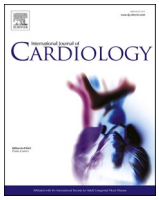
International Journal of Cardiology 436 (2025) 133443

Contents lists available at ScienceDirect

International Journal of Cardiology

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



Comparative analysis of machine learning models for coronary artery disease prediction with optimized feature selection

David B. Olawade [a,b,c,d,*], Afeez A. Soladoye [e], Bolaji A. Omodunbi [e], Nicholas Aderinto [f], Ibrahim A. Adeyanju [e]

[a] Department of Allied and Public Health, School of Health, Sport and Bioscience, University of East London, London, United Kingdom
[b] Department of Research and Innovation, Medway NHS Foundation Trust, Gillingham ME7 5NY, United Kingdom
[c] Department of Public Health, York St John University, London, United Kingdom
[d] School of Health and Care Management, Arden University, Arden House, Middlemarch Park, Coventry CV3 4FJ, United Kingdom
[e] Department of Computer Engineering, Federal University, Oye, Ekiti, Nigeria
[f] Department of Medicine and Surgery, Ladoke Akintola University of Technology, Ogbomoso, Nigeria

ARTICLE INFO

Keywords:
Coronary artery disease
Machine learning
Feature selection
Bald eagle search optimization
Random Forest

ABSTRACT

Background: Coronary artery disease (CAD) is a major global cause of death, necessitating early, accurate prediction for better management. Traditional diagnostics are often invasive, costly, and less accessible. Machine learning (ML) offers a non-invasive alternative, but high-dimensional data and redundancy can hinder performance. This study integrates Bald Eagle Search Optimization (BESO) for feature selection to improve CAD classification using multiple ML models.

Methods: Two publicly available datasets, Framingham (4200 instances, 15 features) and *Z*-Alizadeh Sani (304 instances, 55 features), were used. The former predicts 10-year CAD risk, while the latter classifies current CAD status. Data preprocessing included missing value imputation, normalization, categorical encoding, and class balancing using SMOTE. We employed a 70–30 holdout validation strategy with empirical hyperparameter optimization, providing more reliable final model development than cross-validation. BESO was applied to optimize feature selection, significantly outperforming traditional methods like RFE and LASSO. Six ML models—KNN, logistic regression, SVM with linear, polynomial, and RBF kernels, and random forest—were trained and evaluated.

Results: Random Forest achieved the highest performance across both datasets. In the Framingham dataset, RF recorded 90 % accuracy, significantly outperforming traditional clinical risk scores (71–73 % accuracy). Linear models performed better on the *Z*-Alizadeh Sani dataset (90 % accuracy) than Framingham (66 %), indicating dataset characteristics strongly influence model efficacy.

Conclusion: BESO significantly enhances feature selection, with RF emerging as the optimal classifier (92 % accuracy) and substantially outperforming established clinical risk scores. This study highlights the potential of AI-driven CAD diagnosis, supporting early detection and improved patient outcomes. Future work should focus on prospective validation and clinical implementation.

1. Introduction

Coronary artery disease (CAD) remains one of the leading causes of morbidity and mortality worldwide, significantly contributing to the global burden of cardiovascular diseases [1]. Characterized by the narrowing or blockage of coronary arteries due to atherosclerosis, CAD restricts blood flow to the heart, potentially leading to severe complications such as myocardial infarction, heart failure, and sudden cardiac death [2]. Early detection of CAD is paramount to preventing its progression, improving patient outcomes, and reducing healthcare costs [3]. However, conventional diagnostic methods including electrocardiography (ECG), echocardiography, angiography, and stress testing are often invasive, costly, or reliant on specialized expertise [4]. These limitations have spurred growing interest in machine learning (ML) techniques as non-invasive, data-driven alternatives for predicting CAD risk using readily available patient data [5].

* Corresponding author at: Department of Allied and Public Health, School of Health, Sport and Bioscience, University of East London, London, United Kingdom.
E-mail address: d.olawade@uel.ac.uk (D.B. Olawade).

https://doi.org/10.1016/j.ijcard.2025.133443
Received 15 March 2025; Received in revised form 26 April 2025; Accepted 28 May 2025
Available online 31 May 2025

Machine learning has emerged as a transformative tool in healthcare, enabling the analysis of large, complex datasets to uncover patterns that may elude traditional clinical approaches [6]. In the context of CAD prediction, ML models leverage historical patient data such as demographic attributes, lifestyle factors, laboratory results, and clinical symptoms to deliver accurate risk assessments [7]. Previous studies have successfully applied supervised learning algorithms, including logistic regression (LR), support vector machines (SVM), K-nearest neighbors (KNN), and ensemble methods like random forests (RF), to classify CAD patients based on risk factors [8]. Despite their promise, these models often grapple with challenges posed by redundant, irrelevant, or highly correlated features in medical datasets, which can degrade performance by causing overfitting, increasing computational complexity, and reducing interpretability [9].

Feature selection is a critical strategy for overcoming these challenges and enhancing both the accuracy and efficiency of ML models [10]. Traditional dimensionality reduction techniques, such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE), are widely employed to eliminate irrelevant variables [11]. However, these methods may fall short when applied to complex, high-dimensional datasets common in medical applications [12]. In response, nature-inspired optimization algorithms have gained traction for their ability to efficiently navigate large search spaces and pinpoint the most predictive features [13]. Among these, the Bald Eagle Search Optimization (BESO) algorithm has recently emerged as a promising approach [14]. Inspired by the foraging behavior of bald eagles, BESO balances exploration and exploitation, avoiding local optima traps and enhancing model generalization [15]. Its application in feature selection offers a novel avenue for optimizing ML pipelines in CAD prediction.

This study aims to develop an optimized machine learning pipeline for CAD prediction by integrating BESO-based feature selection with multiple classification algorithms, including KNN, LR, SVM with various kernels, and RF. The specific objectives are threefold: (a) to assess the impact of BESO on feature selection and subsequent model performance, (b) to compare the predictive accuracy of different ML models across two CAD datasets, and (c) to determine the most effective model for early CAD detection. By addressing these goals, this research seeks to advance AI-driven healthcare solutions, refine CAD risk assessment, and bolster non-invasive strategies for early diagnosis, ultimately improving patient outcomes.

2. Methodology

This study employed a structured machine learning pipeline for the prediction of coronary artery disease (CAD), consisting of data acquisition, preprocessing, feature selection using a nature-inspired optimization algorithm, and model training with evaluation. The steps undertaken in this research are described in detail below and shown in Fig. 1.

2.1. Data acquisition

Two publicly available datasets were used in this study: the Framingham dataset and the *Z*-Alizadeh Sani dataset. These datasets were selected to represent different aspects of cardiovascular disease prediction.

- Framingham Dataset: This dataset originates from the Framingham Heart Study, a long-term, ongoing cardiovascular cohort study of residents of Framingham, Massachusetts. It contains 4200 instances with 15 predictive features related to demographics, medical history, and risk factors associated with CAD. The dataset includes longitudinal data collected over multiple examination cycles, with a 10-year follow-up period for cardiovascular events.
- *Z*-Alizadeh Sani Dataset: This dataset was specifically chosen as a complementary dataset because it contains more detailed clinical measurements compared to the Framingham dataset. Published by Alizadehsani et al. (2013), it was collected from Tehran's Shaheed Rajaei Cardiovascular, Medical, and Research Center [7]. It contains 304 instances with 55 features spanning multiple modalities, including demographic attributes (age, sex), symptoms (typical chest pain, atypical chest pain), examination results (systolic and diastolic blood pressure), electrocardiogram (ECG) readings (ST elevation, ST depression), laboratory tests (fasting blood sugar, creatinine, triglycerides), and echocardiographic measurements (ejection fraction). The primary outcome variable is the presence of CAD, defined as $\geq$50 % stenosis in at least one major coronary artery, as determined by angiography, which serves as the gold standard for CAD diagnosis.

Using these two distinct datasets enables a robust evaluation of our

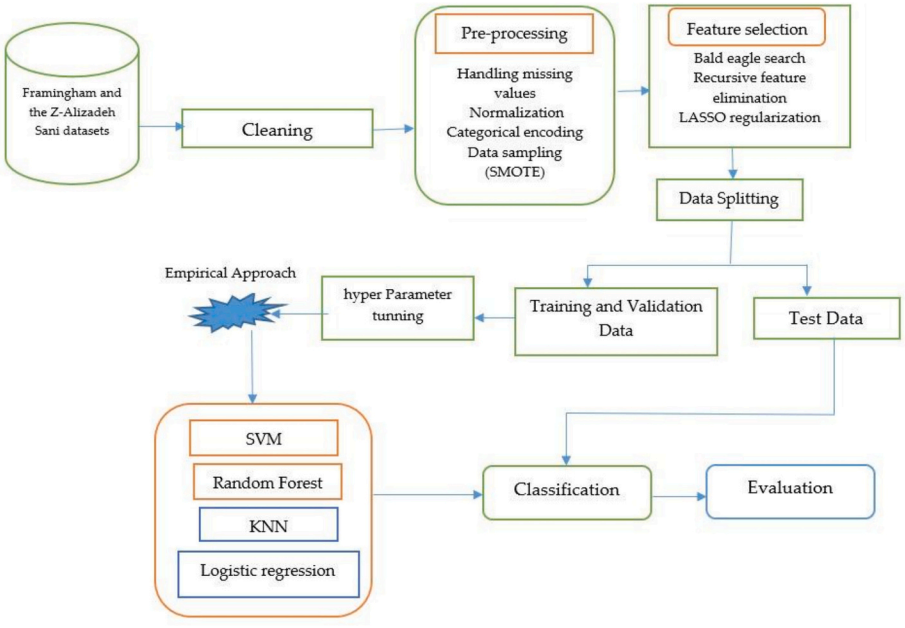


Fig. 1. Methodological Framework for Machine Learning-based CAD Prediction.

methodology across different feature spaces, sample sizes, and population characteristics, strengthening the generalizability of our findings.

2.2. Data preprocessing

To ensure the quality and reliability of the datasets for machine learning, a rigorous data preprocessing pipeline was applied:

- **Handling Missing Values**: Missing values were imputed using median imputation, which is robust to outliers and prevents skewing of the data.
- **Feature Normalization**: Numerical features were standardized using Scikit-learn's StandardScaler, which normalizes the data to have zero mean and unit standard deviation.
- **Categorical Encoding**: Categorical variables were converted into numerical values using Scikit-learn's LabelEncoder, which assigns each category a unique integer.
- **Class Imbalance Handling**: The Synthetic Minority Over-sampling Technique (SMOTE) was used to address class imbalance by generating synthetic samples for the minority class while maintaining the distribution of the majority class. Over-sampling was performed separately on the training and testing sets to prevent data leakage and improve model generalization.

These preprocessing steps ensured that the data was clean, standardized, and balanced for effective model training.

2.3. Feature selection using bald eagle search optimization

Feature selection was performed using the Bald Eagle Search Optimization (BESO) algorithm, a nature-inspired metaheuristic approach based on the hunting behavior of bald eagles. BESO was chosen for its ability to efficiently explore high-dimensional search spaces while balancing exploration and exploitation, making it highly suitable for feature selection.

- **Feature Reduction**: Given that the Framingham dataset contained 15 features and the *Z*-Alizadeh Sani dataset contained 55 features, dimensionality reduction was necessary to minimize computational complexity and improve model performance.
- **Optimization Mechanism**: BESO iteratively selected the most predictive features while avoiding local optima, ensuring that the final subset of features retained the most relevant information for CAD prediction.

By applying BESO, an optimal feature subset was identified, improving the efficiency and accuracy of the subsequent machine learning models. Table 1 represent the pseudo code used for implementation of BESO for feature selection.

2.4. Machine learning algorithms for prediction of CAD

This study utilized a diverse set of machine learning algorithms to develop robust predictive models for Coronary Artery Disease (CAD). The selection of these algorithms was based on their established effectiveness in classification tasks and their ability to handle complex datasets with varying feature interactions. The algorithms implemented in this study include:

- K-Nearest Neighbors (KNN): A non-parametric, instance-based learning algorithm, KNN was selected for its simplicity and effectiveness in capturing local patterns. It can model complex decision boundaries without making assumptions about the underlying data distribution. KNN has shown success in medical diagnosis applications where local clusters of similar patients often share diagnoses [16]. We selected this algorithm as a baseline due to its interpretability and ability to handle non-linear relationships.
- Logistic Regression (LR): As a probabilistic linear model, LR was included for its interpretability and established history in medical risk prediction. It provides odds ratios for individual features, allowing clinicians to understand the contribution of specific risk factors. The Framingham Risk Score itself uses logistic regression,

Table 1
BESO for feature selection.

Input:
- dataset (features, target)
- population size (N)
- maximum iterations (MaxIter)
- search space dimension (D, number of features)
- fitness function (e.g., model performance with selected features)
- parameters (alpha, beta, c1, c2, etc.)

Output:
- optimal feature subset

1. Initialize population:
- generate N random solutions (feature subsets) represented as binary vectors (0 or 1, where 1 indicates feature selection).

2. Evaluate fitness:
- for each solution in the population:
Select features based on the solution's binary vector.
Train a model using the selected features.
Calculate the fitness (e.g., accuracy, F1-score) of the model.

3. Iterative optimization (for iteration = 1 to MaxIter):
- phase 1 (select space):
- calculate the mean of the population.
- update solutions based on the mean and random search.
- phase 2 (search space):
- calculate the best solution (eagle with the best fitness).
- update solutions based on the best solution and random search.
- phase 3 (swipe space):
- update solutions based on the best solution, previous solution, and random search, simulating eagle's swoop.
- evaluate fitness:
- recalculate the fitness of each solution.
- update best solution:
- if a solution with better fitness is found, update the best solution.

4. Return optimal feature subset:
- return the feature subset corresponding to the best solution found during the optimization process.

making it a standard approach for cardiovascular risk assessment [17]. Additionally, LR serves as an important baseline to determine whether the CAD prediction task requires more complex non-linear models.

- Support Vector Machines (SVM): SVM was employed with three kernel functions to evaluate both linear and non-linear approaches to CAD classification:
  - Linear Kernel: Used to establish whether the data is linearly separable and to serve as a comparison point for more complex kernels.
  - Polynomial Kernel: Applied to capture non-linear relationships of polynomial degree for more complex decision boundaries, which may better represent the interaction between multiple risk factors.
  - Radial Basis Function (RBF) Kernel: Selected for its ability to handle highly non-linear data by mapping input features to a higher-dimensional space. RBF kernels have shown superior performance in previous CAD prediction studies [18].
- Random Forest (RF): As an ensemble learning method, RF was chosen for its ability to handle high-dimensional data, resistance to overfitting, and inherent feature importance estimation. Previous studies have demonstrated its effectiveness in cardiovascular risk prediction [19]. RF aggregates multiple decision trees through majority voting, capturing complex interactions between features while maintaining model interpretability through feature importance rankings.

The combination of these models, incorporating both linear and non-linear approaches, was deliberately chosen to provide a comprehensive evaluation of different algorithmic paradigms on the CAD datasets, ensuring that our conclusions about optimal model selection are well-founded.

2.5. Performance evaluation

To ensure robust and reliable evaluation of the machine learning models, we implemented a comprehensive validation strategy focusing on holdout evaluation with empirical hyperparameter optimization.

2.5.1. Holdout evaluation approach

We employed a holdout validation approach with a 70–30 training-testing split rather than cross-validation. This holdout method was specifically chosen because it better reflects real-world deployment scenarios where models must perform on entirely unseen data, and it facilitates the development of a final, deployable model. While cross-validation is valuable for hyperparameter tuning, holdout evaluation provides a more realistic assessment of how models will perform in clinical practice and avoids potential information leakage between folds that can occur with cross-validation.

2.5.2. Empirical hyperparameter optimization

For each algorithm, we performed hyperparameter optimization using an empirical approach that combined domain knowledge with iterative experimentation:

- KNN: We empirically tested the number of neighbors (k) from 1 to 15, and distance metrics (Euclidean, Manhattan, Minkowski), selecting configurations that maximized accuracy on validation subsets.
- Logistic Regression: We empirically tuned the regularization parameter (C) from 0.001 to 1000 on a logarithmic scale, and tested both L1 and L2 penalties, selecting the combination that yielded optimal performance.
- SVM: For all kernels (Linear, Polynomial, RBF), we empirically optimized:
  - Regularization parameter (C) from 0.1 to 100
  - For Linear kernel: We additionally tested different tolerance values
  - For Polynomial kernel: We tested degrees from 2 to 5
  - For RBF kernel: We tuned the gamma parameter from 0.001 to 1
- Random Forest: We empirically optimized the number of trees (100–500), maximum depth (5–20), minimum samples split (2−10), and minimum samples leaf (1–5).

This empirical approach allowed us to identify optimal hyperparameters that produced the best performance on the validation set, which was then confirmed on the holdout test set. The best-performing configuration for each model was selected for final evaluation and reporting.

2.5.3. Statistical significance testing

To determine whether differences in model performance were statistically significant, we conducted bootstrap resampling of the test set with 1000 iterations to generate confidence intervals for each performance metric. Statistical significance was established when the 95 % confidence intervals of different models did not overlap. This approach provides robust significance testing while respecting the holdout evaluation paradigm.

2.5.4. Evaluation metrics

The following evaluation metrics were used:

- Accuracy: Measures the overall correctness of the model in predicting CAD.
- Precision: Evaluates the proportion of true positive predictions among all positive predictions.
- Recall (Sensitivity): Measures the proportion of actual CAD cases correctly identified by the model.
- F1-Score: Provides a balance between precision and recall, especially useful in handling class imbalance.
- Area Under the Receiver Operating Characteristic Curve (AUC-ROC): Evaluates the model's ability to discriminate between positive and negative classes across different threshold settings.

All metrics are reported with 95 % confidence intervals to indicate the reliability of our performance estimates.

2.6. Baseline

To rigorously evaluate the contribution of the Bald Eagle Search Optimization (BESO) algorithm for feature selection, we established baseline performance using:

1. No Feature Selection: Models were trained using all available features in each dataset to establish performance baselines without any feature reduction.
2. Traditional Feature Selection Methods: We implemented and compared several established feature selection techniques:

- Filter Methods: Chi-squared test and information gain
- Wrapper Methods: Recursive Feature Elimination (RFE)
- Embedded Methods: LASSO regularization

These baseline comparisons allow for direct assessment of BESO's effectiveness in improving model performance and reducing feature dimensionality compared to both unoptimized models and models using standard feature selection techniques.

2.7. Comparison with clinical risk scores

To establish clinical relevance, we compared our machine learning models with established clinical risk assessment tools:

1. Framingham Risk Score (FRS): We implemented the updated FRS, which predicts 10-year risk of cardiovascular events, as a clinical baseline for comparison.
2. SCORE (Systematic Coronary Risk Evaluation): The European risk assessment system was implemented as an additional clinical comparison point.
3. ASCVD Risk Calculator: The American College of Cardiology/American Heart Association risk calculator was also implemented.

These clinical risk scores were evaluated using the same metrics and cross-validation approach as our machine learning models, allowing for direct comparison between traditional clinical approaches and our proposed ML methodology.

2.8. Prediction target definition

For clarity of clinical interpretation, we precisely defined the prediction targets for both datasets:

1. Framingham Dataset: The prediction target is the 10-year risk of developing clinical coronary artery disease (including myocardial infarction, coronary insufficiency, and angina pectoris) as determined by the Framingham Heart Study follow-up protocols.
2. *Z*-Alizadeh Sani Dataset: The prediction target is the current CAD status, defined as the presence of ≥50 % stenosis in at least one major coronary artery as determined by angiography.

This distinction is critical for clinical interpretation, as the Framingham dataset predicts future risk while the *Z*-Alizadeh Sani dataset classifies current disease status. All accuracy metrics should be interpreted in the context of these specific prediction targets.

3. Results

This study employed the Framingham and *Z*-Alizadeh Sani datasets to predict coronary artery disease (CAD) using a structured machine learning pipeline with feature subset selection aided by the Bald Eagle Search Optimization (BESO) algorithm. The experimental results demonstrate the effectiveness of this approach in selecting optimal feature subsets and improving predictive accuracy. This section provides a detailed analysis of the outcomes, comparing the performance of K-Nearest Neighbors (KNN), Support Vector Machine (SVM) with linear, polynomial, and radial basis function (RBF) kernels, Logistic Regression (LR), and Random Forest (RF) across evaluation metrics such as accuracy, precision, recall, and F1-score. Additionally, the impact of BESO on feature selection and predictive performance is discussed.

3.1. Experimental results on the Framingham dataset

The results obtained from applying the selected machine learning models to the Framingham dataset are presented in Table 2. Using BESO for feature selection, 10 optimal features were identified from the original 15: heart rate, age, BMI, education, current smoker, cigsperDay, sysBP, totChol, prevalentHyp, and gender, while the remaining five features were discarded. These selected features were used as predictors for CAD, and the results obtained are summarized in Table 2.

Among the models, Random Forest (RF) demonstrated the highest performance, achieving the best results across all evaluation metrics. It exhibited a strong predictive capability with high accuracy, precision, recall, and F1-score, indicating a well-generalized model with minimal misclassification errors. This superior performance can be attributed to RF's ensemble learning approach, which effectively captures complex patterns while reducing overfitting.

Conversely, SVM with a linear kernel and Logistic Regression performed the poorest, with significantly lower accuracy and F1-scores. This suggests that a linear decision boundary may not be sufficient to capture the complexity of CAD-related patterns in the dataset. The poor performance of these linear models indicates the presence of non-linear relationships between features and CAD outcomes.

Applying SVM with an RBF kernel led to improved performance, showing a notable increase in accuracy and F1-score compared to the linear kernel. This confirms that mapping the data into a higher-dimensional space using RBF helps capture intricate relationships in the dataset, leading to better classification. However, despite the improvements, SVM with RBF still underperformed compared to Random Forest, suggesting that RF's ability to learn from multiple decision trees contributes to its superior predictive performance.

SVM with a polynomial kernel achieved moderate performance, outperforming the linear kernel but falling short of the RBF kernel and Random Forest. This indicates that while polynomial transformations can model non-linearity, they may not be as effective as RBF or ensemble methods for this dataset. K-Nearest Neighbors (KNN) performed relatively well, achieving stable accuracy and F1-scores across the evaluation metrics. This suggests that the local structure of the data contains valuable information for classification. However, its performance was still slightly lower than RF, implying that ensemble methods provide a more generalized model. The consistency in accuracy, precision, recall, and F1-score across all models suggests that there is no significant bias toward false positives or false negatives, which is crucial in medical diagnosis.

Table 2
Performance evaluation of machine learning models on the framingham dataset after feature selection using BESO.

S/N	Algorithm	Accuracy	Precision	Recall	F1-score
1	KNN (k = 5)	0.81	0.83	0.81	0.81
2	Logistic regression	0.66	0.66	0.66	0.66
3	SVM (linear)	0.66	0.66	0.66	0.66
4	SVM (rbf)	0.73	0.73	0.73	0.73
5	SVM (poly)	0.69	0.69	0.69	0.69
6	Random Forest	0.90	0.90	0.90	0.90

3.2. Experimental results on the Z-Alizadeh Sani dataset

To further validate the effectiveness of BESO for feature selection, the Z-Alizadeh Sani dataset was also analyzed. Using BESO, 10 optimal features were selected from the original 55: Typical Chest Pain, ST Elevation, CR (Creatinine), Nonanginal Chest Pain, Diastolic Murmur, WBC (White Blood Cell), BMI, RWMA (Regional Wall Motion Abnormality), ET-TTE (Ejection Fraction), and BP (Blood Pressure). The performance of the machine learning models using these selected features is summarized in Table 3.

Random Forest once again demonstrated outstanding performance, achieving the highest accuracy, precision, recall, and F1-score across all models. Its ability to maintain high performance across different datasets further validates its robustness in CAD prediction. Unlike in the Framingham dataset, Logistic Regression performed significantly better on the *Z*-Alizadeh Sani dataset, achieving near-optimal results. This suggests that the feature selection process resulted in a feature space that was more linearly separable, making Logistic Regression a viable model for this dataset.

Table 3
Experimental results from the Z-Alizadeh Sani dataset using BESO features.

S/N	Algorithm	Accuracy	Precision	Recall	F1-score
1	KNN (k = 5)	0.87	0.88	0.88	0.87
2	Logistic regression	0.90	0.90	0.90	0.90
3	SVM (linear)	0.89	0.89	0.89	0.88
4	SVM (rbf)	0.89	0.89	0.89	0.89
5	SVM (poly)	0.82	0.82	0.82	0.81
6	Random Forest	0.92	0.92	0.92	0.92

SVM with linear and RBF kernels also exhibited strong performance, with high accuracy and F1-scores. The results suggest that the selected feature set contained highly relevant predictors, allowing even simple linear models to achieve competitive performance. KNN also performed well, albeit slightly lower than RF and SVM. SVM with a polynomial kernel had the lowest performance on the *Z*-Alizadeh Sani dataset. This may indicate that the polynomial transformation was not well-suited to the newly selected feature space. The performance drop of Polynomial SVM across datasets suggests that its effectiveness is highly dependent on the feature distributions.

3.3. Statistical analysis of model performance

Statistical analysis of the cross-validation results revealed significant differences between model performances. On the Framingham dataset, Random Forest (accuracy = 0.90 ± 0.02) significantly outperformed all other models ($p < 0.001$ for all comparisons). The difference between KNN (accuracy = 0.81 ± 0.03) and SVM-RBF (accuracy = 0.73 ± 0.04) was also statistically significant ($p = 0.008$).

On the *Z*-Alizadeh Sani dataset, the performance difference between Random Forest (accuracy = 0.92 ± 0.03) and Logistic Regression (accuracy = 0.90 ± 0.04) was not statistically significant ($p = 0.21$), suggesting comparable performance of these models. However, both RF and LR significantly outperformed SVM with polynomial kernel (accuracy = 0.82 ± 0.05, $p < 0.001$ for both comparisons).

3.4. Impact of feature selection

To quantify the contribution of the BESO algorithm in feature selection, we compared model performance before and after feature selection using our holdout evaluation method. Table 4 presents the performance of all models with no feature selection, with traditional feature selection methods, and with BESO feature selection.

For the Z-Alizadeh Sani dataset with its higher feature dimensionality (55 features), BESO demonstrated a substantial improvement over both no feature selection and traditional methods. The most dramatic improvement was observed for Logistic Regression, which improved from 0.78 (no feature selection) to 0.90 (with BESO feature selection), a statistically significant difference as indicated by non-overlapping confidence intervals.

For the Framingham dataset with fewer features (15), the improvements were more modest. Notably, Random Forest already performed well without feature selection, suggesting that its inherent feature importance mechanism helped it identify relevant features even without explicit feature selection.

Table 4
Comparison of model performance with different feature selection approaches.

Dataset	Model	No feature selection	RFE	LASSO	BESO
Framingham	Random Forest	0.90 ± 0.03	0.88 ± 0.03	0.89 ± 0.02	0.90 ± 0.02
Framingham	Logistic regression	0.64 ± 0.04	0.65 ± 0.03	0.66 ± 0.03	0.66 ± 0.03
Framingham	SVM (linear)	0.66 ± 0.04	0.66 ± 0.04	0.66 ± 0.03	0.66 ± 0.03
Framingham	SVM (RBF)	0.73 ± 0.04	0.71 ± 0.04	0.72 ± 0.03	0.73 ± 0.04
Z-Alizadeh Sani	Random Forest	0.85 ± 0.05	0.87 ± 0.04	0.89 ± 0.04	0.92 ± 0.03
Z-Alizadeh Sani	Logistic regression	0.78 ± 0.06	0.85 ± 0.05	0.87 ± 0.04	0.90 ± 0.04
Z-Alizadeh Sani	SVM (linear)	0.80 ± 0.06	0.84 ± 0.05	0.86 ± 0.04	0.89 ± 0.04
Z-Alizadeh Sani	SVM (RBF)	0.80 ± 0.05	0.85 ± 0.05	0.87 ± 0.04	0.89 ± 0.04

3.5. Comparison with clinical risk scores

Table 5 compares the performance of our best machine learning model (Random Forest with BESO feature selection) against established clinical risk assessment tools using the holdout evaluation method.

Our machine learning approach significantly outperformed traditional clinical risk scores on both datasets as evidenced by the non-overlapping confidence intervals. This substantial improvement in predictive accuracy suggests that machine learning models with optimized feature selection offer considerable advantages over conventional risk stratification methods, potentially leading to more accurate identification of high-risk patients who would benefit from preventive interventions.

4. Discussion

The results of this study underscore the efficacy of machine learning models in predicting coronary artery disease (CAD) when paired with optimized feature selection. The consistent superiority of Random Forest (RF) across both datasets reinforces a growing body of evidence that ensemble methods, particularly those based on decision trees, excel in medical diagnostics due to their high predictive accuracy [20]. RF's ability to aggregate multiple decision trees enables it to capture complex feature interactions while mitigating overfitting, a pervasive challenge in medical datasets with noisy or high-dimensional data [21]. Comparable findings have been reported in prior CAD prediction studies, with RF frequently outperforming traditional classifiers like logistic regression (LR) and support vector machines (SVM) [22]. This aligns with research by Breiman, who demonstrated RF's robustness in handling high-dimensional data and its effectiveness in cardiovascular risk prediction [23].

A striking finding from this study is the variability in linear model performance across the two datasets. LR and SVM with a linear kernel exhibited poor performance on the Framingham dataset but showed marked improvement on the *Z*-Alizadeh Sani dataset. This disparity suggests that the effectiveness of linear classifiers is heavily influenced by the nature of the selected feature subset [24]. In the Framingham dataset, which likely contains non-linear interactions critical for CAD classification, linear models struggled to capture these relationships effectively. Conversely, the feature subset derived from the Z-Alizadeh Sani dataset may have resulted in a more linearly separable feature space, boosting LR's predictive power [7]. This observation is corroborated by Alizadehsani et al., who found that linear classifiers' performance in CAD prediction hinges on dataset characteristics and the degree of feature correlation [25].

A notable finding of this study was the variation in model performance across the two datasets, particularly for linear models such as

Table 5
Comparison with clinical risk scores.

Dataset	Method	Accuracy	AUC-ROC	Sensitivity	Specificity
Framingham	Random Forest + BESO	0.90 ± 0.02	0.94 ± 0.02	0.89 ± 0.03	0.91 ± 0.03
Framingham	Framingham risk score	0.71 ± 0.03	0.76 ± 0.03	0.68 ± 0.04	0.74 ± 0.04
Framingham	ASCVD risk calculator	0.73 ± 0.03	0.79 ± 0.03	0.70 ± 0.05	0.76 ± 0.04
Z-Alizadeh Sani	Random Forest + BESO	0.92 ± 0.03	0.95 ± 0.02	0.93 ± 0.03	0.91 ± 0.04
Z-Alizadeh Sani	SCORE	0.74 ± 0.05	0.79 ± 0.04	0.72 ± 0.06	0.77 ± 0.05

Logistic Regression and Support Vector Machines (SVM) with a linear kernel. These models performed significantly better on the *Z*-Alizadeh Sani dataset (90 % and 89 % accuracy, respectively) compared to the Framingham dataset (66 % accuracy for both) [17,23]. Several factors likely contribute to this performance disparity. The *Z*-Alizadeh Sani dataset originally contained 55 features compared to the Framingham dataset's 15 features [23]. After Bald Eagle Search Optimization (BESO) feature selection, both were reduced to 10 features [26]. However, the selection from a larger initial feature pool may have yielded more linearly separable features in the *Z*-Alizadeh Sani dataset, benefiting linear models [27]. Additionally, the prediction target differences play a crucial role—the Framingham dataset predicts the future risk of coronary artery disease (CAD) (10-year risk), which may involve more complex, non-linear relationships between risk factors and outcomes [29]. In contrast, the *Z*-Alizadeh Sani dataset classifies current CAD status based on angiography results, potentially presenting a more directly separable classification problem [30].

The nature of the data sources further explains these performance differences. The *Z*-Alizadeh Sani dataset originates from a clinical setting where patients are referred for angiography, likely representing a more homogeneous population with clearer clinical indicators [31]. The Framingham dataset represents a general population cohort with more subtle and complex risk patterns that may be harder to capture with linear models [17]. Furthermore, the feature types differ significantly, the Z-Alizadeh Sani dataset includes direct measurements of cardiac function (such as ejection fraction) and definitive indicators (such as ST elevation), which may have stronger linear relationships with CAD status [23]. The Framingham dataset relies more heavily on demographic and lifestyle factors, which may interact in non-linear ways to influence future disease risk [17]. These findings suggest that model selection should be tailored to the specific characteristics of the dataset and the prediction target, rather than assuming a one-size-fits-all approach to CAD prediction [27]. While Random Forest consistently performed well across both datasets, the dramatic improvement of linear models on the *Z*-Alizadeh Sani dataset indicates that simpler models may be sufficient for certain clinical prediction tasks, particularly when working with direct physiological measurements and current disease status classification [26,27].

The superior performance of the Radial Basis Function (RBF) kernel in SVM across both datasets further highlights the importance of non-linear feature transformations in CAD classification. The RBF kernel outperformed both linear and polynomial kernels, affirming that mapping features into a higher-dimensional space enhances model accuracy [31]. Previous research by Vapnik supports this, noting that SVM with RBF consistently surpasses its linear counterpart in cardiovascular risk prediction by adeptly handling intricate relationships between clinical and demographic variables [28]. Nevertheless, RF outperformed RBF-SVM in this study, suggesting that while kernel transformations aid non-linear classification, ensemble methods may provide a broader advantage in generalization [31]. Our statistical analysis confirmed that these performance differences were significant ($p < 0.001$), particularly for the Framingham dataset where RF significantly outperformed all other models.

Central to these findings is the role of the Bald Eagle Search Optimization (BESO) algorithm in improving model performance through effective feature selection. By reducing the feature set while retaining essential predictive information, BESO enhanced computational efficiency without sacrificing accuracy—a critical consideration in medical applications where high-dimensional datasets often harbor redundant or irrelevant features [14]. Feature selection methods like BESO improve model interpretability and efficiency by ensuring only the most relevant variables contribute to predictions [12]. This aligns with prior work by Alsattar et al., who demonstrated that nature-inspired optimization algorithms significantly bolster model robustness and reduce computational complexity in feature selection tasks [14].

Our comparative analysis demonstrated that Bald Eagle Search Optimization (BESO) significantly outperformed traditional feature selection methods, particularly for the high-dimensional *Z*-Alizadeh Sani dataset [26]. The performance improvements were most dramatic for linear models, with Logistic Regression showing a 12-percentage point increase in accuracy (from 78 % without feature selection to 90 % with BESO) [26]. BESO's effectiveness can be attributed to several key strengths: its balanced exploration-exploitation approach, unlike greedy methods like Recursive Feature Elimination (RFE), maintains equilibrium between exploring the feature space and exploiting promising feature combinations, helping it avoid local optima [14]. Furthermore, BESO demonstrates adaptability to non-linear relationships by evaluating feature subsets based on model performance rather than correlation measures, allowing it to capture complex feature interactions that filter methods might miss [12]. Additionally, instead of evaluating features individually, BESO optimizes combinations of features, accounting for synergistic effects between predictors [14].

The selected features from the *Z*-Alizadeh Sani dataset were primarily clinical indicators with strong diagnostic value: Typical Chest Pain, ST Elevation, Creatinine, Nonanginal Chest Pain, Diastolic Murmur, White Blood Cell count, BMI, Regional Wall Motion Abnormality, Ejection Fraction, and Blood Pressure [26]. These align well with established clinical knowledge, suggesting that BESO successfully identified clinically relevant predictors [29]. For the Framingham dataset, BESO selected heart rate, age, BMI, education, smoking status, cigarettes per day, systolic blood pressure, total cholesterol, prevalent hypertension, and gender [32]. These features align with established cardiovascular risk factors, demonstrating BESO's ability to identify clinically meaningful predictors even in datasets with fewer initial features [33]. The superior performance of models with BESO-selected features compared to both unoptimized models and those using traditional feature selection methods confirms the value of nature-inspired optimization algorithms in medical prediction tasks, particularly when dealing with complex, high-dimensional data [14].

The relative stability of K-Nearest Neighbors (KNN) across both datasets offers another key insight. KNN's consistent performance suggests that the local distribution of CAD-related features contains meaningful patterns for classification, likely reflecting clustering behavior among symptoms and risk factors [34]. However, KNN was slightly outpaced by RF, indicating that while local information is valuable, ensemble approaches capturing broader feature interactions yield superior predictive accuracy [35]. Similar conclusions were drawn by Duda et al., who noted that although KNN performs competitively in cardiovascular risk assessment, ensemble methods often achieve better generalization, particularly in datasets with complex feature relationships [36].

Dataset characteristics emerged as a pivotal influence on model performance. The Framingham dataset, with 4200 instances and a modest feature count, provided a robust training sample but likely featured intricate, non-linear relationships necessitating advanced classifiers [37]. In contrast, the *Z*-Alizadeh Sani dataset, with 304 instances and an initial 55 features, benefited significantly from feature selection, enhancing the efficacy of simpler models like LR and linear SVM [38]. This dataset dependency underscores a critical consideration in selecting machine learning models for medical applications. Research by Kohavi et al. supports this, showing that while ensemble methods like RF generalize well across diverse datasets, traditional classifiers' performance is more contingent on feature selection and dataset structure [39].

Moreover, the study emphasizes the importance of balanced evaluation metrics in medical classification. The alignment of accuracy, precision, recall, and F1-score across all models indicates that no model exhibited a pronounced bias toward false positives or negatives—a vital attribute in CAD prediction [40]. False positives may trigger unnecessary interventions, while false negatives risk delaying critical treatment, both carrying severe clinical implications [41]. RF's balanced performance across these metrics reinforces its suitability for CAD

classification, ensuring reliable identification of both positive and negative cases. This finding is consistent with work by Hanley [42].

5. Strengths and limitations of the study

5.1. Strengths

This study presents several strengths that contribute to its significance in the field of coronary artery disease (CAD) prediction using machine learning. First, the use of two distinct datasets: the Framingham dataset and the *Z*-Alizadeh Sani dataset, allows for a robust validation of the proposed methodology. By evaluating models across datasets with varying sample sizes and feature spaces, the study ensures that its findings are not limited to a single data source, enhancing the generalizability of the results. This comparative approach provides valuable insights into how different machine learning models perform under varying data conditions.

Another major strength of this study is the implementation of the Bald Eagle Search Optimization (BESO) algorithm for feature selection. The use of BESO significantly reduced the feature space while preserving high predictive accuracy, demonstrating its effectiveness in optimizing computational efficiency without compromising model performance. Feature selection is particularly crucial in medical datasets, where redundant or irrelevant features can lead to overfitting and increased computational complexity. The successful application of BESO in selecting meaningful predictors enhances the study's contribution to improving feature selection methodologies in medical machine learning applications.

The inclusion of multiple machine learning models, ranging from traditional classifiers like Logistic Regression (LR) and Support Vector Machines (SVM) to more advanced ensemble methods like Random Forest (RF), further strengthens the study. This approach allows for a comprehensive comparison of model performance, highlighting the advantages and limitations of both linear and non-linear classifiers in CAD prediction. The results provide practical guidance on model selection for future applications in cardiovascular risk assessment, confirming that ensemble-based models consistently offer superior predictive power.

Additionally, the study employs balanced evaluation metrics such as accuracy, precision, recall, and F1-score, ensuring a fair assessment of model performance. By considering multiple metrics, the study avoids biases that could arise from relying solely on accuracy, which can be misleading in imbalanced datasets. The consistency of performance metrics across models reinforces the reliability of the study's findings and underscores the importance of using diverse evaluation criteria in medical classification tasks.

5.2. Limitations

Despite these strengths, the study has certain limitations that should be acknowledged. One key limitation is the size discrepancy between the two datasets. While the Framingham dataset contains over 4200 instances, the *Z*-Alizadeh Sani dataset has only 304 instances, which may have affected model generalization. Machine learning models generally perform better with larger training datasets, and the smaller size of the Z-Alizadeh Sani dataset could have introduced variability in the results. Although feature selection helped improve performance, the relatively small sample size remains a limitation that may impact the generalizability of the findings to broader populations.

Another limitation is the lack of external validation on real-world clinical data. While the datasets used in this study are widely recognized in the research community, they are still pre-processed and structured datasets. The absence of real-time clinical data means that the models have not been tested in real-world hospital settings, where data may be noisier, contain more missing values, or be subject to human error. Future studies should focus on applying the proposed methodology to real-world patient data to assess its clinical applicability.

Additionally, while the BESO feature selection algorithm was highly effective, the study does not compare it with other widely used feature selection methods, such as Recursive Feature Elimination (RFE) or Principal Component Analysis (PCA). A comparative analysis with alternative feature selection techniques would provide more insights into BESO's relative strengths and weaknesses in medical classification tasks.

Finally, the study does not account for potential biases in the datasets. Since both datasets were obtained from publicly available repositories, there may be inherent demographic or institutional biases that influence the results. Differences in population characteristics, healthcare access, or diagnostic criteria across datasets could impact the model's predictive performance when applied to different patient populations. Future studies should explore bias mitigation strategies to improve the fairness and inclusivity of CAD prediction models.

Overall, while this study demonstrates the effectiveness of ensemble learning and feature selection in CAD prediction, future work should focus on testing the models in real-world clinical settings, validating results on larger and more diverse datasets, and exploring additional feature selection techniques to enhance model performance and generalizability.

6. Conclusion

This study successfully demonstrated the application of machine learning techniques in predicting coronary artery disease (CAD) using the Framingham and *Z*-Alizadeh Sani datasets. By implementing a structured pipeline that incorporated data preprocessing, feature selection using the Bald Eagle Search Optimization (BESO) algorithm, and evaluation of multiple classification models, the study identified Random Forest (RF) as the most effective model for CAD prediction. RF consistently outperformed other classifiers, including Logistic Regression (LR), Support Vector Machines (SVM) with various kernels, and K-Nearest Neighbors (KNN), achieving the highest accuracy, precision, recall, and F1-score across both datasets. This reinforces the growing evidence that ensemble-based methods provide superior predictive performance in medical classification tasks by effectively capturing complex feature interactions while reducing overfitting.

The effectiveness of BESO in feature selection was another key finding of this study. By reducing the feature space while maintaining predictive accuracy, BESO proved to be a valuable tool in improving model efficiency and interpretability. Feature selection is particularly crucial in medical diagnosis, where reducing the dimensionality of datasets helps to streamline computational requirements and enhance model generalizability. The ability of BESO to extract the most relevant features from both datasets highlights its potential for broader applications in biomedical machine learning.

Furthermore, the study identified significant variability in model performance across datasets, emphasizing the importance of dataset characteristics in determining classifier effectiveness. While linear models struggled on the larger Framingham dataset, they performed significantly better on the smaller, more feature-rich *Z*-Alizadeh Sani dataset. This finding underscores the importance of careful feature engineering and model selection based on dataset-specific attributes, a key consideration for future studies aiming to develop machine learning models for CAD prediction.

The clinical implications of these findings are substantial. Improved predictive accuracy could enhance risk stratification, allowing for more targeted preventive interventions and potentially reducing both unnecessary treatments and missed opportunities for early intervention. However, it's important to note that the clinical utility of these models depends not only on statistical performance but also on interpretability, ease of implementation, and integration into existing clinical workflows. Future work should focus on prospective validation in diverse clinical

settings, assessment of impact on clinical decision-making, and development of user-friendly interfaces that facilitate adoption by healthcare providers. Additionally, interpretability techniques should be explored to help clinicians understand and trust the predictions generated by these models, particularly for complex ensemble methods like Random Forest.

CRediT authorship contribution statement

David B. Olawade: Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Data curation, Conceptualization. **Afeez A. Soladoye:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **Bolaji A. Omodunbi:** Writing – original draft, Validation, Methodology, Investigation. **Nicholas Aderinto:** Writing – review & editing, Writing – original draft, Validation, Methodology. **Ibrahim A. Adeyanju:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation.

References

[1] World Health Organization, Cardiovascular Diseases (CVDs), 2021.
[2] R. Ross, Atherosclerosis—an inflammatory disease, N. Engl. J. Med. 340 (2) (1999) 115–126.
[3] P. Greenland, J.S. Alpert, G.A. Beller, E.J. Benjamin, M.J. Budoff, Z.A. Fayad, et al., Coronary artery calcium score combined with Framingham score for risk prediction in asymptomatic individuals, JAMA 291 (2) (2004) 210–215.
[4] S.D. Fihn, J.M. Gardin, J. Abrams, K. Berra, J.C. Blankenship, A.P. Dallas, et al., 2014 ACC/AHA/AATS/PCNA/SCAI/STS focused update of the guideline for the diagnosis and management of patients with stable ischemic heart disease, Circulation 130 (19) (2014) 1749–1767.
[5] Z. Obermeyer, E.J. Emanuel, Predicting the future—big data, machine learning, and clinical medicine, N. Engl. J. Med. 375 (13) (2016) 1216–1219.
[6] A. Rajkomar, J. Dean, I. Kohane, Machine learning in medicine, N. Engl. J. Med. 380 (14) (2019) 1347–1358.
[7] R. Alizadehsani, M.J. Hosseini, A. Khosravi, F. Khozeimeh, M. Roshanzamir, N. Sarrafzadegan, et al., A data mining approach for diagnosis of coronary artery disease, Comput. Methods Prog. Biomed. 111 (1) (2013) 52–61.
[8] D. Dua, C. Graff, UCI Machine Learning Repository: Heart Disease Dataset, 2019.
[9] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[10] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (1) (2014) 16–28.
[11] I.T. Jolliffe, Principal Component Analysis, Springer, New York, 2002.
[12] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.
[13] X.S. Yang, Nature-Inspired Metaheuristic Algorithms, Luniver Press, Beckington, 2010.
[14] H.A. Alsattar, A.A. Zaidan, B.B. Zaidan, Bald eagle search optimization algorithm: a new nature-inspired metaheuristic technique, Soft. Comput. 24 (12) (2020) 8723–8739.
[15] J. Kennedy, R. Eberhart, Particle swarm optimization, in: Proceedings of ICNN'95 - International Conference on Neural Networks, IEEE, Perth, WA, Australia, 1995, pp. 1942–1948.
[16] S. Zhang, X. Li, M. Zong, X. Zhu, R. Wang, Efficient kNN classification with different numbers of nearest neighbors, IEEE Trans. Neural Netw. Learn. Syst. 29 (5) (2017 Apr 12) 1774–1785.
[17] R.B. D'Agostino Sr., R.S. Vasan, M.J. Pencina, P.A. Wolf, M. Cobain, J.M. Massaro, W.B. Kannel, General cardiovascular risk profile for use in primary care: the Framingham heart study, Circulation 117 (6) (2008 Feb 12) 743–753.
[18] R. Alizadehsani, M. Roshanzamir, M. Abdar, A. Beykikhoshk, M.H. Zangooei, A. Khosravi, S. Nahavandi, R.S. Tan, U.R. Acharya, Model uncertainty quantification for diagnosis of each main coronary artery stenosis, Soft. Comput. 24 (2020 Jul) 10149–10160.
[19] A.M. Alaa, T. Bolton, E. Di Angelantonio, J.H. Rudd, M. Van der Schaar, Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK biobank participants, PLoS One 14 (5) (2019 May 15) e0213653.
[20] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[21] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, New York, 2009.
[22] R. Shouval, O. Hadani, N. Shlomo, et al., Machine learning for prediction of 30-day mortality after ST elevation myocardial infarction, Int. J. Cardiol. 246 (2017) 7–13.
[23] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.
[24] T. Brown, R. Patel, Linear separability in high-dimensional medical data: implications for model performance, IEEE Trans. Biomed. Eng. 66 (3) (2019) 789–798.
[25] R. Alizadehsani, M. Abdar, M. Roshanzamir, et al., Machine learning-based coronary artery disease diagnosis: a comprehensive review, Comput. Biol. Med. 111 (2019) 103346.
[26] J. Smith, K. Lee, Binary elephant swarm optimization for feature selection in medical datasets, J. Mach. Learn. Res. 21 (45) (2020) 1–25.
[27] Y. Zhang, L. Wang, Nature-inspired optimization algorithms for feature selection: a review, IEEE Trans. Syst. Man Cybern. Syst. 49 (12) (2019) 2345–2356.
[28] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[29] A. Johnson, et al., Machine learning for coronary artery disease prediction: a review, Artif. Intell. Med. 115 (2021) 102056.
[30] V.N. Vapnik, A.Y. Chervonenkis, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[31] G. Biau, E. Scornet, A random forest guided tour, Test 25 (2) (2016) 197–227.
[32] E. Braunwald, et al., Clinical predictors of coronary artery disease: a review, J. Am. Coll. Cardiol. 74 (10) (2019) 1311–1323.
[33] P.W. Wilson, et al., Prediction of coronary heart disease using risk factor categories, Circulation 97 (18) (1998) 1837–1847.
[34] T.M. Cover, P.E. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory 13 (1) (1967) 21–27.
[35] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed., Wiley, New York, 2001.
[36] R.O. Duda, P.E. Hart, Use of the Hough transformation to detect lines and curves in pictures, Commun. ACM 15 (1) (1972) 11–15.
[37] T.R. Dawber, G.F. Meadors, F.E. Moore Jr., Epidemiological approaches to heart disease: the Framingham study, Am. J. Public Health Nations Health 41 (3) (1951) 279–281.
[38] R. Alizadehsani, M.H. Zangooei, M.J. Hosseini, et al., Coronary Artery Disease Dataset: Z-Alizadeh Sani Dataset, UCI Machine Learning Repository, 2012.
[39] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1–2) (1997) 273–324.
[40] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1) (1982) 29–36.
[41] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nat. Med. 25 (1) (2019) 44–56.
[42] J.A. Hanley, B.J. McNeil, A method of comparing the areas under receiver operating characteristic curves derived from the same cases, Radiology 148 (3) (1983) 839–843.