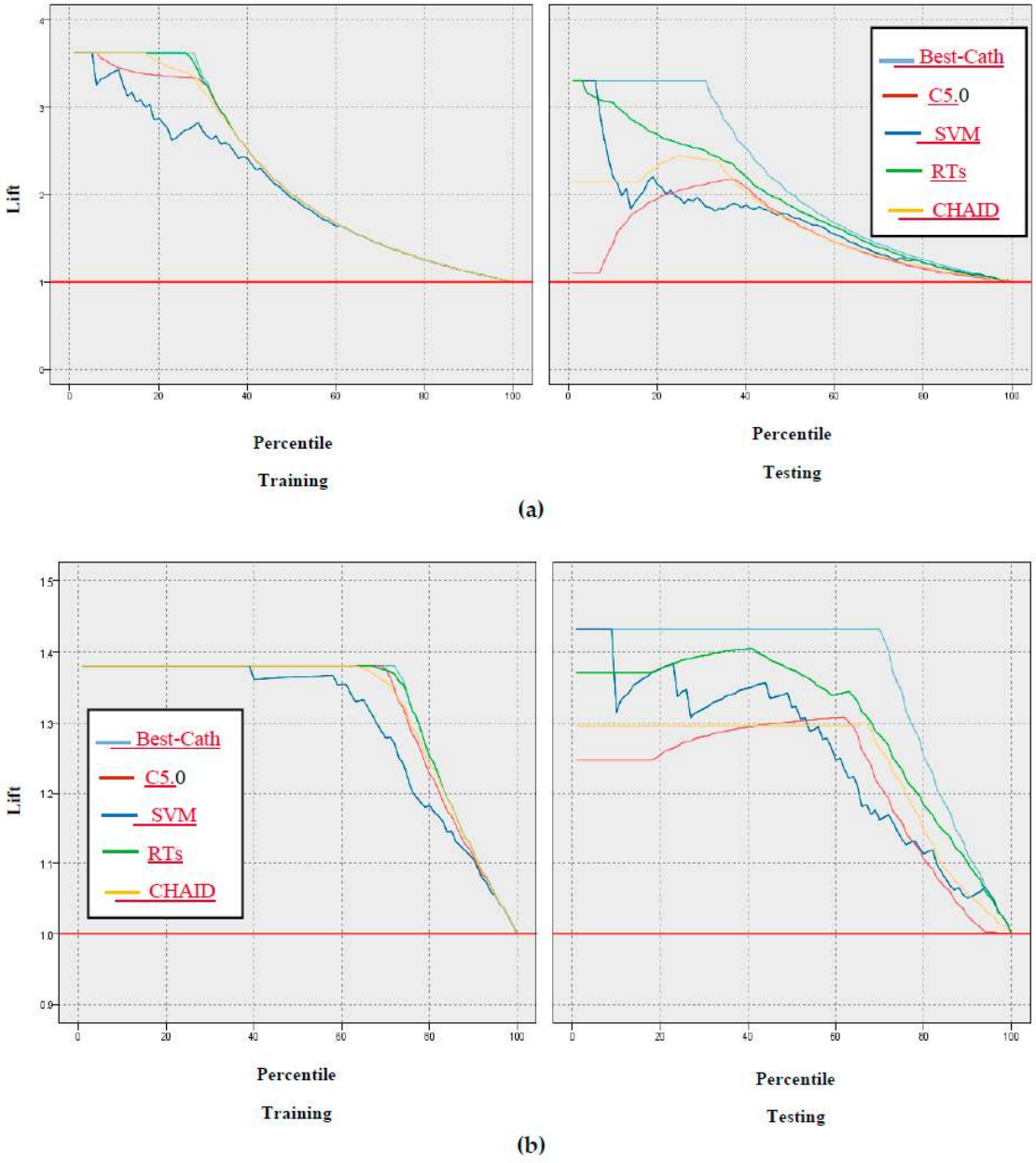

Figure 6. Results based on confidence through the lift chart of models: (**a**) CAD class, (**b**) Normal class.

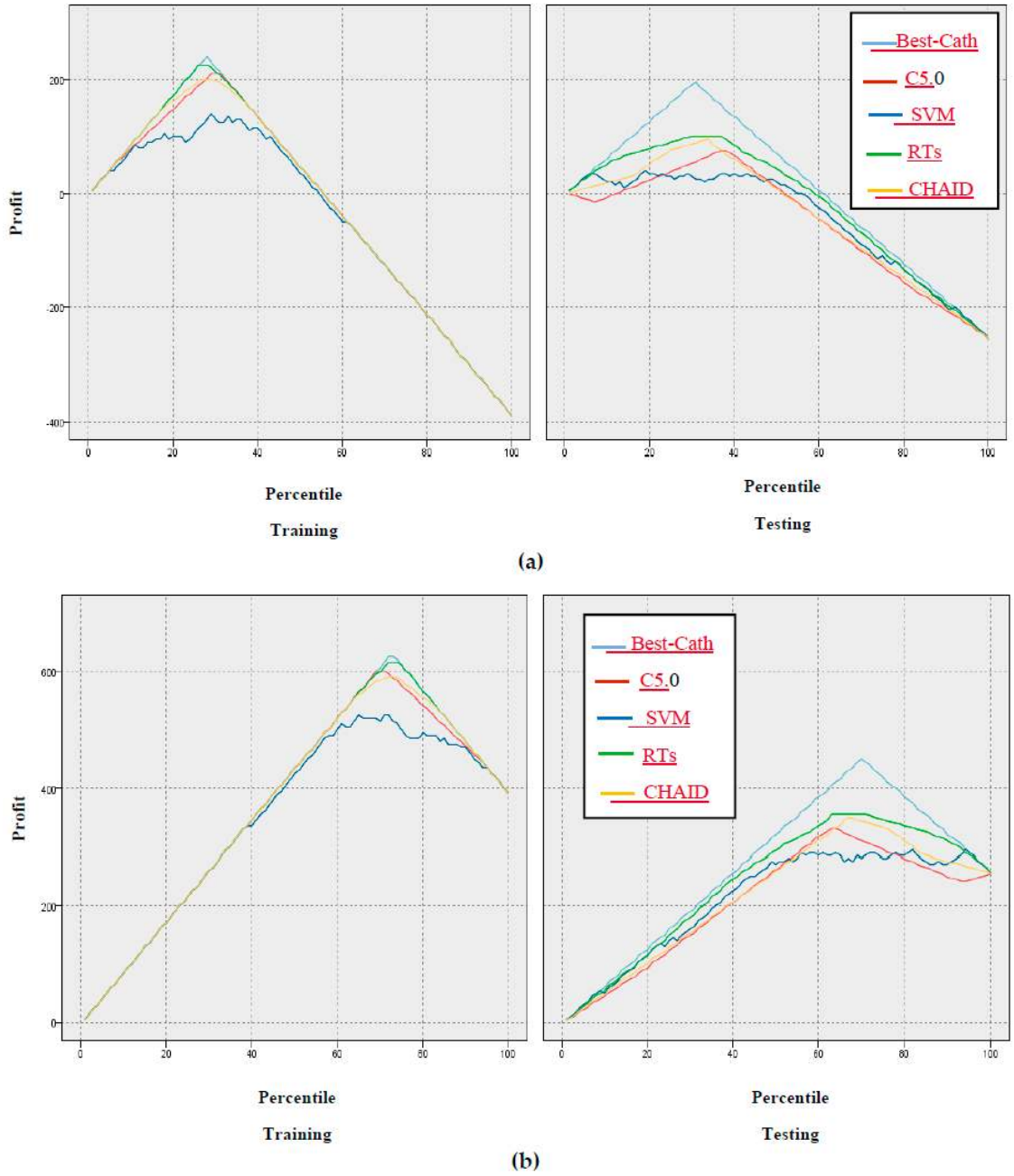

Figure 7. Results based on profit of models: (**a**) Normal class, (**b**) CAD Class.

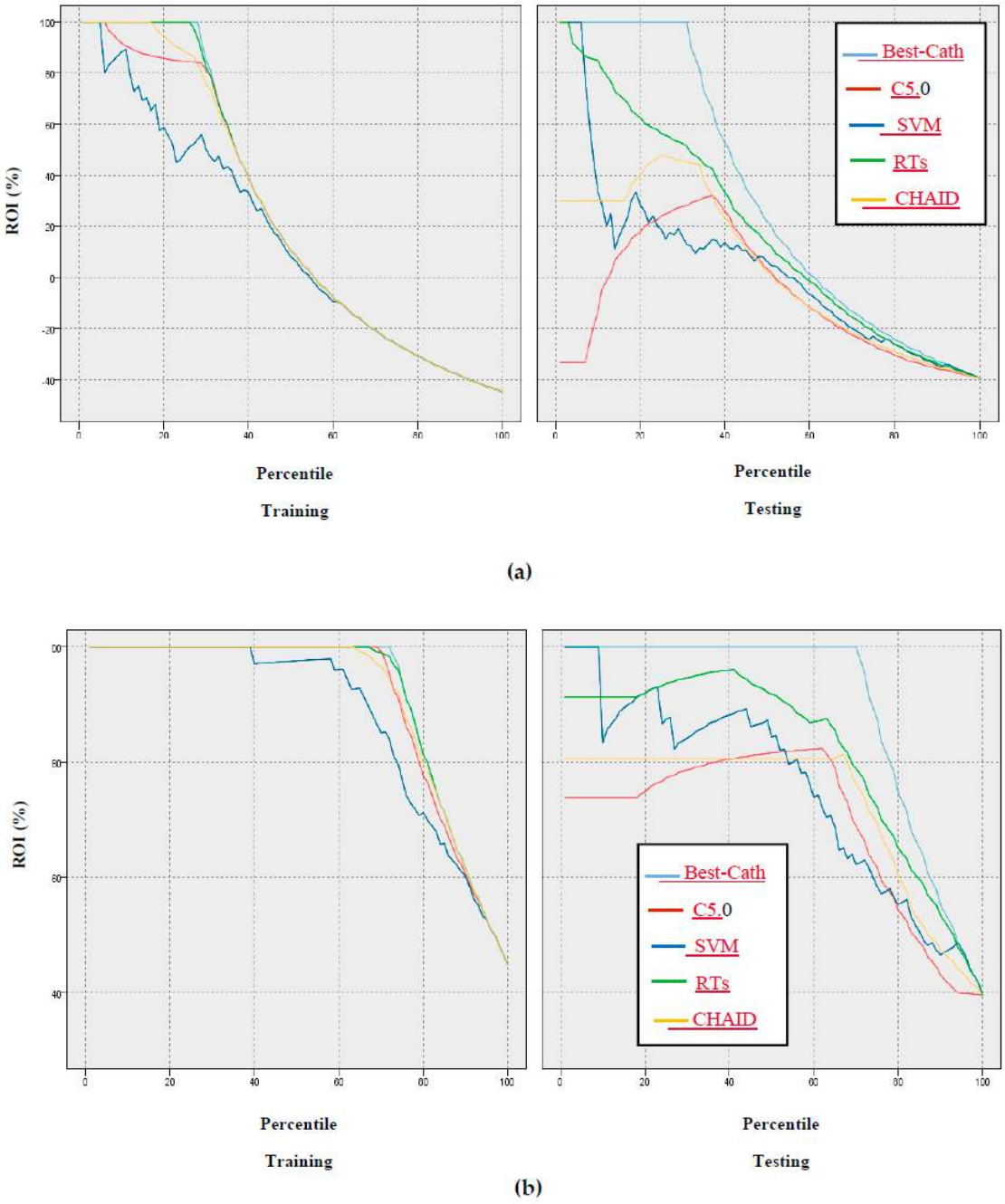

Figure 8. Results based on ROI of models: (**a**) CAD class, (**b**) Normal class.

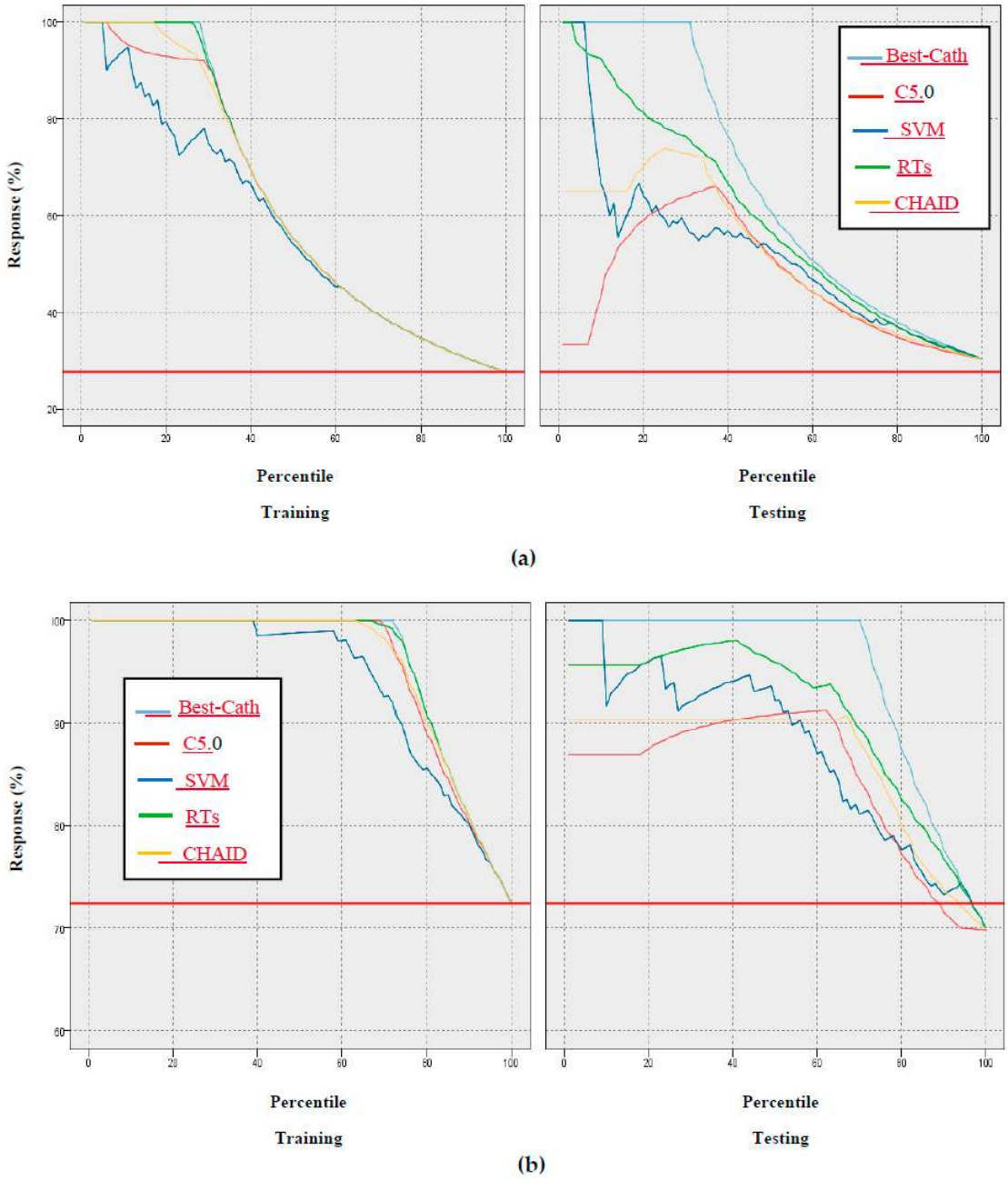

Figure 9. Results based on response of models: (**a**) CAD class, (**b**) Normal class.

According to Figures 5–9 of the criteria in the relevant models for the CAD diagnosis of the Normal class, it can be said that the RTs model has better performance in terms of gain, confidence, ROI, profit, and response criteria than other classification models.

5.2. Evaluation Based on Significant Predictive Features

One of the significant evaluations for comparing classification models for predicting the CAD from Normal is the use of the importance of predictive features. To this end, we examined the models in terms of their importance in the ranking stage of features. In fact, the models were measured according to the weight determined by the predictor features. The weighted importance of the features for the models is shown in Tables 3–5.

Table 3. Predictor significance for features based on ranking for the random trees model.

No.	Feature	Predictor Significance
1	Typical chest pain	0.98
2	TG	0.66
3	BMI	0.63
4	Age	0.58
5	Weight	0.54
6	BP	0.51
7	K	0.48
8	FBS	0.43
9	Length	0.37
10	BUN	0.3
11	PR	0.29
12	HB	0.26
13	Function class	0.25
14	Neut	0.25
15	EF-TTE	0.25
16	WBC	0.24
17	DM	0.23
18	PLT	0.2
19	Atypical	0.19
20	FH	0.18
21	HDL	0.16
22	ESR	0.16
23	CR	0.14
24	LDL	0.14
25	T inversion	0.13
26	DLP	0.13
27	Region RWMA	0.12
28	HTN	0.11
29	Obesity	0.1
30	Systolic murmur	0.09
31	Sex	0.09
32	Dyspnea	0.08
33	Current smoker	0.06
34	BBB	0.05
35	LVH	0.03
36	Edema	0.02
37	Ex-smoker	0.02
38	VHD	0.01
39	St depression	0.01
40	Lymph	0.0

Table 4. Predictor significance for features based on ranking for the support vector machine (SVM) model.

No.	Feature	Predictor Significance
1	Typical chest pain	0.04
2	Atypical	0.03
3	Sex	0.02
4	Obesity	0.02
5	FH	0.02
6	Age	0.02
7	DM	0.02

Table 4. *Cont.*

No.	Feature	Predictor Significance
8	Dyspnea	0.02
9	Systolic murmur	0.02
10	St depression	0.02
11	HTN	0.02
12	LDL	0.02
13	Current smoker	0.02
14	DLP	0.02
15	BP	0.02
16	LVH	0.02
17	Nonanginal	0.02
18	Tin version	0.02
19	Length	0.02
20	Function class	0.02
21	BBB	0.02
22	VHD	0.02
23	CHF	0.02
24	PR	0.02
25	WBC	0.02
26	BUN	0.02
27	FBS	0.02
28	ESR	0.02
29	CVA	0.02
30	Thyroid disease	0.02
31	Lymph	0.02
32	Weight	0.02
33	CR	0.02
34	Airway disease	0.02
35	TG	0.02
36	CRF	0.02
37	Diastolic murmur	0.02
38	Low TH ang	0.02
39	Exertional CP	0.02
40	Weak peripheral pulse	0.02
41	Neut	0.02
42	PLT	0.02
43	St elevation	0.02
44	EF-TTE	0.02
45	K	0.02
46	BMI	0.02
47	Ex-smoker	0.02
48	Lung rates	0.02
49	HDL	0.02
50	Na	0.01
51	Edema	0.01
52	Q wave	0.01
53	HB	0.01
54	Poor R progression	0.01
55	Region RWMA	0.01

Table 5. Predictor significance for features based on ranking for the C5.0 model.

No.	Feature	Predictor Significance
1	Typical chest pain	0.28
2	CR	0.14
3	ESR	0.13
4	T inversion	0.1
5	Edema	0.09
6	Region RWMA	0.08
7	Poor R progression	0.04
8	Sex	0.03
9	DM	0.03
10	BMI	0.02
11	WBC	0.02
12	DLP	0.02
13	Length	0.01
14	Dyspnea	0.0
15	EF-TTE	0.0

6. Results and Discussion

In the modeling process proposed in Section 4, we implemented several data mining models including SVM, CHAID, C5.0, and RTs. The 10-fold cross-validation method was used to build these models so that the data was divided into training (90%) and test (10%) subsets. The results show that the random trees model is the best classification model compared to the other models so that 91.47% accuracy of the RTs model was obtained using the 10-fold cross-validation method. While the accuracy of SVM, CHAID, and C5.0 models were 69.77%, 80.62%, and 82.17%, respectively.

The accuracy was computed using the following formula (TP + TN/(TP + TN + FP + FN)) where TP is true positive, TN is true negative, FP is false positive, and FN is false negative [1].

Another criterion for evaluating the models in this study was the AUC criterion, where 80.90%, 82.30%, 83.00%, and 96.70% was obtained for SVM, CHAID, C5.0, and RTs, respectively.

Furthermore, an achievement of this study was the use of criteria that were not found in previous studies, including gain, confidence, ROI, profit, and response, as shown in Figures 5–9. In terms of this criteria, the random trees model has the best performance compared to the other classification models.

Finally, based on Tables 3–6, it can be found that in each of the four models, the typical chest pain feature was selected as the most significant predictor so that the predictor significance of the typical chest pain feature for the random trees model was equal to 0.98, and the least significant for the lymph feature was equal to zero. Considering this, intervals 1 and 2 were applied in the simulator. In Table 1, typical chest pain was the most significant feature with a value of 0.04 and the region RWMA was the least significant feature with a value of 0.01. According to Tables 4 and 5, typical chest pain was the most significant feature, equal to 0.28 and 0.33, respectively, and the least significant feature according to Table 4 for the EF-TTE feature was equal to zero and the least significant feature according to Table 6 was 0.02. It is therefore confirmed that the RTs model is the best model relative to other classification models according to the above tables.

Table 6. Predictor significance imported for features based on ranking for the Chi-squared automatic interaction detection (CHAID) model.

No.	Feature	Predictor Significance
1	Typical chest pain	0.33
2	Age	0.15
3	T inversion	0.11
4	VHD	0.1
5	DM	0.09
6	HTN	0.04
7	Nonanginal	0.03
8	BP	0.02
9	Region RWMA	0.02
10	HDL	0.02

One of the advantages of the random trees model is the most significant obtained rules of CAD diagnosis that are mentioned in Table 7.

Table 7. The most significant obtained rules for CAD diagnosis using random trees (top decision rules for 'cath' class).

Decision Rule	Most Frequent Category	Rule Accuracy	Forest Accuracy	Interestingness Index
(BP > 110.0), (FH > 0.0), (Neut > 51.0) and (Typical Chest Pain > 0.0)	CAD	1.000	1.000	1.000
(BMI ≤ 29.02), (EF-TTE > 50.0), (CR ≤ 0.9), (Typical Chest Pain > 0.0) and (Atypical = {N})	CAD	1.000	1.000	1.000
(Weight > 8.0), (CR > 0.9), (Typical Chest Pain > 0.0) and (Atypical = {N})	CAD	1.000	1.000	1.000
(K ≤ 4.9), (WBC > 5700.0), (CR < 0.9),	CAD	1.000	1.000	1.000
(DM > 0.0) and (Typical Chest Pain > 0.0)	CAD	1.000	1.000	1.000

According to Table 7, the extracted rules for CAD are described as follows:

If the condition is true of (BP > 110.0), (FH > 0.0), (Neut > 51.0), and (Typical Chest Pain > 0.0), then the CAD exist highly accurate and also interestingness index, otherwise, the person is Normal. If the condition is true of (Typical Chest Pain > 0.0) and (Atypical = {N}), then the result it is like the result of previous conditions. In the following, if the condition is true of (Weight > 8.0), (CR > 0.9), (Typical Chest Pain > 0.0), and (Atypical = {N}), then the person is abnormal or CAD. Finally, if (K ≤ 4.9), (WBC > 5700.0), (CR < 0.9), (DM > 0.0), and (Typical Chest Pain > 0.0) is true, as a result the person is abnormal.

In recent years, several studies have been conducted on the diagnosis of CAD on different datasets using data mining methods. To this end, we review recent research on the updated Z-Alizadeh Sani dataset, as described in Table 8. The results of accuracy, AUC, and Gini criteria for the models were obtained according to the 10-fold cross validation method compared to previous studies.

Taking a look at Table 8, it can be seen that the proposed method based on random trees outperforms other methods in terms of accuracy, AUC, and Gini criteria. It implies that the 40 features extracted by using RTs are the most informative ones about the CAD disease.

Table 8. The performed works for CAD diagnosis on the Z-Alizadeh Sani dataset with the 10-fold cross validation method.

Referense	Methods	No. Features Subset Selection	Accuracy (%)	Auc %	Gini %
[37]	Naïve Bayes-SMO	16	88.52	Not reported	Not reported
[12]	SMO along with information Gain	34	94.08	Not reported	Not reported
[38]	SVM along with average information gain and also information gain	24	86.14 for LAD 83.17 for LCX 83.50 for RCA	Not reported	Not reported
[39]	Neural network-genetic algorithm-weight by SVM	22	93.85	Not reported	Not reported
[36]	SVM along with feature engineering	32	96.40	92	Not reported
[40]	N2Genetic-nuSVM	29	93.08	Not reported	Not reported
In our study	Random trees	40	91.47	96.70	93.40

7. Conclusions

In this study, a computer-aided diagnosis system was used to diagnose CAD as a common heart disease on the Z-Alizadeh Sani dataset [35], and this system was implemented using the IBM Spss Modeler version 18.0 tool. Since angiography is the most common tool of diagnosis of heart disease, it has cost and side effects for individuals. Therefore, artificial intelligence methods, that is, machine learning techniques, can be a solution to the stated challenge. Hence, such classification models including SVM, CHAID, C5.0, and random trees were used for modeling with the 10-fold cross-validation method, that are based on accuracy, AUC, Gini, ROI, profit, confidence, response, and gain, and were examined and evaluated. Finally, based on the criteria stated, the random trees model was found as the best model compared to the other models. The predictive features were selected based on the order of their priority with the highest accuracy, and we concluded that the random trees model with the most significant features of 40 and the accuracy of 91.47% has better performance than the other classification models. Considering this, with this number of features, we had more information gain than the features selected in previous works. Another achievement of this study was the important extraction rules for CAD diagnosis using the random trees model, with these rules shown in Table 6. As future work, the fuzzy intelligent system can be used in combination with artificial intelligence models to diagnose CAD on the Z-Alizadeh Sani dataset and other datasets. Another way to better diagnose CAD disease on this dataset and other real datasets is deep learning models and combining deep learning approaches with distributed design and architecture.

Author Contributions: Conceptualization, S.M.R. and A.M.; Data curation, J.H.J., E.H.J., M.G. and N.N.; Formal analysis, J.H.J., M.G. and N.N.; Funding acquisition, S.S.; Investigation, E.H.J., H.S., M.G. and S.M.R.; Methodology, E.H.J., A.M. and S.S.; Project administration, H.S.; Resources, J.H.J., H.S. and S.M.R.; Software, E.H.J. and H.S.; Supervision, M.G. and L.N.; Validation, N.N.; Visualization, M.G. and S.M.R.; Writing—original draft, J.H.J. and H.S.; Writing—review and editing, J.H.J., L.N., M.G. and A.M.; checking the results, L.N. All authors have read and agreed to the published version of the manuscript.

Funding: We acknowledge the financial support of this work by the Hungarian State and the European Union under the EFOP-3.6.1-16-2016-00010 project.

Acknowledgments: We acknowledge the financial support of this work by the Hungarian State and the European Union under the EFOP-3.6.1-16-2016-00010 project.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Joloudari, J.H.; Saadatfar, H.; Dehzangi, A.; Shamshirband, S. Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection. *Inform. Med. Unlocked* **2019**, *17*, 100255. [CrossRef]
2. Ahmad, G.; Khan, M.A.; Abbas, S.; Athar, A.; Khan, B.S.; Aslam, M.S. Automated Diagnosis of Hepatitis B Using Multilayer Mamdani Fuzzy Inference System. *J. Healthc. Eng.* **2019**, *2019*, 6361318. [CrossRef] [PubMed]
3. Wang, Y.; Kung, L.; Gupta, S.; Ozdemir, S. Leveraging big data analytics to improve quality of care in healthcare organizations: A configurational perspective. *Br. J. Manag.* **2019**, *30*, 362–388. [CrossRef]
4. Amin, M.S.; Chiam, Y.K.; Varathan, K.D. Identification of significant features and data mining techniques in predicting heart disease. *Telemat. Inform.* **2019**, *36*, 82–93. [CrossRef]
5. Zipes, D.P.; Libby, P.; Bonow, R.O.; Mann, D.L.; Tomaselli, G.F. *Braunwal's Heart Disease E-Book: A Textbook of Cardiovascular Medicine*; Elsevier Health Sciences Wiley: San Francisco, CA, USA, 2018.
6. Alizadehsani, R.; Abdar, M.; Roshanzamir, M.; Khosravi, A.; Kebria, P.M.; Khozeimeh, F.; Nahavandi, S.; Sarrafzadegan, N.; Acharya, U.R. Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Comput. Biol. Med.* **2019**, *111*, 103346. [CrossRef]
7. Pace, P.; Aloi, G.; Gravina, R.; Caliciuri, G.; Fortino, G.; Liotta, A. An edge-based architecture to support efficient applications for healthcare industry 4.0. *IEEE Trans. Ind. Inform.* **2018**, *15*, 481–489. [CrossRef]
8. Wah, T.Y.; Gopal Raj, R.; Iqbal, U. Automated diagnosis of coronary artery disease: A review and workflow. *Cardiol. Res. Pract.* **2018**, *2018*, 2016282.
9. World Health Organization (WHO). Cardiovascular diseases (CVD). 2019. Available online: http://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed on 8 April 2019).
10. Centers for Disease Control and Prevention, National Center for Health Statistics. Multiple Cause of Death 1999–2015 on CDC WONDER Online Database, Released December 2016. Data Are from the Multiple Cause of Death Files, 1999–2015, as Compiled from Data Provided by the 57 Vital Statistics Jurisdictions through the Vital Statistics Cooperative Program. Available online: http://wonder.cdc.gov/mcd-icd10.html (accessed on 8 April 2019).
11. Benjamin, E.J.; Virani, S.S.; Callaway, C.W.; Chang, A.R.; Cheng, S.; Chiuve, S.E.; Cushman, M.; Delling, F.N.; Deo, R.; de Ferranti, S.D. Heart disease and stroke statistics-2018 update: A report from the American Heart Association. *Circulation* **2018**, *137*, e67–e492. [CrossRef]
12. Alizadehsani, R.; Habibi, J.; Hosseini, M.J.; Mashayekhi, H.; Boghrati, R.; Ghandeharioun, A.; Bahadorian, B.; Sani, Z.A. A data mining approach for diagnosis of coronary artery disease. *Comput. Methods Programs Biomed.* **2013**, *111*, 52–61. [CrossRef]
13. Cai, J.; Luo, J.; Wang, S.; Yang, S. Feature selection in machine learning: A new perspective. *Neurocomputing* **2018**, *300*, 70–79. [CrossRef]
14. Mojrian, S.; Pinter, G.; Hassannataj Joloudari, J.; Felde, I.; Nabipour, N.; Nadai, L.; Mosavi, A. Hybrid Machine Learning Model of Extreme Learning Machine Radial Basis Function for Breast Cancer Detection and Diagnosis: A Multilayer Fuzzy Expert System. *Preprints* **2019**, 2019100349. [CrossRef]
15. Liu, S.; Motani, M. Feature Selection Based on Unique Relevant Information for Health Data. *arXiv* **2018**, arXiv:1812.00415.
16. Kass, G.V. An exploratory technique for investigating large quantities of categorical data. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1980**, *29*, 119–127. [CrossRef]
17. Abdar, M.; Zomorodi-Moghadam, M.; Das, R.; Ting, I.H. Performance analysis of classification algorithms on early detection of liver disease. *Expert Syst. Appl.* **2017**, *67*, 239–251. [CrossRef]
18. Chaid-analysis. Available online: http://www.statsoft.com/textbook/chaid-analysis/ (accessed on 8 April 2019).
19. Available online: http://www.public.iastate.edu/~{}kkoehler/stat557/tree14p.pdf (accessed on 8 April 2019).
20. Althuwaynee, O.F.; Pradhan, B.; Lee, S. A novel integrated model for assessing landslide susceptibility mapping using CHAID and AHP pair-wise comparison. *Int. J. Remote Sens.* **2016**, *37*, 1190–1209. [CrossRef]
21. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
22. Pramanik, S.; Chowdhury, U.N.; Pramanik, B.K.; Huda, N. A comparative study of bagging, boosting and C4. 5: The recent improvements in decision tree learning algorithm. *Asian J. Inf. Technol.* **2010**, *9*, 300–306.
23. Quinlan, J.R. *Bagging, Boosting, and C4. 5*; in AAAI/IAAI; Wiley: San Fransisco, CA, USA, 1996.

24. Dietterich, T.G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Mach. Learn.* **2000**, *40*, 139–157. [CrossRef]
25. Schapire, R.E. *A Brief Introduction to Boosting*; Springer: Amsterdam, The Netherland, 1999. (In Ijcai)
26. Pang, S.-L.; Gong, J.-Z. C5. 0 classification algorithm and application on individual credit evaluation of banks. *Syst. Eng.-Theory Pract.* **2009**, *29*, 94–104. [CrossRef]
27. Pandya, R.; Pandya, J. C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning. *Int. J. Comput. Appl.* **2015**, *117*, 18–21. [CrossRef]
28. Siknun, G.P.; Sitanggang, I.S. Web-based classification application for forest fire data using the shiny framework and the C5. 0 algorithm. *Procedia Environ. Sci.* **2016**, *33*, 332–339. [CrossRef]
29. Tseng, C.J.; Lu, C.J.; Chang, C.C.; Chen, G.D. Application of machine learning to predict the recurrence-proneness for cervical cancer. *Neural Comput. Appl.* **2014**, *24*, 1311–1316. [CrossRef]
30. Vladimir, V.N.; Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: Heidelberg, Germany, 1995.
31. Vapnik, V. *Statistical Learning Theory*; Wiley: New York, NY, USA, 1998.
32. Navia-Vázquez, A.; Gutierrez-Gonzalez, D.; Parrado-Hernández, E.; Navarro-Abellan, J.J. Distributed support vector machines. *IEEE Trans. Neural Netw.* **2006**, *17*, 1091. [CrossRef] [PubMed]
33. Ali, J.; Khan, R.; Ahmad, N.; Maqsood, I. Random forests and decision trees. *Int. J. Comput. Sci. Issues (IJCSI)* **2012**, *9*, 272–278.
34. Abdoh, S.F.; Rizka, M.A.; Maghraby, F.A. Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. *IEEE Access* **2018**, *6*, 59475–59485. [CrossRef]
35. Set, Z.-A.S.D. UCI Machine Learning Repository, Center for Machine Learning and Intelligent Systems. 2017. Available online: https://archive.ics.uci.edu/ml/machine-learning-databases/00412/ (accessed on 8 April 2019).
36. Alizadehsani, R.; Hosseini, M.J.; Khosravi, A.; Khozeimeh, F.; Roshanzamir, M.; Sarrafzadegan, N.; Nahavandi, S. Non-invasive detection of coronary artery disease in high-risk patients based on the stenosis prediction of separate coronary arteries. *Comput. Methods Programs Biomed.* **2018**, *162*, 119–127. [CrossRef]
37. Alizadehsani, R.; Habibi, J.; Hosseini, M.J.; Boghrati, R.; Ghandeharioun, A.; Bahadorian, B.; Sani, Z.A. Diagnosis of coronary artery disease using data mining techniques based on symptoms and ecg features. *Eur. J. Sci. Res.* **2012**, *82*, 542–553.
38. Alizadehsani, R.; Zangooei, M.H.; Hosseini, M.J.; Habibi, J.; Khosravi, A.; Roshanzamir, M.; Khozeimeh, F.; Sarrafzadegan, N.; Nahavandi, S. Coronary artery disease detection using computational intelligence methods. *Knowl.-Based Syst.* **2016**, *109*, 187–197. [CrossRef]
39. Arabasadi, Z.; Alizadehsani, R.; Roshanzamir, M.; Moosaei, H.; Yarifard, A.A. Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. *Comput. Methods Programs Biomed.* **2017**, *141*, 19–26. [CrossRef]
40. Abdar, M.; Książek, W.; Acharya, U.R.; Tan, R.S.; Makarenkov, V.; Pławiak, P. A new machine learning technique for an accurate diagnosis of coronary artery disease. *Comput. Methods Programs Biomed.* **2019**, *179*, 104992. [CrossRef]
41. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Int. Jt. Conf. Artif. Intell.* **1995**, *14*, 1137–1145.
42. Haq, A.U.; Li, J.P.; Memon, M.H.; Nazir, S.; Sun, R. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mob. Inf. Syst.* **2018**, *2018*, 3860146. [CrossRef]
43. Weng, C.-H.; Huang, T.C.-K.; Han, R.-P. Disease prediction with different types of neural network classifiers. *Telemat. Inform.* **2016**, *33*, 277–292. [CrossRef]
44. Dousthagh, M.; Nazari, M.; Mosavi, A.; Shamshirband, S.; Chronopoulos, A.T. Feature weighting using a clustering approach. *Int. J. Model. Optim.* **2019**, *9*, 67–71.
45. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2005**, *27*, 861–874. [CrossRef]
46. Tan, P.N. *Introduction to Data Mining*; Pearson Education: New Delhi, India, 2018.