Journal of
Clinical Medicine

MDPI

Article

Novel CAD Diagnosis Method Based on Search, PCA, and AdaBoostM1 Techniques

Can Eyupoglu [1] **and Oktay Karakuş** [2,*]

[1] Department of Computer Engineering, Turkish Air Force Academy, National Defence University, Istanbul 34149, Türkiye; caneyupoglu@gmail.com
[2] School of Computer Science and Informatics, Cardiff University, Cardiff CF24 4AG, UK
* Correspondence: karakuso@cardiff.ac.uk

Abstract: Background: Cardiovascular diseases (CVDs) are the primary cause of mortality worldwide, resulting in a growing number of annual fatalities. Coronary artery disease (CAD) is one of the basic types of CVDs, and early diagnosis of CAD is crucial for convenient treatment and decreasing mortality rates. In the literature, several studies use many features for CAD diagnosis. However, due to the large number of features used in these studies, the possibility of early diagnosis is reduced. **Methods:** For this reason, in this study, a new method that uses only five features—age, hypertension, typical chest pain, t-wave inversion, and region with regional wall motion abnormality—and is a combination of eight different search techniques, principal component analysis (PCA), and the AdaBoostM1 algorithm has been proposed for early and accurate CAD diagnosis. **Results:** The proposed method is devised and tested on a benchmark dataset called Z-Alizadeh Sani. The performance of the proposed method is tested with a variety of metrics and compared with basic machine-learning techniques and the existing studies in the literature. The experimental results have shown that the proposed method is efficient and achieves the best classification performance, with an accuracy of 91.8%, ever reported on the Z-Alizadeh Sani dataset with so few features. **Conclusions:** As a result, medical practitioners can utilize the proposed approach for diagnosing CAD early and accurately.

Keywords: AdaBoostM1; cardiovascular diseases; coronary artery disease diagnosis; machine learning; PCA; search techniques

Citation: Eyupoglu, C.; Karakuş, O. Novel CAD Diagnosis Method Based on Search, PCA, and AdaBoostM1 Techniques. *J. Clin. Med.* **2024**, *13*, 2868. https://doi.org/10.3390/jcm13102868

Academic Editor: Andreas A. Kammerlander

Received: 24 March 2024
Revised: 26 April 2024
Accepted: 7 May 2024
Published: 13 May 2024

1. Introduction

Cardiovascular diseases (CVDs) are a class of disorders that include the blood and heart vessels [1,2]. The main types and risk factors of CVDs are shown in Figure 1. Coronary artery disease (CAD) is an illness that influences the blood vessels providing blood to the heart and occurs when coronary arteries are blocked or narrowed. Additionally, it is associated with ischemic heart disease, coronary heart disease, atherosclerotic heart disease, heart failure, heart attack, sudden coronary death, and angina pectoris medical science. A stroke is disease-causing damage to a particular brain area and occurs when blood vessels are ruptured or blocked. Finally, peripheral artery disease affects the blood vessels that supply blood to the feet and legs [2]. Readers are referred to [2] for the main risk factors of CVDs and characteristics that can be associated with CVDs.

CVDs are the main cause of death all over the world. In 2021, the World Health Organization (WHO) [3] reported that nearly 17.9 million people died from CVDs in 2019. This number of mortalities constitutes 32% of the total deaths worldwide, and 85% of these deaths are caused by heart attack and stroke. In addition, the top ten global causes of mortality in 2019 are shown in Figure 2. As shown in the figure, CAD, also known as ischemic heart disease, is the leading reason for deaths [4].

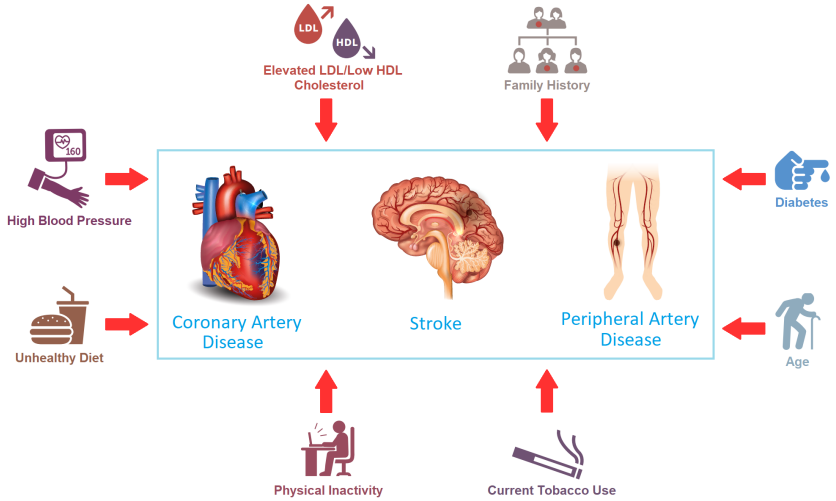

Figure 1. The main types (rectangle at the centre) and risk factors (around the rectangle with arrows) of CVDs [2].

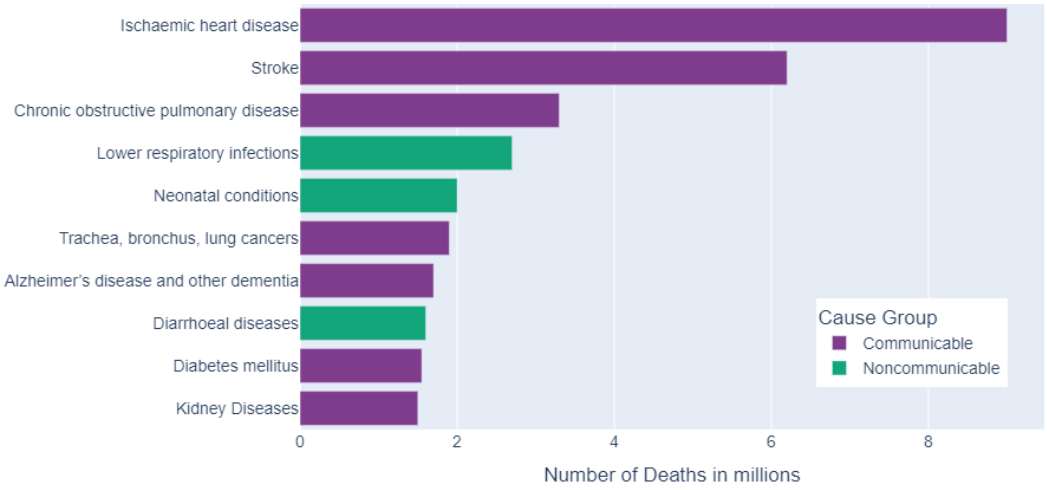

Figure 2. Top ten global causes of mortality in 2019 [4].

Disease diagnosis is a highly complex process in medical science, and many tests are necessary for accurate diagnosis. In order to help medical doctors with the early detection of disease, machine learning and data mining techniques have been widely utilized recently [5]. Especially in CAD, with early detection, the possibility of treatment is greatly increased and patients' lives can be saved.

In the literature, numerous methods have been developed to diagnose CAD on the Cleveland heart disease dataset [6] up to the present [7–17]. The prediction and diagnosis success of the existing studies tested on this dataset is quite satisfying. For this reason, in this study, the performance of the proposed approach is evaluated on a newer dataset called the Z-Alizadeh Sani [18], released in 2017. In the work introduced by Alizadehsani et al. [19], the Z-Alizadeh Sani dataset was collected and utilized for the first time for CAD diagnosis. From 2012 to 2016, Alizadehsani et al. employed various machine learning techniques such as sequential minimal optimization (SMO), artificial neural network (ANN), support vector machine (SVM), Naïve Bayes, bagging, C4.5 decision tree, information gain, and a genetic algorithm for CAD detection [19–25]. In a study using the same dataset, Qin et al. [26] presented a CAD detection method utilizing an ensemble algorithm based on multiple feature selection (EA-MFS) and SVM. Arabasadi et al. [27] proposed a hybrid CAD prediction approach combining a genetic algorithm and multilayer perceptron artificial neural network (MLP-ANN) on a subset of the Z-Alizadeh Sani dataset from which

22 features were selected. In order to diagnose CAD, Babič et al. [28] offered a predictive and descriptive analysis. They used four different classifiers such as decision trees, Naïve Bayes, SVM, and ANN.

In the work of Kılıç and Kaya Keleş [29], the artificial bee colony (ABC) algorithm and SMO technique were utilized for feature selection and classification, respectively. Sixteen features were selected by the ABC algorithm, and SMO was applied to these features. Hu et al. [30] proposed two methods, namely, minimum message length finite inverted Beta-Liouville mixture (MML-IBLMM) and variational finite inverted Beta-Liouville mixture (Var-IBLMM), and then tested the performances of these models on Z-Alizadeh Sani dataset. In the study introduced by Abdar et al. [31], a CAD detection technique called N2Genetic-nuSVM, which is based on a genetic optimizer and nu-support vector classification, was presented. In another work realized by Abdar et al. [32], a nested ensemble nu-support vector classification (NE-nu-SVC) approach was proposed to diagnose CAD accurately. In the feature selection step of the proposed approach, a genetic search method was utilized, and 16 features were selected. In the research of Joloudari et al. [33], the performances of SVM, chi-squared automatic interaction detection (CHAID) decision tree, C5.0 decision tree, and random trees were investigated for CAD diagnosis. The experimental results indicate that the random trees technique is better than the other classifiers. On the other hand, Nasarian et al. [34] presented a hybrid feature selection method named heterogeneous hybrid feature selection (2HFS) that utilizes the synthetic minority over-sampling technique (SMOTE) and adaptive synthetic (ADASYN) for handling the Z-Alizadeh Sani dataset and uses random forests, Gaussian Naïve Bayes, eXtreme Gradient Boosting (XGBoost), and decision tree for CAD classification. In another work proposed by Ashish et al. [35], a CAD detection method based on random forests, SVM, and XGBoost was introduced. In the data-dividing step of the method, the random forests technique was used for training and testing of the Z-Alizadeh Sani dataset. In the classification step, the SVM and XGBoost techniques were utilized together. In a recent study [36], an ensemble feature selection approach and seven classifiers were used, and the best classification accuracy rate was attained with 25 features and the MLP classifier.

The aforementioned studies adopted some combination of feature selection, feature extraction, and classification techniques such as information gain, genetic algorithm, ABC, bagging, decision trees, random trees, Naïve Bayes, SMO, SVM, SVC, ANN, CHAID, and XGBoost to diagnose CAD. Unlike the abovementioned methods, this work proposes a new CAD diagnosis method based on eight different search techniques, principal component analysis (PCA), and AdaBoostM1. To the best of the author's knowledge, there is no other work in the literature utilizing PCA and AdaBoostM1 techniques together for CAD diagnosis in this framework and detecting CAD based on age, hypertension, typical chest pain, t-wave inversion, and region with regional wall motion abnormality features. The major findings and contributions of this research study are as follows:

- Proposes a new method to diagnose CAD based on age, hypertension, typical chest pain, t-wave inversion, and region with regional wall motion abnormality features.
- Explores attribute spaces using eight different search methods, namely, evolutionary, best first, genetic, harmony, particle swarm optimization (PSO), greedy stepwise, rank, and multi-objective evolutionary search.
- Enhances the performance of CAD diagnosis by efficiently taking advantage of using PCA and AdaBoostM1 techniques together.
- The performance of the proposed method is tested in terms of several metrics and compared with basic classifiers and existing studies in the literature.
- Achieves the best classification performance ever reported on the Z-Alizadeh Sani dataset with so few features (five) with an accuracy rate of 91.80%.
- The experimental results demonstrate that the proposed method is promising to be utilized by medical specialists for diagnosing CAD.

The rest of the paper is ordered as follows. The proposed CAD diagnosis method and the dataset used are introduced in Section 2. Section 3 shows the experimental results,

comparing the proposed method's performance to the existing studies in the literature. Finally, conclusions are summarized in Section 4.

2. Materials and Methods

2.1. Dataset Description

In this work, the Z-Alizadeh Sani dataset that is freely available from the University of California—Irvine Machine Learning Repository [18] was used to evaluate the proposed method. The dataset contains 303 records, of which 87 of them are healthy persons and 216 of them are CAD patients. Fifty-five attributes can be classified into four groups: symptom and examination (14 attributes), demographic (17 attributes), electrocardiography (ECG) (7 attributes), and laboratory and echocardiography (echo) (17 attributes). The overview of the Z-Alizadeh Sani dataset, including attribute name, category, and range, is shown in Table 1.

2.2. The Proposed CAD Diagnosis Method

This study presents a new CAD diagnosis method based on age, HTN, typical chest pain, t-wave inversion, and region-RWMA features. The proposed method comprises four basic steps, which are feature selection, feature extraction, data dividing, and classification. The flowchart of the proposed CAD diagnosis method is demonstrated in Figure 3. A correlation-based feature subset selection technique is utilized with evolutionary, best first, genetic, harmony, PSO, greedy stepwise, rank, and multi-objective evolutionary search methods in the feature selection step. Then, the PCA technique transforming the data into another space is used for feature extraction and size reduction on the data obtained after selecting common attributes. In the data-dividing step, the k-fold cross-validation technique is exploited to divide the whole dataset into k separate subsets, in which k-1 subsets are utilized for training and the remaining part is separated for testing. In the classification step, the AdaBoostM1 algorithm is performed for classifying coronary artery disease as healthy or patient. The techniques utilized to perform the proposed diagnostic method are described in the following subsections.

2.2.1. Feature Selection

In the feature selection step of the proposed method, a correlation-based feature subset selection technique [37] was used with eight different search methods, namely, evolutionary [38], best first [39], genetic [40], harmony [41], PSO [42], greedy stepwise [43], rank [44], and multi-objective evolutionary search [45]. To evaluate the worth of a subset of attributes, the feature subset selection technique considers the estimative ability of every feature associated with the redundancy degree between them. The evolutionary search method utilizes an evolutionary algorithm (EA) to discover the attribute space. The best first search method uses a greedy hillclimbing algorithm enhanced with a backtracking ability for searching the space of a subset of attributes. The genetic search method carries out a search utilizing Goldberg's genetic algorithm. The greedy stepwise search method applies a greedy backward/forward search, along with the space of a subset of attributes. To rank all the attributes, the rank search method utilizes a subset or attribute evaluator. Finally, the harmony, PS,O and multi-objective evolutionary search methods explore the attribute space using the harmony, PSO, and multi-objective evolutionary algorithms, respectively. Interested readers can kindly refer to [38–45] for further details about the search methods.

In the initial phase of the feature selection process outlined in the proposed method, eight distinct search methods are employed to explore and identify useful attributes. Table 2 presents the attributes selected through these search methods, along with the respective counts and attribute numbers. Subsequently, in the second stage, attributes common to all search methods are retained. As indicated in Table 2 (common attributes highlighted in bold), the selected attributes are numbered 1, 7, 25, 35, and 54, corresponding to features such as Age, HTN, Typical Chest Pain, T-Wave Inversion, and Region-RWMA.

Table 1. The Overview of Z-Alizadeh Sani Dataset.

#	Attribute Name	Attribute Category	Attribute Range
1	Age	Demographic	30–86
2	Weight	Demographic	48–120
3	Length	Demographic	140–188
4	Sex	Demographic	Female, Male
5	Body Mass Index (BMI, Kg/m^2)	Demographic	18.115–40.901
6	Diabetes Mellitus (DM)	Demographic	Yes, No
7	Hypertension (HTN)	Demographic	Yes, No
8	Current Smoker	Demographic	Yes, No
9	Ex-Smoker	Demographic	Yes, No
10	Family History (FH)	Demographic	Yes, No
11	Obesity (BMI $>$ 25)	Demographic	Yes, No
12	Chronic Renal Failure (CRF)	Demographic	Yes, No
13	Cerebrovascular Accident (CVA)	Demographic	Yes, No
14	Airway Disease	Demographic	Yes, No
15	Thyroid Disease	Demographic	Yes, No
16	Congestive Heart Failure (CHF)	Demographic	Yes, No
17	Dyslipidemia (DLP)	Demographic	Yes, No
18	Blood Pressure (BP, mmHg)	Symptom and examination	90–190
19	Pulse Rate (PR, ppm)	Symptom and examination	50–110
20	Edema	Symptom and examination	Yes, No
21	Weak Peripheral Pulse	Symptom and examination	Yes, No
22	Lung Rales	Symptom and examination	Yes, No
23	Systolic Murmur	Symptom and examination	Yes, No
24	Diastolic Murmur	Symptom and examination	Yes, No
25	Typical Chest Pain	Symptom and examination	Yes, No
26	Dyspnea	Symptom and examination	Yes, No
27	Function Class	Symptom and examination	0, 1, 2, 3
28	Atypical	Symptom and examination	Yes, No
29	Nonanginal Chest Pain	Symptom and examination	Yes, No
30	Exertional Chest Pain	Symptom and examination	Yes, No
31	Low Threshold Angina (Low TH Ang)	Symptom and examination	Yes, No
32	Q-Wave	ECG	Yes, No
33	ST Elevation	ECG	Yes, No
34	ST Depression	ECG	Yes, No
35	T-Wave Inversion	ECG	Yes, No
36	Left Ventricular Hypertrophy (LVH)	ECG	Yes, No
37	Poor R-Wave Progression	ECG	Yes, No
38	Bundle Branch Block (BBB)	ECG	Left, Right, Normal
39	Fasting Blood Sugar (FBS, mg/dL)	Laboratory and echo	62–400
40	Creatine (Cr, mg/dL)	Laboratory and echo	0.5–2.2
41	Triglyceride (TG, mg/dL)	Laboratory and echo	37–1050
42	Low Density Lipoprotein (LDL, mg/dl)	Laboratory and echo	18–232
43	High Density Lipoprotein (HDL, mg/dL)	Laboratory and echo	15.9–111
44	Blood Urea Nitrogen (BUN, mg/dL)	Laboratory and echo	6–52
45	Erythrocyte Sedimentation Rate (ESR, mm/h)	Laboratory and echo	1–90
46	Hemoglobin (HB, g/dL)	Laboratory and echo	8.9–17.6
47	Potassium (K, mEq/lit)	Laboratory and echo	3–6.6
48	Sodium (Na, mEq/lit)	Laboratory and echo	128–156
49	White Blood Cell (WBC, cells/mL)	Laboratory and echo	3700–18,000
50	Lymphocyte (%)	Laboratory and echo	7–60
51	Neutrophil (%)	Laboratory and echo	32–89
52	Platelet (PLT, 1000/mL)	Laboratory and echo	25–742
53	Ejection Fraction (%)	Laboratory and echo	15–60
54	Region-RWMA	Laboratory and echo	0, 1, 2, 3, 4
55	Valvular Heart Disease (VHD)	Laboratory and echo	Mild, Severe, Moderate, Normal

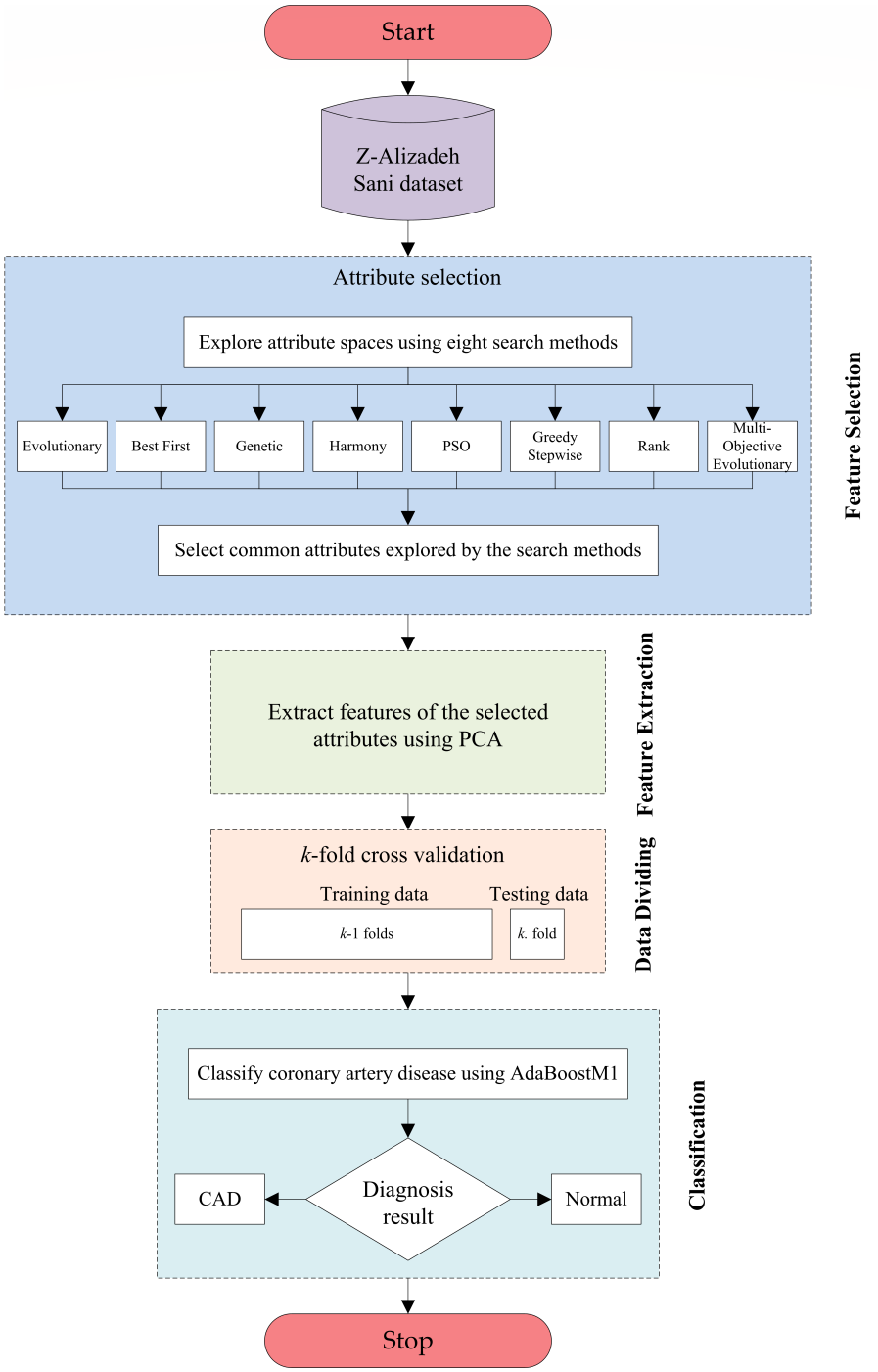

Figure 3. Flowchart of the proposed CAD diagnosis method.

2.2.2. Feature Extraction

The data collected from a system often have dozens of related attributes. However, there may only be a few actual driving forces governing the behavior of a system, even though we have more attributes in the data measuring many system variables that provide redundant information [46]. It is usually possible to simplify problems containing redundancy by taking advantage of dimensionality reduction techniques. PCA is one of the most famous kinds of dimensionality reduction methods and has been widely used in various fields till now. It is intensely used for dimension reduction and feature extraction purposes as it decreases overfitting risk, reduces computational complexity, eliminates distracting noise, and so on [47].

Table 2. The Attributes Chosen Using Search Methods. Bold attribute numbers refer to the common features for all search methods.

Search Method	Number of Chosen Attributes	Attribute No.
Evolutionary	17	**1**, **7**, 9, 14, 15, 18, 24, **25**, 28, 29, 31, **35**, 39, 41, 45, 47, **54**
Best first	12	**1**, 6, **7**, 18, **25**, 28, 29, **35**, 45, 47, 53, **54**
Genetic	15	**1**, 4, 6, **7**, 12, 18, **25**, 28, 29, 32, 34, **35**, 47, 53, **54**
Harmony	17	**1**, **7**, 12, 13, 14, 15, 17, **25**, 27, 28, 29, **35**, 37, 45, 47, 53, **54**
PSO	14	**1**, 6, **7**, 18, **25**, 28, 29, 32, 34, **35**, 45, 47, 53, **54**
Greedy stepwise	10	**1**, 6, **7**, 14, **25**, **35**, 45, 47, 53, **54**
Rank	13	**1**, 6, **7**, 14, **25**, 28, 29, 32, 33, **35**, 45, 53, **54**
Multi-objective evolutionary	10	**1**, 6, **7**, 14, **25**, **35**, 45, 47, 53, **54**

PCA employs orthogonal transformations to condense multiple correlated variables into a reduced set of uncorrelated variables [47,48]. This technique establishes a new orthogonal-basis space where each axis represents a principal component, formed as a linear combination of the original data variables. By rigorously calculating these principal components, PCA ensures no redundancy of information within this new space [46]. Maximizing variance along each axis, PCA aligns the first axis with the highest variance of the data points, while the subsequent axes are orthogonal to the previous ones, sequentially maximizing the remaining variance [46]. Hence, in the transformed space, principal components are arranged in descending order of variance, with the first component explaining the most variance and subsequent components explaining progressively less [47,49].

The mathematical formulations required to compute the principal components are given hereafter. Let $x(t)$ for $t = 1, 2, \ldots, n$ be an arbitrary dataset containing its corresponding instances and features with zero mean. Its covariance matrix R is computed as follows:

$$R = \frac{1}{n-1} \sum_{t=1}^{n} \left[x(t) x(t)^T \right] \tag{1}$$

The next equation can be utilized to compute linear combinations of variables in the original data, i.e., the linear transformation from $x(t)$ to $y(t)$,

$$y(t) = M^T x(t) \tag{2}$$

where M denotes an orthogonal matrix of the size $n \times n$, and the ith column of this matrix, also of the sample covariance matrix R, is essentially equal to the ith eigenvector. At this point, the eigenvalue problem is initially set to be solved by the following equation:

$$\lambda_l q_l = R q_l \tag{3}$$

where q_l represents the corresponding eigenvector, and λ_l stands for an eigenvalue of the covariance matrix R (consider $\lambda_1 > \lambda_2 > \cdots > \lambda_n$). Based on Equation (4), the principal component is computed by

$$y_l(t) = q_l^T x(t), \quad l = 1, \ldots, n \tag{4}$$

where $y_l(t)$ stands for the ith principle component. For additional information and further details, readers can refer to the references [47,48].

For the Z-Alizadeh Sani dataset with selected attributes, Figure 4 illustrates the variance values explained by each principal component generated and depicts only the first eight components, which account for around 95% of the total variance.

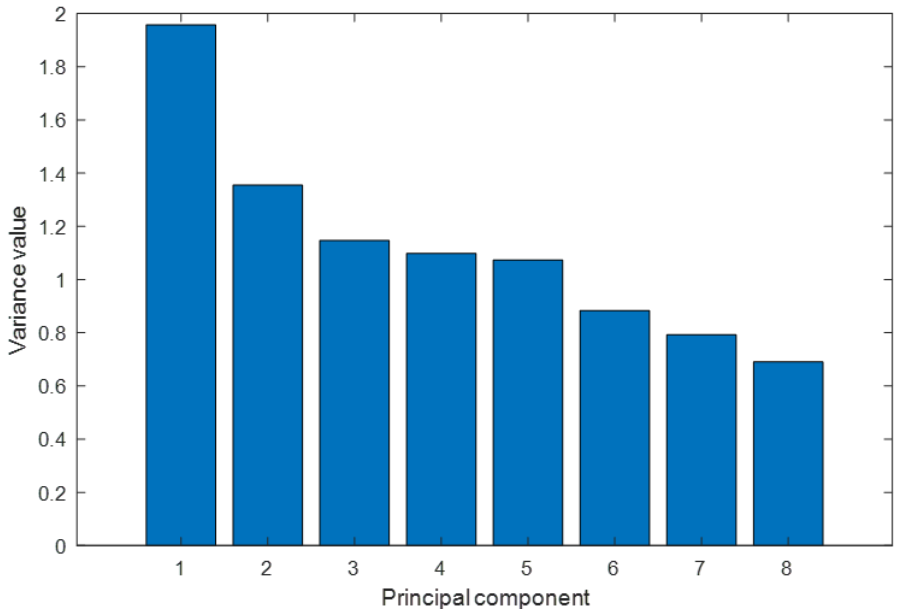

Figure 4. Variance values explained by each principal component.

2.2.3. Data Dividing

A methodology known as k-fold cross-validation can be used to reduce the bias related to a random sampling of the holdout and training data samples when comparing the predicted accuracy of two or more methods. The entire data set is randomly separated into k mutually exclusive subsets of approximately similar size in k-fold cross-validation, also known as rotation estimation. The classification technique is trained and tested k times. k-1 of mutually exclusive subsets are utilized for training, while the remaining one is reserved for testing. With averaging the k individual accuracy measures, the prediction of a technique's overall accuracy is computed as

$$CV\ accuracy = \frac{1}{k}\sum_{i=1}^{k} A_i \quad (5)$$

where A represents the accuracy measure of each fold such as specificity, sensitivity, and hit rate, and k denotes the number of used folds [50,51].

Since it is the most widespread practice for k to have a value of 10, the k-fold cross-validation is also known as 10-fold cross-validation. Empirical studies have shown that the optimal number of folds seems to be 10 [50,51]. For this reason, in this study, 10-fold cross-validation was utilized for evaluating the proposed diagnosis method. Figure 5 shows a visualization of k-fold cross-validation with $k = 10$ [50,51].

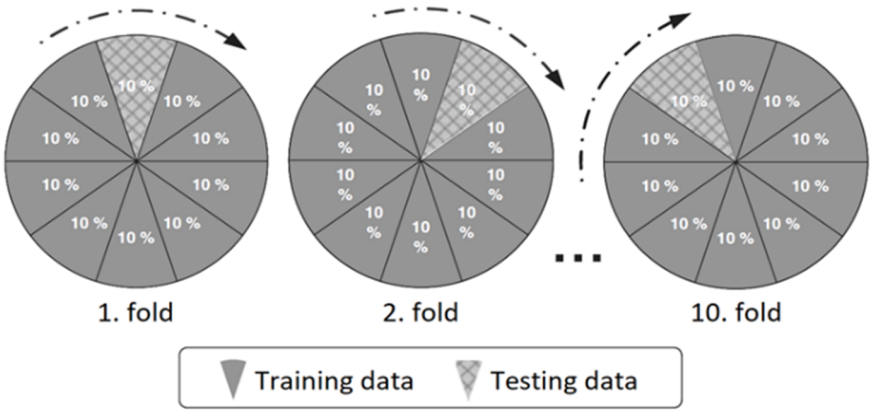

Figure 5. Visualization of 10-fold cross-validation.

2.2.4. Classification

In the classification step of the proposed method, the AdaBoostM1 algorithm [52] is utilized to classify coronary artery disease as patient or normal. The following is a description of the AdaBoostM1 algorithm. $T_n = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ is a training set with Y values in $1, 2, \ldots, k$. Each observation X_i is given a weight $w_b(i)$, which is originally set to $1/n$. After each step, this value is updated. The classifier's error is denoted by ϵ_b and is calculated as follows:

$$\epsilon_b = \sum_{i=1}^{n} [w_b(i)\xi_b(i)] \tag{6}$$

where

$$\xi_b = \begin{cases} 0, & C_b(x_i) = y_i \\ 1, & C_b(x_i) \neq y_i \end{cases} \tag{7}$$

The constant α_b is calculated from the classifier's error in the bth iteration, and this value is utilized for the weight update. Particularly, $\alpha_b = 1/2ln((1-\epsilon_b)/\epsilon_b)$, and for the $b+1$th iteration, the new weight is

$$w_{b+1}(i) = w_b(i) \exp\{\alpha_b \xi_b(i)\} \tag{8}$$

The obtained weights are then normalized to the sum of one. As a result, the weight of incorrectly categorized observations increases while the weight of correctly classified observations reduces, driving the single classifier produced in the next iteration to focus on the most difficult examples. Furthermore, while the single classifier's error is low, differences in weight updates are bigger, since when the classifier gets a high accuracy, the few mistakes become more important. Thus, the alpha constant can be thought of as a learning rate computed as a function of each iteration's mistake. Additionally, this constant is employed in the final decision rule, which gives more weight to the individual classifiers with the lowest error. This process is repeated in each step for $b = 1, 2, 3, \ldots, B$. Finally, the ensemble classifier calculates the weighted sum of each class's votes. As a result, the class with the highest weighted vote receives the assignment. In particular [52,53],

$$C(x) = \arg_{y_j} \max \sum_{b=1}^{B} \left[\alpha_b \delta\left(C_b(x), y_j\right)\right] \tag{9}$$

$$= \arg_{y_j} \max \sum_{b:C_b(x)=y_j} \alpha_b \tag{10}$$

3. Experimental Results and Discussions

3.1. Performance Metrics

In this study, to measure the proposed CAD diagnosis method's effectiveness, various basic metrics, which are accuracy, precision, recall, F_1, AUC, and MCC, are employed, and these metrics are computed as follows [54–57]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{14}$$

$$AUC = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (15)$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)\cdot(TP+FN)\cdot(TN+FP)\cdot(TN+FN)}} \quad (16)$$

where the types of possible outcomes are TP (true positives—correctly labeled as positive tuples), TN (true negatives—correctly labeled as negative tuples), FP (false positives—negative tuples incorrectly labeled as positive), and FN (false negatives—positive tuples mislabeled as negative) in binary estimation [54,58]. The confusion matrix, a summary of the possible outcomes, is demonstrated in Figure 6.

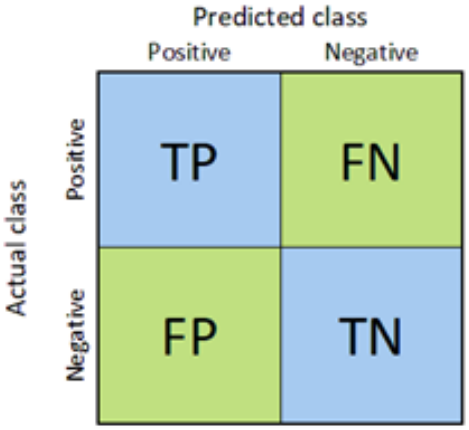

Figure 6. Confusion Matrix.

3.2. Experiments on the Feature Extraction

The initial experiment focused on the feature extraction method employed in the proposed approach. Two distinct techniques were utilized: exclusive feature selection and PCA. The performance metrics results are presented in Table 3, utilizing various data-division methodologies, including an 80% training–20% test split, and 5-fold and 10-fold cross-validation. Notably, employing feature selection alone yielded a classification accuracy of 90.164% in the 80% training–20% test data division. Furthermore, precision, recall, F-measure, and AUC metrics exceeded 0.9, with a Matthews correlation coefficient (MCC) rate of 0.755. Subsequently, integrating PCA with the selected features led to improved accuracy, precision, recall, F-measure, and MCC metrics, resulting in an accuracy rate of 91.803%, as demonstrated in Table 3.

Table 3. Performance Metric Results of Various Feature Extraction Techniques with AdaBoostM1.

Feature Extraction Technique		80/20% Train/Test Split	5-Fold CV	10-Fold CV
Feature selection only	%Acc.	90.164	86.799	86.469
	Precision	0.932	0.893	0.900
	Recall	0.932	0.926	0.912
	F1	0.932	0.909	0.906
	AUC	0.929	0.909	0.907
	MCC	0.755	0.670	0.666
PCA	%Acc.	91.803	88.119	89.109
	Precision	0.933	0.913	0.914
	Recall	0.955	0.921	0.935
	F1	0.944	0.917	0.924
	AUC	0.895	0.888	0.879
	MCC	0.793	0.708	0.730

In the data-dividing methodology of 5-fold cross-validation, an accuracy of 86.799% was attained, and results in the range of 0.893 to 0.926 were obtained for precision, recall,

F-measure, and AUC. When the PCA technique was utilized for feature extraction, some rise was observed in accuracy, precision, recall, F-measure, and MCC metrics. Finally, in 10-fold cross-validation, the use of the PCA technique increased by approximately three percent in the accuracy metric. Moreover, the results of precision, recall, F-measure and M,CC metrics rose. Considering all the results, the best classification accuracy rate of 91.803%, precision rate of 0.933, recall rate of 0.955, F-measure rate of 0.944, and MCC rate of 0.793 were achieved in 80% training–20% test splitting methodology when PCA was used. Additionally, the best AUC rate of 0.929 was achieved with the feature extraction technique of feature selection only.

In addition, the confusion matrices for each feature extraction technique and data-dividing methodologies are given in Figure 7. The confusion matrices acquired for the feature selection only are depicted in Figure 7a–c while Figure 7d–f demonstrate the confusion matrices attained for the PCA technique with the data-dividing methodologies of 80% training–20% test and 5-fold and 10-fold cross-validation, respectively.

Figure 7. Confusion matrices for each feature extraction technique and data-dividing methodology. Figures in (**a**–**c**) refer to "Feature-selection-only" results for 80%-Training-20%Test, 5-fold-CV and 10-fold-CV, respectively. Similarly, (**d**–**f**) refer to the same results respectively but this time for the "PCA" method.

3.3. Comparison with Traditional Methods

This subsection compares the classification results of the proposed method with basic classifiers. On the Z-Alizadeh Sani dataset, several basic techniques were tested in the 10-fold cross-validation. The aforementioned two extraction techniques of feature selection only and PCA were utilized with each basic classifier, and their performance results are shown with regard to the previously mentioned six metrics in Table 4. Along with the proposed approach, this table contains the results of eleven basic classifiers such as Naïve Bayes [59], k-NN (k = 5) [60], C4.5 decision tree [61], locally weighted learning (LWL) [62], K* [63], logistic model trees (LMT) [64], SVM [65], random forests [66], logistic regression [67], Hoeffding tree [68], and deep learning 4J [69]. As can be seen from the table, using PCA to extract the features increased the classification accuracy performance of k-NN, C4.5 decision tree, LWL, SVM, deep learning 4J and the proposed method. LMT and logistic regression with an accuracy of 88.449% are the best classifiers for feature selection only, whereas the proposed method with PCA achieves the best accuracy rate of 89.109%, a recall rate of 0.935, and an F-measure rate of 0.924, surpassing the other techniques.

Table 4. Classification Performance Results Of Basic Classifiers. Bold-faced results refer to the best performing results for each metric.

Feature Extraction Technique		Naïve Bayes [59]	k-NN [60]	C4.5 DT [61]	LWL [62]	K* [63]	LMT [64]	SVM [65]	RF [66]	Log Reg [67]	Hoeff. Tree [68]	DL 4J [69]	Ours
Feature selection only	%Acc.	88.120	85.480	85.150	87.130	83.500	88.450	87.790	81.520	88.450	87.790	85.480	86.470
	Preci.	0.905	0.910	0.890	0.900	0.877	0.910	0.916	0.864	0.910	0.905	0.939	0.900
	Recall	0.931	0.884	0.903	0.921	0.894	0.931	0.912	0.880	0.931	0.926	0.852	0.912
	F_1	0.918	0.897	0.897	0.911	0.885	0.920	0.914	0.872	0.920	0.915	0.893	0.906
	AUC	0.923	0.894	0.830	0.907	0.901	0.919	0.853	0.881	0.922	0.923	0.922	0.907
	MCC	0.705	0.653	0.634	0.681	0.592	0.714	0.703	0.543	0.714	0.697	0.676	0.666
PCA	%Acc.	80.530	86.470	87.790	87.790	81.850	86.800	88.120	81.190	88.450	80.200	86.800	**89.110**
	Preci.	0.920	0.900	0.912	0.905	0.871	0.889	0.917	0.863	0.910	0.919	**0.944**	0.914
	Recall	0.796	0.912	0.917	0.926	0.875	0.931	0.917	0.875	0.931	0.792	0.866	**0.935**
	F1	0.854	0.906	0.915	0.915	0.873	0.910	0.917	0.869	0.920	0.851	0.903	**0.924**
	AUC	0.892	0.878	0.846	0.697	0.858	0.918	0.855	0.874	**0.922**	0.892	0.921	0.879
	MCC	0.581	0.666	0.701	**0.885**	0.555	0.668	0.710	0.536	0.714	0.575	0.703	0.730

3.4. Comparison with Existing Methods in the Literature

In this subsection, the proposed method was compared with the existing studies in the literature using the same dataset, the Z-Alizadeh Sani dataset. The performance comparison of the proposed method with the existing works is presented in Table 5, containing researcher names, years, used method, number of selected features, and accuracy metrics. Between 2012 and 2016, Alizadehsani et al. [19–25] used different numbers of features, such as 16, 20, 24, and 34, and achieved the best accuracy of 94.08% utilizing information gain and SMO. In 2017, Qin et al. [26] applied their CAD detection approach based on EA-MFS and SVM with 34 features and procured an accuracy rate of 93.70%. In the same year, Arabasadi et al. [27] proposed a genetic algorithm and MLP-ANN-based CAD prediction method selecting 22 features, while Babič et al. [28] performed various classifiers such as decision trees, Naïve Bayes, SVM, and ANN and used 27 features to feed these classifiers.

In 2018, Kılıç and Kaya Keleş [29] selected 16 features using the ABC algorithm and then classified CAD utilizing the SMO technique. As a result of their study, an accuracy rate of 89.44% was obtained. In 2019, MML-IBLMM and Var-IBLMM methods introduced by Hu et al. [30] were applied to the Z-Alizadeh Sani dataset and attained an accuracy rate of 81.84%. In the same year, Abdar et al. [31] proposed the N2Genetic-nuSVM approach, selected 29 features, and acquired an accuracy rate of 93.08%. In another work performed by Abdar et al. [32], a CAD diagnosis approach called NE-nu-SVC was presented. In this approach, 16 features were selected and an accuracy of 94.66% was achieved.

In 2020, Joloudari et al. [33] tested the classification performance of various classifiers, selected 40 features, and obtained the best accuracy rate of 91.47% with random trees. In the same year, a hybrid feature selection method called 2HFS was introduced by Nasar-

ian et al. [34], and 38 features were selected using this method. In the sequel, SMOTE and XGBoost techniques were used together and an accuracy rate of 92.58% was reported. In another study presented by Ashish et al. [35], a random forests-, SVM-, and XGBoost-based CAD detection approach was implemented and an accuracy rate of 93.86% was achieved with 10 features. In a recent work [36], 25 features were used, and an accuracy rate of 91.78% with the MLP classifier was obtained.

Unlike these studies, the proposed method in this work utilizes five features, namely, age, hypertension, typical chest pain, t-wave inversion, and region with regional wall motion abnormality. In the dataset with these features, PCA and AdaBoostM1 techniques were used for feature extraction and classification, respectively. The best accuracy of 91.80% was achieved when using these few features on the Z-Alizadeh Sani dataset.

Table 5. Performance Comparison of The Proposed Method with The Existing Studies using the Z-Alizadeh Sani Dataset.

Paper	Year	Method	# of Features	Accuracy (%)
[22]	2012	SMO	16	82.16
[25]	2012	Naïve Bayes–SMO	16	88.52
[21]	2012	SMO	34	92.09
[24]	2012	SMO 1-1	34	92.74
[19]	2013	Information gain + SMO	34	94.08
[20]	2013	Bagging + C4.5	20	79.54 (LAD) 61.46 (LCX) 68.96 (RCA)
[23]	2016	Average and combined information gain + SVM	24	86.14 (LAD) 83.17 (LCX) 83.50 (RCA)
[26]	2017	EA-MFS + SVM	34	93.7
[27]	2017	GA + MLP-ANN	22	93.85
[28]	2017	SVM	27	86.67
[29]	2018	ABC + SMO	16	89.44
[30]	2019	MML-IBLMM and Var-IBLMM	55	81.84
[31]	2019	N2Genetic-nuSVM	29	93.08
[32]	2019	NE-nu-SVC	16	94.66
[33]	2020	Random trees	40	91.47
[34]	2020	2HFS + SMOTE + XGBoost	38	92.58
[35]	2021	Random forests + SVM + XGBoost	10	93.86
[36]	2023	MLP	25	91.78
Proposed		**PCA + AdaBoostM1**	**5**	**91.8**

3.5. Limitations

This work can be considered a retrospective study because it uses a dataset based on past patient records. Researchers conduct this type of study by examining the existing records, historical data, or previous occurrences in order to determine outcomes, relationships, or correlations between variables. In contrast to prospective studies, which follow participants ahead of time, retrospective studies begin with the desired outcome or endpoint and go backwards to investigate the reasons or events that led to it. For example, in a prospective study introduced by Locuratolo et al. [70], patients were evaluated clinically and in the laboratory after 30 days, 3 months, 6 months, and 1 year following the index incident. Various endpoints related to acute coronary syndrome were evaluated. At the

end of the study, the persistence of treatments and the percentage of patients who achieved therapeutic goals were evaluated.

Retrospective studies have some limitations, such as data quality, limited scope, bias, temporal ambiguity, confounding variables, validity of exposure measurement, and causality inference. In spite of these limitations, retrospective studies remain useful in epidemiological research, particularly when prospective investigations are unfeasible or unethical. The method proposed in this study can help medical doctors diagnose CAD early by using a small number of features.

4. Conclusions

This paper introduces an effective approach for diagnosing coronary artery disease (CAD) by leveraging age, hypertension, typical chest pain, T-wave inversion, and regional wall motion abnormality features. The method proposed utilizes eight distinct search techniques, including evolutionary, best first, genetic, harmony, PSO, greedy stepwise, rank, and multi-objective evolutionary search, to perform feature selection on the Z-Alizadeh Sani dataset. Principal component analysis (PCA) and the AdaBoostM1 algorithm are employed for feature extraction and CAD classification, respectively. Through extensive experiments and analyses using various performance metrics, the proposed method achieves the highest prediction performance to date with only five attributes. Notably, it achieves impressive accuracy, precision, recall, F-measure, AUC, and MCC rates of 91.8%, 93.3%, 95.5%, 94.4%, 89.5%, and 79.3%, respectively. These results demonstrate the efficiency of the proposed approach and its potential to serve as a cost-effective tool to aid medical practitioners in CAD diagnosis.

Author Contributions: Conceptualization, C.E. and O.K.; methodology, C.E.; software, C.E.; writing—original draft preparation, C.E. and O.K.; writing—review and editing, C.E. and O.K.; visualization, C.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was funded by Cardiff University Institutional Funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This paper utilised the Z-Alizadeh Sani dataset that is freely available from the University of California—Irvine Machine Learning Repository [18].

Conflicts of Interest: The authors declare that they have no conflicts of interest. This research has not received any specific grant from public funding agencies or commercial or not-for-profit sectors.

References

1. World Health Organization. World Health Organization. Available online: https://www.who.int/ (accessed on 1 December 2023).
2. International Diabetes Federation. Diabetes and Cardiovascular Disease. 2016. Available online: https://idf.org/our-activities/care-prevention/cardiovascular-disease.html (accessed on 1 December 2023).
3. World Health Organization. Cardiovascular Diseases (CVDs). 2021. Available online: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed on 1 December 2023).
4. World Health Organization. The Top 10 Causes of Death. 2020. Available online: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (accessed on 1 December 2023).
5. Eyupoglu, C. Breast cancer classification using k-nearest neighbors algorithm. *Online J. Sci. Technol.* **2018**, *8*, 29–34.
6. Janosi, A.; Steinbrunn, W.; Pfisterer, M.; Detrano, R. Heart Disease Data Set, UCI Machine Learning Repository. Available online: https://archive.ics.uci.edu/ml/datasets/Heart+Disease (accessed on 1 December 2023).
7. Akgül, M.; Sönmez, Ö.E.; Özcan, T. Diagnosis of heart disease using an intelligent method: A hybrid ANN–GA approach. In Proceedings of the International Conference on Intelligent and Fuzzy Systems, Istanbul, Turkey, 21–23 July 2019; pp. 1250–1257.
8. Rajab, W.; Rajab, S.; Sharma, V. Kernel FCM-based ANFIS approach to heart disease prediction. In Proceedings of the Emerging Trends in Expert Applications and Security, Jaipur, India, 17–19 February 2019; pp. 643–650.
9. Uyar, K.; İlhan, A. Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. *Procedia Comput. Sci.* **2017**, *120*, 588–593. [CrossRef]

10. 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]
11. Ali, L.; Niamat, A.; Khan, J.A.; Golilarz, N.A.; Xingzhong, X.; Noor, A.; Bukhari, S.A.C. An optimized stacked support vector machines based expert system for the effective prediction of heart failure. *IEEE Access* **2019**, *7*, 54007–54014. [CrossRef]
12. Burse, K.; Kirar, V.P.S.; Burse, A.; Burse, R. Various preprocessing methods for neural network based heart disease prediction. In *Proceedings of the Smart Innovations in Communication and Computational Sciences*; Springer: Singapore, 2019; pp. 55–65.
13. Paul, A.K.; Shill, P.C.; Rabin, M.; Islam, R.; Murase, K. Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease. *Appl. Intell.* **2018**, *48*, 1739–1756. [CrossRef]
14. 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]
15. Terrada, O.; Cherradi, B.; Raihani, A.; Bouattane, O. Classification and Prediction of atherosclerosis diseases using machine learning algorithms. In Proceedings of the 2019 5th International Conference on Optimization and Applications (ICOA), Kenitra, Morocco, 25–26 April 2019.
16. Gokulnath, C.B.; Shantharajah, S.P. An optimized feature selection based on genetic approach and support vector machine for heart disease. *Clust. Comput.* **2019**, *22*, 14777–14787. [CrossRef]
17. Karayılan, T.; Kılıç, Ö. Prediction of heart disease using neural network. In Proceedings of the 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, 5–7 October 2017; pp. 719–723.
18. Alizadeh Sani, Z.; Alizadehsani, R.; Roshanzamir, M. Z-Alizadeh Sani Data Set, UCI Machine Learning Repository. 2017. Available online: https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani (accessed on 1 December 2023).
19. 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] [PubMed]
20. Alizadehsani, R.; Habibi, J.; Sani, Z.A.; Mashayekhi, H.; Boghrati, R.; Ghandeharioun, A.; Alizadeh-Sani, F. Diagnosing coronary artery disease via data mining algorithms by considering laboratory and echocardiography features. *Res. Cardiovasc. Med.* **2013**, *2*, 133–139.
21. Alizadehsani, R.; Hosseini, M.J.; Sani, Z.A.; Ghandeharioun, A.; Boghrati, R. Diagnosis of coronary artery disease using cost-sensitive algorithms. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining Workshops, Brussels, Belgium, 10–12 December 2012; pp. 9–16.
22. Alizadehsani, R.; Habibi, J.; Sani, Z.A.; Mashayekhi, H.; Boghrati, R.; Ghandeharioun, A.; Bahadorian, B. Diagnosis of coronary artery disease using data mining based on lab data and echo features. *J. Med. Bioeng.* **2012**, *1*, 26–29. [CrossRef]
23. 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]
24. Alizadehsani, R.; Hosseini, M.J.; Boghrati, R.; Ghandeharioun, A.; Khozeimeh, F.; Sani, Z.A. Exerting cost-sensitive and feature creation algorithms for coronary artery disease diagnosis. *Int. J. Knowl. Discov. Bioinform. (IJKDB)* **2012**, *3*, 59–79. [CrossRef]
25. 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.
26. Qin, C.J.; Guan, Q.; Wang, X.P. Application of ensemble algorithm integrating multiple criteria feature selection in coronary heart disease detection. *Biomed. Eng. Appl. Basis Commun.* **2017**, *29*, 1750043. [CrossRef]
27. 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]
28. Babič, F.; Olejár, J.; Vantová, Z.; Paralič, J. Predictive and descriptive analysis for heart disease diagnosis. In Proceedings of the 2017 Federated Conference on Computer Science and Information Systems (fedCSIS), Prague, Czech Republic, 3–6 September 2017; pp. 155–163.
29. Kılıc, Ü.; Kaya Keleş, M. Feature selection with artificial bee colony algorithm on Z-Alizadeh Sani dataset. In Proceedings of the 2018 Innovations in Intelligent Systems and Applications Conference (ASYU), Adana, Turkey, 4–6 October 2018; pp. 1–3.
30. Hu, C.; Fan, W.; Du, J.X.; Bouguila, N. A novel statistical approach for clustering positive data based on finite inverted Beta-Liouville mixture models. *Neurocomputing* **2019**, *333*, 110–123. [CrossRef]
31. 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] [PubMed]
32. Abdar, M.; Acharya, U.R.; Sarrafzadegan, N.; Makarenkov, V. NE-nu-SVC: A new nested ensemble clinical decision support system for effective diagnosis of coronary artery disease. *IEEE Access* **2019**, *7*, 167605–167620. [CrossRef]
33. Joloudari, J.H.; Hassannataj Joloudari, E.; Saadatfar, H.; Ghasemigol, M.; Razavi, S.M.; Mosavi, A.; Nadai, L. Coronary artery disease diagnosis; ranking the significant features using a random trees model. *Int. J. Environ. Res. Public Health* **2020**, *17*, 731. [CrossRef]
34. Nasarian, E.; Abdar, M.; Fahami, M.A.; Alizadehsani, R.; Hussain, S.; Basiri, M.E.; Sarrafzadegan, N. Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach. *Pattern Recognit. Lett.* **2020**, *133*, 33–40. [CrossRef]

35. Ashish, L.; Kumar, S.; Yeligeti, S. Ischemic heart disease detection using support vector machine and extreme gradient boosting method. *Mater. Today Proc.* 2021, *in press*. [CrossRef]
36. Kolukisa, B.; Bakir-Gungor, B. Ensemble feature selection and classification methods for machine learning-based coronary artery disease diagnosis. *Comput. Stand. Interfaces* **2023**, *84*, 103706. [CrossRef]
37. Hall, M.A. Correlation-Based Feature Subset Selection for Machine Learning. Ph.D. Thesis, University of Waikato, Hamilton, New Zealand, 1998.
38. Vikhar, P.A. Evolutionary algorithms: A critical review and its future prospects. In Proceedings of the 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), Jalgaon, India, 22–24 December 2016; pp. 261–265.
39. Pearl, J. *Heuristics: Intelligent Search Strategies for Computer Problem Solving*; Addison-Wesley Longman Publishing Company: Boston, MA, USA, 1984.
40. Goldberg, D.E. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley Longman Publishing Company: Boston, MA, USA, 1989.
41. Fong, S.; Biuk-Aghai, R.P.; Millham, R.C. Swarm search methods in weka for data mining. In Proceedings of the 2018 10th International Conference on Machine Learning and Computing, Macau, China, 26–28 February 2018; pp. 122–127.
42. Moraglio, A.; Chio, C.D.; Poli, R. Geometric particle swarm optimisation. In Proceedings of the European Conference on Genetic Programming, Valencia, Spain, 11–13 April 2007; pp. 125–136.
43. Butterworth, R.; Simovici, D.A.; Santos, G.S.; Ohno-Machado, L. A greedy algorithm for supervised discretization. *J. Biomed. Inform.* **2004**, *37*, 285–292. [CrossRef] [PubMed]
44. Hall, M.A.; Holmes, G. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Trans. Knowl. Data Eng.* **2003**, *15*, 1437–1447. [CrossRef]
45. Jiménez, F.; Sánchez, G.; García, J.M.; Sciavicco, G.; Miralles, L. Multi-objective evolutionary feature selection for online sales forecasting. *Neurocomputing* **2017**, *234*, 75–92. [CrossRef]
46. Statistics and Machine Learning Toolbox. 2018. Available online: https://www.mathworks.com/products/statistics.html (accessed on 1 December 2023).
47. Salo, F.; Nassif, A.B.; Essex, A. Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection. *Comput. Netw.* **2019**, *148*, 164–175. [CrossRef]
48. Jackson, J.E. *A User's Guide to Principal Components*; John Wiley & Sons: Hoboken, NJ, USA, 2005; Volume 587.
49. Yavuz, E.; Eyupoglu, C. An effective approach for breast cancer diagnosis based on routine blood analysis features. *Med. Biol. Eng. Comput.* **2020**, *58*, 1583–1601. [CrossRef] [PubMed]
50. Olson, D.L.; Delen, D. *Advanced Data Mining Techniques*; Springer Science & Business Media: New York, NY, USA, 2008.
51. Eyüpoğlu, C. Büyük Veride Etkin Gizlilik Koruması Için Yazılım Tasarımı /Software Design for Efficient Privacy Preserving in Big Data. Ph.D. Thesis, İstanbul University, Istanbul, Turkey, 2018.
52. Freund, Y.; Schapire, R.E. Experiments with a new boosting algorithm. In Proceedings of the 13th International Conference on Machine Learning, Bari Italy, 3–6 July 1996; pp. 148–156.
53. Cortes, E.A.; Martinez, M.G.; Rubio, N.G. Multiclass corporate failure prediction by Adaboost. M1. *Int. Adv. Econ. Res.* **2007**, *13*, 301–312. [CrossRef]
54. Eyupoglu, C.; Aydin, M.A.; Zaim, A.H.; Sertbas, A. An efficient big data anonymization algorithm based on chaos and perturbation techniques. *Entropy* **2018**, *20*, 373. [CrossRef]
55. Eyüpoğlu, C. Kronik Böbrek Hastalığının Erken Tanısı için Yeni Bir Klinik Karar Destek Sistemi. *Avrupa Bilim Teknol. Derg.* **2020**, *20*, 448–455. [CrossRef]
56. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]
57. Eyüpoğlu, C. Korelasyon Temelli Özellik Seçimi, Genetik Arama ve Rastgele Ormanlar Tekniklerine Dayanan Yeni Bir Rahim Ağzı Kanseri Teşhis Yöntemi. *Avrupa Bilim Teknol. Derg.* **2020**, *19*, 263–271. [CrossRef]
58. Han, J.; Kamber, M.; Pei, J. *Data Mining Concepts and Techniques*, 3rd ed.; Elsevier, Morgan Kaufmann Publishers: San Francisco, CA, USA, 2012.
59. John, G.H.; Langley, P. Estimating continuous distributions in Bayesian classifiers. In Proceedings of the Eleventh conference on Uncertainty in Artificial Intelligence, Montreal, Canada, 18–20 August 1995; pp. 338–345.
60. Aha, D.W.; Kibler, D.; Albert, M. Instance-based learning algorithms. *Mach. Learn.* **1991**, *6*, 37–66. [CrossRef]
61. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers: San Mateo, CA, USA, 1993.
62. Frank, E.; Hall, M.; Pfahringer, B. Locally Weighted Naive Bayes. In Proceedings of the 19th Conference in Uncertainty in Artificial Intelligence, Acapulco, Mexico, 7–10 August 2003; pp. 249–256.
63. Cleary, J.G.; Trigg, L.E. K*: An instance-based learner using an entropic distance measure. In Proceedings of the 12th International Conference on Machine Learning, Tahoe City, CA, USA, 9–12 July 1995; pp. 108–114.
64. Landwehr, N.; Hall, M.; Frank, E. Logistic model trees. *Mach. Learn.* **2005**, *59*, 161–205. [CrossRef]
65. Keerthi, S.S.; Shevade, S.K.; Bhattacharyya, C.; Murthy, K.R. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Comput.* **2001**, *13*, 637–649. [CrossRef]
66. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

67. Le Cessie, S.; Van Houwelingen, H.C. Ridge estimators in logistic regression. *J. R. Stat. Soc. Ser. Appl. Stat.* **1992**, *41*, 191–201. [CrossRef]
68. Hulten, G.; Spencer, L.; Domingos, P. Mining time-changing data streams. In Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 26–29 August 2001; pp. 97–106.
69. Deeplearning4j. Deep Learning for Java. Available online: https://deeplearning4j.org/ (accessed on 1 December 2023).
70. Locuratolo, N.; Scicchitano, P.; Antoncecchi, E.; Basso, P.; Bonfantino, V.M.; Brescia, F.; Carrata, F.; De Martino, G.; Landriscina, R.; Lanzone, S.; et al. Follow-up of patients after an acute coronary event: The Apulia PONTE-SCA program. *G. Ital. Cardiol. (2006)* **2022**, *23*, 63–74.