STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **15**, May 2026, pp 4255–4276.
Published online in International Academic Press (www.IAPress.org)

Artificial Intelligence and Ensemble Learning for Coronary Artery Disease Prediction

Hiba Dhia Jaafar[1], Teba Jabbar Hassan[1], Marwa Hussien Mohamed [2,*],Baesher Abdullateff Abad[1], Sara Salman Qasim[3]

[1] *Medical and Laboratory Techniques Department, Health and Medical Technical College, Imam Jaafar Al-Sadiq University, Baghdad, 10081 IRAQ, hiba.dhia90@gmail.com, tebajabbar9888@gmail.com, bashaeraltimem@gmail.com*
[2] *Computer Technology Engineering Department, Engineering Technologies College, Al-Esraa University, Baghdad, 10081 IRAQ, maraw@esraa.edu.iq,eng_maroo1@yahoo.com*
[3] *Intelligent Medical Systems Department, Science College, Al-Esraa University, Baghdad, 10081 IRAQ, sarah@esraa.edu.iq*

Abstract Coronary artery disease (CAD), being one of the leading causes of death globally, demands efficient and early diagnosis to improve patient outcomes and minimize unnecessary medical procedures. In recent times, machine learning (ML) has been recognized as a promising non-invasive tool to diagnose CAD and develop preventive healthcare measures. In this paper, the effectiveness of various ML algorithms in predicting CAD is thoroughly investigated, including neural networks, decision trees, support vector machines, and ensemble methods such as Random Forest and XGBoost, using publicly available medical datasets. To address class imbalance and improve the reliability of the model, data balancing methods such as SMOTE and ADASYN are employed in the preprocessing stage. The experimental results show that ensemble methods perform better than individual classifiers; among them, XGBoost shows a high accuracy of 94.7%, followed by Random Forest with 92.04%. Furthermore, the application of data balancing methods has improved recall and specificity, thus enhancing the robustness of the model in diagnosis. This shows that well-chosen and well-preprocessed machine learning models can be effective tools in the early detection of CAD, providing a cost-effective alternative to invasive tests while promoting better patient care and the prevention of cardiovascular disease.

Keywords cardiovascular disease (CVD), CAD (coronary artery disease), Heart failure, data mining, Machine learning, classification.

DOI: 10.19139/soic-2310-5070-3354

1. Introduction

*C*VD is defined as a collection of heart conditions affecting the heart, like heart attacks, strokes, and CAD. It was important cause of death worldwide in 2020, according to the World Health Organization (WHO). Also, it caused 173,871 deaths in Egypt, representing 32.4% of all deaths. Egypt's age-standardized mortality rate of 268.11 per 100,000 population ranks the nation 15th globally for CVD mortality. CAD is caused by more than 50% coronary artery blockage, which supplies the heart with oxygenated blood[1]. Figure 1 provides a comparative overview of the major diseases responsible for mortality and their percentages (statistics obtained from https://professional.heart.org/, accessed on 1 January 2025). Cardiovascular diseases lead to the highest percentage of deaths, followed by deaths due to various other diseases such as cancer, respiratory diseases, and infectious diseases. Under cardiovascular diseases, coronary artery disease plays a significant role in causing a high percentage of deaths, highlighting its importance in global mortality statistics. The percentage of various diseases varies greatly in different geographical locations, with a high percentage of deaths due to cardiovascular diseases

*Correspondence to: Marwa Hussien Mohamed (Email: maraw@esraa.edu.iq, eng_maroo1@yahoo.com). Computer Technology Engineering Department, Engineering Technologies College, Al-Esraa University, Baghdad, 10081 IRAQ.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

in Asia, Russia, and the Middle East. The difference in percentages of various diseases in different locations is due to various factors such as lifestyle, socioeconomic status, and preventive measures adopted in different locations. The geographical distribution of various diseases, as depicted in Figure 1, emphasizes the need to adopt effective prevention and early diagnosis of various diseases, which may help reduce the percentage of deaths due to coronary artery disease.

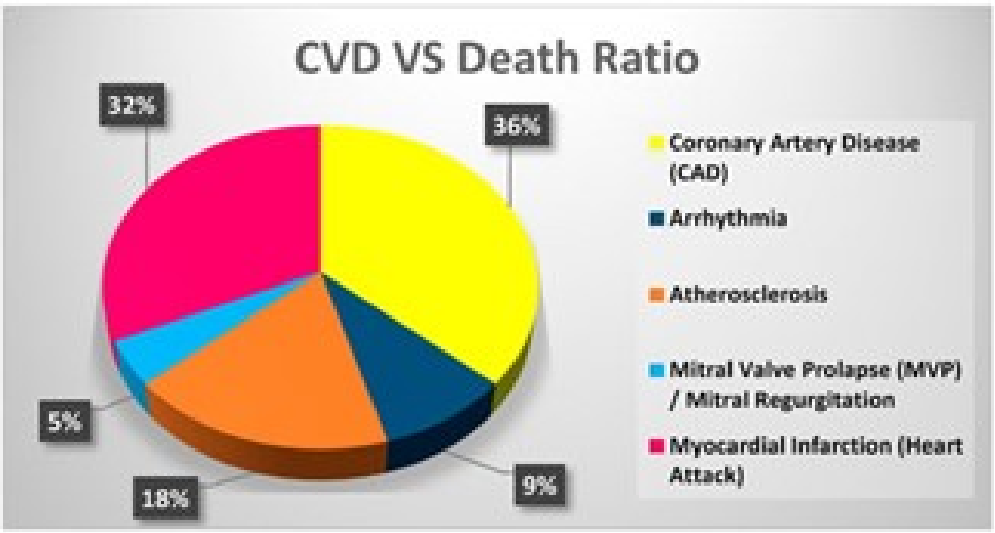

Figure 1. the different diseases and number of deaths.

Early diagnosis of CAD is crucial in an effort to prevent the disease's progression as well as reduce the risks associated with it. Although it is an invasive procedure with risks such as artery dissection, arrhythmia, and even death. Furthermore, image-based diagnostic procedures are costly and inconvenient for large-scale screening, particularly in Third World countries[2]. The drawbacks have motivated the quest for alternative non-invasive, low-cost, and most effective tools for the early diagnosis of CAD, and ML was the best tool to predict this disease with high accuracy at an earlier phase, which makes timely intervention possible. Furthermore, ML-based methods can reduce healthcare spending by enabling mass screening of populations without the requirement of expensive and invasive diagnostic processes. Several risk factors [3]are involved in CAD development, including physical inactivity, obesity, unhealthy diet, smoking, diabetes, high blood pressure, family history of heart disease, and others. Symptoms of CAD include angina (chest pain), shortness of breath, and fatigue, and the majority of heart attacks occur as the initial presentation of the disease. CAD treatment involves lifestyle changes, such as following a healthy diet, being active, and keeping risk factor control, such as managing cholesterol and hypertension. Early detection through non-invasive methods can significantly reduce the incidence of heart attacks and improve the lifestyle of individuals at risk of CAD. In this case, ML can be used to identify patterns and predict the likelihood of disease in individuals who might not yet exhibit noticeable symptoms. Figure 2 shows the CAD.

Usage of machine learning technology in CAD prediction is increasingly emerging as a possibility because it helps to examine bulky amounts of medical data and bring out patterns that one cannot be aware of Machine learning algorithms, especially classification models, can work upon huge amounts of clinical data including age, sex, blood pressure, cholesterol value, electrocardiogram information, and heredity to provide a prediction on the development of CAD. These models are trained to recognize fine patterns in the data that may not be immediately apparent to clinicians. Through this, ML can facilitate early diagnosis and personalized treatment protocols, ultimately leading to better patient outcomes. Decision trees, support vector machines, and neural networks have been applied to classify patients as high-risk or low-risk, aiding in decision-making regarding further testing or intervention. Despite the positive promise of machine learning in CAD prediction, several problems still exist. Quality and amount of medical information form a central problem. Inconsistencies and lacking details may be

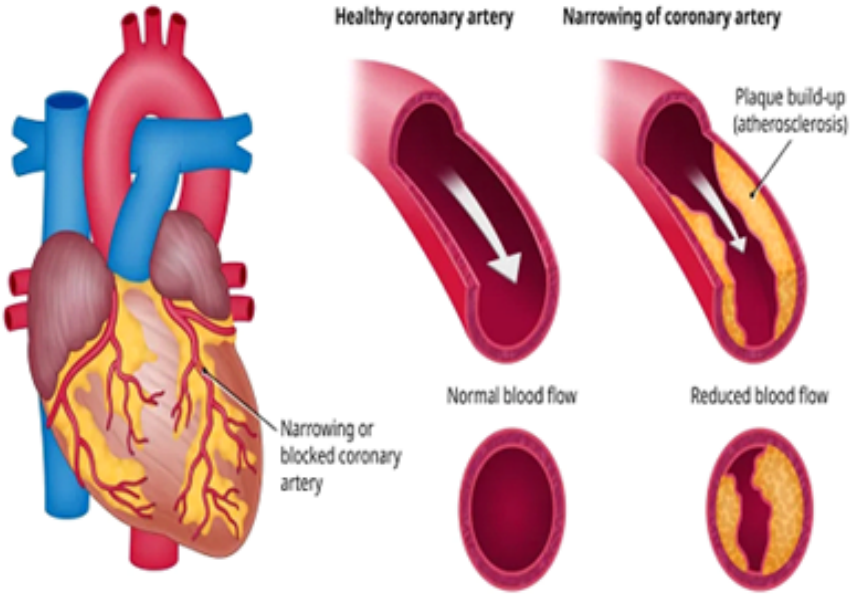

Figure 2. CAD Coronary artery disease [3].

added to data accumulated from various sources, which disrupt the accuracy of forecast models[4]. Apart from that, the interpretability of ML models used in CAD towards clinical acceptance is essential. The majority of machine learning algorithms, particularly complex ones like deep neural networks, are ”black boxes,” and it is difficult for clinicians to get a transparent idea of how predictions are being made. This transparency can limit trust in the recommendation provided by the model. Therefore, explainable and interpretable AI models are more highly desired, and they provide not only good predictions but also explanations regarding the logic behind the model's decision[5].

In recent times, ML algorithms have been extensively used to predict CAD risk based on the analysis of clinical data and the discovery of underlying correlations between various factors and disease outcomes [6]. The most frequently used algorithms include decision trees, support vector machines, k-nearest neighbors, and neural networks, which use patient data like blood pressure, cholesterol levels, and family history to predict the probability of CAD. These algorithms help medical practitioners make informed decisions and provide personalized medical care. However, several issues need to be addressed. For instance, medical data sets are often incomplete and inconsistent, and may impact prediction accuracy. This highlights the importance of well-curated medical data sets that are representative of the population [7]. Additionally, it is vital to enhance the interpretability of ML algorithms to gain the trust of medical practitioners and encourage their use in real-world scenarios. Therefore, transparent and explainable AI models that justify their predictions are necessary to integrate ML algorithms into medical practice [8].

This paper is prepared as follows: Section 2 talks about an overview of CAD and introduces the role of machine learning in its diagnosis. In Section 3, we explore various artificial intelligence (AI) techniques applied in the diagnosis of CVD. Section 4 presents a detailed discussion of the ML algorithms commonly used in CAD prediction. Section 5 outlines the public medical datasets that have been employed in recent studies for predicting CAD, focusing on related works. Section 6 reviews relevant literature and summarizes key related research. Section 7 Discussion and Experimental Results, Section 8 emphasizes the latest progress in artificial intelligence and machine learning for detecting and diagnosing CAD. At the end, Section 9 summarizes the paper and suggests more valuable areas for future work.

2. Background of Cad disease and machine learning

In this Background section, understand the significance of the problem (CAD), the shortcomings of current diagnostic methods, the application of ML in medicine, and the expected benefits and limitations in using ML for CAD prediction.

1. Current Diagnostic Methods for CAD As has been discussed in the Introduction, the traditionally used methods for CAD diagnosis, like coronary angiography, are not suitable for large-scale and early diagnosis due to their invasive nature. Therefore, more emphasis has been placed on the use of additional non-invasive methods for CAD diagnosis and early detection [9, 10]. Modern imaging techniques like CT and MRI have shown their diagnostic potential, although they are costly and not feasible for large-scale use.
2. Risk Factors for CAD Certain risk factors are the cause of the development of CAD. These are fewer features, not highly risk factors such as age, gender, and family history, and modifiable risk factors such as high cholesterol, high blood pressure, diabetes, smoking, and physical inactivity. Obesity, a bad diet, and too much stress are also the main causes of the disease. Figure 3 shows these factors[11]. Having all these risk factors increases the chances of there being the formation of plaque in the coronary arteries, thus atherosclerosis, and ultimately CAD. Screening is required to identify individuals at risk so that prevention of the formation of the disease can be prevented.

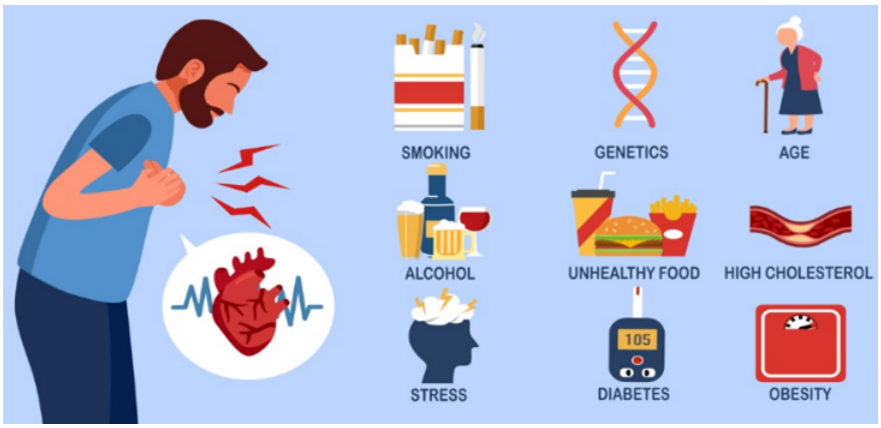

Figure 3. CAD Coronary artery disease Risk Factors

3. Machine Learning in CAD Prediction ML models in CAD have been utilized to predict disease occurrence probability on a large variety of medical data, including clinical features, laboratory results, and demographics[12]. By analyzing this data, ML algorithms can provide insights into an individual's risk profile and enable healthcare professionals to make better decisions regarding further testing and treatment. ML-based approaches can offer non-invasive early diagnosis at low cost with minimal reliance on invasive tests and improved patient outcomes. Moreover, machine learning algorithms can aid in the recognition of high-risk patients who could be treated using preventive interventions or focused interventions before disease development.
4. Opportunities and Challenges of Applying ML for CAD Even though ML is full of possibilities for enhancing CAD diagnosis, some challenges must be addressed. The availability and quality of medical data are one of the significant challenges. Incorrect, incomplete, or biased data can have highly effects on the performance of ML models[13]. Interpretability of machine learning models is also a problem because most algorithms are "black boxes" and the clinicians might not understand how the decisions are being made. To overcome these challenges, greater emphasis is given to developing explainable and transparent AI models that not only provide accurate predictions but also articulate a transparent justification for them[14].

3. MEDICAL DATASETS USED

Datasets for heart disease prediction are collected from a publicly available free online repository. We select these datasets because other field authors commonly use them, and a brief description of each one follows [15]. Table 1 shows the number of features in every dataset. These data sets offer richly documented patient records and accommodate a broad set of attributes that are beneficial in the development of predictive models for CAD. They are available and consistent and have proven to be of great benefit to machine learning researchers and those forecasting cardiovascular disease.

Data Quality and Generalization:

The performance of the machine learning model, specifically for CAD prediction, is largely dependent on the quality, quantity, and representativeness of the dataset [16, 17]. Although publicly accessible datasets such as Cleveland, Z-Alizadeh Sani, and Hungarian contain high-quality clinical and demographic data, they have certain limitations, including differences in sample sizes, feature sets, and class distributions.

This may result in overestimation of the model performance when it is validated using a smaller dataset, which may not generalize well to a larger, more heterogeneous population of patients. Moreover, class imbalance, missing values, and data pre-processing also affect the model's reliability.

Although certain techniques, such as SMOTE, ADASYN, and feature selection methods, have partially addressed these problems, they cannot completely mimic real-world variability. Hence, it is very likely that a model with high accuracy on these datasets may not perform as well when applied to real-world scenarios, especially when dealing with a heterogeneous population of patients from different countries.

It is also very important to curate the dataset, specifically to include a heterogeneous population, and validate it externally, ideally from different centers, to obtain a reliable model.

Table 1. Comparison of commonly used CAD datasets

Dataset	Instances	Features	Class Balance (CHD / Healthy)	Common Uses
z-Alizadeh Sani	303	54	216 / 87	CAD prediction; demographic, clinical, ECG, lab, and echo data; no missing values; used for training predictive models [16]
Statlog	261	75 (often 13 used)	114 / 147	Binary classification for CHD; clean dataset with no missing values; widely used in ML experiments [17]
Cleveland	303	13 (selected from 76)	Simplified to 1 / 0	Most popular dataset for CAD prediction; mixed numerical and categorical features; widely benchmarked [18]
Hungarian	294	13	106 / 188	Small, clean dataset; used for testing and comparing classification models; clearly defined binary classes [19]

4. MACHINE LEARNING ALGORITHMS

We will explore and compare the performance of several machine learning methods for predicting coronary artery disease (CAD) [20]. By utilizing these techniques, we aim to highlight their effectiveness in classifying CAD risk based on various patient features, such as clinical factors and medical history. In the following sections, we will present the prediction results for CAD using these algorithms, showcasing their accuracy, precision, and reliability in detecting patients at risk [21]. Through this comparison, we seek to identify the most promising techniques for

early CAD detection, providing insights into how machine learning can improve cardiovascular disease diagnosis and prevention.

An overview of commonly applied ML algorithms for CAD prediction is given below.

1. Supervised Learning (SL)
   SL is training a model on data with labels, where inputs and their corresponding outputs are known. SL can be applied to regression (predicting continuous values) and classification (predicting categories) [22]. Supervised learning can be used in CAD prediction to classify patients based on their risk factors.
   (a) Linear Regression Linear regression is applied to forecast a continuous target variable from one or more independent variables. It assists in forecasting CAD severity by studying clinical factors and healthy people outcomes' relationships.
   (b) Logistic Regression Logistic regression is applied in binary classification problems, like diagnosing whether a patient has CAD or not. It gives the probability of every potential result (e.g., "Yes" or "No").
2. K-Nearest Neighbor (KNN)
   KNN is a simple, intuitive algorithm that uses classification data points into the majority class of their closest neighbors. In CAD prediction, KNN can find similar patients and predict whether a new patient is at risk based on similarities with others.
3. Decision Trees (DT)
   DT [23] creates a tree-structured model wherein each node maps to a decision based on some attribute. Decision Trees divide the data by applying the most appropriate features to produce predictions or classifications, such as whether a patient suffers from CAD. The method provides easy-to-read results that can be deciphered to enhance decision-making among health workers.
4. Support Vector Machines (SVM)
   SVM is a powerful classifier that separates data into multiple classes depending on the optimal boundary, or hyperplane. SVM can handle linear and non-linear relationships and can be used for CAD prediction in patients with complex, multi-dimensional health data.
5. Artificial Neural Networks (ANN)
   ANNs[24] are computer models of the human brain with the capacity to learn complex patterns. ANNs can accurately predict CAD by learning from large databases and comprehending complex relationships among various patient characteristics and diseases.
6. Random Forest (RF)
   RF [25, 26] is a bagging model that constructs more than one decision tree and then their output is combined. By ensemble, the impact of numerous trees is aggregated so that overfitting is avoided, and predictive accuracy is enhanced. Random Forest has broad applications in CAD prediction due to it being able to handle gigantic and complex data. RF is that it can effectively deal with missing values and noisy data, as the trees are trained on a bootstrap sample of the features and can make splits on the available features. This is important for EHRs, where there are often missing values in the records. RF can effectively deal with nonlinear relationships and interactions between features, which are common in EHRs, and provides feature importance, so that clinicians can identify the features that are most important for the predictions.
7. Extreme Gradient Boosting (XGBoost)
   XGBoost [27] is a gradient boosting algorithm that is used for prediction using an ensemble of decision trees. It has been widely used in disease prediction tasks, including CAD, due to its efficiency and high performance on complicated data.

Comparison of Common Supervised Learning Algorithms for CAD Prediction is shown in Table 2 for all algorithms with computational cost, interpretability, and data requirements. Tree-based models like Random Forest and XGBoost have been demonstrated to work effectively with structured clinical data because they can effectively deal with different data types, handle missing data, and identify non-linear relationships between features with minimal data pre-processing. These models are also robust to noisy data and can naturally incorporate the importance of features, which can provide some basic interpretability, especially in a clinical setting.

Hyperparameter tuning is also an essential factor that can affect the optimal performance of models. For example, the hyperparameters of XGBoost, such as lambda, alpha, and gamma, models from overfitting on small and unbalanced data, while hyperparameters like tree depth, learning rate, and ensemble size can impact the accuracy, recall, and robustness of models on different data folds. If hyperparameter tuning is not properly performed, there can be significant differences in the performance of models reported by different studies.

Table 2. Comparison of Common Supervised Learning Algorithms for CAD Prediction

Algorithm	**Computational Cost**	**Interpretability**	**Data Requirements**
K-Nearest Neighbor (KNN)	Low-Med	High (intuitive)	Moderate dataset; sensitive to feature scaling
Decision Tree	Low	High (transparent rules)	Small-medium datasets; handles mixed data
Support Vector Machine (SVM)	Med-High	Medium (kernel can obscure logic)	Needs normalized data; moderate size; not ideal for very large datasets
Artificial Neural Networks (ANN)	High	Low (black-box)	Large datasets; sensitive to feature scaling
Random Forest (RF)	Med-High	Medium (ensemble reduces transparency)	Handles missing/mixed data; moderate dataset sufficient
Extreme Gradient Boosting (XGBoost)	High	Med-Low (boosted ensemble less interpretable)	Structured datasets; handles missing values; requires careful tuning

5. Literature Reviews and Related Works

Many studies have been conducted to apply the utility of ML models in CAD detection.
Zhang et al. [19] conducted a study to develop a machine learning system for the early detection of Coronary Artery Disease (CAD). To improve the accuracy of CAD prediction, they applied four feature techniques to the Z-Alizadeh Sani dataset: feature smoothing, feature encoding, feature construction, and feature selection. To address class imbalance, they used the SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN techniques for data balancing. For classification, the study employed the XGBoost and Random Forest algorithms. Model performance was assessed using a 10-fold cross-validation method to evaluate stability. The models were further evaluated across several performance metrics, including accuracy, recall, specificity, precision, F1 score, and AUC (Area Under the Curve). The XGBoost algorithm, particularly when combined with feature construction and SMOTE, emerged as the top-performing model. It achieved an impressive classification accuracy of 94.7%, with 96.1% recall, 93.2% specificity, 93.4% precision, 94.6% F1 score, and 98.0% AUC, highlighting its effectiveness in early-stage CAD prediction. A schematic diagram depicting the CAD prediction process is shown in Figure 4.

Muhammad et al. [18] developed a machine learning predictive model for CAD using diagnostic data gathered from two general hospitals in Kano State, Nigeria. The data was used to train several ML algorithms, including KNN, SVM, Random Tree, Naïve Bayes, Gradient Boosting, and Logistic Regression, to create predictive models for CAD diagnosis. The performance of these models was evaluated using various metrics such as Receiver Operating Curve (ROC) analysis, accuracy, specificity, and sensitivity. Among the algorithms tested, the Random Forest-based model achieved the highest accuracy of 92.04%, while the Naïve Bayes model showed the highest specificity at 92.40%. The SVM model demonstrated the highest sensitivity at 87.34%. Additionally, the Random Forest model also produced the highest ROC score of 92.20%. The model's architecture is depicted in Figure 5.

kella & Akella [26] proposed machine learning predictive models for enhancing the accuracy and timeliness of Coronary Artery Disease (CAD) diagnosis. 6 ML algorithms were compared to predict CAD from the Cleveland dataset: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Artificial Neural Network (ANN). Of all the models, the Neural Network model proved to have the

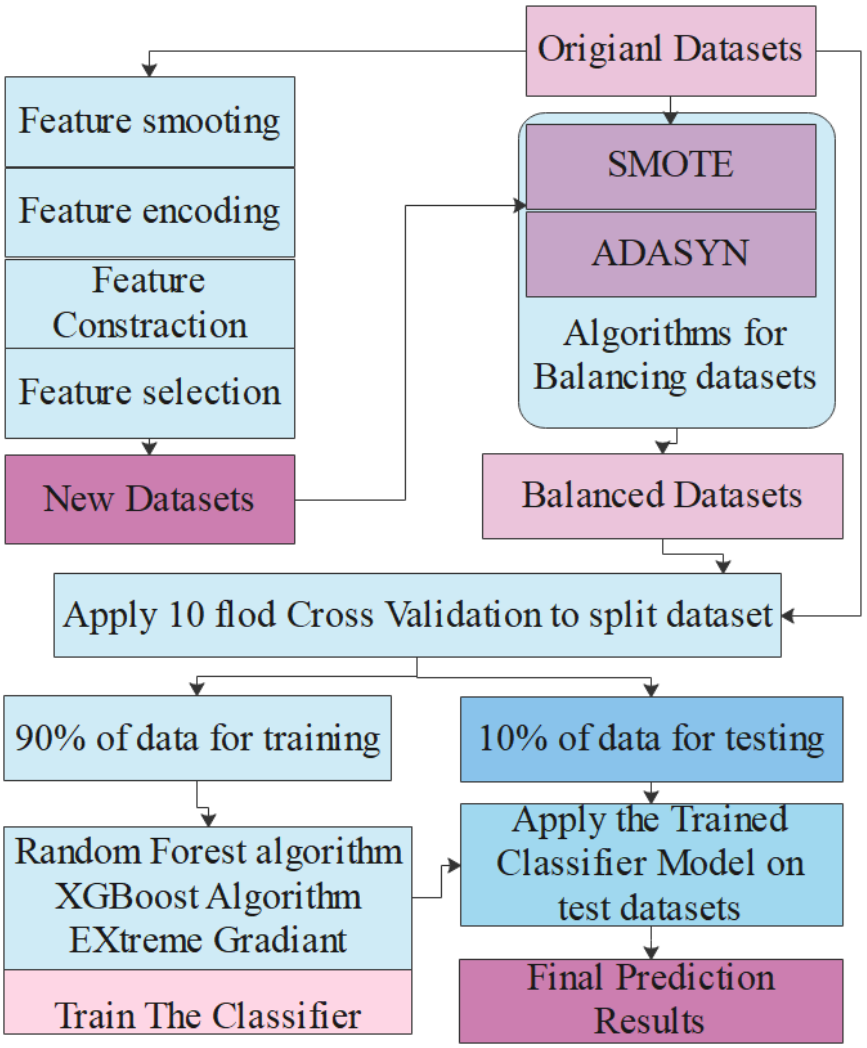

Figure 4. The prediction of CAD algorithm

optimal performance with 93.03% accuracy and a 93.80% recall rate, signifying that it is a highly efficient model to correctly predict CAD cases. From the study, ANN proved superior compared to the conventional techniques such as Logistic Regression and KNN, with the least accuracy levels of about 88% and 85%, respectively. This finding suggests that neural networks, with their ability to model complex data relationships, can be extremely beneficial for CAD detection.

Abdar et al. [28] introduced a new hybrid ensemble learning model for CAD prediction in the context of a CDSS framework. The model's efficacy was assessed utilizing two reputable CAD datasets: the Z-Alizadeh Sani dataset and the Cleveland dataset sourced from the UCI Machine Learning Repository. The suggested model is constructed on the Nested Ensemble (NE) method, which combines several conventional machine learning algorithms, using nu-SVC (nu-Support Vector Classification) as the primary algorithm. The nu-SVC is integrated with various powerful methods such as Stochastic Gradient Descent (SGD), Sequential Minimal Optimization (SMO), Random Forest, Naïve Bayes, stacking, bagging, and voting. To enhance the model's performance, procedures for feature selection and data balancing were implemented. A genetic search algorithm was specifically

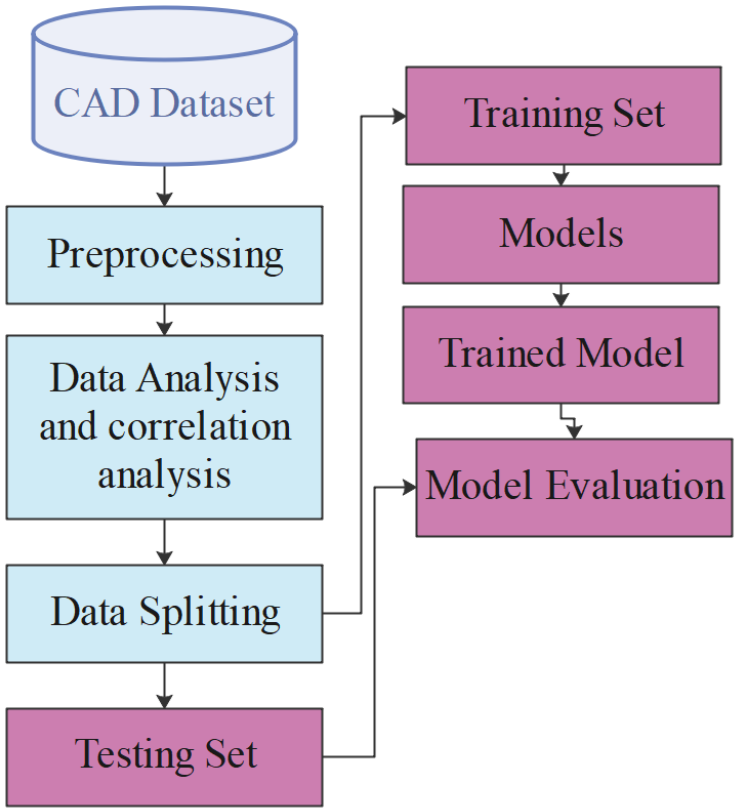

Figure 5. The architecture for detecting CAD [17]

employed for feature selection, and multilevel balancing was achieved through the Class Balancer and Resample techniques. The NE method allows for the combination of various ensemble learning techniques across different layers, which boosts the model's strength and predictive ability. The suggested model attained an accuracy of 94.66% on the Z-Alizadeh Sani dataset and 98.60% on the altered Cleveland dataset, greatly surpassing current machine learning algorithms on these datasets.

Elham Nasarian et al. [29] presented a heterogeneous hybrid feature selection (2HF) approach designed to enhance the detection of Coronary Artery Disease (CAD). Identifying important features from CAD patients is crucial, as diverse features from datasets correlate with different levels of CAD severity. To evaluate the effectiveness of the new 2HFS algorithm, the Nasarian CAD dataset was used. This dataset comprises both clinical characteristics and work-related elements, providing a wider context for CAD prediction. To address the imbalance in the dataset, the authors employed SMOTE and ADASYN methods. In the feature selection phase, various classifiers were utilized, such as Decision Tree, Gaussian Naive Bayes, Random Forest, and XGBoost. The suggested methodology underwent additional validation through its application to three prominent UCI CAD datasets: the Hungarian, Long Beach-VA, and Z-Alizadeh Sani datasets. Experimental findings showed that combining the 2HFS method with SMOTE attained a classification accuracy of 81.23% when employing the XGBoost classifier. These results emphasize the capability of the 2HF feature selection algorithm to enhance CAD detection.

Durgadevi Velusamy et al. [30] introduced an innovative heterogeneous ensemble technique that utilizes three base classifiers: K-Nearest Neighbor (KNN), Random Forest (RF), and Support Vector Machine (SVM). These results were decided through ensemble voting methods, specifically average voting (AVEn), majority voting

(MVEn), and weighted-average voting (WAVEn). To enhance the prediction outcomes of this model, the study employed the Random Forest-based feature selection method and SVM to identify significant features, allowing for the filtering and ranking of related features based on their importance. For the WAVEn algorithm, five key features were chosen, and the model was tested on the Z-Alizadeh Sani dataset. The class imbalance was addressed by balancing the dataset with the Synthetic Minority Over-sampling Technique (SMOTE). The experimental findings indicated that the WAVEn algorithm provided enhanced classification performance on the original dataset, achieving an accuracy of 98.97%, sensitivity of 100%, specificity of 96.3%, and precision of 98.3%. Furthermore, when the WAVEn algorithm was utilized on the balanced dataset, it demonstrated 100% precision, sensitivity, specificity, and accuracy in diagnosing CAD. The suggested hybrid model was developed to accurately diagnose CAD and is illustrated with its structure in Figure 6.

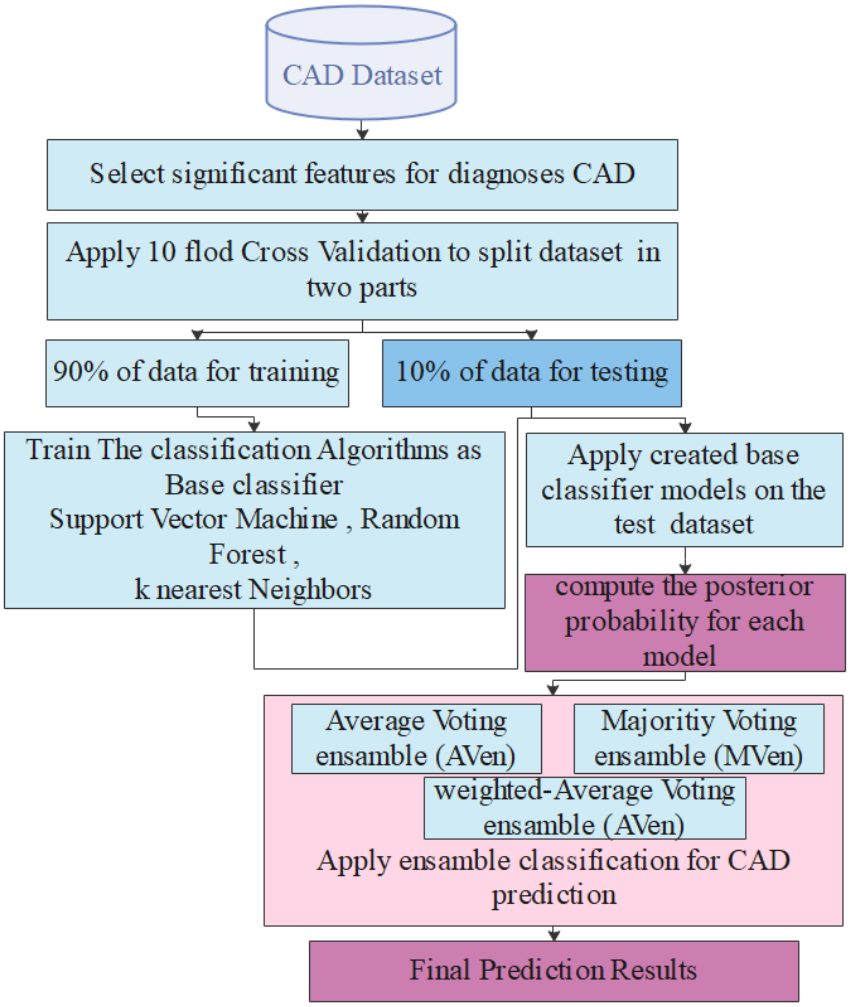

Figure 6. Architecture for ensemble method incorporating three classifiers to detect CAD [28].

Tama et al. [31] introduced a model that employs a two-tier ensemble method. The model combines three ensemble learners: Random Forest (RF), Gradient Boosting Machine (GBM), and Extreme Gradient Boosting Machine (XGBoost). This two-level classifier ensemble model was tested on four public datasets—Z-Alizadeh Sani, Statlog, Cleveland, and Hungarian—to ensure the analysis is entirely comparable with current methods. The researchers utilized a two-step statistical significance test to measure the results. The outcomes confirm the

efficacy of the two-tier classifier ensemble model across various datasets. For instance, when tested with the Z-Alizadeh Sani dataset, the model achieves an accuracy of 98.13%, a F1 score of 96.60%, and an AUC (Area Under the Curve) of 98.70%. In the Statlog dataset, the model achieved an accuracy of 93.55%, an F1 score of 91.67%, and an AUC of 93.42%. In the Cleveland dataset, the model achieved an accuracy of 85.71%, an F1 score of 86.49%, and an AUC of 85.86%. Ultimately, when the model was evaluated on the Hungarian dataset, it achieved an accuracy of 91.18%, an F1 score of 90.91%, and an AUC of 92.98

Table 3 lists all papers' experimental results, the datasets used, and data mining techniques with accuracy, precision, recall, and f- measure.

Table 3. related work papers summary for algorithms used and experimental results

Authors, Year	**Dataset**	**ML Algorithms used**	**Accuracy**	**Precision**	**Recall**	**F-Measure**
Abdar et al., 2019	Z-Alizadeh Sani	NE-nu-SVC + feature selection + multi-step balancing	94.66%	93.40%	96.30%	94.80%
Nasarian et al., 2025	Nasarian CAD dataset	heterogeneous hybrid feature selection (2HF) with Decision Tree, Gaussian Naive Bayes, Random Forest, XGBoost	2HF + SMOTE + XGBoost: 81.23%	80.24%	85.22%	82.07%
Muhammad et al., 2021	Two General Hospitals in Kano, Nigeria	Logistic Regression	80.68%	N/A	83.22%	N/A
		Support Vector Machine	88.68%	87.34%		
		K-nearest neighbor	82.35%	84.3%		
		Random Forest	92.04%	86.5%		
		Naive Bayes	87.5%	83.3%		
		Gradient Boosting	90.9%	87.2%		
Tama et al.,2020	4 datasets (Z-Alizadeh Sani, statlog, Cleveland, Hungarian)	Two-tier ensemble (GBM, XGBoost, RF), PSO-based feature selection	Z-Alizadeh Sani: 98.13%	N/A	N/A	96.6%
			statlog: 93.55%			91.67%
			Cleveland: 85.71%			86.49%
			Hungarian: 91.18%			90.91%
Durgadevi Velusamy et al., 2021	Z-Alizadeh Sani	weighted-average voting (WAVEn) algorithm	96%	N/A	N/A	N/A
Akella	Akella, 2021	Generalized linear model	87.64%	N/A	80.00%	87.86%
		Decision tree	79.78%	74.47%		79.7%
		Random Forest	87.64%	82.61%		87.64%
		Support-vector machine	86.52%	79.59%		86.52%
		Neural network	93.03%	93.38%		89.84%
		K-Nearest Neighbor	84.27%	78.72%		84.19%
S.S. Alotaibi et al., 2020	Z-Alizadeh Sani	Random Forest	81%	72%	86%	78%
		Naïve Bayes	83%	64%	100%	78%
Joloudari et al., 2020	Z-Alizadeh Sani	Random Trees	91.47%	N/A	N/A	N/A
Sayadi et al., 2022	Z-Alizadeh Sani	SVM (SVC) — Logistic Regression with Pearson correlation	95.08%	N/A	95.91%	96.90%
Zhang et al., 2022	Z-Alizadeh Sani dataset	XGBoost, Random Forest with Data processed by feature construction and (ADASYN, SMOTE)	XGBoost with SMOTE: 94.7%	93.4%	96.10%	94.60%

Authors, Year	Dataset	ML Algorithms used	Accuracy	Precision	Recall	F-Measure
			RF with SMOTE: 93.1%	94%	93.30%	93.40%
			XGBoost with ADASYN: 93.6%	93.5%	94.60%	93.80%
			RF with ADASYN: 91.7%	91.6%	93.30%	92.10%

S. S. Alotaibi et al. [32] They used Random Forest (RF) and Naïve Bayes (NB) algorithms. They utilized the Z-Alizadeh-Sani dataset containing clinical and demographic information for CAD patients. The authors carried out feature selection with the correlation coefficient to select the most significant features, which improved the performance of the models. Following 10-fold cross-validation, they determined that the Naïve Bayes model yielded outstanding results, with 100% sensitivity and a 100% negative predictive rate. This suggests that the Naïve Bayes model is extremely efficient at identifying true positive CAD cases and reducing false negatives. On the other hand, the Random Forest algorithm achieved a result of 83%, which, while less than Naïve Bayes, also showed a reasonable rate of performance on only 13 features. The outcome highlighted how this was best for Naïve Bayes to balance the predictive accuracy and sensitivity, and Random Forest performed adequately when it was given fewer features, and so was more effective in real-world applications where feature reduction is necessary.

Joloudari et al. [33] proposed a computer-aided diagnosis (CADx) system to enhance CAD detection using the Z-Alizadeh Sani dataset. The CADx system, which was created using IBM SPSS Modeler version 18.0, was put forward to overcome the limitations of traditional diagnostic methods like angiography, which could be costly and invasive. For improving CAD detection efficiency, the authors applied various machine learning algorithms, including Support Vector Machine (SVM), Chi-squared Automatic Interaction Detection (CHAID), C5.0, and Random Tree[34]. The models were validated using 10-fold cross-validation, and model performances were evaluated using various metrics such as accuracy, AUC (Area Under the Curve), Gini index, ROI, profit, confidence, response, and gain. The Random Tree model was the best, with 91.47% accuracy, 96.70% AUC, and 93.40% Gini index. The results show the achievement of Random Tree in balancing model complexity and predictive capability. Specifically, the 96.70% AUC showed excellent discrimination between CAD-positive and CAD-negative cases, and the 93.40% Gini index confirmed the high confidence of the model in distinguishing between classes. This study illustrates the potential of Random Tree to construct interpretable, efficient CAD diagnostic models.

Sayadi et al. [35]. They used the Pearson correlation feature selection method was employed. Six ML algorithms were tried out: Decision Tree, Deep Learning, Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost. The models were evaluated based on accuracy, sensitivity, specificity, F1 score, and the ROC curve (AUC). Logistic Regression and SVM had the same performance among the models with an accuracy of 95.45%, a sensitivity of 95.91%, a specificity of 91.66%, and an F1 score of 96.90%. Additionally, the ROC curve also confirmed equal performance with 0.98 AUC for both models. The findings demonstrate the power of Logistic Regression and SVM in CAD detection, especially in high-dimensional data[36].

Imran Chowdhury Dipto et al. [37] implemented a prototype CAD detection system on the basis of Logistic Regression, Support Vector Machine (SVM), and Artificial Neural Network (ANN) algorithms. The Shaheed Rajaei Cardiovascular, Medical, and Research Center dataset of patients' medical records with suspected CAD was utilized to train and test the models. Having done statistical analysis, the researchers confirmed that the dataset had no missing values. But during exploratory data analysis, they observed an imbalance in class with a higher number of patients being diagnosed with CAD than without the disease. To alleviate this, they employed the use

of the SMOTE (Synthetic Minority Over-sampling Technique) algorithm in order to level the dataset. In the pre-processing stage, categorical variables were encoded into numbers, and feature and predictor variable matrices were established. The data was divided into training and test sets. Feature Scaling was then applied to normalize the input features, followed by the use of SMOTE again to balance the dataset. The models were trained on both balanced and imbalanced datasets for comparison. The Artificial Neural Network (ANN) had the best performance with a mean accuracy of 93.35% ± 2.56% and an AUC of 0.98 ± 0.02. The Support Vector Machine (SVM) ranked second with an accuracy of 91.37% ± 3.50% and an equal AUC value of 0.98. Logistic Regression also ranked below at an accuracy of 89.61% ± 4.96% and with an AUC of 0.94 ± 0.05. These results show that the ANN and SVM models were effective in CAD prediction, especially when they were trained on the balanced dataset, highlighting the importance of data balancing in improving the accuracy of predictions.

An analysis of some of the recent works on the prediction of CAD using various models indicates certain trends in model selection, feature handling, and balancing techniques. Ensemble models, such as RF and XGBoost, dominate the results of various studies in terms of accuracy. For example, in a study conducted by Zhang et al. [19], it was noted that feature construction and SMOTE resulted in 94.7% accuracy and 98% AUC for XGBoost, while RF performed almost equally well. Another study conducted noted that RF performed better than other models in terms of accuracy (92.04%) and ROC (92.20%).

Moreover, hybrid and nested ensemble models can further improve the predictive capabilities of these models by aggregating various classifiers and utilizing feature selection and balancing approaches. Abdar et al. [28] employed a Nested Ensemble (NE) model by integrating RF, SVM, and various other algorithms to achieve a maximum accuracy of 98.6% on the modified Cleveland dataset. This shows how hybrid models can improve prediction stability and reduce the limitations of various models. Nasarian et al. [29] showed that 2HF feature selection algorithm can improve classification efficiency on heterogeneous datasets by utilizing SMOTE and XGBoost algorithms to achieve 81.23% accuracy.

Two important factors that emerge from the analysis are feature selection and balancing data, which have significant effects on performance variations even when using the same dataset. In general, when SMOTE, ADASYN, or genetic algorithm techniques are used for feature selection, higher accuracy, recall, and AUC values are reported compared to other techniques.

From the methodological point of view, the results show that high-performance CAD models are not only dependent on the selection of the algorithm but also on the quality and preprocessing of the data and evaluation techniques. RF and XGBoost have shown good performance due to their robustness in handling heterogeneous features, their ability to capture non-linear interactions between features, and their overfitting resistance. In contrast, other models like logistic regression and naive Bayes have shown good performance on datasets but have been inconsistent in other datasets.

6. Discussion and Experimental Results

XGBoost, Random Forest, and ANN are overall high performers in most studies, with both high accuracy and AUC scores. Support Vector Machine (SVM) has worked well, particularly in balanced datasets, and Logistic Regression has also shown good performance in several studies. Hybrid ensemble models, such as those used in Abdar et al. [28] and Tama et al. [31], have great potential in improving CAD detection via the combination of multiple classifiers for improving model stability and performance. Feature selection, balancing data (e.g., SMOTE), and using more than one measure of performance (accuracy, sensitivity, specificity, AUC) have been crucial in model performance improvement. Figure 7 and table 3show the results accuracy for different data sets, and only for Z-Alizadeh Sani data sets were shown in Figure 8.

Table 4 lists all paper's experimental accuracy results

Recent advances in artificial intelligence (AI) and machine learning (ML) with CAD) were employed in the AI in Imaging for Personalized Management of CAD trial to coronary CT angiography (CCTA) and optical coherence tomography (OCT) images. These models improved CAD severity prediction with over 90% accuracy in predicting high-risk patients.

Table 4. Comparison of Accuracy results from Different Studies

Study	Accuracy (%)
Zhang et al. [35]	94.7
Muhammad et al. [36]	92.04
Abdar et al. [37]	94.66
Elham Nasarian et al. [38]	81.23
Durgadevi Velusamy et al. [39]	98.97
Tama et al. [32]	98.13
Akella et al. [30]	93.03
S. S. Alotaibi et al. [40]	83
Joloudari et al. [41]	91.47
Sayadi et al. [43]	95.45
Imran Chowdhury Dipto et al. [33]	93.35

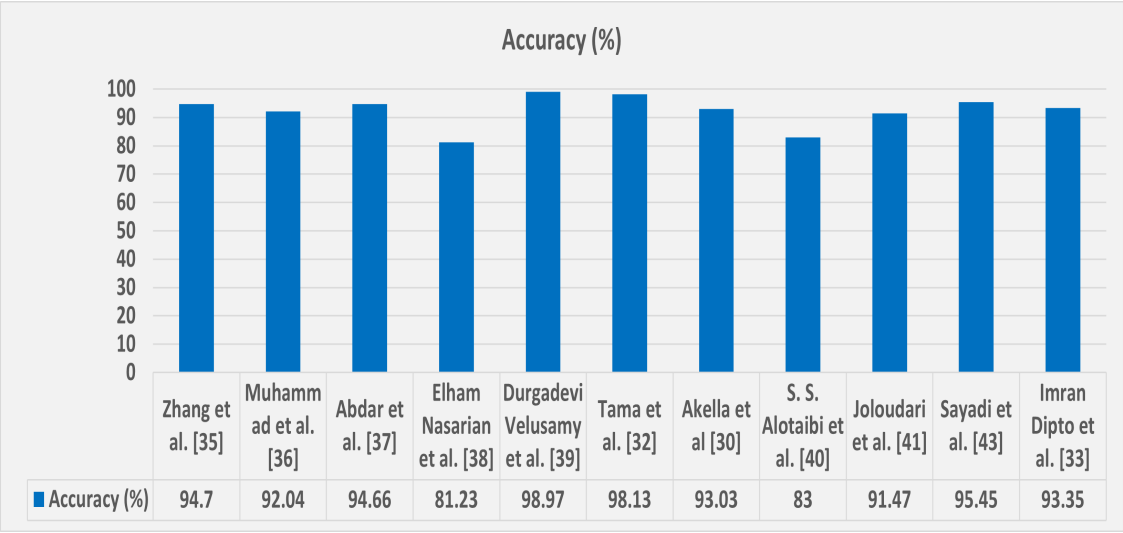

Figure 7. Accuracy results for different datasets

Similarly, in The Heart of Transformation review, random forest classifiers and Gradient Boosting Machines (GBM) were utilized on clinical datasets like the Framingham Heart Study and Cleveland Heart Disease dataset, achieving an accuracy of about 85% in CAD risk prediction [38, 39]. k-Nearest Neighbors (k-NN) and neural networks were combined to analyze clinical and imaging information in Comprehensive Analysis of Cardiovascular Diseases, which proved highly sensitive (92%) and specific (87%) in diagnosing CAD [40]. The study on Transforming Cardiovascular Risk Prediction used logistic regression, random forests, and XGBoost models on datasets like NHANES and Cleveland Heart Disease, suggesting improved prediction, with the models achieving up to 88% accuracy in risk stratification compared to traditional scoring systems like the Framingham Risk Score [41].

Ensemble Learning-Based CAD Detection utilized an ensemble approach combining CatBoost, LightGBM, and Random Forest models on CCTA scans with a classification rate of 94%, surpassing individual models [42] Lastly, in Machine Learning Model Discriminates Ischemic Heart Disease Using Breathome Analysis, Support Vector Machines and Random Forest achieved 91% accuracy in distinguishing patients of ischemic heart disease from healthy individuals by using breathome data, indicating a non-invasive diagnostic application[43, 44]. Collectively, these studies demonstrate that ML models, when applied to both clinical data and imaging, can provide extremely accurate and personalized approaches to the diagnosis and treatment of CAD, leading to improved patient outcomes

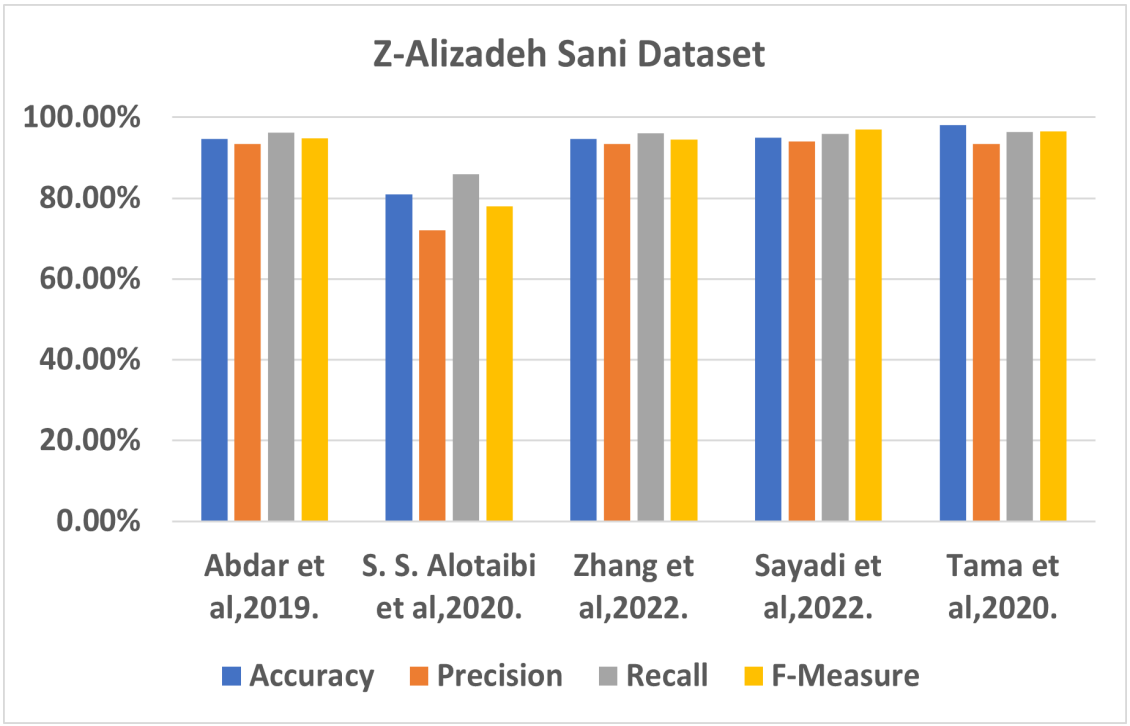

Figure 8. Results for Z-Alizadeh Sani datasets

and more efficient healthcare practices. Collectively, these studies suggest that ML and AI are shifting towards more clinically relevant solutions for CAD detection and management. The strong performers, including XGBoost, Random Forest, and ANN, consistently show strong predictive capabilities across various datasets, and further improvement is seen with ensemble and hybrid techniques, especially when dealing with imbalanced and noisy data. The incorporation of clinical and imaging modalities, including CCTA, OCT, and breathome analysis, facilitates personalized patient risk stratification, enabling early detection of high-risk patients and facilitating intervention strategies. The progressive improvement in accuracy, sensitivity, and specificity of these studies suggests that there is a convergence towards standardized and data-driven solutions for CAD evaluation, which might even surpass traditional scoring systems like the Framingham Risk Score. The trajectory of AI in CAD research is shifting towards multi-modal, interpretable, and scalable solutions, which might redefine diagnostic accuracy and individualized patient care.

Key Findings and Clinical Implications: The findings of the literature synthesis suggest that ensemble and hybrid machine learning approaches, such as Random Forest and XGBoost, have demonstrated excellent performance on clinical datasets, as these approaches are less affected by heterogeneous features and non-linear relationships [19, 18, 28]. Feature selection and balancing (e.g., SMOTE, ADASYN) are essential to improve the accuracy of these approaches, which may explain discrepancies among studies on the same dataset. However, one of the major limitations of these approaches is their lack of interpretability, as “black box” predictions cannot be trusted by clinicians. In fact, studies with ¿90% accuracy also have the risk of false negatives or positives, which may result in delayed or unnecessary clinical decisions, respectively. In conclusion, high-performing ML approaches have significant potential to be applied to CAD prediction, but it is still necessary to address issues related to data curation, interpretability, and real-world validation to incorporate these approaches into clinical decision-making. Although the accuracy of ensemble methods such as Random Forest and XGBoost has consistently shown high performance, the accuracy reported in different studies ranges significantly, i.e., from 81% to over 98%. This variation in accuracy may be attributed to various factors, including differences in dataset characteristics. It has been observed that datasets used in different studies may vary in terms of their size and feature diversity. In some

cases, datasets are small and have limited variability, resulting in overfitting and higher accuracy. In contrast, datasets used in different studies have shown higher variability and accuracy. In addition, preprocessing of datasets using feature selection, normalization, and data augmentation techniques also play an important role in improving model accuracy. Moreover, different validation approaches, including train-test split and cross-validation, have shown different accuracy results. In some cases, accuracy has been reported higher when validation approaches are less robust. Other factors that may have contributed to differences in accuracy include variations in hyperparameter tuning, algorithm configuration, and ensembling methods.

7. The future directions and recommendations

1. Integration of Multi-modal Data for General Risk Prediction Future Direction: The most promising advancement in CAD prediction will be the integration of multi-modal data, including clinical data (e.g., medical history, lab results, risk factors), medical imaging (e.g., CT scans, MRI, echocardiograms), and genetic data. Machine learning algorithms, especially deep learning techniques, will be trained on large datasets to identify complex patterns and predict CAD risk more accurately[44]. Recommendation: Design AI programs able to read multimodal data live and provide patients with individual, dynamic risk estimation. Incorporating genetic data as well as patient lifestyle factors will yield more accurate predictions and intervention prevention tailored to each patient.
2. Explainable AI (XAI) for Clinical Decision Support Future Direction: While AI models, especially deep learning, are very accurate, they are ”black boxes,” and it is difficult for clinicians to trust their predictions. The future of using AI for CAD prediction will need to strive to make the models more interpretable and transparent so that clinicians can better understand why a certain prediction or decision is being made. Recommendation: Invest in Explainable AI (XAI) method development to enhance the transparency of decision-making. This would allow clinicians to understand the justification for CAD predictions and recommendations, making AI instruments more reliable and actionable in clinical practice.
3. Real-Time Monitoring and Forecasting using Wearables. Future Direction: As wearable devices (fitness trackers, smartwatches) are being used more and more, AI-based systems will enable real-time monitoring of heart health factors like blood pressure, heart rate variability, and activity. AI, through continuous data gathering and processing, can detect CAD predictors at an early stage or alert physicians to worsening conditions in those with high-risk factors[42]. Recommendation: Develop AI systems that can handle real-time health data from wearable devices, integrating this information into predictive models for CAD warning signs in early stages. Machine learning needs to be designed to handle real-time streams of data, with continuous monitoring and proactive care for patients.
4. Federated Learning for Data Privacy Future Direction: One of the greatest challenges for CAD prediction is access to quality datasets. Federated learning, where models are trained without sharing raw data, can ensure collaboration across institutions without compromising patient anonymity. This can lead to more and improved datasets used to train AI models with better generalizability of CAD prediction models. Recommendation: Promote the use of federated learning and other privacy-preserving approaches in CAD prediction systems. Collaborative efforts between hospitals, research institutions, and technology companies will help develop heterogeneous datasets while maintaining data privacy and security[39].
5. AI-Driven Population Health Management Future Direction: AI can revolutionize population health management by analyzing large data sets from diverse patient populations. By identifying trends and risk factors at the population level, AI would help public health officials create more focused interventions and prevention programs for CAD . Recommendation: Give priority to developing AI models to analyze big health data in order to discern trends in cardiovascular disease and suggest interventions to reduce the burden of CAD at a population level. This can help public health officials to prioritize prevention strategies and resources.

The focus of future directions in CAD prediction is to ensure that AI models are interpretable and transparent. This can be done with the use of techniques such as SHAP (Shapley Additive Explanations) and LIME (Local

Interpretable Model-agnostic Explanations), which can provide clarity on the contributions made by specific clinical, imaging, and demographic factors to risk predictions. With the use of such XAI techniques, AI models can be made more interpretable and can be incorporated into clinical practice.

Another important aspect that can be considered while developing AI models is ethics and regulations. With the use of AI models, transparency can be ensured, which can be beneficial while making clinical decisions and obtaining patient consent. Similarly, data privacy can be ensured while using data. Moreover, with the use of AI models, it can be ensured that they are compliant with FDA regulations, which can be beneficial while ensuring safety, reliability, and accountability.
An important emerging theme is the interconnection between these future directions, where the advancements in federated learning, multi-modal data integration, and Explainable AI are seen to converge towards the development of better, more secure, and more actionable CAD prediction systems. Federated learning, for instance, enables AI models to be trained on distributed data sets from various hospitals or research centers without sharing the underlying patient data, thus maintaining data privacy while being able to tap into the rich data sources of various data modalities such as clinical, imaging, genetic, and wearable data. Moreover, by incorporating multi-modal data integration, it is possible to allow the AI models to learn complex patterns from various data modalities, thus being able to improve the accuracy of the predictions made by the CAD prediction systems. Furthermore, by incorporating Explainable AI (XAI) techniques, it is possible to ensure that the predictions made by the CAD prediction systems are trustworthy and transparent, thus being able to pave the way towards the development of personalized, privacy-preserving, and interpretable CAD risk assessment systems that are capable of supporting real-time monitoring, dynamic patient risk evaluation, and health management of populations.

8. AI Techniques in Cardiovascular Disease Diagnosis

Machine Learning for Prediction of CVD Risk: Machine learning techniques such as decision trees, random forests, and support vector machines are used to forecast cardiovascular event risk. Describe how AI models integrate variables like age, sex, cholesterol, blood pressure, and family history to provide more accurate risk assessments than traditional scoring algorithms (e.g., Framingham Risk Score) [45]. Deep Learning for Imaging: Deep learning has been highly promising in the image processing of medical images. How AI is used in the analysis of ECGs, echocardiograms, MRI scans, and CT angiograms to detect abnormalities (e.g., coronary artery disease, heart failure). Natural Language Processing (NLP) of Electronic Health Records (EHR) [46] is increasingly used to extract meaningful hidden information from unstructured types of data in EHRs to assist in diagnosing cardiovascular disease.
Artificial Intelligence in Personized Cardiovascular Medicine :

1. AI-Driven Drug Discovery and Development: The use of drug discovery and therapeutic optimization by AI in medicine is an exploratory and developing area in the field of cardiovascular medicine. Machine learning techniques are increasingly being employed for the discovery of new drug targets and for the prediction of drug-drug interactions and optimization of drug therapy. This area is still in its infancy and is facing many challenges in terms of the availability of good biological data and the long validation period. Currently, most drug discovery using AI techniques is still in the preclinical or research phase and has yet to have a significant impact on the field[14, 47].
2. AI-based imaging applications are the most mature and clinically validated domain, with deep learning models showing excellent performance in image interpretation, lesion detection, and disease severity prediction using echocardiography, CT scans, and MRI scans. These models have access to well-structured and labeled data, which has helped them get translated and approved for clinical use. These AI-based imaging tools are now being used as decision support systems in the treatment and management of cardiovascular diseases[48].
3. In contrast, NLP-based applications in cardiology have only achieved a relatively intermediate level of maturity. NLP-based methods have been applied mainly to derive relevant clinical information from

unstructured data sources such as electronic health records, physician notes, or discharge summaries. Although these methods have provided valuable insights, their applicability is limited by factors such as inconsistent clinical documentation, language, or the availability of annotated data sources. Therefore, the clinical evidence base supporting NLP-based cardiovascular applications is still developing, but not as strong as that supporting imaging-based AI systems.

Challenges and Limitations of AI in Cardiovascular Diseases:

1. Data Quality and Integration: Despite the potential of AI, data quality and integration remain challenges. Describe the requirement of high-quality, large-scale datasets for AI systems to properly train models and how incomplete or inconsistent patient data can impact the accuracy of predictions through AI.
2. Interpretability and Trust in AI: AI systems, particularly deep learning models, are ”black boxes,” and it is difficult to understand how they arrive at a specific prediction. This section addresses the importance of interpretability in AI models to convince clinicians to trust the decisions of the AI system.
3. Ethical and Regulatory Concerns: Medical adoption of AI has generated several ethical and regulatory concerns. The issues of patient privacy, consent, and the need for standardized algorithms for AI are discussed. The role of regulatory bodies (e.g., FDA) in ensuring the safety and efficacy of AI solutions is briefly mentioned [49]. The Future of AI in Cardiovascular Disease Management Advances in AI Technology: In the future, this section discusses new AI technologies like reinforcement learning, federated learning, and explainable AI (XAI) used with cardiovascular care.
4. Collaboration of AI with Healthcare Professionals: Rather than replacing healthcare professionals, it is expected that AI will augment clinicians with robust decision-support systems that allow for more accurate diagnoses, enhanced patient care, and efficient cardiovascular disease management.

9. conclusion and Future work

In this paper, we have extensively reviewed the various machine learning (ML) algorithms and techniques applied in CAD prediction and diagnosis. The integration of AI and data mining techniques, such as decision trees, neural networks, support vector machines, and random forests, has shown immense potential for improving CAD diagnosis accuracy and efficiency. Our findings indicate that ML models, particularly those employing ensemble techniques and feature selection, can provide reliable predictions with potential benefits in the early diagnosis and timely treatment of CAD patients. Of all the algorithms under discussion, Random Forest and XGBoost represented the best performance in terms of accuracy, recall, and specificity in CAD prediction, demonstrating their superior capability to handle very complicated medical data. Moreover, the use of datasets such as Cleveland, Z-Alizadeh Sani, and others has played a pivotal role in determining the efficiency of such models, with analysis having captured increasingly high accuracy rates and prediction efficiency. While these advancements have occurred, domains that remain to be enhanced are data imbalances, feature selection, and interpretability of the model. In the future, AI techniques such as federated learning, multi-modal data, and explainability must be explored to obtain accurate CAD risk prediction, which can result in early diagnosis, better outcomes, and cost savings, thus emphasizing the importance of AI in CAD prediction. Also, we should aim at reducing dataset biases by using multi-ethnic and multi-center cohorts. Currently, CAD prediction models are based on datasets from specific geographic or demographic populations. This might limit their generalization and decrease their predictive accuracy in underrepresented populations. By using diverse patient datasets, including differences in genetics, lifestyle, and socioeconomic factors, machine learning models can produce better and more generalizable predictions. This will promote fairness in decision-making and ensure that AI tools are reliable in heterogeneous populations.

REFERENCES

[1] M.H. Bahnasawy, L.Z. Habbak, M.A. Al-Ashry, and M.M. Al-Maie, *Risk Factors for Coronary Artery Disease in Egyptian Women*, The Egyptian Journal of Hospital Medicine, vol. 53, pp. 827–836, 2013.

[2] D.P. Zipes, P. Libby, R.O. Bonow, D.L. Mann, and G.F. Tomaselli, *Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine*, Elsevier Health Sciences, 2018.

[3] RN.com, *Coronary Artery Syndrome*, Available: https://www.rn.com/headlines-in-health/coronary-artery-syndrome/

[4] A. Sayed, M. Khafagy, M. Ali, and M. Mohamed, *Predict student learning styles and suitable assessment methods using click stream*, Egyptian Informatics Journal, vol. 26, Article 100469, 2024. https://doi.org/10.1016/j.eij.2024.100469

[5] R. Alizadehsani et al., *Machine learning-based coronary artery disease diagnosis: A comprehensive review*, Computers in Biology and Medicine, vol. 111, article 103346, 2019.

[6] A.R. Sayed, M.H. Khafagy, M. Ali, and M.H. Mohamed, *Exploring the VAK model to predict student learning styles based on learning activity*, Intelligent Systems with Applications, vol. 25, article 200483, 2025. https://doi.org/10.1016/j.iswa.2025.200483

[7] A. Hassan and M. El-Tekabi, *Advances in Machine Learning for Healthcare Analytics*, in *Comprehensive Data Science and AI Applications*, vol. 12, Springer, 2025, pp. 45–78.

[8] M.A. El Mrabet, K. El Makkaoui, and A. Faize, *Supervised Machine Learning: A Survey*, in Proc. 4th International Conference on Advanced Communication Technologies and Networking (CommNet), 2021, pp. 1–10.

[9] G. Abdallah Radwan, M. Helmy Khafagy, M.T.M. Mabrouk, and M.H. Mohamed, *Coronary artery disease prediction by combining three classifiers*, Journal of Information Hiding and Multimedia Signal Processing, vol. 15, no. 4, pp. 221–235, 2024.

[10] M. Mohamed, M. Khafagy, N. Kamel, and W. Said, *Diabetic mellitus prediction with BRFSS data sets*, Journal of Theoretical and Applied Information Technology, vol. 102, pp. 883–897, 2024.

[11] A. Gupta, A. Sharma, and A. Goel, *Review of Regression Analysis Models*, International Journal of Engineering Research and Technology (IJERT), vol. 11, no. 5, pp. 123-130, 2022.

[12] H. Roopa and T. Asha, *A Linear Model Based on Principal Component Analysis for Disease Prediction*, IEEE Access, vol. 7, pp. 105314–105318, 2019. https://doi.org/10.1109/ACCESS.2019.2931956

[13] D. Maulud and A.M. Abdulazeez, *A review on linear regression comprehensive in machine learning*, Journal of Applied Science and Technology Trends, vol. 1, no. 4, pp. 140–147, 2020.

[14] M. Bounekdja and A. Hakem, *Analyzing and Classifying Coronary Artery Disease Severity Using Statistical Methods and Machine Learning Techniques*, *Statistics, Optimization & Information Computing*, vol. 14, no. 3, pp. 812–829, 2025. https://doi.org/10.19139/soic-2310-5070-2241

[15] K. V. Konathala, H. V. R. Yangoti, S. D. K. Puppala, and N. Pathania, *Cardiovascular Disease Prediction using Machine Learning Algorithms*, SSRN Electronic Journal, 2024, doi:10.2139/ssrn.4833925.

[16] D. Y. Omkari and K. Shaik, *An Integrated Two-Layered Voting (TLV) Framework for Coronary Artery Disease Prediction Using Machine Learning Classifiers*, IEEE Access, vol. 12, pp. 56275–56290, 2024, https://doi.org/10.1109/ACCESS.2024.3389707.

[17] R. Detrano, A. Janosi, W. Steinbrunn, et al., *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.

[18] L.J. Muhammad et al., *Machine learning predictive models for coronary artery disease*, SN Computer Science, vol. 2, no. 5, article 350, 2021. https://doi.org/10.1007/s42979-021-00731-4.

[19] Y. Zhang et al., *Improvement of the Performance of Models for Predicting Coronary Artery Disease Based on XGBoost Algorithm and Feature Processing Technology*, Electronics, vol. 11, no. 3, p. 315, Jan. 2022. https://doi.org/10.3390/electronics11030315

[20] A. Paul, D.P. Mukherjee, P. Das, A. Gangopadhyay, A.R. Chintha, and S. Kundu, *Improved Random Forest for Classification*, IEEE Transactions on Image Processing, vol. 27, no. 8, pp. 4012–4024, 2018. https://doi.org/10.1109/TIP.2018.2834830

[21] S. A. J. Zaidi, A. Ghafoor, J. Kim, Z. Abbas, and S. W. Lee, *HeartEnsembleNet: An Innovative Hybrid Ensemble Learning Approach for Cardiovascular Risk Prediction*, *Healthcare*, vol. 13, no. 5, p. 507, 2025, https://doi.org/10.3390/healthcare13050507.

[22] X. Wu et al., *Top 10 algorithms in data mining*, Knowledge and Information Systems, vol. 14, no. 1, pp. 1–37, Dec. 2007. https://doi.org/10.1007/s10115-007-0114-2

[23] D. Betel, A. Koppal, P. Agius, C. Sander, and C. Leslie, *Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites*, Genome Biology, vol. 11, no. 8, pp. 1–14, Aug. 2010. https://doi.org/10.1186/gb-2010-11-8-r90

[24] X. Wang and I.M. El Naqa, *Prediction of both conserved and nonconserved microRNA targets in animals*, Bioinformatics, vol. 24, no. 3, pp. 325–332, Feb. 2008. doi:10.1093/bioinformatics/btm595

[25] P.S. Linsley et al., *Transcripts Targeted by the MicroRNA-16 Family Cooperatively Regulate Cell Cycle Progression*, Molecular and Cellular Biology, vol. 27, no. 6, p. 2240, Mar. 2007. https://doi.org/10.1128/MCB.02005-06

[26] A. Akella and S. Akella, *Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution*, Future Science OA, vol. 7, no. 6, 2021, https://doi.org/10.2144/fsoa-2020-0206.

[27] R. Detrano, A. Janosi, W. Steinbrunn, et al., *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.

[28] M. Abdar, U.R. Acharya, N. Sarrafzadegan, and V. Makarenkov, *NE-nu-SVC: A New Nested Ensemble Clinical Decision Support System for Effective Diagnosis of Coronary Artery Disease*, IEEE Access, vol. 7, pp. 167605–167620, 2019. https://doi.org/10.1109/ACCESS.2019.2953920

[29] E. Nasarian et al., *Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach*, Pattern Recognition Letters, vol. 133, 2020.

[30] D. Velusamy et al., *Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset*, Computer Methods and Programs in Biomedicine, vol. 198, article 105770, 2021.

[31] B.A. Tama et al., *Improving an Intelligent Detection System for Coronary Heart Disease Using a Two-Tier Classifier Ensemble*, BioMed Research International, article 9816142, 2020. https://doi.org/10.1155/2020/9816142

[32] S.S. Alotaibi et al., *Automated prediction of Coronary Artery Disease using Random Forest and Naïve Bayes*, Proc. 2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia, pp. 109–114, 2020. https://doi.org/10.1109/ICACSIS51025.2020.9263159

[33] J.H. Joloudari et al., *Coronary artery disease diagnosis; ranking the significant features using a random trees model*, International Journal of Environmental Research and Public Health, vol. 17, no. 3, p. 731, 2020.

[34] M.H. Mohamed, M. Elkholy, and M.A. Marzouk, *Innovative Machine Learning Approaches for Identifying Pre-diabetes in Patients*, Journal of Information Hiding and Multimedia Signal Processing, vol. 16, no. 1, pp. 365–378, March 2025.

[35] M. Sayadi et al., *A Machine Learning Model for Detection of Coronary Artery Disease Using Non-invasive Clinical Parameters*, Life, vol. 12, no. 11, p. 1933, 2022. https://doi.org/10.3390/life12111933

[36] M.H. Mohamed, L.F. Ibrahim, K. Elmenshawy, and H.R. Fadlallah, *Adaptive Learning Systems based on ILOS of Courses*, WSEAS Transactions on Systems and Control, vol. 18, pp. 1–17, 2023. https://doi.org/10.37394/23203.2023.18.1

[37] I.C. Dipto, T. Islam, H.M. Rahman, and M.A. Rahman, *Comparison of Different Machine Learning Algorithms for the Prediction of Coronary Artery Disease*, Journal of Data Analysis and Information Processing, April 2020.

[38] S.S. Qasim et al., *An Intelligent System Using Deep Learning for Healthcare Monitoring in Light of the COVID-19 and Future Pandemics Based on IoT*, Al-Esraa University College Journal for Engineering Sciences, vol. 6, no. 9, pp. 48–67, Dec. 2024. https://doi.org/10.70080/2790-7732.1004

[39] Fadlallah, H. R., Mohamed, M. H., Alayash, W., Stephan, J. J., Qasim, S. S., Alqabany, T., & Ali, M., 'SECURE IOT COMMUNICATIONS USING SCRP-DRIVEN DYNAMIC QUASIGROUP CRYPTOGRAPHY', Journal of Theoretical and Applied Information Technology, vol. 104, no. 1, Jan. 2026, https://doi.org/10.5281/zenodo.18259402.

[40] W.K. AlSaraj, L.A.G. Zghair, and A.H. Mohammed, *Capacity of Self Compact Concrete Walls Using Attapulgite as a Partial Replacement of Cement Under One Way and Two Way Action Restriction*, Al-Esraa University College Journal for Engineering Sciences, vol. 6, no. 9, pp. 16–30, Dec. 2024. https://doi.org/10.70080/2790-7732.1002

[41] D.-I. Kasartzian and T. Tsiampalis, *Transforming Cardiovascular Risk Prediction: A Review of Machine Learning and Artificial Intelligence Innovations*, Life, vol. 15, no. 1, p. 94, 2025. https://doi.org/10.3390/life15010094

[42] A.R.W. Sait and A.M.A.B. Awad, *Ensemble Learning-Based Coronary Artery Disease Detection Using Computer Tomography Images*, Applied Sciences, vol. 14, no. 3, p. 1238, 2024. https://doi.org/10.3390/app14031238

[43] N.A. Jaafar, *A Study on Improving the Accuracy and Effectiveness of Similarity Detection Processes in Text Files Using NLP Techniques*, Al-Esraa University College Journal for Engineering Sciences, vol. 6, no. 9, pp. 1–15, Dec. 2024. https://doi.org/10.70080/2790-7732.1001

[44] B.A. Marzoog et al., *Machine Learning Model Discriminate Ischemic Heart Disease Using Breathome Analysis*, Biomedicines, vol. 12, no. 12, p. 2814, 2024. https://doi.org/10.3390/biomedicines12122814

[45] M. Almutairi and S. Dardouri, *Intelligent Hybrid Modeling for Heart Disease Prediction*, Information, 2025, 16, 869, https://doi.org/10.3390/info16100869.

[46] L.M. Gladence, M. Karthi, and V.M. Anu, *A statistical comparison of logistic regression and different Bayes classification methods for machine learning*, ARPN Journal of Engineering and Applied Sciences, vol. 10, no. 14, pp. 5947–5953, 2015.

[47] E. Aniq, F.-A. El Ghanaoui, and M. Chakraoui, *Vision Transformers for Breast Cancer Mammographic Image Classification*, Statistics, Optimization & Information Computing, 2025, 15(2), 1087–1098, https://doi.org/10.19139/soic-2310-5070-2539.

[48] El-Aziem, Ayman H. Abd, Marwa Hussien Mohamed, and Ahmed Abdelhafeez. High-Security Image Encryption Using Baker Map Confusion and Extended PWAM Chaotic Diffusion. *Computers*, 15(2):106, 2026. https://doi.org/10.3390/computers15020106

[49] J. Silva, A. Alves, P. Santos, and L. Matioli, *A new SVM solver applied to Skin Lesion Classification*, Statistics, Optimization & Information Computing, 2024, 12(4), 1149–1172, https://doi.org/10.19139/soic-2310-5070-2005.