Prediction of Heart Disease using an Ensemble Learning Approach

Ghalia A. Alshehri, Hajar M. Alharbi
Department of Computer Science
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

***Abstract*—The ability to predict diseases early is essential for improving healthcare quality and can assist patients in avoiding potentially dangerous health conditions before it is too late. Various machine learning techniques are used in the medical field. Nonetheless, machine learning is critical in determining the future of pharmaceuticals and patients' health. This is because the various classification techniques provide a high level of accuracy. However, because so much data are being gathered from patients, it becomes harder to find meaningful cardiac disease predictions. A vital research task is to identify these characteristics. Individual classification algorithms in this situation cannot generate flawless models capable of reliably predicting heart disease. As a result, higher performance might be achieved by using ensemble learning approaches (ELA), producing accurate cardiac disease predictions. In the present research work, we utilized an ELA for the early prediction of heart disease, using a new combination including four machine learning algorithms—adaptive boosting, support vector machine, decision tree, and random forest—to increase the accuracy of the prediction results. We used two wrapper methods for feature selection: forward selection and backward elimination. We used the proposed model with three datasets: the StatLog UCI dataset, the Z-Alizadeh Sani dataset, and the Cardiovascular Disease (CVD) dataset. We obtained the highest accuracy when using our proposed model with the Z-Alizadeh Sani dataset, where it was 0.91, while the StatLog UCI dataset was 0.83. The CVD dataset obtained the lowest accuracy, 0.73.**

Keywords*—Machine learning; ensemble learning; classification; disease prediction; heart disease

I. Introduction

Heart disease is a devastating illness that kills more people worldwide than other diseases. According to the annual statistical books of the Ministry of Health and the World Health Organization, heart disease caused 42% of deaths from non-communicable diseases in the Kingdom of Saudi Arabia (KSA) in 2010 [1]. Mortality from heart disease can be reduced if an accurate diagnosis is made early on. Modern medical science has demonstrated significant and effective ways of dealing with heart-related issues. Moreover, medical difficulties can now be addressed using artificial intelligence. Electrocardiogram (ECG), angiography screening, and blood tests are the most popular methods for detecting heart disease [2]. High cholesterol, blood pressure, and hypertension can all increase the risk of heart disease, but such signs may go unnoticed by the average person. Chest pain, breathlessness, and heart palpitations are frequent symptoms of heart disease. Angina, also known as angina pectoris, is a form of cardiac disease wherein the heart receives insufficient oxygen. Breathlessness can occur due to heart failure when the heart becomes too weak to circulate blood. Some cardiac problems have no symptoms, particularly in the elderly and those with diabetes. When considering these factors, the healthcare industry must keep additional information about patients and their medications to generate diagnostic reports.

The advancement of computing and storage technology has allowed the healthcare industry to collect and retain routine medical data, allowing for more consistent and reliable support in medical choices. Patients' data are collected and maintained digitally in many developed countries. The information is then analyzed to make the required medical judgments regarding prediction, diagnosis, and treatment options [3]. Machine learning (ML) methods have been quite helpful in solving complicated classification and prediction problems [4]. One ML technique that can be used to predict future outcomes is classification. ML is crucial in recognizing cardiac illness from extensive data. ML aids in the decision-making process based on historical data. Classification, usually called supervised ML, predicts future events based on historical data. Medical ML employs techniques such as classification to generate insights and provide medical outcomes depending on the data [5]. In its most basic form, ML uses preprogrammed algorithms that learn and improve their operations by analyzing input data and making reasonable predictions. These algorithms tend to produce more accurate predictions as additional data are fed. Despite variations in classification, ML algorithms can be divided into three groups based on their objectives and how the underlying machine is trained: supervised, unsupervised, and semi-supervised [6]. A labeled training dataset trains the underlying algorithm in supervised ML techniques. The unlabeled test dataset is then assigned to the trained algorithm, categorizing it into similar categories [7]. It is feasible to gain insight into a patient's medical history and to provide clinical support through such an analysis. The risk that a person will develop heart disease can be predicted by training and testing classification algorithms. However, because the medical problem is so severe, the remedy necessitates greater classification accuracy, which is not provided by traditional classification algorithms.

Ensemble methods could be employed in this situation. More specifically, ensemble classification algorithms that integrate two or more classification techniques and generate the best prediction results are used to identify cardiac disease. Several ML techniques are used to treat heart illnesses due to their superior performance and capacity to comprehend

the relationships between features' input and output variables compared to experienced physicians or doctors. The values of different tests performed on a person are typically used as input features. Many classifications, clustering, and deep learning algorithms have been implemented by researchers worldwide. Despite this, given the rise in heart disease rates each year, newer ML methods should be implemented with significant features to improve the results of existing classification algorithms.

This research used an ELA containing four ML algorithms—adaptive boosting (AdaBoost), support vector machine (SVM), decision tree (DT), and random forest (RF)—to obtain the best results for predicting heart disease. We used two wrapper methods, forward selection, and backward elimination, for the feature selection step and analyzed unbalanced data.

II. Background

A. Literature Review

We chose studies in the same field from 2017 to 2022, summarized them in Table I, and arranged them from oldest to newest. Yekkala et al. [3] studied using particle swarm optimization (PSO) and an ensemble classifier to predict cardiac disease. PSO was used as a feature selection method to eliminate the least-rated features, while ensemble methods were used to lower the misclassification rate and increase classification performance. The experiments showed that applying the bagged tree ensemble classifier to the PSO can significantly enhance learning accuracy. Dinh et al. [8] developed supervised ML models to detect individuals with cardiovascular, prediabetes, and diabetes diseases using the NHANES dataset [9]. Multiple ML models (logistic regression, SVMs, RF, and gradient boosting) were assessed for their classification performance and integrated into a weighted ensemble model to increase detection accuracy. David [5] used the StatLog dataset from the University of California–Irvine (UCI) data repository [10] to compare three algorithms—AdaBoost, Bagging, and Stacking—to identify the top ensemble classification method for predicting heart disease. According to one study, AdaBoost has been experimentally demonstrated to offer ideal results compared to its competitors. Liu et al. [11] suggested a unique ensemble learning method for medical diagnosis using imbalanced data. Using data preprocessing, training-based classifiers, and a final ensemble, they presented the SMOTE-CVCF integrated filter technique, C-SVM, and V-SVM with five kernel functions, a weighted fusion approach, and a SAGA method to optimize the weight vector. According to the empirical findings, the suggested ensemble learning method could outperform other cutting-edge categorization models. By randomly partitioning the dataset into smaller categories and employing a classification and regression tree (CART), Mienye et al. [12] improved an ML technique for forecasting the risk of heart disease. A modified version of the weighted aging classifier ensemble was employed to ensure the best performance, and a modified version of the weighted aging classifier ensemble was used to create a homogenous ensemble from several CART models. A novel coronary heart disease detection technique based on ML, such as classifier ensembles, was proposed by Tama et al. [13]. As a result, a two-tier ensemble was built, with certain ensemble classifiers serving as the foundation for another ensemble. The model was evaluated using several heart disease datasets, and the proposed approach performed better than any base classifier in the ensemble. Yadav and Pa [14] proposed four algorithms for classifying data using trees and evaluated their accuracy, precision, and sensitivity. The M5P, random tree, reduced error pruning, and random forest ensemble approaches were employed in the first of the three experimental setups used for the analysis. The second experiment employed four tree-based techniques using recursive feature elimination, while Lasso regularization was used on top of the tree-based methods in the third trial. Predicting heart problems is just one of the many uses of this derivation process. Velusamy and Ramasamy [15] developed an ensemble algorithm with five features chosen based on feature importance, which was assessed using the Z-Alizadeh Sani dataset [16] and balanced using synthetic minority oversampling. When used on the balanced dataset, the weighted average voting (WAVEn) algorithm diagnosed coronary artery disease (CAD) with 100% accuracy, specificity, sensitivity, and precision. Tuncer et al. [17] proposed an ECG signal detection approach involving preprocessing, feature extraction, concatenation, selection, and classification. Fifteen sub-bands of ECG signals were generated during the preprocessing step. A maximum classification percentage of 96.60% was achieved for the MIT-BIH Arrhythmia dataset [18] using K-NN, and 97.80% accuracy was achieved using SVM for the St. Petersburg ECG dataset [19].

A summary of the literature review is shown in Table I. We abbreviated the labels of some metrics: Acc = accuracy, Sens = sensitivity, Spec = specificity, AUC = the area under the ROC curve, PPV = positive predictive value, NPV = negative predictive value, Prec = precision, MCC = Matthew's correlation coefficient, and Kappa = Cohen's kappa. As a reminder, a positive predictive value refers to precision, and recall refers to sensitivity. However, it is worth noting that the terminology may vary among different studies.

TABLE I: Summary of the Literature Review

Ref	Year	Method	Dataset	Best Result
[3]	2017	Ensemble methods (Bagged Tree, RF and AdaBoost) along with PSO	StatLog [20]	Bagged Tree Acc=100% Sens=100% Spec=100% PPV=100% NPV=100%
[8]	2019	A weighted ensemble model contains (Logistic Regression, SVM, RF, Gradient Boosting)	NHANES [9]	Prec=76% Recall=76% F1=76% AUC=83.9%

Continued on next page

TABLE I: Summary of the Literature Review (Continued)

[5]	2020	AdaBoost, Bagging and Stacking	StatLog [20]	AdaBoost Prec=81.2% Recall=80.6% F1=80.2%
[11]	2020	C-SVM and V-SVM with 5 kernel functions	UCI repository [21] + KEEL [22]	SPECTF heart dataset Prec=91.17% Recall=100% F1=95.38% AUC=95.98%
[12]	2020	Ensemble CART models	Cleveland [23] + Framingham [24]	Cleveland Acc=93% Framingham Acc=91%
[13]	2020	RF, Gradient Boosting, Extreme Gradient Boosting	Z-Alizadeh Sani [16] + StatLog [20], Cleveland [23], and Hungarian from UCI [10]	Z-Alizadeh Sani Acc=98.31% F1=96.60% AUC=98.70%
[14]	2020	M5P tree, Random tree and Error Reduced Pruning tree with RF Ensemble method	UCI repository [21]	Pearson Correlation feature selection on RF Acc=99.9% Sens=99.6% Spec=91.6%
[15]	2021	Heterogeneous ensemble method (K-NN, RF and SVM), with WAVEn	Z-Alizadeh Sani [16]	WAVEn method Acc=100% Kappa=100% Sens=100% Spec=100% Prec=100% F1=100% MCC=100%

Continued on next page

TABLE I: Summary of the Literature Review (Continued)

[17]	2022	LDA, K-NN, and SVM	MIT-BIH Arrhythmia [18] + St. Petersburg ECG [19]	MIT-BIH Arrhythmia with K-NN Acc=96.60% St.Petersburg ECG with SVM Acc=97.80%

B. Dataset

We used three datasets for this study: the StatLog UCI dataset [10] and [20], the Z-Alizadeh Sani dataset [16], and the CVD dataset [25]. The StatLog dataset [20] from the UCI repository is commonly used for various cardiac illnesses. It contains 13 attributes and 270 cases. Information is included about the following attributes: Age; Sex; Chest pain type (Chp); Resting blood pressure (Bp); Serum cholesterol (Sch) in mg/dl; Fasting blood sugar (Fbs) greater than 120 mg/dL; Resting electrocardiographic result (Ecg); Maximum heart rate (Mhrt) achieved; Exercise induced angina (Exian); Old peak (Opk) = ST depression induced by exercise relative to rest; Slope of the peak exercise ST segment (Slope); Number of major vessels colored by fluoroscopy (Vessel); and Defect type (Thal). The target field "Class" indicates whether the patient has heart disease, with a value of 0 for no disease and 1 for disease.

The Z-Alizadeh Sani dataset [16] from the UCI repository contains 303 patient records, each with 54 features. The attributes are classified into four categories: (i) demographic, (ii) symptom and examination, (iii) ECG, and (iv) laboratory and echo features. Each patient falls into one of two categories: CAD or normal. If a patient's diameter narrowing is greater than or equal to 50%, they are classified as having CAD; otherwise, they are classified as normal.

The CVD dataset [25] contains 70,000 patient records with the following different features: Age, Height, Weight, Gender, Systolic blood pressure (Ap_hi), Diastolic blood pressure (Ap_lo), Cholesterol, Glucose (Gluc), Smoking, Alcohol intake (Alco), and Physical activity (Active). The target class "Cardio" determines whether a patient is suffering from a cardiovascular illness (expressed as 1) or is healthy (shown as 0).

III. Methodology

A. Data Preprocessing

1) Normalization: In this step, we normalize the data. To enhance machine performance, an algorithm for learning data normalization is a preprocessing step that alters the attribute value in accordance with a standard scale or range. Examples of normalization methods include min–max, z-score, and decimal scaling [26]. There are many ML frameworks in the Python environment, such as sklearn [27]. This framework includes several helpful normalization algorithms, such as MinMaxScaler, MaxAbsScaler, StandardScaler, RobustScaler,

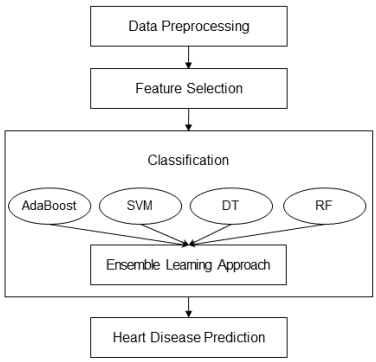

Fig. 1. Methodology framework.

and Normalizer. We used MinMaxScaler for normalization in this research.

2) Imbalanced Data: Imbalanced data are datasets with a highly uneven proportion of classes. Random Over Sampler, Random Under Sampler, Synthetic Minority Over-sampling Technique (SMOTE), and Tomek links are all examples of imbalanced data treatment [28]. To deal with the imbalanced data, we used SMOTE; the oversampling method involves producing synthetic instances rather than replacing oversampling for the minority class [29].

B. Data Splitting

When data is split, it is divided into two or more subsets. A two-part split is typically used to evaluate or test the data, and the other to train the model. Data splitting is a crucial feature of data science, especially for constructing data-collected models. This approach aids in the accuracy of data models and processes that employ data models, such as ML. We used cross-validation to split the dataset. Cross-validation is the most commonly used data-splitting approach in model selection. It separates the data into k distinct sections (k-folds) [30]. The validation set consists of one component (fold). The model is trained on the remaining k-1 portions (or folds), then applied to the validation set, and its prediction performance is recorded. This method was performed k times, resulting in each portion being utilized as a validation set just once. After averaging the recorded predicted performances, the optimal model parameter was selected with the best average predictive performance.

C. Feature Selection

Among practitioners, feature selection is a popular strategy for decreasing dimensionality. It seeks to choose a small subset of essential characteristics from the original collection based on specified criteria. Enhanced learning performance (e.g., increased learning accuracy for classification), reduced computation costs, and enhanced model interpretability are common outcomes of assessment criteria. Feature selection examples include filter, wrapper, and embedded methods [31]. Wrapper models assess the quality of features selected using a particular classifier and provide a simple and robust solution to the feature selection problem independent of the learning machine used [32]. We used two wrapper methods: forward selection method and backward elimination. The forward selection method begins with no features. In each iteration, the feature that enhances the model performance is added until the model's performance is not improved by adding a new one. In contrast, the backward elimination method begins with the entire set of features and then gradually eliminates the least promising ones.

D. Classification

Classification is an ML approach to predicting data, such as group membership [33]. We used four classification models: AdaBoost, SVM, DT, and RF.

1) AdaBoost: One ensemble method for ML is called AdaBoost, or adaptive boosting. Decision trees of one level, or those with only one split [34], are AdaBoost's most frequently employed estimator. Decision stumps are another name for these trees.

2) SVM: SVM is one of the most renowned and practical techniques for dealing with data classifications, learning, and prediction challenges. The data points nearest the decision surface are support vectors [35]. It uses a hyperplane to classify data vectors in infinite dimensional space. The simplest type of SVM is the maximal margin classifier, which aids in determining the most basic classification problem of linearly separable training data with binary classification [36]. The maximal margin classifier determines the hyperplane with the most significant margin in real-world complexities. SVMs employ a variety of kernel methods. In this work, we used a linear kernel.

3) DT: The categorization process was improved by the straightforward DT modeling approach. All decision tree algorithms are typically built in two stages: (i) tree growth, where the training set is divided repeatedly based on local optimal criteria until the majority of the records in the partition have the same class label, and (ii) tree pruning, where the size of the tree is reduced to make it more comprehensible [35].

4) RF: A classification system using several decision trees is called the random forest approach. It uses bagging and feature randomization to create each tree, resulting in an uncorrelated forest of trees whose forecast by the committee is more accurate than any one tree [37].

E. Ensemble Learning Approach

Combining data fusion, data modeling, and data mining into a unified framework is the aim of ensemble learning. A set of features is first extracted from the ensemble learning data using various transformations [38]. Based on these learned attributes, a few learning algorithms produce mediocre predictions. Finally, utilizing voting systems, ensemble learning combines the valuable data from the quick findings to provide knowledge discovery and enhanced prediction performance.

F. Evaluations

One of the best ways to evaluate how well the proposed model performs is to examine its accuracy, PPV, NPV, sensitivity, specificity, AUC, MCC, and Kappa. Accuracy evaluates

how often the classifier guesses accurately [39]. The accuracy of a forecast can be defined as the ratio of correct predictions to total predictions, and is defined as

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

The PPV reveals how many of the accurately anticipated cases were positive. Precision is functional when False positives are more of a worry than false negatives. The proportions of true positive and true negative results in diagnostic tests and statistics are PPV and NPV [40]. PPV and NPV describe the effectiveness of a diagnostic test or a similar statistical metric. A high result indicates that the statistic is accurate and can be determined, and is defined as

$$PPV(Precision) = \frac{TP}{TP+FP}$$

$$NPV = \frac{TN}{FN+TN}$$

According to Miao and Miao [41], sensitivity is the probability of successfully diagnosing the presence of cardiac disease in individuals, and is defined as

$$Sensitivity = Recall = \frac{TP}{TP+FN}$$

The probability of successfully identifying patients without cardiac disease is known as specificity, and is defined as

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

F1 provides a synthesis of the PPV and sensitivity measurements. It reaches its optimum when PPV and sensitivity are equal, and is defined as

$$F1 = \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*TP}{2*TP+FP+FN}$$

The measure of a classifier's ability to discriminate across classes is called the area under the curve (AUC) [40], and is defined as

$$AUC = Sensitivity - (1 - Specificity)$$

$$AUC = TPR - FPR$$

A statistical tool for assessing models is the MCC. It is accountable for determining the difference between anticipated and actual values, and is defined as

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

Kappa statistic is an excellent tool for handling difficulties involving multiple and unbalanced classes [42], and is defined as

$$Kappa = \frac{Po-Pe}{1-Pe}$$

IV. Experiments and Results

A. Experiment 1

In this experiment, we used three datasets without applying the selection features. First, our proposed approach used the Statlog dataset. Using the Python environment AdaBoost classifier (with random state = 50), RF classifier (with n estimators = 600), SVM classifier (with linear kernel), and DT classifier (with random state = 500), we obtained the results mentioned in Table II.

In addition, we used the Z-Alizadeh Sani dataset with our proposed approach using the Python environment AdaBoost classifier (with random state = 400), RF classifier (with n estimators = 100), SVM classifier (with linear kernel), and DT classifier (with random state = 22); we obtained the results mentioned in Table III.

TABLE II. Experiment 1 - Results of Statlog UCI Dataset Without Selection

Model	Acc	Sens	Spec	PPV	NPV	AUC	F1	Kappa	MCC
AdaBoost	0.80	0.8	0.8	0.82	0.81	0.89	0.80	0.62	0.62
RF	0.84	0.82	0.82	0.87	0.83	0.92	0.84	0.69	0.70
SVM	0.82	0.8	0.8	0.84	0.81	0.91	0.81	0.65	0.66
DT	0.76	0.76	0.76	0.76	0.77	0.76	0.76	0.53	0.53
ELA	0.80	0.80	0.81	0.81	0.81	0.90	0.81	0.62	0.62

TABLE III. Experiment 1 - Results of Z-Alizadeh Sani Dataset Without Selection

Model	Acc	Sens	Spec	PPV	NPV	AUC	F1	Kappa	MCC
AdaBoost	0.91	0.92	0.92	0.90	0.93	0.96	0.91	0.82	0.83
RF	0.92	0.92	0.91	0.91	0.92	0.97	0.91	0.84	0.86
SVM	0.87	0.89	0.89	0.85	0.89	0.93	0.87	0.74	0.74
DT	0.84	0.76	0.87	0.83	0.83	0.84	0.84	0.69	0.70
ELA	0.89	0.92	0.90	0.87	0.88	0.96	0.89	0.77	0.77

Furthermore, we used the CVD dataset without applying any selection feature methods using the Python environment AdaBoost classifier (with random state = 50), RF classifier (with n estimators = 500), SVM classifier (with linear kernel), and DT classifier (with random state = 200); we obtained the results mentioned in Table IV.

B. Experiment 2

In this experiment, we used the features obtained from three studies: Yekkala et al. [3], Velusamy and Ramasamy [15], and Chintan et al. [43]. The feature selection step was carried out by the three studies. In the first study, Yekkala et al. [3] used a PSO feature selection method to extract features from the Statlog dataset and extract seven features: Chp, Ecg, Mhrt, Exian, Opk, Vessel, and Thal. We took the same seven features they obtained and used them with the proposed model: the AdaBoost classifier (with random state = 1), RF classifier (with n estimators = 10), SVM classifier (with linear kernel), and DT classifier (with random state = 2), and obtained the results mentioned in Table V.

In the second study, Velusamy and Ramasamy [15] used feature selection from the Z-Alizadeh Sani dataset based on SVM. It is based on model information in which the model is trained to integrate the relationship between predictors for computing variable importance. They chose the top 12 features, namely: Atypical, Typical Chest pain, Age, Region with Regional wall motion abnormality (Region RWMA), Ejection Fraction (EF-TTE), Nonanginal Chest Pain (Nonanginal), Hypertension (HTN), FBS, Tinversion, BP, Diabetes Mellitus (DM), and TG. Then chose the following top five significant features: Typical Chest pain, Atypical, Age, Region RWMA, and EF-TTE. We used these features with the proposed model:

TABLE IV. Experiment 1 - Results Of CVD Dataset Without Selection

Model	Acc	Sens	Spec	PPV	NPV	AUC	F1	Kappa	MCC
AdaBoost	0.72	0.63	0.81	0.77	0.56	0.79	0.70	0.45	0.46
RF	0.72	0.70	0.73	0.72	0.71	0.78	0.71	0.44	0.44
SVM	0.72	0.60	0.83	0.79	0.67	0.78	0.68	0.44	0.45
DT	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.28	0.28
ELA	0.68	0.69	0.67	0.68	0.68	0.77	0.68	0.37	0.37

TABLE V. EXPERIMENT 2 - RESULTS WITH 7 FEATURES FROM STATLOG UCI DATASET

Model	Acc	Sens	Spec	PPV	NPV	AUC	**F1**	Kappa	MCC
AdaBoost	0.77	0.74	0.74	0.80	0.76	0.85	0.76	0.55	0.56
RF	0.80	0.78	0.77	0.83	0.80	0.90	0.80	0.62	0.66
SVM	0.84	0.82	0.82	0.86	0.83	0.91	0.83	0.68	0.68
DT	0.76	0.80	0.80	0.76	0.80	0.76	0.77	0.53	0.55
ELA	0.80	0.81	0.82	0.82	0.83	0.91	0.80	0.59	0.61

TABLE VI. EXPERIMENT 2 - RESULTS WITH 12 FEATURES FROM Z-ALIZADEH SANI DATASET

Model	Acc	Sens	Spec	PPV	NPV	AUC	**F1**	Kappa	MCC
AdaBoost	0.89	0.89	0.89	0.89	0.91	0.96	0.88	0.78	0.79
RF	0.90	0.86	0.89	0.91	0.88	0.96	0.89	0.76	0.83
SVM	0.89	0.92	0.92	0.87	0.92	0.94	0.89	0.79	0.79
DT	0.85	0.87	0.87	0.83	0.88	0.84	0.85	0.69	0.70
ELA	0.91	0.93	0.91	0.90	0.92	0.96	0.90	0.82	0.81

the AdaBoost classifier (with random state = 1), RF classifier (with n estimators = 10), SVM classifier (with linear kernel), and DT classifier (with random state = 2). We obtained the results mentioned in Table VI and Table VII.

In the third study, Chintan et al. [43] used feature selection from the CVD dataset and estimated the mean arterial pressure (MAP) from the diastolic blood pressure (Ap_lo) and systolic blood pressure (Ap_hi) values for each instance. Patients' ages were initially given in days. Nonetheless, it was changed to years by dividing it by 365 to improve the analysis and prediction. They transformed the attributes of height and weight into body mass index (BMI), which may increase the performance of the heart disease prediction model. The nine features they obtained from the selection were Age, Gender, BMI, MAP, Cholesterol, Gluc, Smoke, Alco, and Active. We used these features with the proposed model: the AdaBoost classifier (with random state = 50), RF classifier (with n estimators = 500), SVM classifier (with linear kernel), and DT classifier (with random state = 200). We obtained the results mentioned in Table VIII.

C. Experiment 3

In this experiment, we used two wrapper methods for selecting features: forward selection and backward elimination. An iterative process called forward selection starts with the model having no features. The feature that best enhances our model is added in each iteration until the model's performance is not improved by adding a new variable. Backward elimination helps the model perform better by starting with all the features and removing the least important aspects one at a time. We keep doing this until we see no improvement when we remove features. We used four algorithms: AdaBoost, SVM, DT, and RF. We used R-squared as a measure of the performance of the feature-selection models. We applied it to the Statlog UCI, the Z-Alizadeh Sani, and the CVD datasets.

TABLE VII. EXPERIMENT 2 - RESULTS WITH 5 FEATURES FROM Z-ALIZADEH SANI DATASET

Model	Acc	Sens	Spec	PPV	NPV	AUC	**F1**	Kappa	MCC
AdaBoost	0.89	0.89	0.89	0.89	0.91	0.96	0.88	0.78	0.79
RF	0.90	0.86	0.89	0.91	0.88	0.96	0.89	0.76	0.83
SVM	0.83	0.87	0.87	0.81	0.86	0.92	0.84	0.67	0.67
DT	0.83	0.82	0.82	0.85	0.84	0.84	0.83	0.67	0.68
ELA	0.86	0.84	0.84	0.86	0.85	0.94	0.85	0.71	0.71

TABLE VIII. EXPERIMENT 2 - RESULTS WITH 9 FEATURES FROM CVD DATASET

Model	Acc	Sens	Spec	PPV	NPV	AUC	F1	Kappa	MCC
AdaBoost	0.72	0.64	0.64	0.76	0.69	0.78	0.69	0.44	0.44
RF	0.70	0.74	0.66	0.70	0.66	0.78	0.71	0.32	0.32
SVM	0.71	0.59	0.59	0.78	0.67	0.77	0.67	0.43	0.44
DT	0.63	0.60	0.60	0.63	0.62	0.63	0.62	0.26	0.26
ELA	0.66	0.65	0.67	0.67	0.66	0.74	0.66	0.33	0.33

The result of feature selection with the Statlog UCI dataset is shown in Table IX, and it had the highest impact when using SVM with forward and backward methods. With the Statlog UCI dataset, we obtained the same result using both methods, but we chose the SVM with the forward method, as the number of significant features is less than 10. The results of the proposed model with the Statlog UCI dataset after using the SVM with the forward selection method are shown in Table XII. With the Z-Alizadeh Sani dataset, the highest R-squared result of the feature selection method was the SVM with the backward method using 26 features, as shown in Table X. The results of the proposed model with the Z-Alizadeh Sani dataset after using the SVM with the backward elimination method are shown in Table XIII.

With the CVD dataset, the result of feature selection is shown in Table XI, and it had the highest impact when using DT with backward, RF with forward, AdaBoost with forward, and AdaBoost with backward. We chose AdaBoost with the forward method. The results of the proposed model with the CVD dataset after using AdaBoost with the forward selection technique are shown in Table XIV.

We compared the results of the proposed model with previous studies that used one of the three datasets. When comparing the results of the proposed model with David's model [5], which used the Statlog UCI dataset, we found that the results of the proposed model outperformed their obtained results. In contrast, we obtained PPV = 0.83, which is higher than the [5] model (PPV = 0.812), and the proposed model got a higher sensitivity = 0.83, whereas the [5] model got a lower result (0.806 sensitivity). In addition, the [5] model had a lower F1 (F1 = 0.802) compared to the F1 of the developed model (F1 = 0.83). Unfortunately, David [5] was content with only three matrices to evaluate their results. It would have been better if they had used more matrices to comprehensively view the results. While Yekkala et al. [3] got 100% accuracy, sensitivity, specificity, PPV, and NPV. Some previous studies [13] and [15], which used the Z-Alizadeh Sani dataset obtained better results than the proposed model. Tama et al. [13] got an accuracy = 98.31%, F1 = 96.60% and AUC = 98.70%, while the proposed model obtained an accuracy = 91%, F1 = 89%, and AUC = 97%. Velusamy and Ramasamy [15] got 100% in accuracy, sensitivity, specificity, Kappa, PPV, F1, and MCC. However, this does not necessarily mean that their results are as good as in ML when 100% accuracy is achieved, which may indicate data overfitting.

TABLE IX: Expermint 3 - Best Wrapper Methods Results with Reduced Features Sets using Statlog UCI Dataset

Method	Number of features	Name of features	R2
Forward+DT	3	Exian, Vessel, Thal	0.34

Continued on next page

TABLE IX: Expermint 3 - Best Wrapper Methods Results with Reduced Features Sets using Statlog UCI Dataset (Continued)

Backward+DT	10	Age, Sex, Chp, Bp, Sch, Fbs, Ecg, Exian, Opk, Vessel	0.14
Forward+RF	3	Exian, Vessel, Thal	0.36
Backward+RF	12	Age, Sex, Chp, Bp, Sch, Fbs, Mhrt, Exian, Opk, Slope, Vessel, Thal	0.34
Forward+SVM	10	Age, Sex, Chp, Bp, Sch, Ecg, Mhrt, Exian, Vessel, Thal	0.44
Backward+SVM	12	Age, Sex, Chp, Bp, Sch, Fbs, Ecg, Mhrt, Exian, Opk, Vessel, Thal	0.44
Forward+AdaBoost	7	Sex, Chp, Fbs, Exian, Slope, Vessel, Thal	0.36
Backward+AdaBoost	8	Sex, Chp, Bp, Sch, Fbs, Opk, Vessel, Thal	0.33

TABLE X: Experiment 3 - Best Wrapper Methods Results with Reduced Features Sets using Z-Alizadeh Sani Dataset

Method	Number of features	Name of features	R2
Forward+DT	6	DM, EX-Smoker, Typical Chest Pain, EF-TTE, Region RWMA, VHD	0.42
Backward+DT	22	Age, Length, DM, HTN, Typical Chest Pain, Function Class, Q Wave, Tinversion, TG, LDL, ESR , K, Na, Region RWMA, Sex, CRF, CHF, DLP, Weak Peripheral Pulse, Dyspnea, LowTH Ang, LVH	0.31
Forward+RF	12	DM, EX-Smoker, Edema, Typical Chest Pain, St Elevation, Na, CHF, Region RWMA, Airway disease, Lung rales, LowTH Ang, Poor R Progression	0.42
Backward+RF	16	Age, HTN, BP, PR, Typical Chest Pain, St Elevation, Tinversion, CR, HDL, ESR, HB, Region RWMA, Obesity, CRF, Exertional CP, VHD	0.52
Forward+SVM	33	Age, Length, DM, HTN, Current Smoker, FH, PR, Typical Chest Pain, Q Wave, St Elevation, St Depression, Tinversion, FBS, CR, TG, HDL, BUN, HB, PLT, Lymph, Region RWMA, CRF, CVA, Thyroid Disease, CHF, DLP, Weak Peripheral Pulse, Lung rales, Dyspnea, Exertional CP, LowTH Ang, Poor R Progression, VHD	0.625

Continued on next page

TABLE X: Experiment 3 - Best Wrapper Methods Results with Reduced Features Sets using Z-Alizadeh Sani Dataset (Continued)

Backward+SVM	26	Age, Length, DM, HTN, Current Smoker, FH, PR, Typical Chest Pain, Function Class, Q Wave, Tinversion, CR, HDL, HB, K, WBC, Lymph, EF-TTE, Region RWMA, Sex, DLP, Airway disease, Lung rales, Dyspnea, Atypical, Nonanginal	0.626
Forward+AdaBoost	25	DM, HTN, Current Smoker, Edema, Typical Chest Pain, Q Wave, St Elevation, St Depression, CR, Region RWMA, Sex, CRF, CVA, CHF, Lung rales, Airway disease, Weak Peripheral Pulse, Systolic Murmur, Diastolic Murmur, Dyspnea, Exertional CP, LowTH Ang, Poor R Progression, VHD, LVH	0.50
Backward+AdaBoost	14	Age, BMI, DM, HTN, PR, ESR, Typical Chest Pain,CR, Tinversion, HB, WBC, EF-TTE, Region RWMA, Nonanginal	0.46

TABLE XI: EXPERMINT 3 - BEST WRAPPER METHODS RESULTS WITH REDUCED FEATURES SETS USING CVD DATASET

Method	Number of features	Name of features	R2
Forward+DT	8	Gender, Ap-hi, Ap-lo, Cholesterol, Gluc, Smoke, Alco, Active	-0.08
Backward+DT	7	Gender, Ap-hi, Cholesterol, Gluc, Smoke, Alco, Active	-0.07
Forward+RF	6	Ap-hi, Cholesterol, Gluc, Smoke, Alco, Active	-0.07
Backward+RF	8	Age, Height, Weight, Ap-hi, Ap-lo, Cholesterol, Gluc, Smoke	-0.16
Forward+SVM	5	Age, Height, Ap-hi, Cholesterol, Gluc	-0.08
Backward+SVM	10	Age, Gender, Height, Weight, Ap-hi, Cholesterol, Gluc, Smoke, Alco, Active	-0.08
Forward+AdaBoost	10	Age, Gender, Weight, Ap-hi, Ap-lo, Cholesterol, Gluc, Smoke, Alco, Active	-0.07
Backward+AdaBoost	10	Age, Gender, Weight, Ap-hi, Ap-lo, Cholesterol, Gluc, Smoke, Alco, Active	-0.07

V. Conclusion

The main goal of this study was to predict cardiac disease utilizing ELA, which included four ML algorithms: AdaBoost, SVM, DT, and RF. We applied it to three datasets: the StatLog UCI dataset, the Z-Alizadeh Sani dataset, and the CVD dataset.

TABLE XII. EXPERIMENT 3 - RESULTS OF PROPOSED METHOD WITH WRAPPER SELECTION USING STATLOG UCI DATASET

Model	Acc	Sens	Spec	PPV	NPV	AUC	F1	Kappa	MCC
AdaBoost	0.81	0.81	0.81	0.81	0.83	0.88	0.80	0.62	0.63
RF	0.84	0.84	0.84	0.85	0.85	0.91	0.85	0.69	0.71
SVM	0.82	0.79	0.79	0.85	0.81	0.90	0.81	0.65	0.66
DT	0.78	0.77	0.77	0.79	0.78	0.78	0.77	0.56	0.57
ELA	0.83	0.83	0.82	0.85	0.83	0.90	0.83	0.67	0.68

TABLE XIII. EXPERIMENT 3 - RESULTS OF PROPOSED METHOD WITH WRAPPER SELECTION USING Z-ALIZADEH SANI DATASET

Model	Acc	Sens	Spec	PPV	NPV	AUC	F1	Kappa	MCC
AdaBoost	0.90	0.90	0.90	0.91	0.91	0.95	0.90	0.81	0.82
RF	0.91	0.93	0.93	0.91	0.93	0.98	0.92	0.84	0.84
SVM	0.89	0.90	0.90	0.89	0.90	0.95	0.89	0.78	0.79
DT	0.85	0.86	0.86	0.84	0.88	0.85	0.85	0.70	0.71
ELA	0.91	0.92	0.90	0.93	0.92	0.97	0.89	0.81	0.84

We used two wrapper methods for the feature selection step, forward selection and backward elimination, and we dealt with the data imbalance using SMOTE. When using the proposed model with the StatLog UCI dataset, we obtained accuracy = 0.83, sensitivity = 0.83, specificity = 0.82, AUC = 0.90, PPV = 0.85, NPV = 0.83, F1 = 0.83, Kappa = 0.67, MCC = 0.68. When we used the Z-Alizadeh Sani dataset, we obtained accuracy = 0.91, sensitivity = 0.92, specificity = 0.90, AUC = 0.97, PPV = 0.93, NPV = 0.92, F1 = 0.89, Kappa = 0.81, MCC = 0.84. When using the CVD dataset, we obtained accuracy = 0.73, sensitivity = 0.63, specificity = 0.82, AUC = 0.77, PPV = 0.78, NPV = 0.69, F1 = 0.70, Kappa = 0.45, MCC = 0.46. In future work, we aim to collect a local dataset from King Abdullah Hospital - Bisha in the KSA, apply the proposed model to it, and improve the accuracy of the model. We will also use PSO and Gray Wolf Optimizer feature selection techniques that have shown promising results in disease prediction to further improve the model's performance.

TABLE XIV. EXPERIMENT 3 - RESULTS OF PROPOSED METHOD WITH WRAPPER SELECTION USING CVD DATASET

Model	Acc	Sens	Spec	PPV	NPV	AUC	F1	Kappa	MCC
AdaBoost	0.72	0.63	0.82	0.78	0.69	0.77	0.70	0.45	0.46
RF	0.72	0.63	0.82	0.78	0.69	0.77	0.70	0.45	0.46
SVM	0.72	0.65	0.79	0.76	0.69	0.77	0.70	0.45	0.45
DT	0.72	0.63	0.82	0.78	0.69	0.77	0.70	0.45	0.46
ELA	0.73	0.63	0.82	0.78	0.69	0.77	0.70	0.45	0.46

REFERENCES

[1] M. G. T. health, "Heart disease is the cause of 42% of deaths from non-communicable diseases in the kingdom," Oct 2013. [Online]. Available: https://www.moh.gov.sa/Ministry/MediaCenter/News/Pages/News-2013-10-30-002.aspx

[2] A. Rath, D. Mishra, G. Panda, and S. C. Satapathy, "Heart disease detection using deep learning methods from imbalanced ecg samples," *Biomedical Signal Processing and Control*, vol. 68, p. 102820, 2021.

[3] I. Yekkala, S. Dixit, and M. Jabbar, "Prediction of heart disease using ensemble learning and particle swarm optimization," in *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*. IEEE, 2017, pp. 691–698.

[4] G. Ranganathan, J. Chen, and Á. Rocha, "Inventive communication and computational technologies," Ph.D. dissertation, Springer, 2021.

[5] H. B. F. David, "Impact of ensemble learning algorithms towards accurate heart disease prediction," *ICTACT Journal on Soft Computing*, vol. 10, no. 3, pp. 2084–2089, 2020.

[6] U. N. Dulhare, K. Ahmad, and K. A. B. Ahmad, *Machine learning and big data: concepts, algorithms, tools and applications*. John Wiley & sons, 2020.

[7] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–16, 2019.

[8] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–15, 2019.

[9] N. health and nutrition examination survey, "National health and nutrition examination survey," Dec. 2022, accessed: 2022-12-10.

[10] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Uci machine learning repository-heart disease data set," *School Inf. Comput. Sci., Univ. California, Irvine, CA, USA*, 1988.

[11] N. Liu, X. Li, E. Qi, M. Xu, L. Li, and B. Gao, "A novel ensemble learning paradigm for medical diagnosis with imbalanced data," *IEEE Access*, vol. 8, pp. 171 263–171 280, 2020.

[12] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Informatics in Medicine Unlocked*, vol. 20, p. 100402, 2020.

[13] B. A. Tama, S. Im, and S. Lee, "Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble," *BioMed Research International*, vol. 2020, 2020.

[14] D. C. Yadav and S. Pal, "Prediction of heart disease using feature selection and random forest ensemble method," *International Journal of Pharmaceutical Research*, vol. 12, no. 4, pp. 56–66, 2020.

[15] D. Velusamy and K. Ramasamy, "Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset," *Computer Methods and Programs in Biomedicine*, vol. 198, p. 105770, 2021.

[16] Z. Alizadehsani, R. Alizadehsani, and M. Roshanzamir, "Z-alizadeh sani data set," 2017. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani

[17] T. Tuncer, S. Dogan, P. Plawiak, and A. Subasi, "A novel discrete wavelet-concatenated mesh tree and ternary chess pattern based ecg signal recognition method," *Biomedical Signal Processing and Control*, vol. 72, p. 103331, 2022.

[18] G. B. Moody and R. G. Mark, "The impact of the mit-bih arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001.

[19] T. Viktor and A. Khaustov, "St.-petersburg institute of cardiological technics 12–lea d arrhythmia database," *Circulation-Electronic*, vol. 101, no. i23, pp. e215–e220, 2000.

[20] S. Sumbria, "Statlog (Heart) Data Set — kaggle.com," https://www.kaggle.com/datasets/shubamsumbria/statlog-heart-data-set, 2019, [Accessed 16-Jun-2023].

[21] A. Asuncion and D. Newman, "Uci machine learning repository," 2007.

[22] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework." *Journal of Multiple-Valued Logic & Soft Computing*, vol. 17, 2011.

[23] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *The American journal of cardiology*, vol. 64, no. 5, pp. 304–310, 1989.

[24] A. Bhardwaj, "Framingham heart study dataset," Apr 2022. [Online]. Available: https://www.kaggle.com/aasheesh200/framingham-heart-study-dataset

[25] S. Ulianova, "Cardiovascular Disease dataset — kaggle.com," https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset, 2019, [Accessed 16-Jun-2023].

[26] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*. Springer, 2015.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[28] R. Longadge and S. Dongre, "Class imbalance problem in data mining review," *arXiv preprint arXiv:1305.1707*, 2013.

[29] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[30] D. Krstajic, L. J. Buturovic, D. E. Leahy, and S. Thomas, "Cross-validation pitfalls when selecting and assessing regression and classification models," *Journal of cheminformatics*, vol. 6, pp. 1–15, 2014.

[31] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," *Data classification: Algorithms and applications*, p. 37, 2014.

[32] N. Hasan and Y. Bao, "Comparing different feature selection algorithms for cardiovascular disease prediction," *Health and Technology*, vol. 11, pp. 49–62, 2021.

[33] T. N. Phyu, "Survey of classification techniques in data mining," in *Proceedings of the international multiconference of engineers and computer scientists*, vol. 1. Citeseer, 2009, pp. 727–731.

[34] R. E. Schapire, "Explaining adaboost," *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pp. 37–52, 2013.

[35] A. A. Soofi and A. Awan, "Classification techniques in machine learning: applications and issues," *J. Basic Appl. Sci*, vol. 13, pp. 459–465, 2017.

[36] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu *et al.*, "Top 10 algorithms in data mining," *Knowledge and information systems*, vol. 14, pp. 1–37, 2008.

[37] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert systems with applications*, vol. 134, pp. 93–101, 2019.

[38] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers of Computer Science*, vol. 14, no. 2, pp. 241–258, 2020.

[39] J. A. Swets, "Measuring the accuracy of diagnostic systems," *Science*, vol. 240, no. 4857, pp. 1285–1293, 1988.

[40] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings 27*. Springer, 2005, pp. 345–359.

[41] K. H. Miao, G. J. Miao *et al.*, "Mammographic diagnosis for breast cancer biopsy predictions using neural network classification model and receiver operating characteristic (roc) curve evaluation," *Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Bioinformatics*, vol. 3, no. 9, pp. 1–10, 2013.

[42] D. Chicco, M. J. Warrens, and G. Jurman, "The matthews correlation coefficient (mcc) is more informative than cohen's kappa and brier score in binary classification assessment," *IEEE Access*, vol. 9, pp. 78 368–78 381, 2021.

[43] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective heart disease prediction using machine learning techniques," *Algorithms*, vol. 16, no. 2, p. 88, 2023.