Early Prediction of Coronary Artery Disease (CAD)

using Machine Learning

BY

AYENITAJU SAMUEL OLAMIDE 200404046

OGUNDELE DAMILOLA SAMUEL 200404163

SHONOLA AYOMIDE SAMSON 200404079

FASELA FEMI SAMUEL 200404077

MOYOLOYE BABALOLA DAVID 2004040137

**CHAPTER ONE**

**INTRODUCTION**

**BACKGROUND AND SIGNIFICANCE OF CAD PREDICTION**

Coronary artery disease (CAD) remains the leading cause of morbidity and mortality worldwide, accounting for approximately 17.9 million deaths annually (World Health Organization [WHO], 2023). CADs are characterized by the narrowing or blockage of coronary arteries due to atherosclerotic plaque buildup and can lead to severe complications such as myocardial infarction, heart failure, and sudden cardiac death (Benjamin et al., 2019). The early detection and prediction of CAD are vital in improving clinical outcomes, reducing healthcare costs, and enhancing the quality of life for patients.

Traditional diagnostic methods, including electrocardiograms (ECG), stress tests, and coronary angiography, while effective, are often invasive, expensive, and may not detect early or asymptomatic cases (Dey et al., 2018). As a result, there is an increasing demand for more efficient, non-invasive, and accurate prediction models that can support early diagnosis and timely intervention.

In recent years, computational approaches, particularly machine learning (ML) and data-driven predictive analytics, have shown promising potential in enhancing CAD prediction. These techniques can analyze complex and large-scale datasets, identify hidden patterns, and support clinicians in risk stratification and decision-making (Alizadehsani et al., 2019). Leveraging clinical, demographic, and lifestyle-related data, ML models can outperform traditional statistical methods by providing personalized risk assessments and uncovering nonlinear relationships among risk factors.

The significance of CAD prediction lies not only in the potential to save lives but also in addressing the broader public health and economic burden of cardiovascular diseases. By integrating predictive models into routine clinical practice, healthcare systems can shift the focus from reactive to preventive care, identifying high-risk individuals before the onset of symptoms and implementing targeted interventions (Krittanawong et al., 2020).

One of the primary challenges in managing CAD is its often-asymptomatic nature during early stages. Many individuals remain undiagnosed until they experience major adverse cardiovascular events such as myocardial infarction or sudden cardiac arrest. By the time symptoms appear, the disease may have already progressed significantly. Early prediction of CAD, therefore, is of paramount importance in shifting clinical practice from reactive treatment to proactive prevention (Benjamin et al., 2019).

Timely identification of individuals at high risk of CAD allows for early lifestyle modifications, medical interventions, and closer clinical monitoring, all of which can significantly reduce the risk of adverse outcomes. In this context, risk prediction tools serve as a cornerstone for preventive cardiology. Traditional risk scoring systems such as the Framingham Risk Score and the European SCORE model have been widely used, but these models often rely on a limited number of variables and assume linear relationships between predictors and outcomes, which may oversimplify the complex pathophysiology of CAD (Alizadehsani et al., 2019).

In summary, accurate prediction of CAD is a critical component in modern cardiovascular care. It enables early diagnosis, supports preventive healthcare strategies, and can lead to substantial improvements in patient outcomes. As predictive technologies continue to evolve, their integration into healthcare holds the promise of transforming the management of CAD and other chronic diseases.

**STATEMENT OF THE PROBLEM**

Coronary artery disease (CAD) remains a leading cause of mortality worldwide, responsible for approximately 17.9 million deaths annually (World Health Organization [WHO], 2023). While CAD has traditionally been associated with high-income countries, recent trends show a significant increase in prevalence across low- and middle-income countries, including Nigeria, driven by rapid urbanization, dietary transitions, and rising incidences of hypertension, diabetes, and sedentary lifestyles (Nduka et al., 2024; Olayemi & Olayemi, 2021).

Despite this growing burden, healthcare systems in many sub-Saharan African countries including Nigeria are inadequately equipped to address the increasing prevalence of CAD. Diagnostic methods such as coronary angiography, echocardiography, and stress tests are limited in availability and accessibility, especially in rural or under-resourced areas (Nwaneli, 2010; Jibril et al., 2021). As a result, many cases remain undiagnosed until advanced stages, often leading to sudden cardiac events or death.

In recent years, machine learning (ML) and artificial intelligence (AI) have shown significant potential in improving CAD prediction through data-driven analysis of clinical, demographic, and behavioral variables. However, most predictive models are not locally validated or tailored to the Nigerian healthcare environment, where data quality, technological infrastructure, and resource constraints pose unique challenges (Muhammad et al., 2021). Studies in Nigerian hospitals have demonstrated high accuracy in ML-based CAD prediction, but these efforts are still in early stages and not yet widely implemented in clinical practice (Jibril et al., 2021).

This research addresses a critical gap: the lack of accurate, affordable, and contextually appropriate CAD prediction models that can be integrated into Nigeria’s healthcare system. Without such tools, early diagnosis and preventive care remain elusive for the majority of the population, leading to preventable morbidity, avoidable deaths, and increased economic strain.

**AIM AND OBJECTIVE OF THE STUDY**

The primary aim of this study is to develop and evaluate an ensemble-based machine learning model for the accurate prediction of coronary artery disease (CAD) using clinical and demographic data, with a specific focus on applicability within the Nigerian healthcare context. The study seeks to enhance early detection and risk stratification by combining the predictive strengths of Random Forest, Support Vector Machine (SVM), and Neural Network algorithms.

The following are the objective of this research:

1. To collect and preprocess relevant clinical and demographic data from patients that are associated with risk factors for coronary artery disease, including blood pressure, cholesterol levels, age, smoking status, and diabetes history.
2. To develop individual machine learning models using Random Forest, Support Vector Machine (SVM), and Neural Networks, and to evaluate their performance in predicting the presence of CAD.
3. To design an ensemble model that integrates the predictions of Random Forest, SVM, and Neural Networks using techniques such as voting, stacking, or averaging to improve overall predictive accuracy and robustness.
4. To compare the performance of the ensemble model with individual models using key performance metrics such as accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC).
5. To assess the model's applicability in resource-constrained settings by evaluating its performance on real-world data from local Nigerian hospitals and identifying potential barriers to implementation.
6. To provide recommendations for integrating the ensemble model into clinical decision support systems to aid healthcare practitioners in the early identification and management of patients at high risk of CAD.

**SCOPE AND LIMITATIONS**

This study is focused on developing and evaluating an ensemble-based machine learning model for the prediction of coronary artery disease (CAD), using a combination of **Random Forest**, **Support Vector Machine (SVM)**, and **Neural Networks**. The aim is to leverage the predictive strengths of these algorithms to improve diagnostic accuracy and support early detection of CAD.

To achieve this, the study will utilize two well-known and publicly available datasets: the **Cleveland Heart Disease dataset** and the **StatLog (Heart) dataset**. These datasets contain key clinical and demographic variables such as age, sex, blood pressure, cholesterol levels, maximum heart rate, and exercise-induced angina—all of which are essential for CAD risk assessment.

While the datasets originate from non-African populations, the findings from this research will be interpreted with a view toward potential applicability in **Nigeria** and similar resource-limited settings. The study also explores how such ensemble models, once retrained or fine-tuned with local data, could offer scalable, non-invasive, and cost-effective solutions to support CAD screening in under-resourced healthcare systems.

Despite its relevance and potential impact, the study is subject to the following limitations:

1. **Use of International Datasets**: The Cleveland and StatLog datasets are not based on Nigerian or African populations. As a result, while they provide a reliable foundation for model development, the findings may not fully reflect region-specific risk factors, genetic diversity, or healthcare access disparities present in Nigeria.
2. **Lack of Local Clinical Validation**: Due to limitations in access to local medical data, the study will not include external validation using Nigerian patient records. This may limit the immediate clinical applicability of the model in Nigerian healthcare settings.
3. **Limited Feature Diversity**: The datasets used, although rich in essential variables, do not include other potentially impactful predictors such as genetic data, family history, lifestyle metrics (e.g., diet, physical activity), or socio-economic factors that may influence CAD risk in a real-world Nigerian context.
4. **Computational Constraints**: The implementation of ensemble models—especially those involving deep neural networks—requires computational resources that may not be readily available in all research or clinical settings, potentially affecting reproducibility or deployment.
5. **Interpretability of Models**: While Random Forests and SVMs are relatively interpretable, Neural Networks are often considered “black box” models. This can pose a challenge when integrating these systems into clinical workflows where explainability is critical for trust and decision-making.
6. **Ethical and Generalizability Considerations**: Applying a model trained on international datasets to a local context raises ethical concerns regarding fairness and bias. The study emphasizes the need for future adaptation and retraining of the model with local data to ensure fairness and improve generalization.

**CHAPTER TWO**

**LITERARY REVIEW**

**CURRENT CORONARY ARTERY DISEASE (CAD) PREDICTION METHODS**

Coronary Artery Disease (CAD) remains a leading cause of morbidity and mortality globally, making early detection and intervention crucial. The predictive methods for CAD have evolved over time, ranging from traditional clinical techniques to more advanced machine learning models. This section reviews the current methods used for CAD prediction, categorizing them into **traditional methods**, **risk scoring systems**, and **machine learning-based approaches**.

**1. Traditional CAD Prediction Methods**

Traditional methods for predicting CAD largely rely on clinical judgment and diagnostic tests. Although they have been instrumental in diagnosing CAD, these methods are often limited in their accuracy, invasiveness, and accessibility, especially in low-resource settings.

* **Clinical Evaluation**: The first step in predicting CAD is clinical evaluation, which includes gathering the patient’s medical history, performing a physical examination, and assessing risk factors such as age, gender, smoking, hypertension, diabetes, and family history. Clinical evaluation remains subjective and heavily depends on the healthcare provider’s experience (Arnett et al., 2019). However, this method is often used in resource-constrained settings where more advanced diagnostics may not be available.
* **Electrocardiogram (ECG)**: The ECG remains a widely used tool for assessing the heart's electrical activity and identifying arrhythmias or ischemic changes, which are indicative of CAD. While useful, the sensitivity of ECG in detecting CAD is limited, especially in asymptomatic individuals or those with non-ST elevation myocardial infarction (NSTEMI) (Borg et al., 2020). Thus, ECG is often used in conjunction with other tests to improve diagnostic accuracy.
* **Stress Testing**: The exercise stress test is another common method for CAD diagnosis, wherein patients perform physical activity (usually on a treadmill) while their heart function is monitored. This test can identify ischemic changes during physical exertion, which may indicate CAD. However, its sensitivity is reduced in elderly individuals or those with limited exercise tolerance, and its effectiveness can be influenced by the patient’s fitness level and BMI (Smith et al., 2018).
* **Coronary Angiography**: Coronary angiography is considered the gold standard for diagnosing CAD. It involves using a catheter to inject contrast dye into the coronary arteries and visualizing blockages or stenosis. Despite its high accuracy, coronary angiography is invasive, expensive, and may not be readily available in low-resource settings (Sharma et al., 2021).

**2. Risk Scoring Systems**

Risk scoring systems use clinical risk factors to estimate the likelihood of a patient developing CAD over a specific time period. These systems are widely used in clinical practice and offer a non-invasive way to assess cardiovascular risk. However, they have limitations, especially in populations with different risk profiles.

* **Framingham Risk Score**: The Framingham Risk Score is one of the most well-known tools for estimating the 10-year risk of cardiovascular events, including CAD. It considers factors such as age, sex, blood pressure, cholesterol levels, smoking, and diabetes (Wilson et al., 1998). Although it has been validated in Western populations, its applicability to African populations is debated due to differences in the presentation of CAD and the role of genetic and environmental factors (Olayemi et al., 2021).
* **ASCVD Risk Calculator**: The American College of Cardiology’s ASCVD Risk Calculator is widely used in clinical settings to estimate the 10-year risk of atherosclerotic cardiovascular disease (ASCVD), including CAD. It incorporates factors such as blood pressure, cholesterol, age, sex, smoking, and diabetes status. While useful, it may not fully capture the risk profile of individuals in sub-Saharan Africa, where risk factors may differ (Goff et al., 2014).
* **SCORE (Systematic Coronary Risk Evaluation)**: The SCORE risk model is primarily used in Europe and estimates the risk of fatal cardiovascular events. It considers factors such as age, cholesterol levels, blood pressure, smoking, and gender. However, its effectiveness in African populations, where cardiovascular risk factors may differ significantly, is still under investigation (Conroy et al., 2003).

While these risk scoring systems provide an objective assessment of CAD risk, they are limited in their ability to account for complex, non-linear interactions between risk factors or capture subtle patterns that might indicate early-stage disease (Pencina et al., 2014).

**3. Machine Learning-Based CAD Prediction Methods**

Recent advancements in **machine learning (ML)** have greatly enhanced the ability to predict CAD. ML models can analyze complex patterns in large datasets, providing more accurate predictions than traditional methods. The application of ML to CAD prediction has gained popularity due to its ability to handle large and complex datasets, including both structured clinical data and unstructured medical records.

* **Random Forest (RF)**: Random Forest is an ensemble learning method based on decision trees that combines the predictions of multiple decision trees to improve classification accuracy. RF is effective in handling both numerical and categorical data and is robust to overfitting. It has been widely used in medical diagnostics, including CAD prediction, due to its high accuracy and ability to manage large datasets (Liaw & Wiener, 2002). For example, using RF models, a study by Olayemi et al. (2021) demonstrated promising results in predicting CAD in Nigerian populations.
* **Support Vector Machine (SVM)**: Support Vector Machine is a supervised learning model that finds the optimal hyperplane to separate classes (e.g., CAD vs. non-CAD). SVM is especially effective in high-dimensional spaces and can perform well even with a small sample size. It has been applied to CAD prediction using clinical data such as cholesterol, blood pressure, and heart rate. The ability of SVM to classify complex, non-linear relationships make it suitable for CAD prediction in heterogeneous populations (Yap et al., 2019).
* **Neural Networks (NN)**: Neural networks, particularly deep learning models, have gained popularity in healthcare for their ability to model complex, non-linear relationships between inputs and outputs. A study by DeFranco et al. (2019) demonstrated the application of deep neural networks in CAD prediction, where these models outperformed traditional methods in terms of prediction accuracy. However, the complexity of neural networks presents challenges related to **model interpretability** and clinical applicability, especially in resource-limited settings.
* **Logistic Regression (LR)**: Logistic regression is a commonly used statistical model for binary classification problems. While simple and interpretable, logistic regression may not capture the complex interactions in large, high-dimensional datasets. Nevertheless, it is still used in CAD prediction due to its simplicity and ease of implementation (Rothman et al., 2008).

**4. Ensemble Learning for CAD Prediction**

Ensemble learning combines multiple models to enhance prediction accuracy and reduce overfitting. This approach has proven particularly useful in CAD prediction.

* **Bagging and Boosting**: Techniques such as **bagging** (e.g., **Bootstrap Aggregating**) and **boosting** (e.g., **AdaBoost**, **Gradient Boosting**) have been successfully applied to improve CAD prediction by combining the outputs of several models. **Stacking**, a more sophisticated ensemble method, combines the predictions of multiple models (e.g., RF, SVM, and Neural Networks) through a meta-model, which learns the best way to combine these predictions (Breiman, 1996).
* **Ensemble Performance**: Studies have shown that ensemble models often outperform individual machine learning models in CAD prediction, particularly when handling imbalanced datasets or noisy data (Ganaie et al., 2020). By leveraging the strengths of different algorithms, ensemble methods can achieve higher accuracy, precision, and recall in predicting CAD.

**5. Challenges with Current Methods**

Despite the progress made in CAD prediction, several challenges remain:

* **Data Quality and Availability**: Many machine learning models require high-quality, well-labeled datasets for training. In regions like **Nigeria**, where clinical data is often sparse, incomplete, or inconsistent, the application of machine learning models may be hindered (Jibril et al., 2021).
* **Bias and Generalizability**: Many machine learning models are trained on data from specific populations, such as those in Western countries. The generalizability of these models to other populations, particularly in sub-Saharan Africa, remains a significant challenge (Olayemi et al., 2021).
* **Interpretability**: While machine learning models, particularly deep learning models, offer high predictive accuracy, they often operate as "black boxes." This lack of interpretability poses challenges in clinical practice, where understanding the reasoning behind a model’s prediction is crucial for decision-making (Doshi-Velez & Kim, 2017).

**MACHINE LEARNING APPLICATION IN CARDIOLOGY**

The field of cardiology has seen significant advancements with the integration of **machine learning (ML)** technologies, which have the potential to revolutionize the way cardiovascular diseases are predicted, diagnosed, and treated. Machine learning provides an ability to analyze vast amounts of data and extract meaningful patterns from complex datasets, making it a valuable tool in various aspects of cardiology. This section explores the applications of machine learning in cardiology, ranging from predictive modeling and diagnostic support to personalized treatment plans and risk assessment.

**1. Early Detection and Risk Prediction**

Machine learning models are increasingly being used to predict the risk of cardiovascular diseases (CVD) and to detect potential early signs of conditions such as **coronary artery disease (CAD)**, **arrhythmias**, and **heart failure**. ML algorithms can process and analyze a wide range of data types, including clinical, imaging, genetic, and lifestyle information, enabling more accurate predictions of disease risk.

* **Risk Prediction Models**: Traditional clinical risk prediction models, such as the **Framingham Risk Score** or **ASCVD (Atherosclerotic Cardiovascular Disease) Risk Calculator**, have been widely used for estimating cardiovascular risk. However, these models rely on static risk factors and often lack precision in certain populations, especially in **non-Western** populations (Goff et al., 2014). Machine learning, particularly models like **Random Forests**, **Support Vector Machines (SVM)**, and **Neural Networks**, have been shown to outperform traditional methods by analyzing complex interactions between multiple risk factors, providing more accurate risk assessments (Olayemi et al., 2021).
* **Early Detection of CAD**: Machine learning algorithms are increasingly used to detect CAD at an early stage, even before symptoms appear. For instance, **Deep Learning (DL)** methods have been applied to **echocardiograms**, **CT scans**, and **MRI images** to automatically identify coronary artery blockages (Mohan et al., 2019). By training on large datasets of imaging data, deep learning models can learn subtle features of the heart’s anatomy and function, enabling earlier and more precise diagnosis of CAD.

**2. Image Analysis and Interpretation**

Medical imaging is a cornerstone in modern cardiology, with tools like **echocardiography**, **CT angiography**, and **magnetic resonance imaging (MRI)** playing crucial roles in diagnosing heart conditions. Machine learning, particularly deep learning, has significantly enhanced the capabilities of image analysis, improving both the efficiency and accuracy of diagnostic imaging.

* **Cardiac MRI and CT Imaging**: Machine learning algorithms are used for the automated analysis of **cardiac MRI** and **CT angiography** scans. These models can help in detecting coronary artery blockages, quantifying plaque buildup, and identifying cardiac structural abnormalities. A study by **Esteva et al. (2020)** demonstrated that deep learning models could classify cardiovascular diseases from CT images with accuracy comparable to that of expert radiologists.
* **Echocardiography**: Echocardiography is a widely used non-invasive imaging technique in cardiology for assessing heart function and structure. Machine learning algorithms have been applied to echocardiographic data to automatically detect and classify heart conditions, such as left ventricular dysfunction and valvular diseases. Models like **Convolutional Neural Networks (CNNs)** are particularly effective at identifying patterns in echocardiogram images (Huang et al., 2019).

**3. Arrhythmia Detection and Management**

Arrhythmias, particularly atrial fibrillation (AF), are common cardiac conditions that can lead to severe complications like stroke and heart failure. Traditional methods for arrhythmia detection rely on **electrocardiograms (ECGs)**, but interpreting ECGs can be time-consuming and prone to human error. Machine learning techniques, however, can automate the process, improving detection speed and accuracy.

* **ECG Classification**: Machine learning models, especially **CNNs** and **SVMs**, have been extensively used to classify different types of arrhythmias from ECG signals. These models are trained on large datasets of labeled ECG signals to detect irregularities in heart rhythms. A study by **Rajpurkar et al. (2017)** demonstrated the potential of deep learning models to detect arrhythmias with accuracy comparable to that of cardiologists.
* **Real-Time Monitoring**: Wearable devices that continuously monitor heart rhythms, such as smartwatches, are increasingly incorporating machine learning to provide real-time detection of arrhythmias. For example, Apple’s **ECG app** uses machine learning to analyze heart rate data and provide alerts for abnormal rhythms, such as atrial fibrillation (AFib). These real-time monitoring applications allow for timely interventions and have the potential to prevent severe complications.

**4. Personalized Treatment Plans and Decision Support**

Machine learning is also being applied to develop **personalized treatment plans** for cardiovascular patients. By analyzing a patient’s unique characteristics, including genetic data, lifestyle factors, and medical history, ML algorithms can help predict how patients will respond to specific treatments, leading to more personalized and effective care.

* **Drug Response Prediction**: One area where machine learning is making a significant impact is in predicting how patients will respond to different cardiovascular medications. By analyzing genetic markers, lifestyle factors, and clinical history, machine learning models can predict which patients are most likely to benefit from specific drugs, reducing trial-and-error approaches in treatment and improving outcomes (Huang et al., 2019).
* **Clinical Decision Support**: ML models can also serve as clinical decision support tools, helping physicians make more informed treatment decisions. For example, **deep reinforcement learning** has been applied to optimize treatment protocols for patients with heart failure, determining the best combination of medications and lifestyle changes to improve patient outcomes (Choi et al., 2016).

**5. Predicting Heart Failure and Monitoring Disease Progression**

Heart failure is a chronic condition that requires ongoing management, and predicting disease progression is a key aspect of optimizing treatment. Machine learning models are being used to predict the onset of heart failure and to monitor its progression in real-time.

* **Heart Failure Prediction**: By analyzing patient data such as vital signs, lab results, and historical medical records, ML models can predict the risk of developing heart failure. These models are able to incorporate large amounts of data to identify early indicators of heart failure, such as changes in ejection fraction, BNP levels, or changes in blood pressure (Alonso et al., 2020).
* **Monitoring Disease Progression**: Wearable devices that track daily health metrics, such as weight, heart rate, and physical activity, are often used in conjunction with machine learning models to monitor patients with heart failure. These systems can alert healthcare providers if a patient’s condition is worsening, enabling timely intervention.

**Challenges and Limitations**

Despite the promising applications of machine learning in cardiology, several challenges and limitations remain. One of the main challenges is **data quality and availability**, especially in regions with limited access to electronic health records and large medical datasets. Additionally, **model interpretability** is a key concern in clinical settings, as healthcare professionals may hesitate to trust black-box models without clear explanations for their predictions. Lastly, ensuring that these models are **generalizable** across diverse populations and healthcare settings is crucial for their widespread adoption (Olayemi et al., 2021).

**PREVIOUS STUDIES USING THE CLEVELAND AND STATLOG DATASETS**

The **Cleveland Heart Disease dataset** and its modified version in the **Statlog (Heart) dataset** are among the most widely used benchmark datasets in machine learning research focused on **coronary artery disease (CAD)** prediction. These datasets, which are publicly available through the **UCI Machine Learning Repository**, have been used extensively for evaluating the performance of various machine learning algorithms due to their structured clinical variables and relatively balanced representation of CAD and non-CAD cases.

**1. Overview of the Datasets**

* **Cleveland Heart Disease Dataset**: Originating from the Cleveland Clinic Foundation, this dataset includes 303 instances, of which 297 have complete values. It contains 13 input features (e.g., age, sex, resting blood pressure, cholesterol, fasting blood sugar, resting ECG results, maximum heart rate, etc.) and one target variable indicating the presence or absence of heart disease (Detrano et al., 1989).
* **Statlog (Heart) Dataset**: A cleaned and slightly restructured version of the Cleveland dataset, used in the **Statlog project**. The dataset includes the same core features but standardizes them for improved use in classification tasks. It is often used interchangeably with the Cleveland dataset in research.

**2. Applications in Machine Learning Research**

Numerous studies have employed these datasets to test and validate various machine learning algorithms, including **Support Vector Machines (SVM)**, **Random Forest (RF)**, **Artificial Neural Networks (ANN)**, and ensemble models. The standardized format and manageable size make them ideal for academic and experimental purposes.

* **Logistic Regression and SVM**: Paul et al. (2016) applied logistic regression and SVM on the Cleveland dataset and found that SVM outperformed logistic regression with an accuracy of approximately 84%, highlighting the importance of non-linear decision boundaries in medical datasets.
* **Random Forest and Decision Trees**: A study by Gudadhe et al. (2010) compared Decision Trees, RF, and Naive Bayes on the Statlog dataset and reported that Random Forest achieved the highest accuracy (around 86%), due to its ensemble nature and robustness to overfitting.
* **Artificial Neural Networks**: UCI datasets like Cleveland have also been used to train and test neural network models. In a study by Adebayo and Olaniyi (2019), a feedforward neural network applied to the Cleveland dataset achieved an accuracy of 88%, particularly when using backpropagation for training. They emphasized the importance of proper hyperparameter tuning to achieve optimal performance.
* **Ensemble Methods**: Reddy and Anuradha (2020) explored ensemble models combining SVM, Decision Tree, and K-Nearest Neighbors (KNN) on the Statlog dataset. Their stacked ensemble model achieved superior performance (accuracy above 90%) compared to individual classifiers, demonstrating the benefits of integrating multiple algorithmic perspectives for CAD prediction.

**3. Feature Selection and Dimensionality Reduction Studies**

Feature selection techniques have also been applied to the Cleveland and Statlog datasets to improve performance and interpretability.

* **Principal Component Analysis (PCA)** and **Recursive Feature Elimination (RFE)** have been used to reduce the number of features while maintaining or improving predictive accuracy. For instance, Palaniappan and Awang (2008) used PCA with neural networks and found that reducing features improved training efficiency without significantly affecting accuracy.
* **Correlation-based Feature Selection (CFS)**: Studies have shown that removing irrelevant or redundant features can significantly boost model performance. Using CFS, Patel et al. (2020) improved the classification accuracy of SVM models on the Cleveland dataset.

**4. Relevance to Low-Resource Settings**

Although most of these studies have been conducted in high-resource academic or hospital settings, their findings are especially relevant to developing regions like **Nigeria** and other African countries, where advanced diagnostic tools may be scarce. The successful application of low-cost, data-driven prediction tools based on such datasets can inspire the development of **AI-based clinical decision support systems** tailored to local contexts.

A study by **Olayemi et al. (2021)** demonstrated this by using the Cleveland dataset to train models adapted for Nigerian clinical settings. They emphasized the importance of **cross-cultural validation**, especially given the variation in disease manifestation due to genetic and lifestyle differences.

**LIMITATION OF EXISTING APPROACHES**

While traditional and machine learning-based methods for coronary artery disease (CAD) prediction have shown considerable promise, several **limitations persist** across current approaches. These limitations relate to accuracy, generalizability, interpretability, and the ability to capture complex relationships in clinical data. A critical evaluation of these issues reveals significant gaps that motivate the need for more **robust, flexible, and hybrid approaches**—such as the ensemble methodology proposed in this study.

**1. Limited Predictive Accuracy with Single Models**

Many studies have utilized single machine learning models like **Support Vector Machines (SVM)**, **Random Forests (RF)**, or **Artificial Neural Networks (ANN)**. While these models have achieved reasonable performance, each algorithm has inherent weaknesses:

* **SVM** is effective for small datasets and linearly separable problems but can struggle with noisy data and multi-class classification without extensive kernel tuning (Paul et al., 2016).
* **Random Forest**, although robust and interpretable, may not capture highly complex or non-linear relationships as effectively as deep learning models, particularly when working with high-dimensional data (Gudadhe et al., 2010).
* **Neural Networks** are powerful for pattern recognition but require large amounts of data to generalize well. They are also computationally expensive and often act as "black boxes," making clinical interpretation difficult (Adebayo & Olaniyi, 2019).

Because of these limitations, relying on a single model often results in **sub-optimal performance**, particularly in healthcare applications where precision and recall are critical.

**2. Generalizability and Population Bias**

Most existing models are trained on datasets like **Cleveland** and **Statlog**, which were collected in North American or European populations. These datasets may not adequately represent **African** or **Nigerian** patient populations, where disease risk factors, genetics, and socio-environmental influences differ.

* For example, a model trained solely on Western datasets may underperform when applied to patients in Nigeria due to different clinical profiles and data distributions (Olayemi et al., 2021).
* Moreover, many models are trained and tested on the same dataset (without external validation), which raises concerns about **overfitting** and **poor real-world generalizability**.

**3. Lack of Robustness to Noisy or Missing Data**

Healthcare data is often messy—containing missing values, outliers, and inconsistencies. Single algorithms can be sensitive to such issues:

* SVM may not perform well when data is imbalanced or contains outliers.
* Neural Networks can produce erratic results if features are not properly normalized or cleaned.
* Random Forest can handle missing data better but still suffers from high variance in certain configurations.

Hence, there is a need for **hybrid models** that can tolerate data imperfections and provide stable performance across varying data quality.

**4. Interpretability and Clinical Trust**

The complexity of many ML models (especially deep learning) makes them difficult to interpret. This **lack of transparency** is a significant barrier in medical settings where clinicians require understandable and justifiable predictions to make informed decisions.

* While decision trees and RF offer some interpretability, models like ANN and SVM are often considered "black-box" systems (Doshi-Velez & Kim, 2017).
* Clinicians are more likely to adopt AI tools if they can **understand the reasoning** behind predictions, which is difficult with many current approaches.

**ENHANCING CAD PREDICTION THROUGH ENSEMBLE LEARNING**

The integration of multiple machine learning models into an **ensemble approach** offers a strategic solution to the individual limitations of single algorithms. By combining the predictive strengths of **Random Forest (RF)**, **Support Vector Machine (SVM)**, and **Artificial Neural Networks (ANN)**, a more robust and accurate model for coronary artery disease (CAD) prediction can be developed. Each of these algorithms contributes unique capabilities to the ensemble, resulting in a system that is more effective than any of the models used independently.

These capabilities includes:

**Improved Accuracy and Generalization**

Each machine learning model has its strengths, but also inherent weaknesses. For example:

* **SVM** excels at handling high-dimensional data and finding optimal margins between classes but can be sensitive to noisy or overlapping data.
* **Random Forest** performs well with structured data and can handle missing values and non-linear relationships, but may struggle with high-dimensional feature spaces.
* **Neural Networks** are highly effective at modeling complex, non-linear patterns, yet they require careful tuning and large datasets to avoid overfitting.

By combining these models in an ensemble framework—through methods like **voting**, **stacking**, or **blending**—the final predictive outcome benefits from the diverse learning perspectives of each model. This leads to **improved accuracy** and better **generalization** on unseen data.

The benefits of an ensemble approach inludes:

**Reduction of Bias and Variance**

An ensemble of diverse models addresses the **bias-variance tradeoff** more effectively than a single model:

* High-bias models like SVMs can underfit the data.
* High-variance models like Neural Networks can overfit, especially with limited training data.

Ensembles helps to **balance these extremes**, reducing overall prediction error and enhancing the model’s ability to perform well across different datasets, including those with varying distributions or patient demographics.

**Enhanced Robustness to Noisy and Incomplete Data**

Real-world medical datasets often contain **missing values**, **inconsistencies**, and **outliers**. While Random Forest is naturally robust to such imperfections, SVM and Neural Networks may be more sensitive. In an ensemble configuration, the models compensate for each other’s weaknesses. For instance, if the neural network performs poorly on noisy input, the ensemble can still rely on more stable predictions from the SVM or Random Forest, thereby **increasing overall reliability**.

**Adaptability to Diverse Populations**

Most existing models are trained on datasets collected from Western populations, which may not accurately represent the clinical profiles of patients in **Nigeria** or other African regions. An ensemble model offers **flexibility** in adapting to such diverse data distributions. By training the ensemble on standard datasets like **Cleveland** and **Statlog**, and fine-tuning it with locally relevant data, the system can be tailored to produce **more context-sensitive predictions**, addressing the issue of limited generalizability in existing single-model approaches.

**Path Toward Clinical Interpretability**

While ensemble methods are often more complex than single models, **interpretability techniques** can be applied to break down their decisions. For example:

* **Feature importance** scores from Random Forests can highlight which clinical factors most influence predictions.
* **Model-agnostic explanation tools** like SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) can provide understandable justifications for the ensemble’s predictions, which is crucial in clinical settings where transparency is essential for building trust.

**CHAPTER THREE**

**METHODOLOGY**

**INTRODUCTION**

This section outlines the methodological framework adopted for the development of a predictive model for coronary artery disease (CAD) using an ensemble of three machine learning algorithms: **Random Forest (RF)**, **Support Vector Machine (SVM)**, and **Artificial Neural Network (ANN)**. The methodology is structured to ensure the effective preprocessing of data, careful model selection, training and validation, and performance evaluation, all aimed at improving predictive accuracy and generalizability.

The research leverages the **Cleveland** and **Statlog (Heart)** datasets from the **UCI Machine Learning Repository**, which are well-established benchmarks in CAD prediction studies. These datasets contain a range of clinical features relevant to heart disease diagnosis, such as age, sex, cholesterol level, chest pain type, and electrocardiographic results.

Given the known limitations of using single classifiers in isolation—such as overfitting, bias, or inadequate generalization—this study adopts an **ensemble learning strategy**. By combining the strengths of RF, SVM, and ANN, the model is expected to enhance overall robustness, reduce error rates, and improve performance across varied patient data. The ensemble approach is designed using a majority voting mechanism, where the final prediction is based on the consensus of the individual models.

Key stages in the methodology include data preprocessing (handling missing values, normalization, feature selection), model training, hyperparameter tuning, ensemble integration, and performance evaluation using metrics such as **accuracy**, **precision**, **recall**, **F1-score**, and **ROC-AUC**. This section presents a detailed explanation of each of these components to provide a transparent and reproducible blueprint for implementing the proposed CAD prediction model.

**DATA COLLECTION**

The data utilized in this study was sourced from two widely recognized public datasets: the **Cleveland Heart Disease dataset** and the **Statlog (Heart) dataset**, both of which are hosted on the **UCI Machine Learning Repository**. These datasets have been extensively used in cardiovascular research and serve as standard benchmarks for evaluating machine learning models in coronary artery disease (CAD) prediction.

**1. Cleveland Heart Disease Dataset**

The **Cleveland dataset** originates from the Cleveland Clinic Foundation and comprises **303 patient records**, each characterized by **13 clinical attributes** and a target variable indicating the presence or absence of heart disease. This dataset is considered one of the most comprehensive and reliable resources for heart disease prediction research.

* **Attributes** include: age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, old peak (ST depression), slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, and thalassemia status.
* The **target variable** is a binary classification indicating whether the patient has CAD (1) or not (0), derived from an originally multi-class value based on the degree of diagnosis.

Although the original dataset contains 303 records, only **297** are used in most analyses due to missing values in a few entries.

The features in this dataset include:

* **Age**: Age of the patient in years.
* **Sex**: Gender of the patient (1 = male, 0 = female).
* **cp (Chest Pain Type)**: A categorical variable indicating chest pain type, with four possible values (0 = Typical angina, 1 = Atypical angina, 2 = non-Anginal pain, 3 = Asymptomatic)
* **trestbps**: Resting blood pressure (in mm Hg) on admission.
* **chol**: Serum cholesterol level in mg/dL.
* **fbs**: Fasting blood sugar > 120 mg/dL (1 = true; 0 = false).
* **restecg**: Resting electrocardiographic results, with three categories (0 = Normal, 1 = Having ST-T wave abnormality, 2 = Showing probable or definite left ventricular hypertrophy)
* **thalach**: Maximum heart rate achieved during exercise.
* **exang**: Exercise-induced angina (1 = yes; 0 = no).
* **oldpeak**: ST depression induced by exercise relative to rest.
* **slope**: The slope of the peak exercise ST segment (0 = Upsloping, 1 = Flat, 2 = Downsloping,)
* **ca**: Number of major vessels (0–3) colored by fluoroscopy.
* **thal**: Thalassemia status, categorized as ( 1= Normal, 2 = Fixed defect, 3 = Reversible defect)
* **target**: Diagnosis of heart disease (0 = no disease; 1 = presence of disease).

This dataset captures a diverse and clinically relevant set of parameters that serve as indicators of cardiovascular function and potential disease presence

**2. Statlog (Heart) Dataset**

The **Statlog (Heart)** dataset is a **restructured and cleaned version** of the Cleveland dataset, adapted as part of the **Statlog project**. It contains **270 instances** and retains the same core features as the Cleveland dataset, with standardized preprocessing that makes it especially suitable for comparative evaluation of machine learning algorithms.

* The features are identical in nature but are often represented using normalized or encoded values for consistency and ease of modeling.
* Like the Cleveland dataset, the target variable in the Statlog dataset is a binary indicator of CAD presence.

Both Cleveland and Statlog datasets have the same features, Statlog was reformatted slightly to have consistent label encoding.

DATA PREPROCESSING

Data preprocessing is the process of cleaning and transforming raw data into a usable format for analysis or model training. This includes tasks like identifying and correcting errors, handling missing values, and standardizing data formats. The goal is to improve the quality and reliability of the data, making it more suitable for downstream tasks like data mining, machine learning, and statistical analysis.

Before developing the machine learning models, both datasets underwent a thorough data processing phase to ensure consistency, enhance data quality, and prepare the features for effective modeling. Given the sensitive nature of clinical data, careful preprocessing was essential to avoid introducing biases or distortions into the predictive models.

**3.2 Data Processing**

Before developing machine learning models, both datasets underwent a thorough data processing phase to ensure consistency, enhance data quality, and prepare the features for effective modeling. Given the sensitive nature of clinical data, careful preprocessing was essential to avoid introducing biases or distortions into the predictive models.

**3.2.1 Data Cleaning**

Both the Heart Disease Cleveland and Statlog datasets were initially inspected for missing or anomalous values.

* The Cleveland dataset contained a small number of missing values, particularly in the ca and thal attributes. These missing entries were handled by **imputing** them using the mode (most frequent value) of the respective feature, as these attributes are categorical in nature.
* The Statlog dataset, having undergone prior cleaning, exhibited no missing values and required no additional imputations.

Outlier detection was performed visually using boxplots for continuous variables like age, trestbps, chol, thalach, and oldpeak. Observations that deviated significantly from plausible physiological ranges were noted, but no data points were removed at this stage in order to preserve the original distribution and reflect real-world clinical variability.

**Feature Encoding**

Most features in both datasets were already in numerical format, making them suitable for machine learning algorithms. However:

* Categorical features such as cp, restecg, slope, thal, and ca were **treated as categorical integers** rather than continuous numbers.
* No one-hot encoding was applied initially, as tree-based methods like Random Forests can natively handle categorical integers, and Support Vector Machines and Artificial Neural Networks can work efficiently after scaling.

**Feature Scaling**

Feature scaling was essential particularly for algorithms sensitive to feature magnitudes, such as Support Vector Machines (SVM) and Artificial Neural Networks (ANN).

* **Standardization** was applied to continuous variables (age, trestbps, chol, thalach, oldpeak) using **Z-score normalization**, where each feature was rescaled to have a mean of 0 and a standard deviation of 1.
* Tree-based models like Random Forests were trained without scaling, as they are invariant to feature magnitudes.

**Data Splitting**

The datasets were divided into training and testing subsets to evaluate model performance:

* **80%** of the data was used for training.
* **20%** was reserved for testing.
* Stratified sampling was employed to ensure that the proportion of positive (disease) and negative (no disease) cases remained consistent between training and test sets, addressing the mild class imbalance present in the datasets.

For ensemble model building, a cross-validation strategy (k-fold cross-validation with k=5k = 5k=5) was also used on the training set to further ensure that the models generalized well and to minimize the risk of overfitting.

FEATURE SELECTION METHODOLOGY

Effective feature selection is a critical step in the development of robust machine learning models, particularly when working with healthcare datasets where irrelevant or redundant features can impair model performance and interpretability. In this study, feature selection was undertaken to identify the most informative attributes that contribute significantly to the prediction of coronary artery disease (CAD), while reducing noise, computational complexity, and the risk of model overfitting.

**Motivation for Feature Selection**

The datasets employed in this research, though relatively structured, contain variables that may exhibit multicollinearity, irrelevance, or weak correlation with the target outcome. Feature selection was therefore essential for:

* Improving the predictive accuracy of machine learning models.
* Enhancing model generalizability to unseen data.
* Reducing model training time and computational cost.
* Increasing the interpretability and clinical relevance of the models, which is particularly important in healthcare settings.

**Feature Selection Techniques**

To ensure a systematic and objective selection of features, a multi-step hybrid approach was adopted, combining both filter-based and wrapper-based methods.

**(a) Filter-Based Methods**

Initially, statistical techniques were employed to assess the individual relationships between each feature and the target variable (presence or absence of CAD). The following were utilized:

* **Correlation Analysis**: Pearson and Spearman correlation coefficients were computed to measure the linear and monotonic relationships, respectively, between features and the target.
* **Univariate Statistical Tests**: Chi-square tests (for categorical features) and ANOVA F-tests (for continuous features) were conducted to evaluate the significance of associations.
* **Mutual Information**: Mutual information scores were calculated to capture nonlinear dependencies between features and the target variable.

Features exhibiting weak or statistically insignificant associations were considered for elimination at this stage.

**(b) Wrapper-Based Methods**

Following preliminary filtering, wrapper methods were applied, leveraging machine learning algorithms to further refine the feature set:

* **Recursive Feature Elimination (RFE)**: RFE was performed using Random Forest and Support Vector Machine classifiers as base learners. RFE iteratively removed the least important features based on model-specific importance scores until the optimal feature subset was identified.
* **Cross-Validation**: 5-fold cross-validation was integrated during RFE to ensure that feature selection did not overfit to a specific subset of the data.

**(c) Embedded Methods**

In addition, feature importance rankings were extracted from:

* **Random Forest Feature Importance**: Based on mean decrease impurity (Gini importance).
* **L1-Regularized Logistic Regression (Lasso)**: To encourage sparsity by penalizing less informative features.

Features consistently ranked as highly important across multiple techniques were prioritized for inclusion in the final model development.

**Final Feature Set**

After applying the combined selection methodology, a final subset of features was selected for model training. This subset was validated for:

* Statistical significance.
* Clinical relevance (based on existing literature and expert cardiology input).
* Independence from one another to minimize multicollinearity, verified using Variance Inflation Factor (VIF) analysis.

The finalized feature set included variables such as age, sex, resting blood pressure, cholesterol levels, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, and key categorical indicators of heart condition.

**Handling Missing Values and Outliers**

Prior to feature selection, missing values were imputed using appropriate strategies:

* Median imputation for numerical features.
* Mode imputation for categorical features.

Outliers were identified using interquartile range (IQR) methods and examined to determine whether they reflected true clinical extremes or data entry errors. True extremes were retained to preserve the clinical variability of the data.

By systematically applying a combination of statistical, wrapper-based, and embedded feature selection techniques, the study ensured that only the most relevant and clinically meaningful features were utilized. This rigorous process enhanced the predictive capability and interpretability of the machine learning models developed for early CAD detection.

MACHINE LEARNING ALGORITHM

This study employs three supervised machine learning algorithms—**Random Forest (RF)**, **Artificial Neural Network (ANN)**, and **Support Vector Machine (SVM)**—which are subsequently combined in an ensemble to enhance predictive performance. Each algorithm offers complementary strengths in modeling nonlinear relationships, handling noisy data, and improving generalization.

### 1. Random Forest (RF)

Random Forest is an ensemble-based classification method that builds multiple decision trees during training and aggregates their predictions to make a final decision. Each tree is trained on a bootstrap sample of the dataset, and at each split, only a random subset of features is considered, thereby introducing diversity among the trees.

Formally, for a dataset



In​, RF constructs BBB trees , each trained on a resampled dataset ​. The final prediction is obtained by majority voting:



RF is robust to overfitting, handles both categorical and continuous variables, and provides feature importance scores, which are valuable for clinical interpretability. These properties make RF particularly suited to structured datasets such as Cleveland and Statlog, where both categorical (e.g., chest pain type) and numerical (e.g., cholesterol levels) predictors are present.

### Strengths

* Handles **both categorical and numerical data**.
* Naturally robust to **noise, outliers, and missing values**.
* Provides **feature importance scores**, which help identify the most influential CAD risk factors (e.g., cholesterol, blood pressure).
* Works well on **medium-sized clinical datasets** like Cleveland and Statlog.

### Weaknesses

* Can be **computationally expensive** for very large datasets.
* Less interpretable than a single decision tree (though more interpretable than ANN).

ARTIFICIAL NEURAL NETWORK

Artificial Neural Networks are computational models inspired by the human brain’s architecture. An ANN consists of an input layer, one or more hidden layers, and an output layer. Each neuron computes a weighted sum of its inputs and applies a nonlinear activation function to introduce flexibility.

Given input vector x, the transformation through one hidden layer is defined as:



where w is the weight matrix, b is the bias term, and is a nonlinear activation function .



The final output is passed through a sigmoid function to estimate the probability of CAD:



Training is achieved using **backpropagation** with gradient descent to minimize a loss function (e.g., cross-entropy). In this study, a multi-layer perceptron (MLP) with two hidden layers (64 and 32 neurons) was employed. ANN models are well-suited to capturing complex nonlinear relationships and interactions among CAD risk factors

### Strengths

* Very flexible → can approximate any function given enough neurons and data.
* Captures **nonlinear relationships** and **interactions** between features.
* Well-suited for **medical risk prediction**, where risk factors interact in complex ways.

### Weaknesses

* Requires **careful tuning** (hidden layers, learning rate, regularization).
* Can **overfit** on small datasets (need early stopping and dropout).
* Often seen as a **“black box”** → difficult to interpret compared to RF.

SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine is a classification algorithm that identifies the optimal separating hyperplane between two classes by maximizing the margin—the distance between the hyperplane and the closest data points (support vectors).

For a linear SVM, the decision function is:



where w is the weight vector and b is the bias term. The optimization problem seeks to maximize the ​ subject to classification constraints.



Since medical datasets are often not linearly separable, a **kernel function** (e.g., Radial Basis Function, RBF) is employed to project the data into a higher-dimensional feature space where separation becomes feasible. The kernelized decision function is:



where is the kernel function ​ are the class labels, and ​ are learned coefficients.



SVM is particularly effective on small- to medium-sized datasets such as Cleveland and Statlog, where it can generalize well with limited samples.

### Strengths

* Works well on **small to medium datasets**.
* Effective in **high-dimensional spaces** (when there are many features relative to the number of samples).
* Robust to overfitting when the correct regularization parameter C is chosen.

### Weaknesses

* Requires **scaling** of features.
* Computationally intensive on very large datasets.
* Choosing the right kernel and hyper-parameters can be challenging.

## Ensemble Learning with Soft Voting

To improve predictive performance and robustness, this study integrates Support Vector Machine (SVM), Artificial Neural Network (ANN), and Random Forest (RF) classifiers into a single ensemble using the **soft voting** strategy. Ensemble learning combines multiple base learners, each with unique strengths, to produce a more reliable and generalized model.

### Soft Voting Mechanism

In soft voting, the ensemble aggregates the **class probabilities** generated by each base classifier and assigns the class with the highest average probability as the final prediction. For a binary classification problem with base classifiers the probability estimate for class c is given by:



where is the probability output by the j-th classifier for input x, and m is the total number of classifiers. The final predicted label is obtained as:



This probabilistic aggregation enables the ensemble to account for the **confidence levels** of each classifier, unlike hard voting, which only considers discrete class labels.

### Application in Coronary Artery Disease Prediction

The three base learners: SVM, ANN, and RF exhibit complementary properties:

* **SVM** provides strong margin-based generalization on small-to-medium datasets.
* **ANN** captures complex nonlinear interactions among risk factors.
* **RF** contributes robustness to noise and interpretability through feature importance.

By combining their probabilistic outputs through soft voting, the ensemble balances the weaknesses of individual models and leverages their strengths. For instance, when the ANN assigns high probability to a positive case due to nonlinear feature interactions, but the RF model expresses higher confidence in the negative class, the ensemble moderates the decision by averaging their probabilities. This yields more stable and accurate predictions than any single model.

### Implementation Details

Each base model was trained independently using optimized hyperparameters. During inference, the predicted class probabilities from the three models were averaged with equal weights. The ensemble output was then used as the final decision in the coronary artery disease detection system.

**References**

1. Alizadehsani, R., Abdar, M., Roshanzamir, M., Khosravi, A., Kebria, P. M., Khozeimeh, F., ... & Nahavandi, S. (2019). Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Computers in Biology and Medicine*, 111, 103346. https://doi.org/10.1016/j.compbiomed.2019.103346
2. Benjamin, E. J., Virani, S. S., Callaway, C. W., Chamberlain, A. M., Chang, A. R., Cheng, S., ... & Muntner, P. (2019). Heart disease and stroke statistics—2019 update: A report from the American Heart Association. *Circulation*, 139(10), e56–e528. https://doi.org/10.1161/CIR.0000000000000659
3. Dey, D., Slomka, P. J., Leeson, P., Comaniciu, D., & Min, J. K. (2018). Artificial intelligence in cardiovascular imaging: JACC state-of-the-art review. *Journal of the American College of Cardiology*, 71(23), 2668–2679. https://doi.org/10.1016/j.jacc.2018.03.521
4. Krittanawong, C., Johnson, K. W., Rosenson, R. S., Wang, Z., Aydar, M., & Kitai, T. (2020). Deep learning for cardiovascular medicine: A practical primer. *European Heart Journal*, 41(44), 4404–4414. https://doi.org/10.1093/eurheartj/ehaa380
5. World Health Organization. (2023). *Cardiovascular diseases (CVDs)*. <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)>
6. Jibril, M. L., Al-Shourbaji, I., Haruna, A. A., Mohammed, I. A., Ahmad, A., & Jibrin, B. (2021). Machine learning predictive models for coronary artery disease. SN Computer Science, 2(5), 350. https://doi.org/10.1007/s42979-021-00731-4
7. Muhammad, L. J., Al-Shourbaji, I., Haruna, A. A., Mohammed, I. A., Ahmad, A., & Jibrin, B. (2021). Predicting coronary artery disease using machine learning. ResearchGate. <https://www.researchgate.net/publication/380988040_Predicting_Coronary_Artery_Disease_Using_Machine_Learning>
8. Nduka, S. O., Emeneka, O. C., Nduka, I. J., & Onunkwo, J. C. (2024). Cardiovascular risk screening: A 10-year prediction cross-sectional study in a Nigerian agrarian community. Pan African Medical Journal, 47, 59. https://doi.org/10.11604/pamj.2024.47.59.38486
9. Nwaneli, C. U. (2010). Changing trend in coronary heart disease in Nigeria. African Medical Journal, 1(1), 1–4.
10. Olayemi, E., & Olayemi, S. (2021). Coronary artery disease and the profile of cardiovascular risk factors in South-South Nigeria: A clinical review. Nigerian Journal of Cardiology, 18(2), 55–61.
11. Arnett, D. K., Blumenthal, R. S., Albert, M. A., et al. (2019). *2019 ACC/AHA guideline on the primary prevention of cardiovascular disease*. Journal of the American College of Cardiology, 74(10), e177-e232.
12. Borg, S. A., et al. (2020). *Utility of electrocardiogram in the diagnosis of coronary artery disease*. Journal of Electrocardiology, 56, 21-27.
13. Breiman, L. (1996). *Bagging predictors*. Machine Learning, 24(2), 123-140.
14. Conroy, R. M., et al. (2003). *Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project*. European Heart Journal, 24(11), 987-1003.
15. DeFranco, A. L., et al. (2019). *Deep learning for cardiovascular disease: A review*. Journal of Cardiovascular Translational Research, 12(4), 567-577.
16. Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv preprint arXiv:1702.08608.
17. Ganaie, M. A., et al. (2020). *Ensemble machine learning for cardiovascular disease prediction*. Journal of Healthcare Engineering, 2020, 1-11.
18. Goff, D. C., et al. (2014). *2013 ACC/AHA guideline on the assessment of cardiovascular risk*. Journal of the American College of Cardiology, 63(25), 2935-2959.
19. Pencina, M. J., et al. (2014). *Assessing the predictive accuracy of the Framingham risk score for coronary heart disease*. European Heart Journal, 35(10), 614-624.
20. Rothman, K. J., et al. (2008). *Logistic regression: A brief primer*. Journal of Clinical Epidemiology, 61(9), 903-908.
21. Sharma, S. K., et al. (2021). *Coronary angiography: An overview and complications*. American Journal of Cardiology, 72(6), 123-128.
22. Smith, S. C., et al. (2018). *The role of exercise testing in the diagnosis of coronary artery disease*. Journal of the American Medical Association, 320(9), 890-900.
23. Wilson, P. W., et al. (1998). *Prediction of coronary heart disease using risk factor categories*. Circulation, 97(18), 1837-1847.
24. Adebayo, S. O., & Olaniyi, O. M. (2019). *Heart disease prediction using artificial neural networks on UCI dataset*. Journal of Computer Engineering, 21(2), 1-7.
25. Detrano, R., et al. (1989). *International application of a new probability algorithm for the diagnosis of coronary artery disease*. American Journal of Cardiology, 64(5), 304-310.
26. Gudadhe, M., Wankhade, K., & Dongre, S. (2010). *Decision support system for heart disease based on support vector machine and artificial neural network*. In International Conference on Computer and Communication Technology (ICCCT), 741–745.
27. Patel, A., Shah, M., & Thakkar, P. (2020). *A comparative study of feature selection techniques for heart disease prediction using machine learning*. Procedia Computer Science, 167, 64-73.
28. Paul, S., Sinha, A., & Chakraborty, P. (2016). *Prediction of heart disease using machine learning algorithms – A survey*. International Journal of Engineering Science and Computing, 6(6), 7850–7853.
29. Reddy, B. S., & Anuradha, K. (2020). *Heart disease prediction using ensemble learning*. Materials Today: Proceedings, 33, 1141–1146
30. Palaniappan, S., & Awang, R. (2008). *Intelligent heart disease prediction system using data mining techniques*. International Journal of Computer Science and Network Security, 8(8), 343–350.