Computer Standards & Interfaces 84 (2023) 103706

Contents lists available at ScienceDirect

Computer Standards & Interfaces

journal homepage: www.elsevier.com/locate/csi

Ensemble feature selection and classification methods for machine learning-based coronary artery disease diagnosis

Burak Kolukisa, Burcu Bakir-Gungor *

Department of Computer Engineering, Faculty of Engineering, Abdullah Gul University, Kayseri, Turkey

ARTICLE INFO

Keywords:
Machine learning
Classification
Ensemble feature selection
Domain knowledge-based feature selection
Coronary artery disease diagnosis

ABSTRACT

Coronary artery disease (CAD) is a condition in which the heart is not fed sufficiently as a result of the accumulation of fatty matter. As reported by the World Health Organization, around 32% of the total deaths in the world are caused by CAD, and it is estimated that approximately 23.6 million people will die from this disease in 2030. CAD develops over time, and the diagnosis of this disease is difficult until a blockage or a heart attack occurs. In order to bypass the side effects and high costs of the current methods, researchers have proposed to diagnose CADs with computer-aided systems, which analyze some physical and biochemical values at a lower cost. In this study, for the CAD diagnosis, (i) seven different computational feature selection (FS) methods, one domain knowledge-based FS method, and different classification algorithms have been evaluated; (ii) an exhaustive ensemble FS method and a probabilistic ensemble FS method have been proposed. The proposed approach is tested on three publicly available CAD data sets using six different classification algorithms and four different variants of voting algorithms. The performance metrics have been comparatively evaluated with numerous combinations of classifiers and FS methods. The multi-layer perceptron classifier obtained satisfactory results on three data sets. Performance evaluations show that the proposed approach resulted in 91.78%, 85.55%, and 85.47% accuracy for the Z-Alizadeh Sani, Statlog, and Cleveland data sets, respectively.

1. Introduction

According to a report published by the World Health Organization (WHO), coronary artery disease (CAD) causes 32% of the total deaths in the world [1]. The WHO estimates that CAD will cause the death of approximately 23.6 million people in 2030. CAD is a condition in which large arterial vessels that supply the heart become atherosclerosis. Atherosclerosis is the accumulation of a fatty substance called atheroma in the walls of the arteries, which causes narrowing and congestion in the vessels. Since CAD develops over time, it is difficult to diagnose at the early stages. It is usually diagnosed when the veins are clogged or when a person has a heart attack. CAD and heart attacks (heart disease) are the most common types of cardiovascular diseases (CVD). CVD is the most general term that embraces all kinds of diseases that affect the heart or blood vessels. In recent machine learning (ML) studies, the terms CAD, CVD, and heart disease are used interchangeably. Diagnosis of this disease requires special devices and medical specialists. The most commonly used methods for the diagnosis of CAD are blood tests and electrocardiograms. Each method has its own advantages and disadvantages, but in general, these methods are costly. Generally, the testing phase of CADs may not be economically feasible, and unfortunately, more than three-quarters of CADs occur in low- and middle-income countries [2]. In these countries that are adversely affected by the economic situation, people die from CAD at a young age because the diagnosis is not made on time [3]. In the last decades, researchers have attempted to develop computer-aided systems in healthcare [4,5]. Along this line, effective methodologies that check some physical and biochemical values at a lower cost can be developed for CAD diagnosis. Especially with the development of information technologies, it is possible to diagnose CADs by examining specific parameters rather than the outputs of these expensive devices.

Researchers have worked on the applications of different data mining (DM) and ML algorithms in various fields [6–10]. As recently reviewed in [11], up to 2019, there have been 149 studies on ML-based CAD diagnosis. Many different CAD datasets have been analyzed. Although existing studies provide valuable insights and foundations for CAD diagnosis, the performance results vary significantly depending on sample size, the number and type of features, and the ethnicity of the patients in the data set. Due to these factors, there is no internationally recognized standard approach for CAD diagnosis. In this respect, this study aims to create a generic heart disease diagnosis model that can

* Corresponding author.
E-mail addresses: burak.kolukisa@agu.edu.tr (B. Kolukisa), burcu.gungor@agu.edu.tr (B. Bakir-Gungor).

https://doi.org/10.1016/j.csi.2022.103706
Received 8 February 2022; Received in revised form 19 September 2022; Accepted 1 November 2022
Available online 3 November 2022

work for different CAD data sets. The main contributions of this study can be summarized as follows:

- An exhaustive ensemble feature selection (FS) and a probabilistic ensemble FS method have been proposed.
- Seven computational FS methods and one domain knowledge based (DK) FS method are used in the proposed approaches. The DK-based FS method ranks the features based on the expertise of the cardiologists.
- The proposed approaches use six different single classifiers and four different ensemble voting classifiers. A grid-search parameter optimization method is used to automatically find the best values for the ML methods' hyper-parameters.
- The proposed methods are evaluated independently on different publicly available data sets (The Cleveland, Statlog, and Z-Alizadeh Sani data sets).
- We tested the proposed approaches on complex vs. noncomplex, linear vs. nonlinear, balanced vs. imbalanced, and sparse vs. nonsparse data sets; and we have shown that the proposed methods are effective for different data sets with different characteristics.
- Performance results show that the accuracy of the proposed model is 91.78%, 85.55%, and 85.47% on the Z-Alizadeh Sani, Statlog, and Cleveland data sets, respectively.

The rest of this article is organized as follows: Section 2 provides an overview of the current ML-based CAD diagnosis literature. In Section 3, available data sets, FS, and classification methods are described. In Section 4, the proposed ensemble approaches are explained in detail. In Section 5, the performance evaluations of different models are presented. The last section concludes the paper, and future research directions are discussed.

2. Related works

In recent years, different ML and DM methods have been applied to CAD data sets for different purposes. The reason behind the popularity of ML techniques for CAD diagnosis stems from their low complexity, easy implementation, and ability to generate high performance metrics. In a recent review paper, Z-Alizadeh stated that as of 2019, 149 research articles have been published on ML-based CAD diagnosis, and more than 90% of the CAD detection studies used supervised algorithms [11]. As noted in [11], the top techniques used on almost all CAD data sets are decision tree (DT), support vector machine (SVM), and logistic regression (LR). The same review paper [11] reported that there are 67 different CAD data sets that could be used in the ML and DM domains. These existing CAD data sets were collected from 3 continents and differed from each other in terms of ethnicity, feature size, sample size, and feature characteristics. In our earlier studies [12,13], we illustrated commonly used CAD data sets based on their sample size. These publicly available data sets are Heart Disease, Z-Alizasdeh Sani, and South Africa. The Heart Disease and Z-Alizadeh Sani data sets are available on the UCI machine learning repository. The Heart Disease data set collection contains four data sets: Cleveland, VA Long Beach, Hungary, and Switzerland, which are collected in different countries and regions. In the literature, while some studies use one of these data sets, others use a combination of these UCI Heart Disease data sets.

Anbarasi et al. [14] used four of the Heart Disease data sets, which contained 13 features and 909 samples. Using the genetic algorithm, the best six features affecting CAD are obtained. Three different algorithms are used with k-fold cross-validation (CV), where the value of k is not mentioned in the original study. The study was conducted using Weka 3.6.0. El-Bialy et al. [15] applied pruned DT and fast DT techniques to the same data set. In their study, characteristic features are extracted from the Heart Disease data sets. Using a fast DT with a 10-fold CV in Weka, they obtained 78.06% accuracy on the Heart Disease data set. Reddy et al. [16] proposed to combine four Heart Disease data sets and the Statlog data set. They filled the missing values by taking the average of the available data and thereby obtained a data set including 1190 samples. The authors used three different percentage splits for classification methods. They achieved consistent performance results after reducing the number of features to eight. The performance results decreased when the feature size was reduced to six. The authors report that the best result of 94.96% accuracy is obtained with a random forest (RF) [17] classifier using the eight features with a 10-fold CV (the percentage split for training is 80%, for testing 20%). In their study, the R tool was used to conduct the experiments.

Shouman et al. [18] used the Cleveland data set, which contains 13 features and 303 samples. The Cleveland data set is the most popular data set since other data sets have so many missing values. Only six samples from the Cleveland data set were missing, and they were removed from the experiment. The authors investigated different DT techniques in order to achieve high performance results for diagnosing heart diseases. The best accuracy of 84.1% is obtained with the frequency discretion gain rate DT method using a 10-fold CV. In another study [19], 85% accuracy was obtained when the Statlog data set was used with the cascade neural network algorithm. From the 270 samples, 150 samples are used for training, and the remaining samples are used for testing.

Alizadehsani et al. [20] used the Z-Alizadeh Sani data set, which contains 54 features and 303 samples. The authors have applied a cost-sensitive CAD diagnosis algorithm named MetaCost. A FS method is applied to reduce the number of features to 34, and three new features are generated using the feature extraction method. Five different classification algorithms are used in MetaCost using a 10-fold CV. The proposed solution is reported to generate a high sensitivity of 97.22% and an accuracy of 92.09%, which makes the Sequential Minimal Optimization (SMO) algorithm better than the other different alternative classifiers, and the study was conducted with the tool Rapid Miner. In order to increase the diagnosis rate of CAD, Joloudari et al. [21] selected significant predictive features for ML methods. A max–min scaler is applied to the data set, and ML is applied to the data set with a 10-fold CV. They obtained the best results with random tree models, with an accuracy of 91.47%. Bhatnagar [22] proposes a hybrid model that consists of an optimization process based on particle swarm optimization and a firefly nature-inspired classification based on discriminant analysis. In their study, first the data is split into training and test sets, and then normalization and feature extraction operations are applied separately. They achieved a high accuracy score of 95%.

Babič et al. [23] analyzed three publicly available CAD data sets: the Heart Disease, South African Heart Disease, and Z-Alizadeh Sani data sets. The authors performed both predictive and descriptive analyses. In the predictive analysis, DT, Naive Bayes (NB), SVM, and neural network (NN) classifiers are used to decide whether a person has heart disease. In the descriptive analysis, the association and decision rules are used to extract steps to support decisions during the diagnosis process. The authors achieved 89.93% accuracy with NN on the Heart Disease data set, 73.87% accuracy with DT on the South African heart disease data set, and 86.67% accuracy with NN on the Z-Alizadeh Sani data set.

In our previous work, we analyzed different heart disease data sets using linear discriminant analysis (LDA), and a new hybrid FS method [12,24]. The goal was to reduce the computational cost by reducing the number of features, and to generate a model has a satisfactorily performance for each data set. Using an ensemble FS method with a MLP classifier, we achieved 88.11% and 82.50% accuracy values on the Z-Alizadeh Sani and Cleveland data sets, respectively. The best accuracy value of 92.74% is obtained with fisher linear discriminant analysis (FLDA) via the SVM classifier on the Z-Alizadeh Sani data set [12,24]. In addition, in a previous study, we proposed a novel self-optimized and adaptive ensemble ML algorithm for CAD diagnosis. In [25], the system automatically selects the most effective ML models without any preprocessing or FS method. We achieved 88.38% and

83.43% accuracy values on the Z-Alizadeh Sani and Cleveland data sets, respectively [25].

In all these studies, different FS techniques, parameter optimizations, and several ML techniques are used to obtain a good result on the CAD data sets. Although these models perform well for a specific data set, sample size, feature size, or the ethnicity of patients affect results dramatically. In this study, using the proposed methods, we aim to obtain a generic classification model that generates satisfactory performance metrics for different CAD data sets.

3. Materials & methods

3.1. Data sets

The review paper by Alizadehsani [11] reported that there are 70 different CAD data sets that could be used in the ML and DM domains. These existing CAD data sets were collected from 3 continents and differed from each other in terms of ethnicity, feature size, sample size, and feature characteristics. When analyzing these data sets, it has been observed that the number of features is low in large datasets. Whereas in medium or small data sets, the number of features is higher compared to large data sets. In general, the small number of samples or features affects performance results adversely, and the obtained performance results are difficult to generalize. Therefore, in order to compare our model with existing studies, three widely used CAD diagnosis data sets (i.e., Z-Alizadeh Sani, Statlog, and Cleveland data sets) were selected. As reported in [11], among 70 CAD data sets, the Cleveland data set is the most widely used one with 3770 citations; the Stadlog data set is the third-most popular one with 239 citations; and the Z-Alizadeh Sani data set is the eighth-most popular one with 131 citations. Also, in terms of completeness, the Z-Alizadeh Sani and Statlog data sets are complete, and the Cleveland data set has fewer missing values compared to other existing data sets. In this study, these three publicly available CAD data sets are used.

The Cleveland [26] data set is one of the four data sets included in the Heart Disease data set collection. Among these data sets, we preferred to use the Cleveland data set, since other data sets have too many missing values. There are 76 features in this data set, but we used 13 of them, which do not have missing values. The Cleveland and Statlog data sets contain 303 and 270 samples, respectively. Only the Cleveland data set has six missing values. These samples are ignored in this study instead of applying data correction. Each sample in these data sets is labeled as either (i) healthy if their vessels narrowed less than 50%, or (ii) CAD, otherwise. The Z-Alizadeh Sani [27] data set has 303 samples and 55 features, which are categorized into one of the four following groups: “Demographics”, “Symptom and Examination”, “ECG”, and “Laboratory and Eco”. Each sample in this data set is labeled as either (i) healthy or (ii) unhealthy. The characteristics of the Cleveland and Z-Alizadeh Sani data sets are provided in detail in our earlier studies [12,13].

As reported by Dedeturk et al. [28], analyzing the data set in terms of complexity, linearity, sparsity, and imbalance ratio is important. Along this line, we have comparatively analyzed the Cleveland, Statlog, and Z-Alizadeh Sani data sets, as shown in Fig. 1. Fisher’s discriminant ratio (F1) computes the overlaps between the features in different classes. A high F1 score indicates that there is at least one feature that overlaps between the classes, while a low F1 score indicates a more complex problem where no individual feature separates the classes. To this end, the high F1 value of the Z-Alizadeh Sani data set implies that this data set is more linear than the Cleveland and Statlog problems, whereas the Cleveland data set represents more of a non-linear problem. Complexity metrics give an empirical score that does not tell enough about the size or strength of the measured data set on its own. The score is valuable when it can be compared to other data sets [29]. To measure the complexity (N1) of the data set, we applied the methods proposed by Ho et al. [30], where N1 measures the

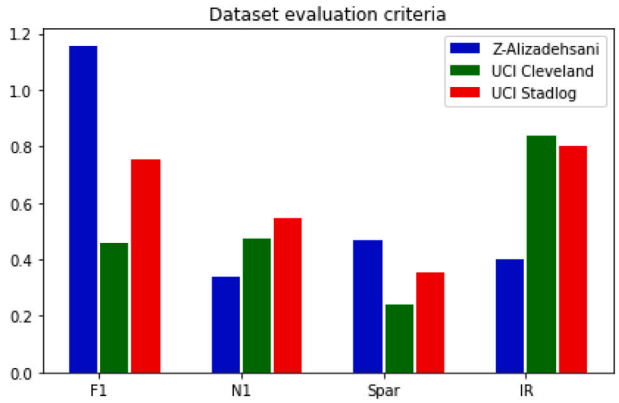

Fig. 1. Evaluation of Z-Alizadeh Sani, Cleveland, and Statlog data sets based on Fisher’s discriminant ratio (F1), complexity (N1), imbalance ratio (IR), and sparsity (Spar) measures.

separability of classes’ distributions. Higher N1 values demonstrate that more complex boundaries are needed to discriminate between classes. Based on the computed N1 values, we can state that the Statlog data set is more complex than the Cleveland and Z-Alizadeh Sani data sets. The imbalance ratio is calculated by dividing the normal samples by the CAD samples. The calculated imbalance ratios (Fig. 1) show that the Cleveland and Statlog data sets are balanced, whereas the Z-Alizadeh Sani data set is unbalanced. Lastly, the sparsity value of a data set is computed by dividing the number of zero elements by the total number of elements. While the high sparsity value of the Z-Alizadeh Sani data set implies that this data set is sparser than the Cleveland and Statlog data sets, the lower sparsity value of the Cleveland data set indicates more of a dense problem. Taken together, one can observe from Fig. 1 that the Cleveland, Statlog, and Z-Alizadeh Sani data sets have different characteristics.

3.2. Feature selection methods

In ML processes, FS methods are effective auxiliary methods that are used to generate a valid and reliable model. FS methods determine the predictive power of each feature according to the target variable, and they propose new subsets of features. The objectives of the FS methods are: (i) to reduce the complexity of the generated model; (ii) to eliminate noise; (iii) to prevent overfitting; (iv) to enable ML algorithms to work faster; and (v) to improve the performance results [31].

FS methods can be grouped into three categories (filter-based, wrapper-based, and embedding methods) based on how they combine the selection algorithm and the model-building process. Filter-based methods select variables regardless of the model, which is independent of the classifier. The filter-based methods operate based on correlations, and they are efficient in determining unrelated features. These methods are effective in terms of reducing computational time and preventing overfitting. The wrapper-based method allows the detection of interactions between variables, but it increases the computational time and the risk of overfitting. The embedding method combines the benefits of both approaches by performing FS and classification concurrently [32,33].

In general, each FS method has its own advantages and disadvantages. Among different FS methods, it becomes difficult to decide which one is more suitable and which one works better for a specific data set. In order to analyze CAD data sets, FS methods from each one of the three categories can be used. In this study, different FS methods, i.e., chi-square (CS), gain ratio (GR), information gain (IG), relief F (RFF), support vector machine feature selection (SVM FS), bee search (BS), conditional mutual information maximization (CMIM), and a DK-based FS method, are used.

The CS is a statistical hypothesis test, which is a univariate filter that organizes each feature independently by class, and it is used for the FS technique [34]. IG is an entropy-based FS technique. IG is a symmetrical measure that ranks characteristics based on the entropy criteria. IG has a bias toward the features with high value, even though these features are less informative than the other features. GR method is the modified version of IG such that it reduces its bias toward high-value features that have no predictive power. The RF algorithm estimates the quality of features according to the relationships between them instead of acting independently. The RFF method is an extension of the relief algorithm that is more resilient and capable of handling incomplete and noisy data. RFF FS chooses the best features from the data set by giving each one a different weight based on how it compares to the other features in the set. All four of these FS methods are filter-based [35]. The SVM FS method was introduced by Guyon [36], who tried to refine the optimal feature set using SVM FSs in a wrapper approach. It assigns the scores of the features by using the square of the weights obtained by the SVM classifier and removes the irrelevant features by training an SVM classifier iteratively. The CMIM [37] FS approach ranks the features based on their conditional entropy and mutual information with the target class. Then, it permits the addition of a new feature to the collection of chosen features if and only if the new feature contains additional information. The artificial bee colony (ABC) is a swarm-based meta-heuristic algorithm that emulates the behavior of honeybees in search of food. Dervis Karaboga [38] was the first to model these behaviors in order to solve optimization problems in 2005. There are different variations of the ABC algorithm, and it can be simulated as a neighbor search algorithm in its simplest form. There is a package in the Weka software called metaphor search techniques that includes BS [39], which is used to choose the best feature.

In our previous study, we proposed a DK-based FS method for CAD disease diagnosis [24]. This method ranks the features of the CAD data set according to the expertise of the cardiologists. The features are divided into the following two groups: (i) Framingham Heart Study (FHS), and (ii) Clinically Important Findings (CIF). In practice, patients receive a diagnosis of CAD via referring to the FHS, which is supported by the National Institute of Heart Lung and Blood (NHLBI). In this study, 5209 men and women were observed to determine the main factors that cause CAD. These participants went through physical examinations and lifestyle interviews to assess the relationship between CADs and other factors [40]. More than 1200 articles, which refer to FHS while diagnosing CAD, have already been published in well-known medical journals. Therefore, the FHS study is considered as a fundamental and leading resource for CAD diagnosis. When our cardiologist collaborators examined the features of the Cleveland and Z-Alizadeh Sani data sets, they determined the essential features according to their medical expertise. We refer to the features that are selected by cardiologists as "CIF" [24]. When scoring the features in the CAD data set according to the DK, the features that are contained in the CIF and FHS are scored high, and the features that are not included in the CIF and FHS are scored low. Therefore, while diagnosing CAD using ML, it could be evaluated whether the factors used in the computational model are compatible with cardiovascular medical literature. The DK-based scores that indicate the importance of the features for the Z-Alizadeh Sani and the Cleveland data sets are presented in [24]. Further details of the DK-based FS method for CAD diagnosis can be found in our previous publication [24].

The advantages of filter-based FS approaches are as follows: (i) they are computationally efficient, (ii) they have fast processing capability, and (iii) they are less prone to overfitting. Unfortunately, in some cases, they are incapable of finding features and could give low precision. The primary benefit of the wrapper-based FS approaches is their high precision rate. Some disadvantages of these methods are (i) they are prone to overfitting, (ii) they are computationally costly, and (iii) they are processed slowly due to recursion. The embedded method, which is a hybrid of filter- and wrapper-based methods, inherits both the advantages and the disadvantages of the two categories. Considering the pros and cons of the FS approaches, in this study, combinations of eight FS methods are used to obtain robust features for classification.

Table 1
An example ensemble score calculation for five different features using the feature importance scores obtained via three different FS methods (FS1, FS2, and FS3).

Features	FS 1	FS 2	FS 3	ES of features
Feature 1	5	4	2	(5 + 4 + 2)/3 = 3.66
Feature 2	1	5	5	(1 + 5 + 5)/3 = 3.66
Feature 3	4	1	4	(4 + 1 + 4)/3 = 3.00
Feature 4	3	3	3	(3 + 3 + 3)/3 = 3.00
Feature 5	2	2	1	(2 + 2 + 1)/3 = 1.66

ES: Ensemble Score.

4. Proposed method

One of the main goals of FS is to determine important scores of the features in the data set and select features accordingly. While high scores are assigned to the essential features, low scores are given to the uninformative and redundant features. The FS method is beneficial for improving performance results and reducing computational time [41]. However, the effects of each FS method are different from each other, and this causes differences in classification performance [31]. A feature that is scored high by one FS method can be scored low by another FS method. It is hard to define a consistent feature score at the final step with a certain level of confidence [42]. For example, in our preliminary analysis with different FS methods on the Z-Alizadeh Sani data set, the "Tinversion" feature is found to be important by seven FS methods, but it is identified as irrelevant by the DK-based FS. While the "Pulse Rate" feature is detected as important by the SVM FS technique, it is denoted as irrelevant by other FS methods. In order to address this inconsistency, in this study we proposed two ensemble FS methods. The first method (exhaustive ensemble FS) searches through all possible combinations of FS methods and aims to compensate for the differences in feature scores among different FS methods. In the exhaustive ensemble FS approach, low-scoring features are ignored in the classification process, and the chance of achieving better performance results may be overlooked. To address this issue, a second method (probabilistic ensemble FS approach) has been proposed. In the probabilistic ensemble FS method, a probabilistic score for each feature is calculated using different FS methods. These probabilistic scores are further used to determine the selection rate of the features. Following this trend, different numbers of features are selected and tested in the classification process. The details of these methods are presented in the next two subsections.

4.1. Exhaustive ensemble feature selection approach

It has been reported in different domains that by taking the average of the feature importance scores that are obtained by different FS methods, high classification performance can be achieved [43]. In our earlier study, we applied this idea to the CAD diagnosis problem [12]. In order to take advantage of different FS methods, in [25], we have utilized the following hybrid FS methodology for CAD diagnosis. We obtained seven different rankings of features by applying seven different FS techniques individually. To calculate the final ranking of a feature, we take the average of seven rankings obtained via different FS methodologies. However, there are many different combinations of different FS methods, and one needs to examine all those possibilities.

Table 1 shows an example ensemble score calculation for five different features using the feature importance scores obtained via three different FS methods (FS1, FS2, and FS3). Here, the importance score refers to the reverse of the ranking of a feature among all other features. For example, if a feature is identified as the most important feature, it gets a score of k, the total number of features. The second most important feature gets a score of $k - 1$. On the other hand, if a feature is identified as the least important feature, it gets a score of 1. Some FS methods do not assign a score to each feature. In such cases, an average score is assigned to those features that are not

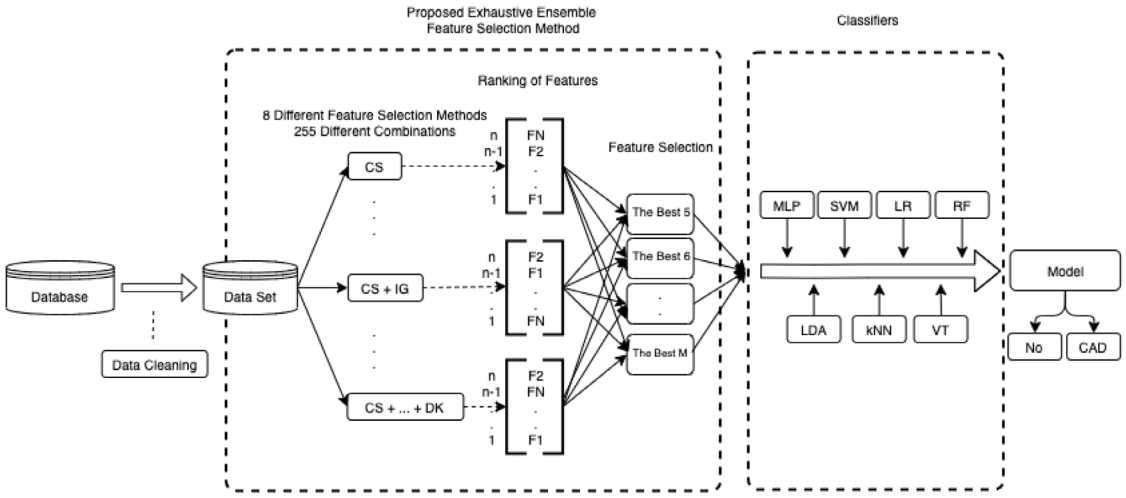

Fig. 2. Schematic representation of the proposed exhaustive ensemble FS model.

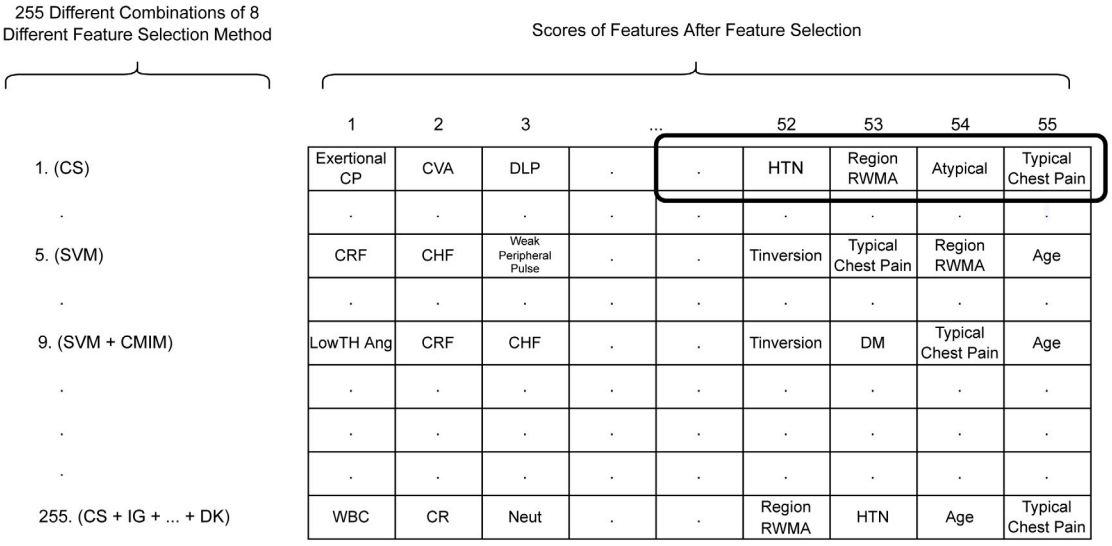

Fig. 3. Selection of the features using the proposed exhaustive ensemble FS model, visualized for the Z-Alizadeh Sani data set. For each of the 255 different combinations of 8 different FS methods, the best five features with the highest scores are selected.

scored by the FS methods. For each feature, the average of the scores obtained using different FS algorithms is calculated and named as the ensemble score of the feature. As shown in Table 1, if a single FS method had been applied, then features 2 and 3 could be evaluated as insignificant, whereas this feature gets a high score when an ensemble FS method is applied. On the other hand, for the 5th feature, in all three FS methods, the scores of this feature are low, and the score of the ensemble FS of this feature is also low, indicating that there is no need to hesitate when removing this feature from the data set. Along this line, an ensemble FS approach that considers all possible combinations (2^n-1) of n different FS methods has been developed. The schematic representation of our exhaustive ensemble FS is presented in Figs. 2 and 3. As illustrated in Fig. 2, the exhaustive ensemble FS method performs an exhaustive search and analyzes all possible sub-sets of different FS methods. As shown in Fig. 2, in our experiments, we utilized 8 different FS methods ($n = 8$) and hence, 255 different FS lists are generated for each of the possible combinations (e.g., CS + IG). For each sub-set of the FS method (e.g., CS + IG), an ensemble score of a feature is computed as the average importance score of that feature among the tested FS methods (as depicted in Fig. 2). Then, for each sub-set of the FS method (e.g., CS + IG), the features are ranked from highest score to lowest score based on the calculated ensemble scores. Hence, a total of 255 ranked feature lists have been generated. For each ranked feature list, in which features are ranked from the highest score to the lowest score based on the calculated ensemble scores, different numbers of top features are selected (as presented in Fig. 3), and tested with 10 different classifiers (as shown in Fig. 2). The pseudocode of the proposed exhaustive ensemble FS approach is also presented in Supplementary Table 1.

4.2. Probabilistic ensemble feature selection approach

The goal of the probabilistic ensemble FS approach is to reveal the features that computational FS methods deem insignificant, but the biological DK states that these features could improve the performance results. In other words, in the probabilistic ensemble FS approach, we aim to overcome the problem of getting stuck in the local maximum and to incorporate the features into the classification processes if the medical specialists consider these to be important features but the computational FS methods do not consider them significant features.

Table 2 shows an example of probabilistic score calculation for five different features using the ensemble scores of the features that are computed in Table 1. For each feature, the ensemble score refers to the average of the importance scores assigned to that feature by

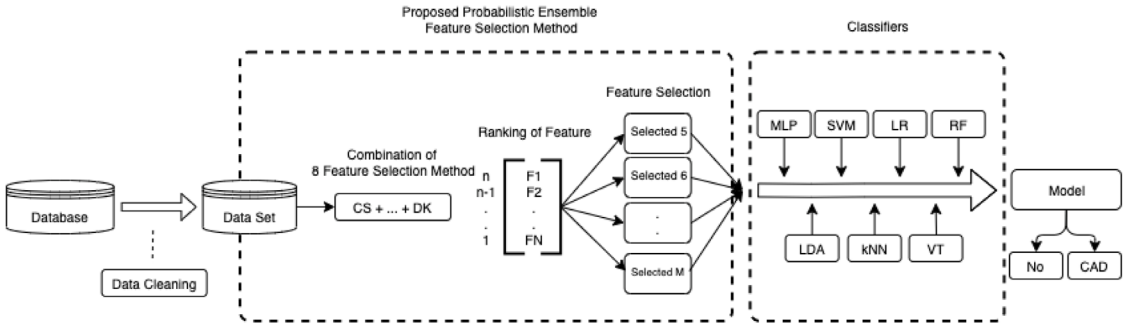

Fig. 4. Schematic representation of the proposed probabilistic ensemble FS model.

Table 2
An example of a probabilistic score calculation for five different features.

Features	ES of features	PS of features
Feature 1	3.66	3.66/Total = 0.244
Feature 2	3.66	3.66/Total = 0.244
Feature 3	3.00	3.00/Total = 0.200
Feature 4	3.00	3.00/Total = 0.200
Feature 5	1.66	1.66/Total = 0.110

ES: Ensemble Score, PS: Probabilistic Score.

Table 3
The hyper-parameter values of the classification methods.

Methods	Parameters of classifiers
SVM	C = 1, 10, 100, 1000 Gamma = 1e−3, 1e−4 Kernel = linear, rbf
MLP	Alpha = 10-e2, 10-e3, 10-e4, 10-e5, 10-e6 Hidden layer size = (100, 50) max iter = 500, 1000 activation = relu, tanh
RF	estimators = 100
KNN	Neighbors = 1, 3, 5, 7, 9, 11
LR	solver = lbfgs max_iter = 2000
LDA	Default

different FS methods (shown as FS1, FS2, and FS3 on Table 1). As shown in Table 2, the probabilistic scores of the features are calculated by dividing the ensemble feature score by the total ensemble score. Hence, the probabilistic scores of the features are considered as a selection probability for each feature. The schematic representation of our probabilistic ensemble FS method is shown in Fig. 4. In order to select a feature, we randomly generate a number, and based on the probabilistic scores of the features, the one that corresponds to that random number is selected. Following this trend, different numbers of features are selected and tested in the classification process (as shown in Fig. 4). The pseudocode of the proposed probabilistic ensemble FS approach is also presented in Supplementary Table 2.

4.3. Hyper-parameter optimization

Parameter optimization ensures the use of the most appropriate parameter for the classification model. Parameters must be determined in the classification methods before training a model, and the model is trained according to these parameters. For each hyper-parameter configuration, models are trained on a training set and tested on a test set. In this study, the hyper-parameter values of some classifiers are obtained by a grid search optimization algorithm, which is a common method for finding these hyper-parameters. The implementation is done in Python using the Scikit-Learn library. The hyper-parameters that are obtained by a grid search optimization algorithm and used in the classification algorithms are shown in Table 3.

4.4. Voting classifier

Voting is one of the ensemble ML methods that combines the predicted class labels from multiple models. One of the goals of the ensemble approach is to minimize bias and variation [9]. The voting ensemble method predicts the class as the one that has the largest sum of the votes from different models, or by averaging each model's weight. There are mainly two types of voting methods: (i) hard voting (H) and (ii) soft voting (S). While hard voting works based on the majority rule, soft voting assigns final class labels by combining probabilistic scores calculated by different classifiers. Here we ensembled the three and five classifiers using soft and hard voting ensemble methods separately. So, we have four different ensemble classifiers: S3, S5, H3, and H5, where S and H stand for soft voting and hard voting, respectively, and the numbers after S and H stand for the number of classifiers used in the ensemble classifier. MLP, SVM, and LR classifiers are used in S3 and H3. MLP, SVM, LR, RF, and LDA classifiers are used in S5 and H5 ensemble classifiers.

5. Experiments

In our experiments, we utilized three publicly available CAD data sets to diagnose heart disease to test our two proposed ensemble FS approaches that use seven different computational FS methods, a DK-based FS method, six single classifiers, and one ensemble classifier with four different variations. In our study, the samples with missing features are removed from the data sets instead of filling them with synthetic data.

Firstly, the importance scores of the features are calculated using eight different FS methods on the Cleveland, Statlog, and Z-Alizadeh Sani data sets, as shown in Supplementary Tables 3, 4, and 5, respectively. The last column in these tables indicates the average importance scores (ensemble scores) of each feature. These features are presented in decreasing order of the ensemble scores. The scores presented in these tables are used for the calculation of ensemble and probabilistic feature scores. Secondly, we perform the exhaustive ensemble FS approach (shown in Fig. 2, Supplementary Table 1). To this end, the ensemble scores of the features are computed 255 times (for each one of the combinations of eight FS methods). Hence, as illustrated in Figs. 2 and 3, 255 different lists showing the scores of the features are generated. As shown in Figs. 2 and 3, different numbers of features (t) are tested for each of the 255 feature-ranking lists. Each classifier is run using the top t features from each list. The t value has been increased one by one, from 5 to M (the maximum number of selected features), and tested separately in each classifier. In our previous experiments [24], we observed that on the Z Alizadeh Sani data set, when the number of features exceeded 25, the computational time increased significantly and the performance results decreased. Therefore, for the Z-Alizadeh Sani data set, M is set to 25, and 21 $(25 - 5 + 1)$ different combinations of top t features are generated for each combination of FS methods. For the Cleveland and Statlog data sets, M is set to 12,

Table 4
The best performance results of each classifier on three different CAD data sets using all features.

Data set	Classifier	SN (%)	PRE (%)	FM	AUC	ACC (%)
Z-Alizadeh Sani	SV5	85.00	88.40	0.882	0.933	88.40
	SVM	85.98	88.77	0.880	0.931	88.08
	HV5	85.43	88.32	0.877	–	87.77
	SV3	84.57	87.41	0.869	0.937	87.09
	RF	80.98	87.89	0.863	0.920	86.77
	HV3	84.43	87.33	0.866	–	86.77
	LR	82.11	86.33	0.855	0.930	86.10
	LDA	83.77	86.68	0.858	0.905	85.76
	MLP	79.77	83.83	0.833	0.918	83.48
	KNN	54.34	63.75	0.646	0.509	68.67
Cleveland	LDA	82.72	83.50	0.829	0.900	83.09
	HV5	81.78	82.60	0.819	–	82.09
	LR	81.44	82.56	0.815	0.901	81.74
	SVM	81.20	82.04	0.813	0.905	81.42
	SV5	81.02	81.71	0.809	0.897	81.09
	HV3	80.50	81.47	0.806	–	80.73
	RF	82.01	81.54	0.805	0.893	80.71
	SV3	80.46	80.97	0.802	0.908	80.37
	MLP	79.91	81.00	0.799	0.877	80.05
	KNN	67.63	68.08	0.677	0.699	68.00
Statlog	LR	83.58	85.06	0.838	0.901	84.07
	LDA	83.58	85.04	0.838	0.906	84.07
	SVM	83.08	84.92	0.834	0.903	83.70
	SV5	83.08	84.64	0.834	0.893	83.70
	HV3	83.16	84.57	0.834	–	83.70
	SV3	82.91	84.19	0.831	0.905	83.33
	HV5	82.75	84.21	0.831	–	83.33
	RF	80.91	82.92	0.816	0.898	82.96
	MLP	79.41	80.74	0.798	0.887	80.00
	KNN	65.25	66.32	0.658	0.699	66.29

SN: Sensitivity, PRE: Precision, FM: F-Measure, AUC: Area Under Curve, ACC: Accuracy.

since the total number of features is 13, and 8 $(12 - 5 + 1)$ different combinations of top t features are generated for each combination of FS methods.

Thirdly, the probabilistic ensemble FS approach (shown in Fig. 4, Supplementary Table 2) is conducted only for the Z-Alizadeh Sani data set. Since the Cleveland and Statlog data sets only have 13 features, the probabilistic ensemble FS approach is not performed on these data sets. On the Z-Alizadeh Sani data set, we selected 5 to M features and ran each set of selected features separately in each classification algorithm 50 times (as shown in Fig. 4).

In our experiments, 10 different classifiers are used, including six single classifiers and one ensemble classifier with different variations. Single classifiers include k-nearest neighbor (KNN), LDA, LR, MLP, SVM, and RF. The ensemble classifiers include four different variations (S3, S5, H3, H5, where S and H indicate soft and hard voting, respectively, and the numbers following S and H indicate the number of classifiers used in the ensemble classifier). For KNN, MLP, and SVM classifiers, we perform parameter optimization. For the classification task, we used stratified 10-fold cross-validation. All FS methods have been carried out using Weka [44], and all classification methods have been realized using Python [45] with the Scikit-Learn library [46].

When the proposed exhaustive ensemble FS method is applied to the Z-Alizadeh Sani data set, 53.550 models (255 * 21 * 10) are generated. For the Cleveland and Statlog data sets, in total, 40.800 models (255 * 8 * 10 * 2) are generated using the exhaustive ensemble FS method. The probabilistic ensemble FS approach generated 10.500 models (50 * 21 * 10) for the Z-Alizadeh Sani data set. In total, in our experiments, 104.850 models were generated.

6. Performance results

In this study, three publicly available CAD data sets are experimented with using the proposed exhaustive ensemble FS approach, the probabilistic ensemble FS approach and ten different classifiers. In Table 4, for each classifier, the performance results are shown when all features are used. In Table 5, for each classifier, the best performance results that are obtained for three different CAD data sets using the exhaustive ensemble FS approach are shown. Table 5 also presents the names of the combinations of FS methods (or the number of combinations that give the same accuracy) and the number of features used in the model. Among different classifiers for the Cleveland and Z-Alizadeh Sani data sets, the best performance results are obtained with the MLP classifier. For the Statlog data set, the best result was achieved with the KNN classifier. The best accuracy scores are obtained as 91.78%, 85.47%, and 86.66% for the Z-Alizadeh Sani, Cleveland, and Statlog data sets, respectively. These performance metrics are obtained with 9, 5, and 21 features on the Cleveland, Statlog, and ZAlizadeh Sani data sets, and the lists of these selected features can be found in Supplementary Tables 6, 7, and 8, respectively. In addition to the single classifiers, the exhaustive ensemble FS approach is also tested using ensemble classifiers (both hard and soft voting methods). As shown in Table 5, the best performing ensemble classifier resulted in 91.45% accuracy on the Z-Alizadeh Sani data set using hard voting and 22 features obtained as an ensemble of SVM and BS FS methods. The list of these 22 selected features in the Z-Alizadeh Sani data set can be found in Supplementary Table 8. This accuracy value was the second highest accuracy value obtained for the Z-Alizadeh Sani data set, as presented in Table 5. Similarly, for the Cleveland and Statlog data sets, the second highest accuracy values (84.80) and (86.29) are obtained using an ensemble classifier, which applies soft voting with 9 and 10 features, respectively. As shown in Table 6, the performance results of the probabilistic ensemble FS approach are also comparable. In the probabilistic ensemble FS approach, the highest accuracy (91.14) is also obtained using an MLP classifier using 25 selected features from the Z-Alizadeh Sani data set. The list of these 25 selected features can be found in Supplementary Table 9. Using both the exhaustive ensemble and probabilistic FS approaches, the MLP classifier had one of the best performance results in all three CAD data sets. This model resulted in 91.78% accuracy, 93.50% sensitivity, 95.14% precision, 0.941 F-measure, and 0.956 area under the curve (AUC) on the Z-Alizadeh Sani data set. The same classifier achieved 85.47% accuracy, 82.96% sensitivity, 86.22% precision, 0.839 F-Measure, and 0.911 AUC on the Cleveland data set. Lastly, the same classifier achieved 85.55% accuracy, 85.00% sensitivity, 86.25% precision, 0.853 F-measure, and 0.888 AUC on the Statlog data set. In order to allow the reproducibility of our results, we present the hyper-parameters of the best-performing classifiers in Table 7.

When we analyze our results on the Z-Alizadehsani data set via comparing the list of selected features among the top scoring exhaustive ensemble FS approaches, we observed that the three features (Typical Chest Pain, Region RWMA, and HTN) are commonly selected and used by all classifiers (presented in Supplementary Table 8). Also, age, tinversion, nonanginal, DM, and EF-TTE features are used by almost all classifiers. Taken together, these eight features correspond to the first eight features listed in Supplementary Table 5. Similarly, when comparing the selected features in the probabilistic ensemble FS approach applied to the Z-Alizadeh Sani data set, we noticed in Supplementary Table 9 that the Typical Chest Pain and Region RWMA features are used by all classifiers. Also, age and K features are used by almost all classifiers. On the other hand, tinversion, nonanginal, DM, EF-TTE, and K features are not selected as important based on the medical expertise of the doctors. However, using different ML models, these features are identified as critical for CAD diagnosis.

7. Discussion

Through the development of ensemble FS approaches and the incorporation of DK, this study aims to create a computational CAD diagnosis model that can work accurately for different CAD data sets.

Table 5
The best performance results of each classifier on three different CAD data sets using the exhaustive ensemble FS method.

Data set	Classifier (Combined FS methods)	NF	SN (%)	PRE (%)	FM	AUC	ACC (%)
Z-Alizadeh Sani	MLP (SVM)	21	93.50	95.14	0.941	0.956	91.78
	HV3 (SVM + BS)	22	93.48	94.88	0.940	–	91.45
	SV3 (SVM + BS)	22	93.93	93.75	0.938	0.955	91.13
	SVM (SVM)	19	93.48	94.37	0.937	0.950	91.13
	LR (SVM)	19	94.39	93.65	0.938	0.954	91.11
	HV5 (SVM + BS)	22	93.48	94.01	0.935	–	90.80
	SV5 (SVM)	22	93.48	93.65	0.933	0.959	90.79
	RF (CS + SVM + BS)	12	92.59	92.59	0.927	0.937	90.49
	LDA (3 different FS Combination)	24	92.57	93.83	0.930	0.924	90.15
	KNN (2 different FS Combination)	5	92.87	85.57	0.841	0.891	84.86
Cleveland	MLP (CMIM)	9	82.96	86.22	0.839	0.911	85.47
	SV3 (IG + GR + RF + BS + DK)	9	81.16	85.21	0.843	0.916	84.79
	RF (IG + RF + SVM)	10	80.00	81.67	0.812	0.903	84.74
	SV5 (GR + SVM + CMIM)	9	83.39	84.07	0.839	0.917	84.44
	KNN (CS+RF+SVM)	5	81.37	85.93	0.825	0.903	84.41
	HV5 (CS + IG + SVM + BS + DK)	12	82.96	84.32	0.832	–	83.77
	HV3 (RF + CMIM + DK)	9	83.17	84.05	0.833	–	83.43
	LDA (24 different FS Combination)	11	78.51	84.52	0.809	0.901	83.41
	LR (2 different FS Combination)	10	79.34	83.78	0.811	0.902	83.11
	SVM (13 different FS Combination)	10	80.05	83.13	0.811	0.911	83.10
Statlog	KNN (IG + GR + BS)	5	86.75	87.65	0.869	0.899	87.03
	SV3 (CS + RF)	10	85.91	87.37	0.864	0.883	86.66
	SVM (4 different Combination)	10	86.08	87.39	0.864	0.892	86.66
	HV3 (CS + IG + RF + CMIM)	9	85.75	87.04	0.861	–	86.29
	SV5 (IG + GR + RF + BS + CMIM + DK)	9	84.91	86.32	0.846	0.896	86.29
	HV5 (IG + BS + CMIM)	6	85.33	87.04	0.857	–	85.92
	LDA (CS + RF + CMIM + DK)	11	85.92	86.74	0.857	0.892	85.92
	LR (64 different Combination)	7	85.08	86.41	0.853	0.896	85.55
	MLP (IG + RF + BS + DK)	6	85.00	86.25	0.853	0.887	85.55
	RF (IG + SVM + DK)	11	82.00	83.68	0.828	0.881	85.51

NF: Number of Features included, SN: Sensitivity, PRE: Precision, FM: F-Measure, AUC: Area Under Curve, ACC: Accuracy.

Table 6
The best performance results of six classifiers on the Z-Alizadeh Sani data set using the probabilistic ensemble FS method.

Data set	Classifier	NF	SN (%)	Pre (%)	FM	AUC	Acc (%)
Z-Alizadeh Sani	MLP	25	94.41	93.53	0.937	0.941	91.14
	RF	13	94.93	90.09	0.924	0.916	90.13
	SVM	25	92.96	93.32	0.929	0.929	90.01
	LR	23	93.50	92.29	0.927	0.947	89.51
	LDA	24	90.71	94.51	0.924	0.920	89.47
	KNN	8	87.40	91.13	0.890	0.894	84.77

NF: Number of Features included, SN: Sensitivity, Pre: Precision, FM: F-Measure, AUC: Area Under Curve, Acc: Accuracy.

In the last decade, in order to diagnose CAD, several DM and ML studies have been conducted, and the number of those studies keeps increasing. In our study, one of the most emphasized points is that there is no internationally recognized approach for CAD diagnosis using ML. Most of these studies present good performance results. However, these models did not perform well when they were applied to different CAD data sets. Hence, this study aims to achieve satisfying results with a single classifier on different CAD data sets instead of excellent results on a single data set. As can be seen in Table 5, the MLP classifier has achieved satisfactory results in three CAD data sets.

In this study, two novel ensemble FS approaches are proposed: The exhaustive ensemble FS method incorporates all possible combinations of different FS methods; the probabilistic ensemble FS approach assigns a probability score to each feature, indicating a chance for being included. Hence, the probabilistic FS approach gives a chance to lower scoring features. In order to test our proposed approach, we have utilized the Cleveland, Statlog, and Z-Alizadeh Sani data sets. When we compare the performance of our models (Table 5) with the performance of the models that do not use FS (Table 4), we observe that our proposed approaches have higher performance metrics. It is noteworthy to mention that our proposed approach also takes into account the models that use a single FS method in addition to the ensembles of FS methods. In our experiments using 10 different classifiers with stratified 10-fold CV, the accuracy of the proposed approach is comparable with the existing models, as shown in Table 8. However, most of the existing studies do not give adequate information on the data preprocessing, CV procedure, and data-splitting processes that might drastically affect the performance results and limit the reproducibility of the findings. Also, none of the existing studies present a detailed performance evaluation. Whereas in this study, the performance results of the proposed approach are shown in Tables 5 and 6 using several evaluation metrics, such as accuracy, sensitivity, precision, F-measure, and AUC.

A total of 94.350 models are generated using the proposed ensemble FS approach. Among these, we listed the top 1000, top 500, and top 100 accuracy values and checked the details of the models that gave rise to these performance metrics. To this end, we analyzed the frequencies of classification algorithms and the frequencies of FS methodologies in these top 1000, top 500, and top 100 lists, as shown in the Supplementary Figs 1, 2, and 3 for the Cleveland, Statlog, and Z-Alizadeh Sani data sets, respectively. In terms of classifiers, Supplementary Fig. 1 A shows that MLP and S3 classifiers generally gave better performance metrics than other classifiers for the Cleveland data set. As shown in Supplementary Fig. 2 A, among other classifiers, the SVM classifier resulted in better performance metrics for the Statlog data set. For the Z-Alizadeh Sani data set, as shown in Supplementary Fig. 3 A, MLP and SVM classifiers resulted in better performance metrics compared to other classifiers. In the exhaustive ensemble FS approach, especially for the Statlog data set, different classification models generate the same accuracy levels. Therefore, we plot the composition of the classifiers that generated the top five accuracies, as shown in Figs. 5, 6, and 7 for the Z-Alizadeh Sani, Cleveland, and Statlog data sets, respectively. Similar to our finding in Table 5, Figs. 5, 6, and 7 show that the MLP classifier works well on the three CAD data sets used in this study. We have also demonstrated the frequencies of the FS methods in Supplementary Figs. 1B, 2B, and 3B for the UCI Cleveland, UCI Stadlog, and Z-Alizadeh Sani data sets, respectively. The frequencies of each FS method are very close to each other, as shown in Supplementary Figs.

Table 7
The optimum hyper-parameters for ML methods are found by grid-search optimization.

Methods	Cleveland	Statlog	Z-Alizadeh Sani
SVM	C = 100, kernel = rbf	C = 1, kernel = linear	C = 1 kernel = linear
MLP	alpha = 10-e2, layer size = 50, max iter = 1000, activation = relu	alpha = 1.0, layer size = 100, max iter = 1000, activation = relu	alpha = 10-e2, layer size = 50, max iter = 1000, activation = tanh
RF	estimators = 100	estimators = 100	estimators = 100
KNN	neighbors = 11	neighbors = 9	neighbors = 9
LR	solver = lbfgs, max iter = 2000	solver = lbfgs, max iter = 2000	solver = lbfgs, max iter = 2000
LDA	default	default	default

Table 8
Performance results of the proposed method and existing studies.

Study, Year	Data set	Method	Accuracy (%)
[47], 2008	Cleveland	KNN	85.55
[48], 2011	Cleveland	SVM	80.06
[18], 2011	Cleveland	DT	84.10
[20], 2012	Z-Alizadeh Sani	SMO	92.09
[49], 2013	Z-Alizadeh Sani	Bagging	79.54
[19], 2013	Cleveland	SVM	82.00
[15], 2015	Cleveland	DT	78.54
[23], 2017	Z-Alizadeh Sani	DT	86.67
[16], 2019	Cleveland	RF	92.16
[21], 2020	Z-Alizadeh Sani	RT	91.47
[22], 2021	Z-Alizadeh Sani	DA	95.00
[50], 2021	Z-Alizadeh Sani	DT	93.00
This study	Cleveland	MLP	85.47
This study	Statlog	MLP	85.55
This study	Z-Alizadeh Sani	MLP	91.78

1B and 2B for the UCI Cleveland and UCI Stadlog data sets. One can observe from Supplementary Fig. 3B that the SVM and the DK-based FS methods yielded better performance metrics compared to the other FS methods for the Z-Alizadeh Sani data set. Based on these findings, we can state that, in terms of the frequencies of FS methods in the top 100, top 500, and top 1000 results, SVM and DK FS methods achieve satisfactory results on the three CAD data sets used in this study.

On the other hand, with the development of new technologies, it becomes possible to build an online system that can diagnose CAD disease all around the world using edge computing [51]. Such systems can provide closed-to-user and ultra-low latency computation [52,53]. Hereby, a better model can be obtained, and the accuracy of the diagnosis can be improved.

8. Conclusions

With the development of ML techniques, it becomes possible to diagnose CAD at a lower cost by using biochemical values. However, there is no internationally recognized standard ML approach for CAD diagnosis. Although some studies have reported satisfactory performance results for the CAD diagnosis on a particular CAD data set, these models do not perform well on different CAD data sets. In this study, for a CAD diagnosis problem, two ensemble FS approaches have been proposed, and different classification algorithms have been experimented. Our exhaustive ensemble FS approach analyzed all possible combinations of seven different computational FS methods and one DK-based FS method for different numbers of features. Our probabilistic FS approach assigns a probabilistic score to each feature, indicating the chance that feature will be included in the ensemble FS method. Hence, the probabilistic FS approach gives a chance to lower scoring features. For the classification task, six single classifiers and one ensemble classifier with four variations are utilized. Although none of the existing studies present a detailed performance evaluation, in this study, the performance results of the proposed approach are presented with several evaluation metrics, including accuracy, sensitivity, precision, F1-measure, and AUC. In our

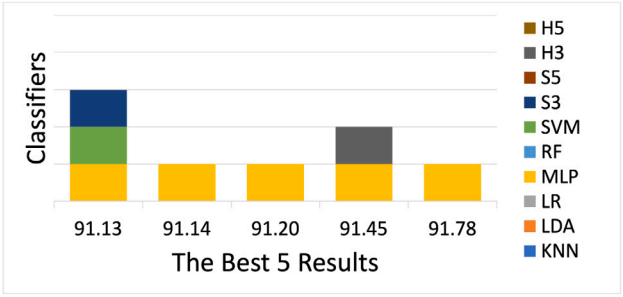

Fig. 5. The best five performing classifiers on the Z-Alizadeh Sani data set.

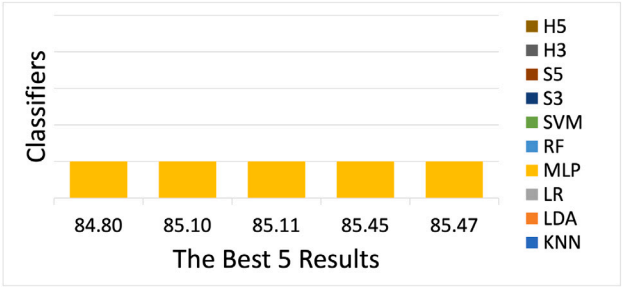

Fig. 6. The best five performing classifiers on the Cleveland data set.

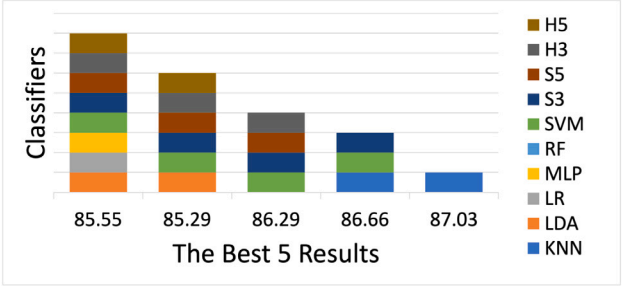

Fig. 7. The best five performing classifiers on the Statlog data set.

experiments with stratified 10-fold CV, the MLP classifier achieved one of the best performance results in three CAD data sets. This model resulted in 91.78% accuracy, 93.50% sensitivity, 95.14% precision, 0.941 F-measure, and 0.956 AUC on the Z-Alizadeh Sani data set. The same classifier achieved 85.47% accuracy, 82.96% sensitivity, 86.22% precision, 0.839 F-measure, and 0.911 AUC on the Cleveland data set. Lastly, the same classifier achieved 85.55% accuracy, 85.00% sensitivity, 86.25% precision, 0.853 F-measure, and 0.888 AUC on the Statlog data set.

The main contribution of this paper is the proposed ensemble-based FS approaches. It is noteworthy to state that our primary goal is to offer

an adaptive approach that can work on several CAD data sets without additional analysis. The proposed approaches are tested on three different CAD data sets and shown to generate decent performance in terms of accuracy, sensitivity, precision, F-measure, and AUC. Our proposed model could be adapted to current practice easily. In future work, using our proposed FS methodology, we plan to study various deep learning algorithms with different CAD data sets.

CRediT authorship contribution statement

Burak Kolukisa: Conceptualization, Methodology, Software, Visualization, Validation, Writing – original draft. **Burcu Bakir-Gungor:** Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.csi.2022.103706.

Data availability

Publicly available data was used for the research described in the article.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csi.2022.103706.

References

[1] World health organization, 2022, URL https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). (Accessed 02 September 2022).

[2] Coronary heart disease, 2022, URL https://www.bhf.org.uk/informationsupport/conditions/coronary-heart-disease. (Accessed 02 September 2022).

[3] A. Mandal, Cardiovascular disease diagnosis, 2019, URL https://www.news-medical.net/health/Cardiovascular-Disease-Diagnosis.aspx. (Accessed 02 September 2022).

[4] G. Rabottino, A. Mencattini, M. Salmeri, F. Caselli, R. Lojacono, Performance evaluation of a region growing procedure for mammographic breast lesion identification, Comput. Stand. Interfaces 33 (2) (2011) 128–135.

[5] H. Ku, W. Susilo, Y. Zhang, W. Liu, M. Zhang, Privacy-preserving federated learning in medical diagnosis with homomorphic re-encryption, Comput. Stand. Interfaces (80) (2022) 103583.

[6] R. Espinosa, D. García-Saiz, M. Zorrilla, J.J. Zubcoff, J.N. Mazón, S3Mining: A model-driven engineering approach for supporting novice data miners in selecting suitable classifiers, Comput. Stand. Interfaces 65 (2019) 143–158.

[7] Y. Song, Web service reliability prediction based on machine learning, Comput. Stand. Interfaces (73) (2021) 103466.

[8] S.H. Chin, C. Lu, Y.F. P. T. Ho, T. J., Commodity anti-counterfeiting decision in e-commerce trade based on machine learning and internet of things, Comput. Stand. Interfaces (76) (2021) 103504.

[9] A.A. Afuwape, Y. Xu, J.H. Anajemba, G. Srivastava, Performance evaluation of secured network traffic classification using a machine learning approach, Comput. Stand. Interfaces (78) (2021) 103545.

[10] P.K. Roy, A.K. Tripathy, T.H. Weng, K.C. Li, Securing social platform from misinformation using deep learning, Comput. Stand. Interfaces (84) (2023) 103674.

[11] R. Alizadehsani, M. Abdar, M. Roshanzamir, A. Khosravi, P.M. Kebria, F. Khozeimeh F, et al., Machine learning-based coronary artery disease diagnosis: A comprehensive review, Comput. Biol. Med. 111 (2019).

[12] B. Kolukısa, H. Hacılar, G. Goy, M. Kuş, B. Bakır-Güngör, A. Aral, et al., Diagnosis of coronary heart disease via classification algorithms and a new feature selection methodology, Int. J. Data Min. Sci. 1 (1) (2019) 8–15.

[13] B. Kolukısa, Development of Data Mining Methodologies and Machine Learning Models to Understand Cardiovascular Disease Mechanisms, Abdullah Gül University, Kayseri, Turkey, 2020.

[14] M. Anbarasi, E. Anupriya, N.C.S.N. Iyengar, Enhanced prediction of heart disease with feature subset selection using genetic algorithm, Int. J. Eng. Sci. Technol. 2 (10) (2010) 5370–5376.

[15] R. El-Bialy, M.A. Salamay, O.H. Karam, M.E. Khalifa, Feature analysis of coronary artery heart disease data sets, Procedia Comput. Sci. 65 (2015) 459–468.

[16] N.S.C. Reddy, S.S. Nee, L.Z. Min, C.X. Ying, Classification and feature selection approaches by machine learning techniques: Heart disease prediction, Int. J. Innov. Comput. 9 (1) (2019).

[17] T.K. Ho, M. Basu, Complexity measures of supervised classification problems, IEEE Trans. Pattern Anal. Mach. Intell. 24 (3) (2002) 289–300.

[18] M. Shouman, T. Turner, R. Stocker, Using decision tree for diagnosing heart disease patients, in: Proceedings of the Ninth Australasian Data Mining Conference-Volume 121, 2011, pp. 23–30.

[19] R. Chitra, V. Seenivasagam, Heart disease prediction system using supervised learning classifier, Bonfring Int. J. Software Eng. Soft Comput. 3 (1) (2013) 1–7.

[20] R. Alizadehsani, M.J. Hosseini, Z.A. Sani, A. Ghandeharioun, R. Boghrati, Diagnosis of coronary artery disease using cost-sensitive algorithms, in: IEEE 12th International Conference on Data Mining Workshop, 2012, pp. 9–16.

[21] J.H. Joloudari, E. Hassannataj Joloudari, H. Saadatfar, M. Ghasemigol, S.M. Razavi, A. Mosavi, et al., Coronary artery disease diagnosis; Ranking the significant features using a random trees model, Int. J. Environ. Res. Public Health 17 (3) (2020) 731.

[22] A. Bhatnagar, An effectual machine learning based coronary artery disease classification for low error rates, Turkish J. Comput. Math. Educ. 12 (6) (2021) 5433–5442.

[23] F. Babič, J. Olejár, Z. Vantová, J. Paralič, Predictive and descriptive analysis for heart disease diagnosis, in: 2017 IEEE Federated Conference on Computer Science and Information Systems, 2017, pp. 155–163.

[24] B. Kolukisa, H. Hacilar, G. Goy, M. Kus, B. Bakir-Gungor, A. Aral, et al., Evaluation of classification algorithms, linear discriminant analysis and a new hybrid feature selection methodology for the diagnosis of coronary artery disease, in: 2018 IEEE International Conference on Big Data, Big Data, 2018, pp. 2232–2238.

[25] B. Kolukisa, L. Yavuz, A. Soran, B. Bakir-Gungor, D. Tuncer, A. Onen, et al., Coronary artery disease diagnosis using optimized adaptive ensemble machine learning algorithm, Int. J. Biosci., Biochem. Bioinform. 10 (1) (2020) 58–65.

[26] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.J. Schmid, S. Sandhu, et al., International application of a new probability algorithm for the diagnosis of coronary artery disease, Am. J. Cardiol. 64 (5) (1989) 304–310.

[27] R. Alizadehsani, J. Habibi, M.J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, et al., A data mining approach for diagnosis of coronary artery disease, Comput. Methods Programs Biomed. 111 (1) (2013) 52–61.

[28] B.K. Dedeturk, B. Akay, Spam filtering using a logistic regression model trained by an artificial bee colony algorithm, Appl. Soft Comput. 91 (2020).

[29] G. Polančič, B. Cegnar, Complexity metrics for process models–A systematic literature review, Comput. Stand. Interfaces 51 (2017) 104–117.

[30] T.K. Ho, M. Basu, Complexity measures of supervised classification problems, IEEE Trans. Pattern Anal. Mach. Intell. 24 (3) (2002) 289–300.

[31] Z.M. Hira, D.F. Gillies, A review of feature selection and feature extraction methods applied on microarray datas, Adv. Bioinform. (2015).

[32] I. Guyon, S. Gunn, M. Nikravesh, L.A. Zadeh, Feature Extraction: Foundations and Applications, Vol. 207, Springer, 2008.

[33] R. Jensen, O. Shen, Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches, Vol. 24, no. 3, John Wiley & Sonse, 2008, pp. 289–300.

[34] Y. Zhai, W. Song, X. Liu, X. Zhao, A chi-square statistics based feature selection method in text classification., in: 2018 IEEE 9th International Conference on Software Engineering and Service Science, ICSESS, 2018, pp. 160–163.

[35] K. Igor, E. Šimec, M. Robnik-Šikonja, Overcoming the myopia of inductive learning algorithms with RELIEFF, Appl. Intell. 7 (1) (1997) 39–55.

[36] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2002) 389–422.

[37] G. Brown, A. Pocock, M.J. Zhao, M. Luján, Conditional likelihood maximisation: A unifying framework for information theoretic feature selection, J. Mach. Learn. Res. 13 (2012) 27–66.

[38] D. Karaboga, An Idea Based on Honey Bee Swarm for Numerical Optimization, Technical Report-Tr06, Erciyes University, Engineering Faculty, Computer Engineering Department, 2005.

[39] D.T. Pham, M. Castellani, The bees algorithm: Modelling foraging behaviour to solve continuous optimization problems, Proc. Inst. Mech. Eng., Part C: J. Mech. Eng. Sci. 223 (12) (2009) 2919–2938.

[40] Framingham heart study (FHS), 2011, URL https://www.nhlbi.nih.gov/science/framingham-heart-study-fhs. (Accessed 02 September 2022).

[41] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (1) (2014) 16–28.

[42] M.J. Tsai, C.S. Wang, J. Liu, J.S. Yin, Using decision fusion of feature selection in digital forensics for camera source model identification, Comput. Stand. Interfaces 34 (3) (2012) 292–304.

[43] C. Mohan, S. Nagarajan, An improved tree model based on ensemble feature selection for classification, Turk. J. Electr. Eng. Comput. Sci. 27 (2) (2019) 1290–1307.

[44] F. Eibe, M.A. Hall, I.H. Witten, J. Pal, The WEKA workbench, 2016, Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques, 4.

[45] G. Van Rossum, F.L. Drake, Python 3 reference manual CA: CreateSpace, 2009.
[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[47] S. Palaniappan, R. Awang, Intelligent heart disease prediction system using data mining techniques, in: 2008 IEEE/ACS International Conference on Computer Systems and Applications, 2008, pp. 108–115.
[48] M. Kumari, S. Godara, Comparative study of data mining classification methods in cardiovascular disease prediction 1, 2011.
[49] R. Alizadehsani, J. Habibi, M.J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, et al., Diagnosing coronary artery disease via data mining algorithms by considering laboratory and echocardiography features, Res. Cardiovasc. Med. 2 (3) (2013) 133.
[50] G. Sharma, G. Rani, V.S. Dhaka, Efficient predictive modelling for classification of coronary artery diseases using machine learning approach, IOP Conf. Ser.: Mater. Sci. Eng. 1099 (1) (2021).
[51] L. Zhao, W. Zhao, A. Hawbani, A.Y. Al-Dubai, G. Min, A.Y. Zomaya, C. Gong, Novel online sequential learning-based adaptive routing for edge software-defined vehicular networks, IEEE Transactions on Wireless Communications 20 (5) (2020) 2991–3004.
[52] L. Zhao, H. Li, N. Lin, M. Lin, C. Fan, J. Shi, Intelligent content caching strategy in autonomous driving toward 6g, IEEE Transactions on Intelligent Transportation Systems 23 (7) (2021) 9786–9796.
[53] L. Zhao, K. Yang, Z. Tan, X. Li, S. Sharma, Z. Liu, A novel cost optimization strategy for sdn-enabled uav-assisted vehicular computation offloading, IEEE Transactions on Intelligent Transportation Systems 22 (6) (2020) 3664–3674.

Burak Kolukisa received the B.S in Computer Engineering From Erciyes University, Kayseri, in 2016 and M.S. degrees in Electrical and Computer Engineering from Abdullah Gul University, Kayseri, Turkey, in 2020. Currently, he is working as a Research Assistant at the Department of Computer Engineering, Abdullah Gul University. His current research interests are Data Mining, Machine Learning, and Deep Learning. He is also a Ph.D. candidate Electrical and Computer Engineering program at Abdullah Gul University, Kayseri, Turkey.

Burcu Bakir Gungor received her B.Sc. degree from Sabanci University; her M.Sc. degree in Bioinformatics from Georgia Institute of Technology; and her PhD degree from Georgia Institute of Technology/Sabanci University. She worked at the Bioinformatics Research Center, Medical College of Wisconsin, from 2007–2009. From 2009 to 2011, she worked at the Department of Computer Engineering, Bahcesehir University. Then, she worked as an Assistant Professor at the Department of Genetics and Bioinformatics, at the same university. From 2012 to 2013, she was part of the Advanced Genomics and Bioinformatics Research Center, UEKAE, BILGEM, TUBITAK. Now, she works as an Assistant Professor at the Department of Computer Engineering at Abdullah Gul University. In 2022, she was deemed worthy of an award in the L'ORÉAL – UNESCO National Fellowship for Women in Science Programme. She is the recipient of "Best Paper" awards at the UBMK 2020 and 4th EvoBIO Conferences. She is an Editorial Board member of PeerJ journal; she is the reviewer of several prestigious international journals including Bioinformatics, Machine Learning, Journal of Computational Biology, Frontiers in Genetics; and she is a Technical Program Committee member of SIU, UBMK and HIBIT conferences. Her research interests include bioinformatics, computational genomics, applications of machine learning and data mining in bioinformatics.