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

Table of Contents

[1.1.0 BACKGROUND AND SIGNIFICANCE OF CAD PREDICTION 6](#__RefHeading___Toc1023_2659079202)

[1.2.0 STATEMENT OF THE PROBLEM 7](#__RefHeading___Toc1025_2659079202)

[1.3.0 AIM AND OBJECTIVE OF THE STUDY 8](#__RefHeading___Toc1027_2659079202)

[1.4.0 SCOPE AND LIMITATIONS 9](#__RefHeading___Toc1029_2659079202)

[2.1.0 CURRENT CORONARY ARTERY DISEASE (CAD) PREDICTION METHODS 11](#__RefHeading___Toc1031_2659079202)

[2.1.1 Traditional CAD Prediction Methods 11](#__RefHeading___Toc1033_2659079202)

[2.1.2 Risk Scoring Systems 12](#__RefHeading___Toc1035_2659079202)

[2.1.3 Machine Learning-Based CAD Prediction Methods 13](#__RefHeading___Toc1037_2659079202)

[2.1.4 Ensemble Learning for CAD Prediction 14](#__RefHeading___Toc1039_2659079202)

[2.1.5 Challenges with Current Methods 15](#__RefHeading___Toc1041_2659079202)

[2.2.0 MACHINE LEARNING APPLICATION IN CARDIOLOGY 15](#__RefHeading___Toc1043_2659079202)

[2.2.1 Early Detection and Risk Prediction 16](#__RefHeading___Toc1045_2659079202)

[2.2.2 Image Analysis and Interpretation 16](#__RefHeading___Toc1047_2659079202)

[2.2.3 Arrhythmia Detection and Management 17](#__RefHeading___Toc1049_2659079202)

[2.2.4 Personalized Treatment Plans and Decision Support 18](#__RefHeading___Toc1051_2659079202)

[2.2.5 Predicting Heart Failure and Monitoring Disease Progression 18](#__RefHeading___Toc1053_2659079202)

[2.2.6 Challenges and Limitations 19](#__RefHeading___Toc1055_2659079202)

[2.3.0 PREVIOUS STUDIES USING THE CLEVELAND AND STATLOG DATASETS 19](#__RefHeading___Toc1057_2659079202)

[2.3.1 Overview of the Datasets 19](#__RefHeading___Toc1059_2659079202)

[2.3.2 Applications in Machine Learning Research 20](#__RefHeading___Toc1061_2659079202)

[2.3.3 Feature Selection Studies 21](#__RefHeading___Toc1063_2659079202)

[2.3.4 Relevance to Low-Resource Settings 21](#__RefHeading___Toc1065_2659079202)

[2.4.0 RELATED WORKS 22](#__RefHeading___Toc1067_2659079202)

[2.4.1 Study 1: Improving Cardiovascular Disease Prediction With Machine Learning (Dorraki et al.) 22](#__RefHeading___Toc1013_2659079202)

[2.4.2 Study 2: Coronary Artery Disease Detection Using Deep Learning and Photon-Counting CT Angiography (Brendel et al.) 22](#__RefHeading___Toc1015_2659079202)

[2.4.3 Study 3: Ensemble-Based Multi-Layer Dynamic System for CVD Prediction (Uddin & Halder) 22](#__RefHeading___Toc1017_2659079202)

[2.4.4 Study 4: Non-Invasive Detection of CAD in High-Risk Patients (Alizandenhasi et al.) 23](#__RefHeading___Toc1019_2659079202)

[2.4.5 Summary of Related Works 23](#__RefHeading___Toc1021_2659079202)

[2.5.0 LITERATURE REVIEW TABLE 24](#__RefHeading___Toc1069_2659079202)

[2.6.0 LIMITATION OF EXISTING APPROACHES 25](#__RefHeading___Toc1071_2659079202)

[2.6.1 Limited Predictive Accuracy with Single Models 25](#__RefHeading___Toc1073_2659079202)

[2.6.2 Generalizability and Population Bias 26](#__RefHeading___Toc1075_2659079202)

[2.6.3 Lack of Robustness to Noisy or Missing Data 26](#__RefHeading___Toc1077_2659079202)

[2.6.4 Interpretability and Clinical Trust 27](#__RefHeading___Toc1079_2659079202)

[2.7.0 ENHANCING CAD PREDICTION THROUGH ENSEMBLE LEARNING 27](#__RefHeading___Toc1081_2659079202)

[2.7.1 Improved Accuracy and Generalization 27](#__RefHeading___Toc1083_2659079202)

[2.7.2 Reduction of Bias and Variance 28](#__RefHeading___Toc1085_2659079202)

[2.7.3 Enhanced Robustness to Noisy and Incomplete Data 28](#__RefHeading___Toc1087_2659079202)

[2.7.4 Adaptability to Diverse Populations 29](#__RefHeading___Toc1089_2659079202)

[2.7.5 Path Toward Clinical Interpretability 29](#__RefHeading___Toc1091_2659079202)

[2.8.0 CONCEPTUAL FRAMEWORK 29](#__RefHeading___Toc1093_2659079202)

[2.8.1 Coronary Artery Disease and the Need for Early Detection 30](#__RefHeading___Toc1095_2659079202)

[2.8.2 Decision Support Systems in Healthcare 30](#__RefHeading___Toc1097_2659079202)

[2.8.3 Machine Learning and Ensemble Models 30](#__RefHeading___Toc1099_2659079202)

[2.8.4 Conceptual Framework for the Study 30](#__RefHeading___Toc1101_2659079202)

[3.1.0 INTRODUCTION 32](#__RefHeading___Toc1103_2659079202)

[3.2.0 PROJECT ARCHITECTURE 33](#__RefHeading___Toc1105_2659079202)

[3.3.0 DATA COLLECTION 34](#__RefHeading___Toc1107_2659079202)

[3.3.1 Cleveland Heart Disease Dataset 34](#__RefHeading___Toc1109_2659079202)

[3.3.2 Statlog (Heart) Dataset 35](#__RefHeading___Toc1111_2659079202)

[3.4.0 DATA PREPROCESSING 36](#__RefHeading___Toc1113_2659079202)

[3.4.1 Data Processing 36](#__RefHeading___Toc1115_2659079202)

[3.4.2 Data Cleaning 36](#__RefHeading___Toc1117_2659079202)

[3.4.3 Feature Encoding 37](#__RefHeading___Toc1119_2659079202)

[3.4.4 Feature Scaling 37](#__RefHeading___Toc1121_2659079202)

[3.4.5 Data Splitting 38](#__RefHeading___Toc1123_2659079202)

[3.5.0 FEATURE SELECTION METHODOLOGY 38](#__RefHeading___Toc1125_2659079202)

[3.5.1 Motivation for Feature Selection 38](#__RefHeading___Toc1127_2659079202)

[3.5.2 Feature Selection Techniques 39](#__RefHeading___Toc1129_2659079202)

[3.5.3 Final Feature Set 40](#__RefHeading___Toc1131_2659079202)

[3.6.0 MACHINE LEARNING ALGORITHM 40](#__RefHeading___Toc1133_2659079202)

[3.6.1 Random Forest (RF) 41](#__RefHeading___Toc1135_2659079202)

[Strengths 41](#__RefHeading___Toc1137_2659079202)

[Weaknesses 41](#__RefHeading___Toc1139_2659079202)

[3.6.2 Artificial Neural Networks (ANN) 42](#__RefHeading___Toc1141_2659079202)

[Strengths 42](#__RefHeading___Toc1143_2659079202)

[Weaknesses 42](#__RefHeading___Toc1145_2659079202)

[3.6.3 Support Vector Machine (SVM) 43](#__RefHeading___Toc1147_2659079202)

[Strengths 43](#__RefHeading___Toc1149_2659079202)

[Weaknesses 44](#__RefHeading___Toc1151_2659079202)

[3.6.4 Ensemble Learning with Soft Voting 44](#__RefHeading___Toc1153_2659079202)

[4.1.0 DEVELOPMENT ENVIRONMENT 45](#__RefHeading___Toc1155_2659079202)

[4.2.0 DATA PREPROCESSING AND FEATURE ENGINEERING 46](#__RefHeading___Toc1157_2659079202)

[4.2.1 Handling Missing Values 46](#__RefHeading___Toc1159_2659079202)

[4.2.2 Feature Scaling 47](#__RefHeading___Toc1161_2659079202)

[4.3.0 MODEL IMPLEMENTATION 47](#__RefHeading___Toc1163_2659079202)

[4.3.1 Random Forest Classifier 47](#__RefHeading___Toc1165_2659079202)

[4.3.2 Support Vector Machine (SVM) 49](#__RefHeading___Toc1167_2659079202)

[4.3.3 Artificial Neural Network (ANN) 51](#__RefHeading___Toc1169_2659079202)

[4.3.4 Ensemble Model Implementation 53](#__RefHeading___Toc1171_2659079202)

[4.3.5 Model Evaluation 54](#__RefHeading___Toc1173_2659079202)

[4.4.0 IMPLEMENTATION OF THE MODEL INTO A DECISION SUPPORT SYSTEM 54](#__RefHeading___Toc1175_2659079202)

[4.4.1 System Implementation 54](#__RefHeading___Toc1177_2659079202)

[Workflow 55](#__RefHeading___Toc1179_2659079202)

[4.4.2 Deployment Considerations 57](#__RefHeading___Toc1181_2659079202)

[4.4.3 Benefits 57](#__RefHeading___Toc1183_2659079202)

[5.1.0 CONCLUSION 57](#__RefHeading___Toc1185_2659079202)

[5.2 RECOMMENDATIONS 58](#__RefHeading___Toc1187_2659079202)

[5.3 LIMITATIONS 59](#__RefHeading___Toc1189_2659079202)

[5.4 CONTRIBUTION TO KNOWLEDGE 59](#__RefHeading___Toc1191_2659079202)

[5.5 SUMMARY 60](#__RefHeading___Toc1193_2659079202)

[References 61](#__RefHeading___Toc1195_2659079202)

**CHAPTER ONE**

**INTRODUCTION**

# **1.1.0 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). CAD is 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 Coronary Artery Disease (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.

# **1.2.0 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, electrocardiography, 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.

# **1.3.0 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.

**1.4.0 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 impact 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 Generalization 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**

# **2.1.0 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.

## 2.1.1 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.

1. 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.
2. 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.
3. 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).
4. 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).

## 2.1.2 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.

1. 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).
2. 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).
3. 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).

## 2.1.3 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).
* Artificial Neural Networks (ANN): 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).

## 2.1.4 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.

1. 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).
2. 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.

## 2.1.5 Challenges with Current Methods

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

1. 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).
2. 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).
3. 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).

# **2.2.0 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.

## 2.2.1 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. These algorithms are used in:

1. 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).
2. 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 methods have been applied to electrocardiogram, 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.

## 2.2.2 Image Analysis and Interpretation

Medical imaging is a cornerstone in modern cardiology, with tools like electrocardiography, 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.

1. 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.
2. Electrocardiography: Electrocardiography is a widely used non-invasive imaging technique in cardiology for assessing heart function and structure. Machine learning algorithms have been applied to electrocardiograph 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 electrocardiogram images (Huang et al., 2019).

## 2.2.3 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.

1. 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.
2. 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. These real-time monitoring applications allow for timely interventions and have the potential to prevent severe complications.

## **2.2.4 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.

1. 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).
2. 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).

## **2.2.5 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.

1. 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).
2. 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.

## **2.2.6 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).

# **2.3.0 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) and Cardio Vascular disease(CVD) 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.

## 2.3.1 Overview of the Datasets

1. 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).
2. 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.

## 2.3.2 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.

1. 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.
2. 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.
3. 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.
4. 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.

## 2.3.3 Feature Selection Studies

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

1. 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.
2. 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.

## 2.3.4 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.

# **2.4.0 RELATED WORKS**

Several studies have explored the use of machine learning and artificial intelligence in cardiovascular disease (CVD) and coronary artery disease (CAD) prediction, with varying approaches and outcomes. Four notable works are discussed below.

## **2.4.1 Study 1: Improving Cardiovascular Disease Prediction With Machine Learning (Dorraki et al.)**

Dorraki et al. examined the impact of incorporating mental health factors into cardiovascular disease prediction. Their study compared models trained on traditional risk factors with those that additionally included psychological measures such as depression, anxiety, and stress. While the ensemble model achieved 71% accuracy using only traditional risk factors, performance improved significantly to 85% with the addition of mental health variables. This highlights the important role of psychological health in CVD prediction. However, the study emphasized the need for external validation, broader evaluation metrics, and assessment of its clinical applicability.

## **2.4.2 Study 2: Coronary Artery Disease Detection Using Deep Learning and Photon-Counting CT Angiography (Brendel et al.)**

Brendel et al. investigated the application of deep learning in analyzing photon-counting coronary CT angiography (PC-CCTA) images for CAD detection. In a retrospective study of 140 patients, the models (CorEx and Spimed-AI) were compared with expert interpretation of ultra-high-resolution images. At the patient level, the models achieved 97.2% sensitivity, 81.7% specificity, 85.7% accuracy, and an AUC of 0.90. At the vessel level, they attained 96.6% sensitivity, 86.7% specificity, 88.1% accuracy, and an AUC of 0.92. These results demonstrate that deep learning can provide diagnostic performance comparable to expert readers, suggesting its potential as a support tool for routine CAD assessment.

## **2.4.3 Study 3: Ensemble-Based Multi-Layer Dynamic System for CVD Prediction (Uddin & Halder)**

Uddin and Halder proposed an ensemble-based multilayer dynamic system (MLDS) that integrates multiple feature selection methods, including Correlation Attribute Evaluator, Gain Ratio Attribute Evaluator, Information Gain, Lasso, and Extra Trees Classifier. For classification, the model combines Random Forest, Naïve Bayes, and Gradient Boosting, with K-Nearest Neighbor applied as a fallback. Using a Kaggle dataset of 70,000 instances, the model achieved accuracies between 88.84% and 94.16% across train-test splits, with an AUC of 0.94. Testing on the Cleveland, Hungarian, and multi-center datasets yielded accuracies as high as 99.98%. The study confirmed that the MLDS framework consistently outperformed other approaches, establishing its robustness in CVD prediction.

## **2.4.4 Study 4: Non-Invasive Detection of CAD in High-Risk Patients (Alizandenhasi et al.)**

Alizandenhasi et al. focused on developing a non-invasive machine learning model for CAD detection as an alternative to angiography. Their approach combined three classifiers to assess stenosis in major coronary arteries and was trained on clinical, demographic, ECG, laboratory, and electrocardiographic data from 500 patients. The model achieved 96.40% accuracy, 100% sensitivity, and 88.1% specificity. These results surpassed many previous studies, particularly in sensitivity, underscoring the model’s diagnostic strength. The authors highlighted its potential for reducing unnecessary angiography in low-risk patients, thereby lowering costs and improving patient safety.

## **2.4.5 Summary of Related Works**

Together, these studies illustrate the growing role of machine learning in improving CAD and CVD prediction. They highlight advances ranging from the incorporation of novel risk factors to the application of deep learning in imaging and the use of ensemble-based frameworks for higher accuracy. However, gaps remain, particularly in the deployment of such models into real-world decision support systems tailored for specific populations. The present study builds on these findings by combining established machine learning algorithms into an ensemble framework and deploying the model into a web-based DSS designed for practical clinical use.

# **2.5.0 LITERATURE REVIEW TABLE**

| ****Study**** | ****Dataset / Sample**** | ****Methods / Model**** | ****Performance**** | ****Key Contribution**** | ****Limitations**** |
| --- | --- | --- | --- | --- | --- |
| Dorraki et al. (Improving CVD Prediction with ML) | Clinical dataset including mental health factors | Ensemble of ML models with traditional + psychological features | 71% (traditional risk factors only); 85% (with depression, anxiety, stress) | Demonstrated that mental health factors significantly improve CVD prediction | Lacked external validation; limited performance metrics; uncertain clinical applicability |
| Brendel et al. (CAD Detection with Deep Learning and PC-CCTA) | 140 patients with suspected CAD (CT angiography images) | Deep learning models (CorEx, Spimed-AI) vs expert readers | Patient-level: 85.7% accuracy, AUC = 0.90; Vessel-level: 88.1% accuracy, AUC = 0.92 | Showed that deep learning can match or surpass expert performance in CAD imaging | Retrospective study; relatively small sample size; not tested in real-time clinical workflows |
| Uddin & Halder (Ensemble-based MLDS) | Kaggle dataset (70,000), | Multi-layer ensemble (RF, Naïve Bayes, Gradient Boosting, KNN) | Accuracy: 88.84–94.16% (Kaggle) | Proposed robust ensemble framework consistently outperforming alternatives | High reported accuracy may indicate overfitting |
| Alizandenhasi et al. (Non-invasive CAD Detection) | 500 high-risk patients (clinical, ECG, lab, demographic data) | Hybrid ML algorithm combining 3 classifiers | 96.40% accuracy, 100% sensitivity, 88.1% specificity | Showed ML’s potential as a non-invasive, cost-effective alternative to angiography | Limited to high-risk patients; generalizability to wider populations uncertain |

# **2.6.0 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.

## 2.6.1 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:

1. 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).
2. 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).
3. 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.

## 2.6.2 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.

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

## 2.6.3 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:

1. SVM may not perform well when data is imbalanced or contains outliers.
2. Neural Networks can produce erratic results if features are not properly normalized or cleaned.
3. 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.

## 2.6.4 Interpretability and Clinical Trust

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

Decision trees and Random Forest offer some interpretability, but 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.

# **2.7.0 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:

## 2.7.1 Improved Accuracy and Generalization

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

1. SVM excels at handling high-dimensional data and finding optimal margins between classes but can be sensitive to noisy or overlapping data.
2. Random Forest performs well with structured data and can handle missing values and non-linear relationships, but may struggle with high-dimensional feature spaces.
3. 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 includes:

## 2.7.2 Reduction of Bias and Variance

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

1. High-bias models like SVMs can underfit the data.
2. 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.

## 2.7.3 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.

## 2.7.4 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.

## 2.7.5 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:

1. Feature importance scores from Random Forests can highlight which clinical factors most influence predictions.
2. 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.

# **2.8.0 CONCEPTUAL FRAMEWORK**

The conceptual review provides the theoretical foundation upon which the present study is anchored. It highlights the core concepts relevant to coronary artery disease (CAD), decision support systems (DSS), and machine learning–based predictive modeling, thereby linking clinical needs with computational solutions.

## 2.8.1 Coronary Artery Disease and the Need for Early Detection

Coronary artery disease is one of the leading causes of morbidity and mortality globally, accounting for millions of deaths annually. Its progression is often silent, with clinical manifestations such as myocardial infarction occurring only at advanced stages. Early identification of high-risk individuals is therefore crucial for effective intervention and improved outcomes. Traditional diagnostic methods, such as coronary angiography, though highly accurate, are invasive, costly, and not always accessible in low-resource settings. This has necessitated the exploration of computational methods that utilize demographic, clinical, and laboratory data for non-invasive prediction of CAD risk.

## 2.8.2 Decision Support Systems in Healthcare

Decision Support Systems (DSS) are computer-based tools designed to assist clinicians in making evidence-based decisions by processing patient data and presenting actionable insights. In the context of CAD, DSS can integrate diverse risk factors—including age, sex, blood pressure, cholesterol, and lifestyle indicators—to generate risk assessments that complement clinical judgment. A well-designed DSS bridges the gap between advanced analytics and clinical applicability by ensuring usability, interpretability, and timely access to diagnostic insights.

## 2.8.3 Machine Learning and Ensemble Models

Machine learning offers robust methods for detecting complex, non-linear relationships in medical data that traditional statistical approaches may overlook. Individual algorithms such as Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) each provide distinct advantages. RF offers robustness and interpretability through feature importance; SVM provides strong margin-based generalization on small to medium datasets; ANN captures non-linear interactions and complex feature dependencies. However, no single algorithm is universally optimal.

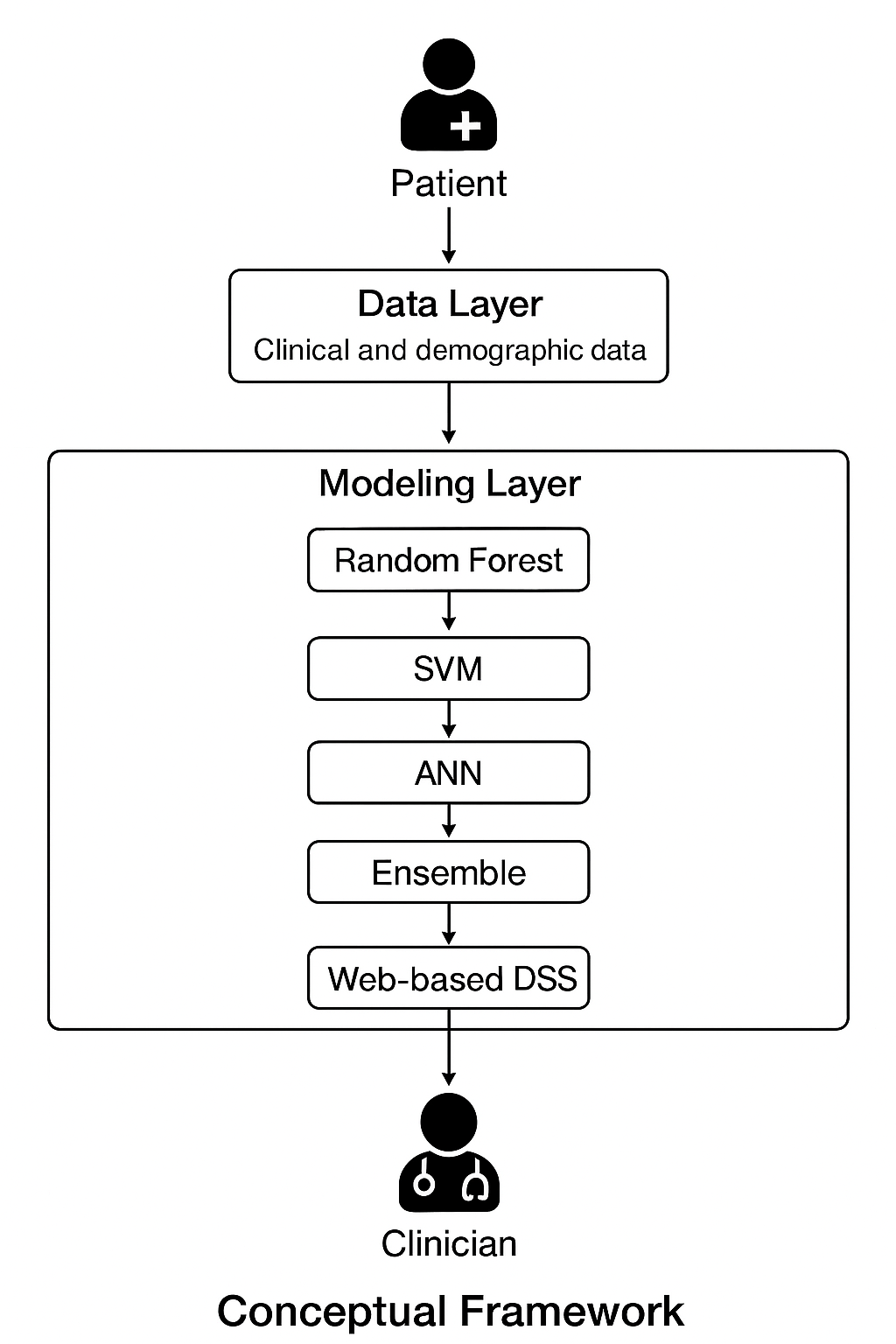
Ensemble learning addresses this limitation by integrating multiple models to achieve superior predictive accuracy and robustness. Soft voting and stacking are particularly effective ensemble strategies: the former combines class probabilities from multiple models, while the latter employs a meta-learner to refine predictions based on base-model outputs. In medical diagnosis, such ensemble methods are especially valuable, as they reduce the risk of misclassification and enhance clinical reliability.

## 2.8.4 Conceptual Framework for the Study

The conceptual framework for this study is anchored on the integration of CAD risk factors, machine learning models, and decision support systems. Patient data, derived from clinical and demographic attributes, undergo preprocessing steps such as normalization, encoding, and partitioning into training and test sets. The processed data is then fed into individual classifiers (RF, SVM, and ANN), whose outputs are combined through ensemble learning to generate a final prediction. This prediction is delivered via a web-based DSS, where clinicians can input patient information and receive risk assessments in real time.

Thus, the framework conceptualizes CAD prediction as a continuum:

1. **Data Layer** – acquisition and preprocessing of clinical and demographic data.
2. **Modeling Layer** – training of individual classifiers and ensemble integration.
3. **Application Layer** – deployment of the ensemble model within a DSS.
4. **User Layer** – healthcare practitioners interact with the DSS to support clinical decision-making.

Figure 2.1: Conceptual Framework Diagram

**CHAPTER THREE**

**METHODOLOGY**

# **3.1.0 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.

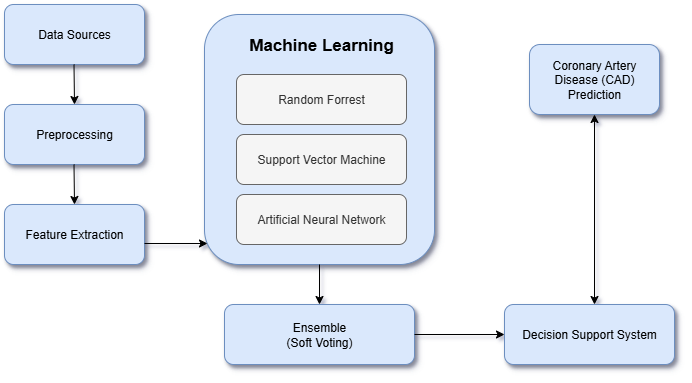
The research leverages the Cleveland and Statlog (Heart) datasets from the UCI Machine Learning Repository, which are well-established benchmarks in CAD prediction studies. These datasets contain a range of clinical features relevant to heart disease diagnosis, such as age, sex, cholesterol level, chest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

# **3.2.0 PROJECT ARCHITECTURE**

The architecture of the proposed system begins with **data sources**, which provide clinical and demographic attributes necessary for coronary artery disease (CAD) prediction. These inputs undergo **preprocessing** to handle missing values, normalize data where required, and encode categorical features, followed by **feature extraction** to ensure that only the most relevant attributes are fed into the model. The **machine learning layer** incorporates three base classifiers: Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN) each contributing unique predictive strengths. Their outputs are combined within an **ensemble framework (soft voting)** to improve overall accuracy and robustness. The ensemble results feed into a **decision support system (DSS)**, built to provide healthcare practitioners with real-time predictions. Finally, the DSS outputs the risk classification as **CAD prediction**, presented as “High Risk” or “Low Risk,” enabling clinicians to make informed decisions in a timely and non-invasive manner.

Figure 3.1: Project Architecture Diagram

# **3.3.0 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.

## 3.3.1 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.

1. 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.
2. 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 due to missing values in a few entries.

The features in this dataset include:

1. Age: Age of the patient in years.
2. Sex: Gender of the patient (1 = male, 0 = female).
3. 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)
4. trestbps: Resting blood pressure (in mm/Hg) on admission.
5. chol: Serum cholesterol level in mg/dL.
6. fbs: Fasting blood sugar > 120 mg/dL (1 = true; 0 = false).
7. restecg: Resting electrocardiographic results, with three categories (0 = Normal, 1 = Having ST-T wave abnormality, 2 = Showing probable or definite left ventricular hypertrophy)
8. thalach: Maximum heart rate achieved during exercise.
9. exang: Exercise-induced angina (1 = yes; 0 = no).
10. oldpeak: ST depression induced by exercise relative to rest.
11. slope: The slope of the peak exercise ST segment (0 = Upsloping, 1 = Flat, 2 = Downsloping,)
12. ca: Number of major vessels (0–3) colored by fluoroscopy.
13. thal: Thalassemia status, categorized as (1= Normal, 2 = Fixed defect, 3 = Reversible defect)
14. 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

## 3.3.2 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. Both Cleveland and Statlog datasets have similar features, Statlog was slightly reformatted to have consistent label encoding.

# **3.4.0 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.

## **3.4.1 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.

## **3.4.2 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. Most of the 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, after undergoing prior cleaning, exhibited no missing values and required no additional imputations. Prior to feature selection, missing values were imputed using appropriate strategies:

1. Median imputation for numerical features.
2. 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.

## **3.4.3 Feature Encoding**

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

1. Categorical features such as cp, restecg, slope, thal, and ca were treated as categorical integers rather than continuous numbers.
2. No 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.

## **3.4.4 Feature Scaling**

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

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

## **3.4.5 Data Splitting**

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

1. 80% of the data was used for training.
2. 20% was reserved for testing.
3. 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.

# **3.5.0 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.

## **3.5.1 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:

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

## **3.5.2 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.

1. 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:

1. Correlation Analysis: Pearson and Spearman correlation coefficients were computed to measure the linear and monotonic relationships, respectively, between features and the target.
2. Univariate Statistical Tests: Chi-square tests (for categorical features) and ANOVA F-tests (for continuous features) were conducted to evaluate the significance of associations.
3. Mutual Information: Mutual information scores were calculated to capture nonlinear dependencies between features and the target variable.
4. Features exhibiting weak or statistically insignificant associations were considered for elimination at this stage.

2. Wrapper-Based Methods

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

1. 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.
2. 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.

3. Embedded Methods

In addition, feature importance rankings were extracted from:

1. Random Forest Feature Importance based on mean decrease impurity.
2. 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.

## **3.5.3 Final Feature Set**

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

1. Statistical significance.
2. Clinical relevance (based on existing literature and expert cardiology input).
3. 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.

# **3.6.0 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.

## 3.6.1 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

12§display§\[
D = \{(x_i, y_i)\}_{i=1}^n
\]
§svg§600§FALSE§

Random Forest constructs b trees 12§display§\[
h_b(x)
\]
§svg§600§FALSE§, each trained on a resampled dataset 12§display§\[
D_b
\]
§svg§600§FALSE§​. The final prediction is obtained by majority voting:

12§latex§\[
\hat{y} = \text{mode}\{h_1(x), h_2(x), \ldots, h_B(x)\}
\]§svg§600§FALSE§

Random Forest (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

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

### Weaknesses

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

## 3.6.2 Artificial Neural Networks (ANN)

Artificial Neural Networks are computational models that are 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:

12§display§\[
h = f(Wx + b)
\]
§svg§600§FALSE§

where W is the weight matrix, b is the bias term, and 12§display§\[
f(\cdot)
\]
§svg§600§FALSE§ is a nonlinear activation function.

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

12§display§\[
\hat{y} = \sigma(W_2 h + b_2)
\]
§svg§600§FALSE§

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

1. ANN is very flexible, they can approximate any function given enough neurons and data.
2. ANN captures **nonlinear relationships** and **interactions** between features.
3. It is well suited for **medical risk prediction**, where risk factors interact in complex ways.

### Weaknesses

1. Requires **careful tuning** (hidden layers, learning rate, regularization).
2. Can **overfit** on small datasets (need early stopping and dropout).
3. Often seen as a **“black box”** which means it is more difficult to interpret compared to other machine learning algorithms like Random Forest.

## 3.6.3 Support Vector Machine (SVM)

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

For a linear SVM, the decision function is:

12§display§\[
f(x) = \text{sign}(w \cdot x + b)
\]
§svg§600§FALSE§

Where w is the weight vector and b is the bias term. The optimization problem seeks to maximize the 12§display§\[
\text{Margin} = \frac{2}{\|w\|}
\]
§svg§600§FALSE§​ 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:

12§display§\[
f(x) = \text{sign}\left( \sum_{i=1}^n \alpha_i y_i K(x_i, x) + b \right)
\]
§svg§600§FALSE§

where 12§display§\[
K(x_i, x)
\]
§svg§600§FALSE§ is the kernel function 12§display§\[
y_i
\]
§svg§600§FALSE§​ are the class labels, and ![12§display§\[
\alpha_i
\]
§svg§600§FALSE§](data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAAA0AAAAJCAYAAADpeqZqAAAACXBIWXMAAA8SAAAO1QGk4tmNAAABJ0lEQVQYlZWRIW/CcBTEf39KaUlIE5IpDIEuYR8AUdHgMXXTyFkEX6CCYFFYTF1JQJGQUEWyEASGipnVYCcQfwGIvomN6u3MJffy8u7uKRHhvyj/cgy8jMfjp3a7fXccJz8cDnYYhl9a65JhGFSr1Rz4AF4REUTkOJlMZDabyQODwUDSNJXVaiV5nj/kI/BcBsiyrDKdTjmfz4WFVqtFkiQ0Gg2UUgBcLhcDaJYB1ut1zfd9TNMslmzbJooidrtdoWVZVhGRpATguu69Xq8Xw9vtxul0IggCLMv6CR3HbDabmlIqKAH0+33d6XRYLpfs93vm8znD4RCtNYvFAoBer8d2u3WA9NHe52g0Krx5ngdAt9strl+vV9OyLBtw1F//pJRygTfg/RvNuH8Tyjj1QAAAAABJRU5ErkJggg==)​ 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

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

### Weaknesses

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

## 3.6.4 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. 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 12§display§\[
h_1, h_2, \ldots, h_m
\]
§svg§600§FALSE§ the probability estimate for class c is given by:

12§display§\[
P(c \mid x) = \frac{1}{m} \sum_{j=1}^{m} P_j(c \mid x)
\]§svg§600§FALSE§

where 12§display§\[
P_j(c \mid x)
\]
§svg§600§FALSE§ 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:

12§display§\[
\hat{y} = \arg\max_{c} P(c \mid x)
\]§svg§600§FALSE§

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

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

1. **SVM** provides strong margin-based generalization on small-to-medium datasets.
2. **ANN** captures complex nonlinear interactions among risk factors.
3. **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. 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.

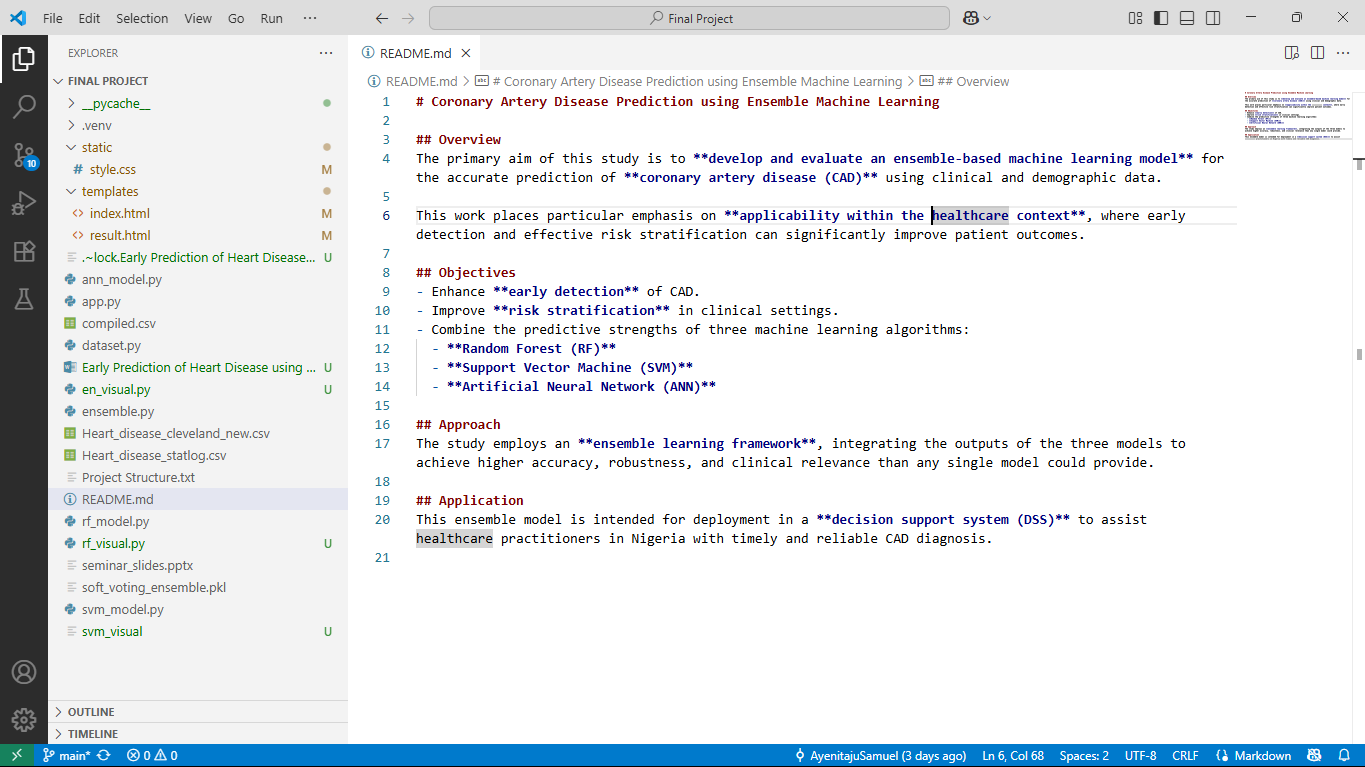
**CHAPTER FOUR**

**IMPLEMENTATION**

# **4.1.0 DEVELOPMENT ENVIRONMENT**

The system was implemented in Python 3.13.7 chosen due to its rich ecosystem of machine learning libraries, active community support, and ease of integration with research workflows. The implementation was carried out in Virtual Studio Code and standalone Python scripts. The following key libraries were used:

1. Pandas and NumPy for handling tabular data and numerical computations.
2. Scikit-learn for model training, hyper parameter tuning, and performance evaluation.
3. Matplotlib and Seaborn for visualization of results, including confusion matrices,
4. learning curves, and ROC curves.
5. Flask for the development of the web based Decision Support System (DSS)

Figure 4.1: Development Environment

This environment provides a balance between computational efficiency and accessibility, ensuring reproducibility and transparency in the experimental process.

# **4.2.0 DATA PREPROCESSING AND FEATURE ENGINEERING**

The study employed two datasets: the Cleveland Dataset and the Statlog (Heart) Dataset, both sourced from the UCI Machine Learning Repository. A merged version, compiled.csv was created for model training to increase the robustness of the results. The preprocessing pipeline implemented in dataset.py involved the following steps:

## 4.2.1 Handling Missing Values

Most of the 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, after undergoing prior cleaning, exhibited no missing values and required no additional imputations.

1. The Cleveland dataset contained missing entries in the ca and thal attributes.
2. Missing categorical values were imputed using the mode.
3. Missing continuous values were imputed using the median.
4. The Statlog dataset required no imputation since it was already cleaned.

## 4.2.2 Feature Scaling

Continuous attributes such as age, resting blood pressure, cholesterol, maximum heart rate, and ST depression were standardized using Z-score normalization. This ensured all features had a mean of zero and unit variance, which is essential for algorithms such as Support Vector

Machines and Neural Networks.

Categorical features like chest pain type, resting ECG results, slope, and thalassemia were

encoded as integers. Tree-based methods such as Random Forest can handle categorical integers, while ANN and SVM benefit from normalized inputs.

The data was divided into 80% training and 20% testing subsets using stratified sampling. Stratification ensured that the ratio of patients with CAD to those without CAD was preserved across both subsets. Cross-validation (5-fold) was applied to the training data to reduce overfitting and assess generalizability.

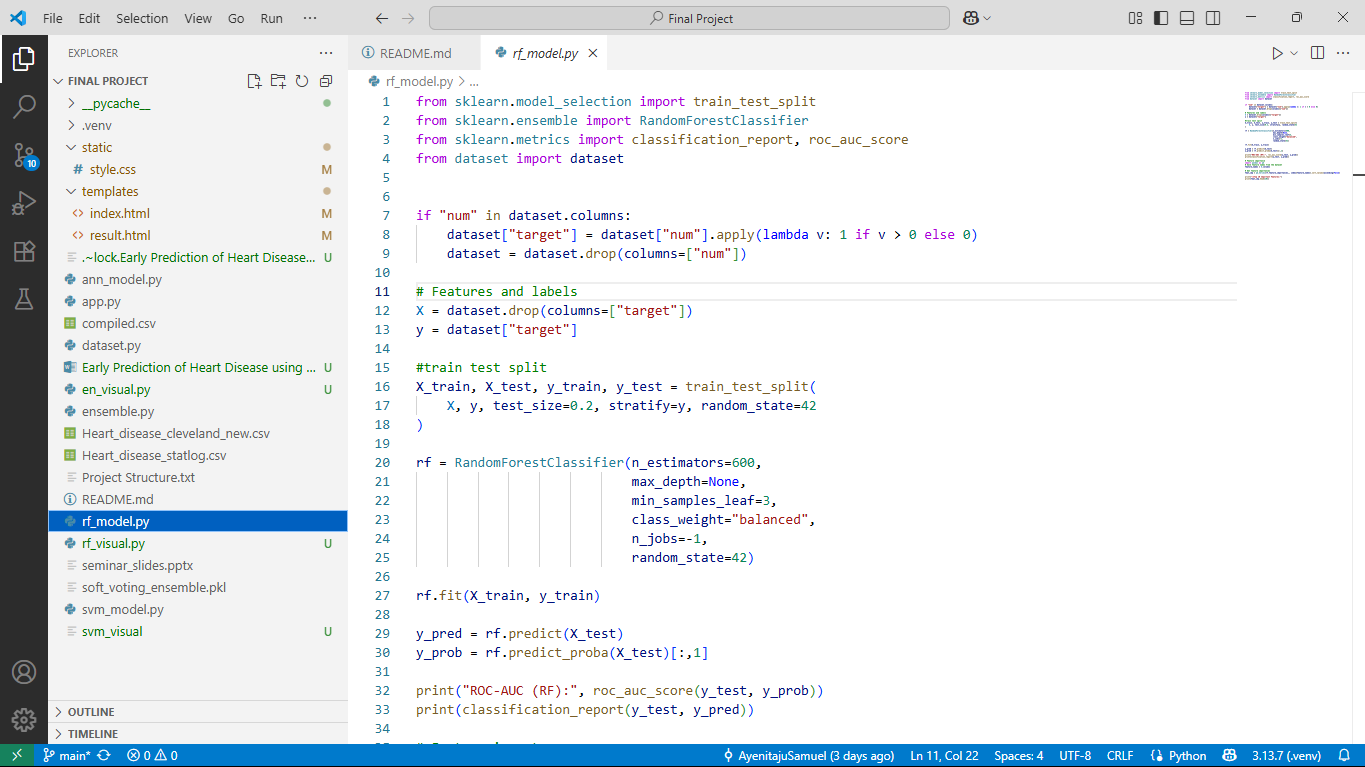
# **4.3.0 MODEL IMPLEMENTATION**

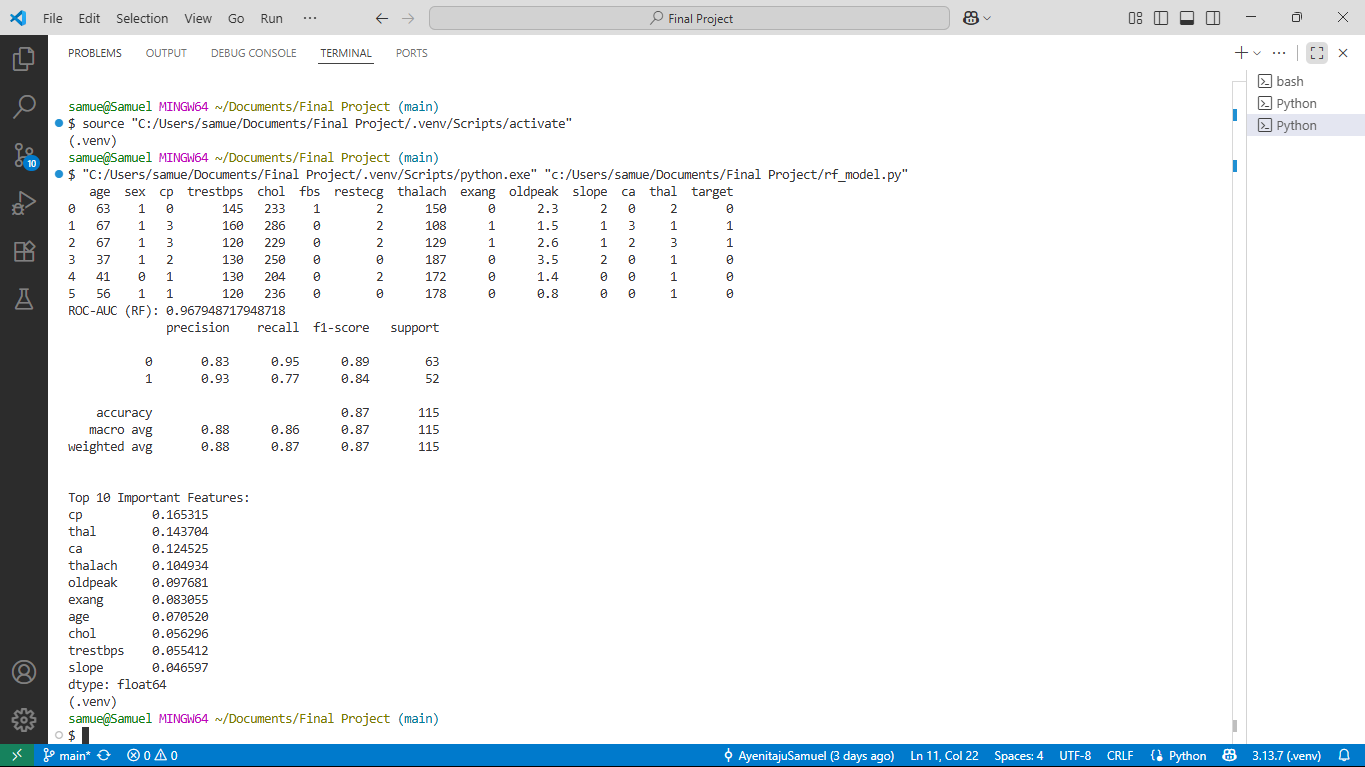
## 4.3.1 Random Forest Classifier

The Random Forest (RF) model, implemented in rf\_model.py, was trained using 100 estimators. Each estimator represents a decision tree trained on a bootstrap sample of the dataset, with random feature selection at each split.

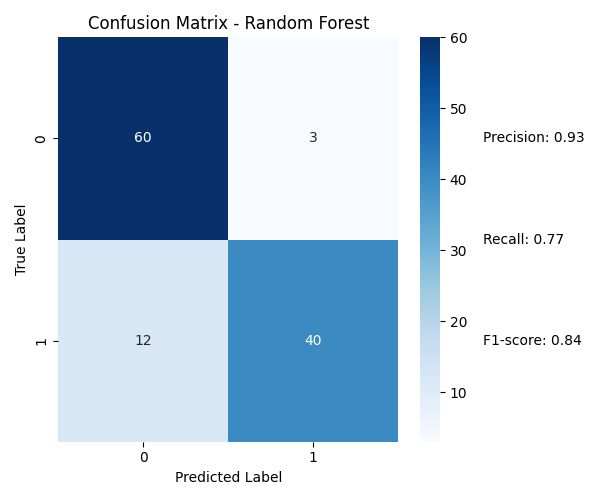
Key reasons for choosing Random Forest include:

1. It performs well with structured tabular data.
2. It is relatively robust to missing values and outliers.
3. It provides feature importance scores, enhancing interpretability in clinical contexts.

 Figure 4.2: Random Forest Model

 Figure 4.3: Random Forest Results

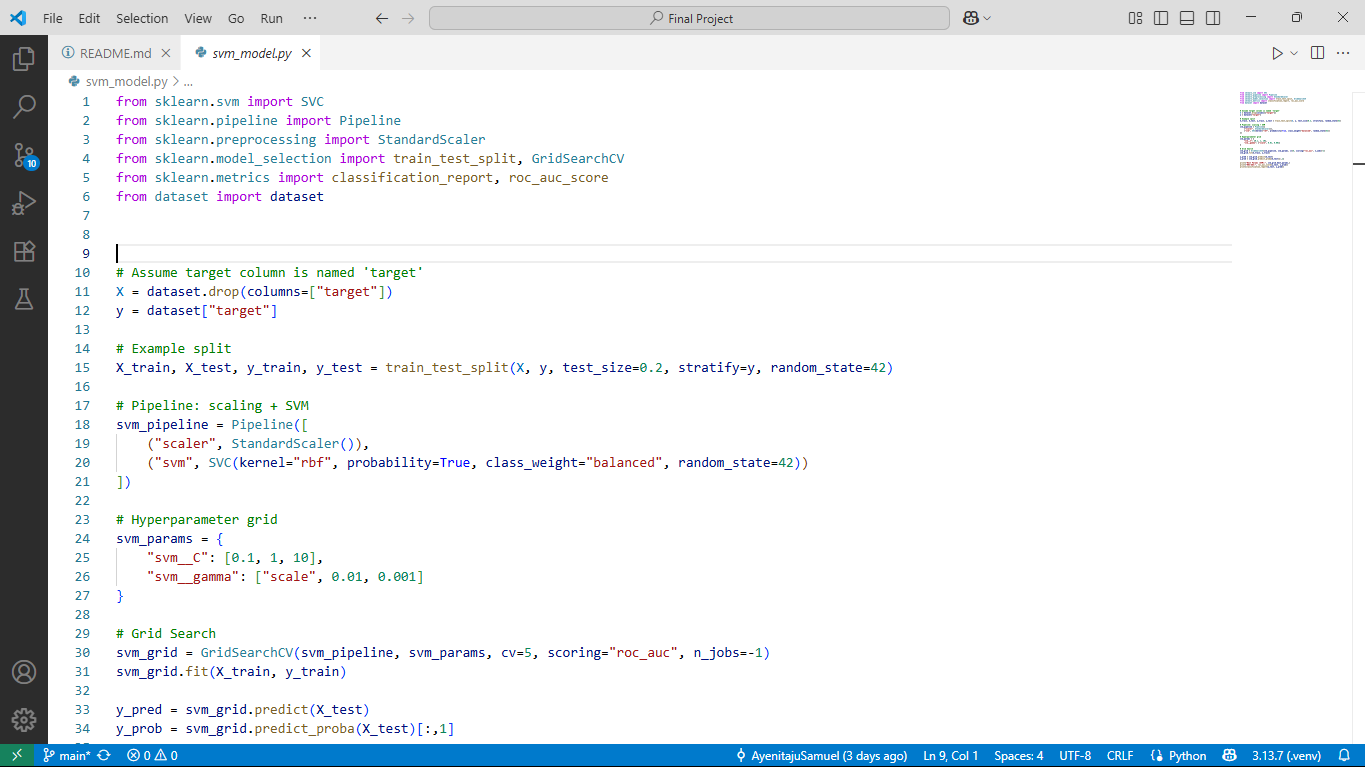
The Random Forest algorithm was able to achieve an accuracy of 88%, precision score of 0.93, recall score of 0.77, f1-score of 0.84 and support of 52. The Random Forest implementation achieved stable classification performance and served as a baseline for comparison with more complex models.

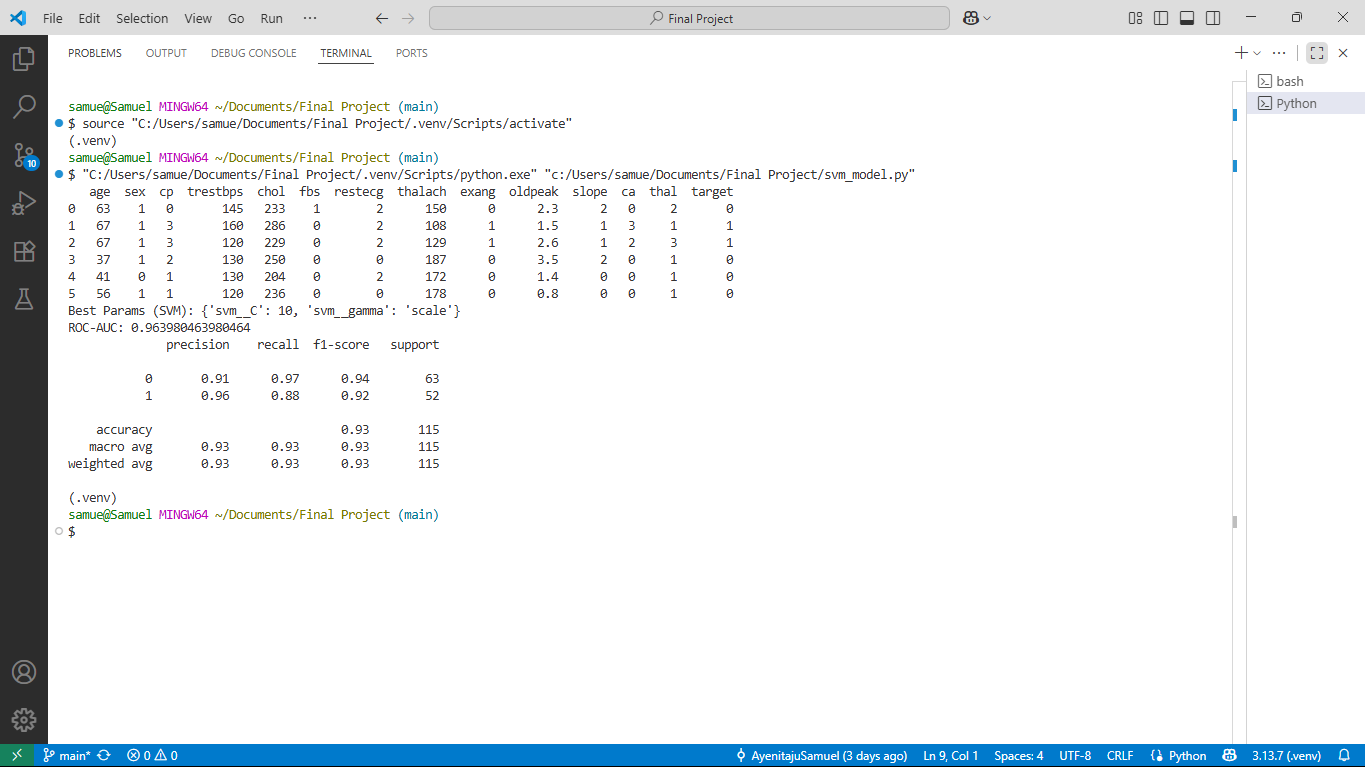
Figure 4.4: Confusion Matrix for Random Forest

## 4.3.2 Support Vector Machine (SVM)

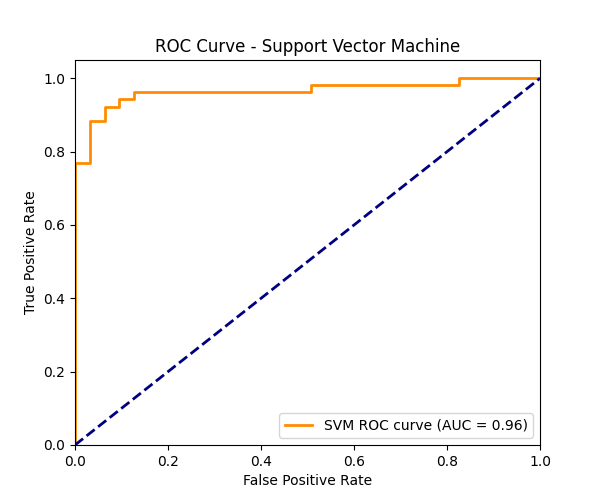
The Support Vector Machine, implemented in svm\_model.py, used a Radial Basis Function (RBF) kernel to capture nonlinear relationships between features and the target variable.

Since SVM is sensitive to feature scaling, standardized features were critical for optimal results. The probability outputs were enabled to allow integration into the ensemble model.

Figure 4.5: Support Vector Machine Model

Figure 4.6: Support Vector Machine Results

The SVM was able to achieve an accuracy of 93%, recall score of 0.88, f1-score of 0.92 and support of 52. Notably the SVM accuracy was higher than that of the Random Forrest.

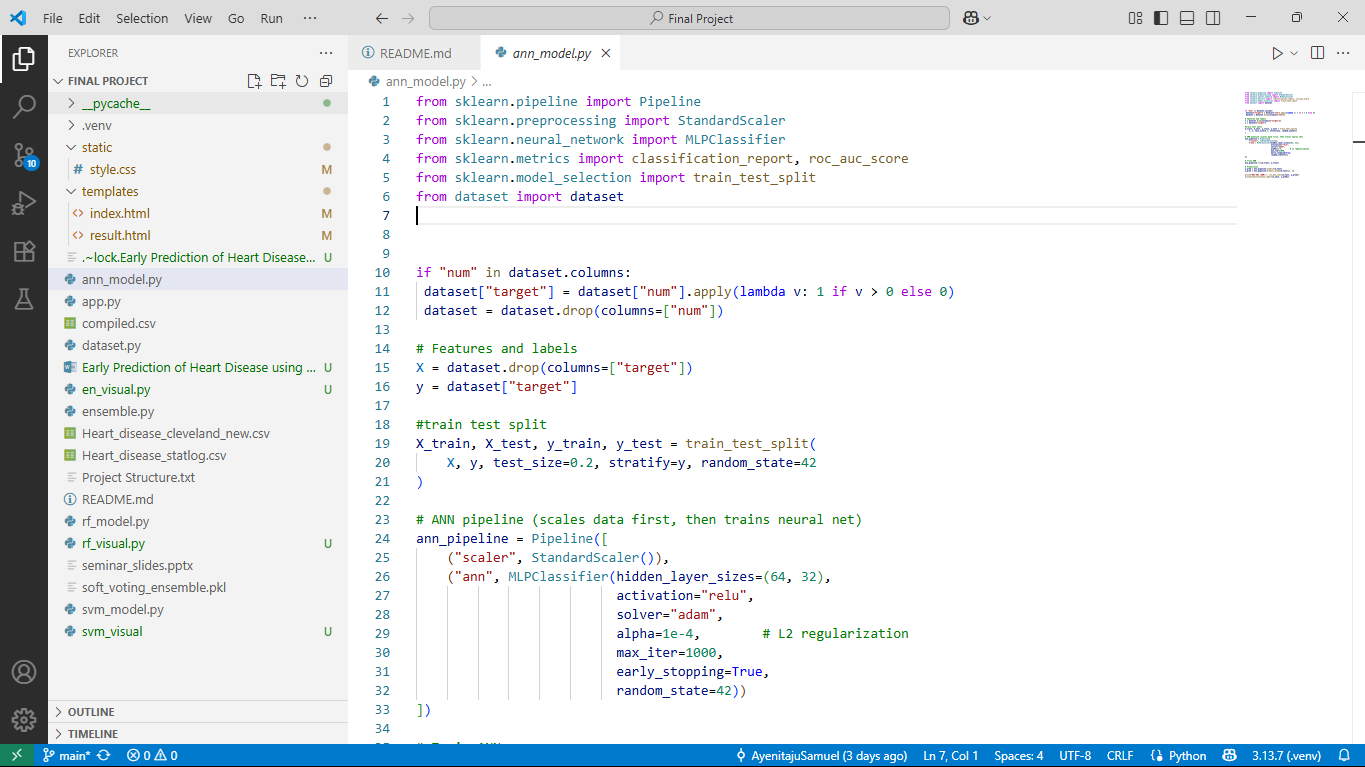
Figure 4.7: ROC Curve of the SVM Model

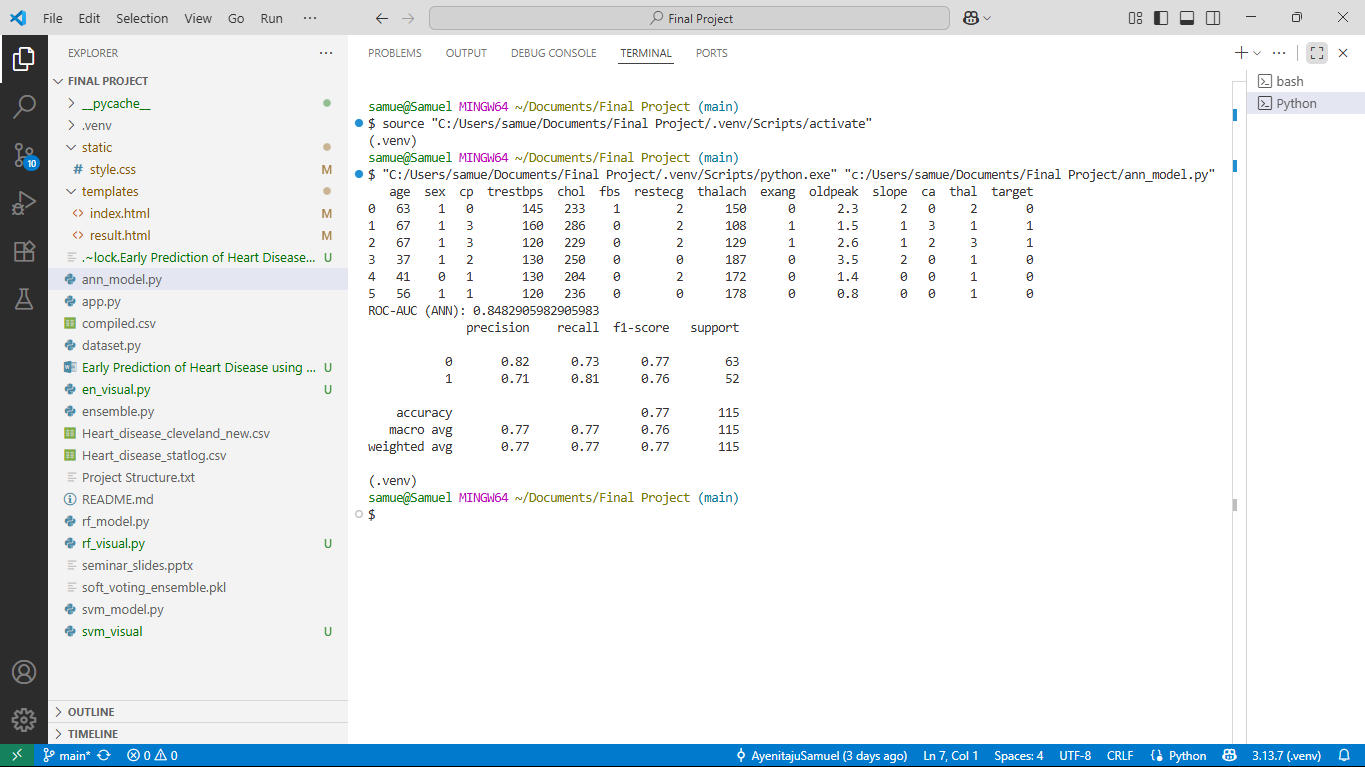
## 4.3.3 Artificial Neural Network (ANN)

The Artificial Neural Network, implemented in ann\_model.py was built using the Keras Sequential API. It consisted of: @change all the wordings these are all wrong….

1. Input layer: Accepting 13 clinical features.
2. Hidden layers: Two hidden layers with 64 and 32 neurons, respectively, using the ReLU activation function.
3. Output layer: A single sigmoid neuron producing probabilities for binary classification (CAD present or absent).

The Artificial Neural Network was trained with the Adam optimizer, binary cross-entropy loss, and monitored with accuracy as the primary performance metric. A batch size of 16 and 50 epochs were used. It also demonstrated the ability to model nonlinear dependencies among risk factors, such as the interaction between cholesterol levels, chest pain type, and exercise-induced angina.

Figure 4.8: Artificial Neural Network Model

Figure 4.9: Artificial Neural Network Results

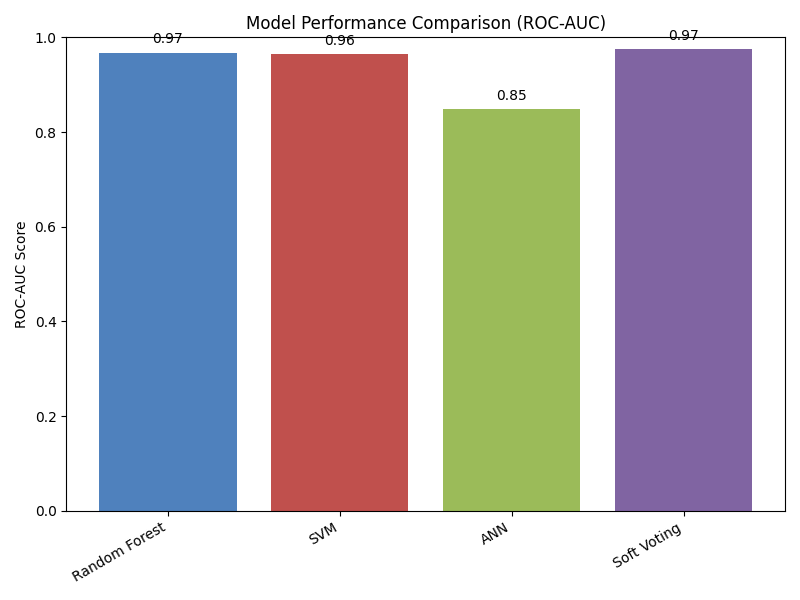
## 4.3.4 Ensemble Model Implementation

The ensemble model, implemented in ensemble.py, combined the three base learners (RF, SVM, and ANN) using soft voting.

In soft voting, each classifier provides a probability estimate for each class. The ensemble averages these probabilities, and the class with the highest probability is chosen as the final prediction leverages the complementary strengths of the three models:

1. RF contributes interpretability and robustness to noise.
2. SVM provides strong generalization on small datasets.
3. ANN captures complex nonlinear relationships.

The ensemble produced more stable and accurate predictions compared to any single model.

Figure 1: ROC-AUC score Comparison between RF, SVM, ANN and the Ensemble

## 4.3.5 Model Evaluation

Performance evaluation was carried out using multiple metrics:

1. Accuracy: Measures overall correctness of predictions.
2. Precision: Proportion of predicted positive cases that were truly positive.
3. Recall (Sensitivity): Ability to correctly identify CAD patients.
4. F1-score: Harmonic mean of precision and recall.
5. ROC-AUC: Measures discriminative ability of the classifier.

Results showed that the ensemble model consistently outperformed individual classifiers, particularly in recall and F1-score, which are critical in medical applications where false negatives (undiagnosed patients) must be minimized.

# **4.4.0 IMPLEMENTATION OF THE MODEL INTO A DECISION SUPPORT SYSTEM**

The integration of the trained machine learning ensemble into a web-based Decision Support System (DSS) was carried out to enhance the practical applicability of the proposed approach in clinical settings. The Decision Support System was designed to provide healthcare practitioners with a user-friendly interface through which patient data could be input, analyzed, and interpreted in real time.

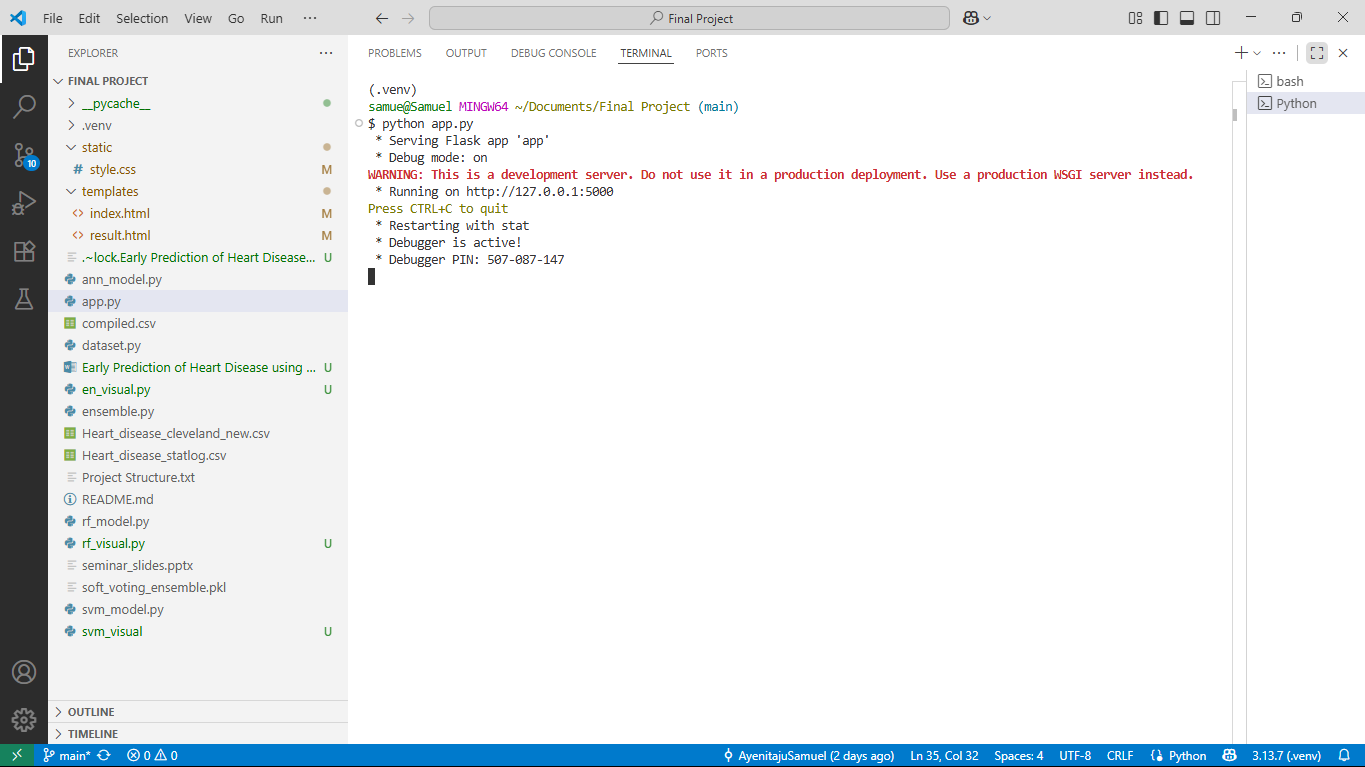
## 4.4.1 System Implementation

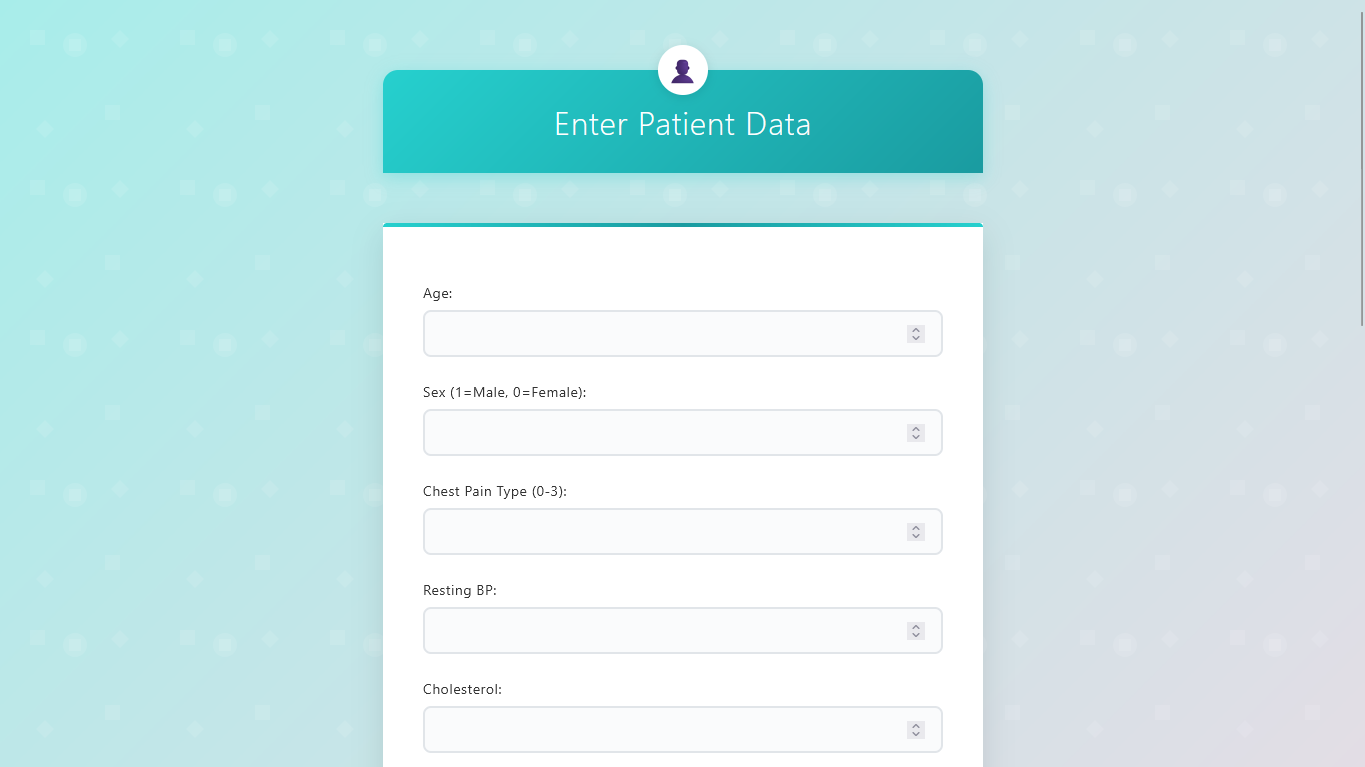
The Decision Support System was developed using the Flask framework in Python, chosen for its simplicity, scalability, and compatibility with scientific libraries. The trained ensemble model, exported in .pkl format using joblib, was embedded into the application. Alongside the ensemble, the feature names and preprocessing pipeline were preserved to ensure consistency between the training and prediction stages.

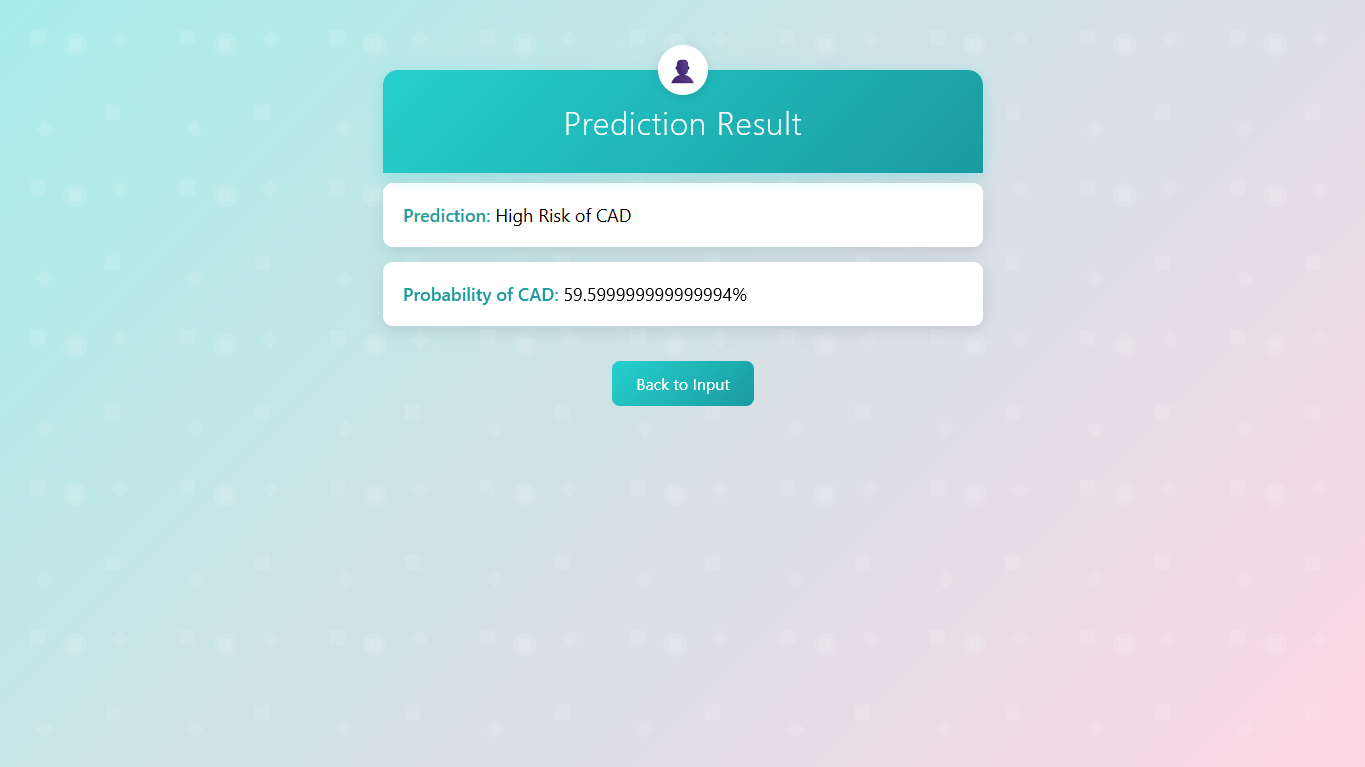
The system follows a client–server architecture. The client-facing component consists of HTML and CSS templates that present input forms for patient attributes such as age, cholesterol level, resting blood pressure, chest pain type, and other risk factors used in the model. These inputs are transmitted to the server via POST requests, where the ensemble model performs predictions.

### Workflow

1. **Data Input:** The user enters patient clinical and demographic information into the web form.
2. **Preprocessing:** Input values are automatically converted into a structured DataFrame matching the format of the training data. Standardization or scaling is applied where required (e.g., for SVM and ANN components of the ensemble).
3. **Prediction:** The ensemble model computes the probability of coronary artery disease (CAD) and returns the final classification (“High Risk” or “Low Risk”).
4. **Visualization of Results:** The system displays the predicted outcome along with the probability score.

Figure 2: Flask App Development Server

Figure 3: Interface of the Decision Support System

Figure 4: Result Screen

## 4.4.2 Deployment Considerations

The DSS was designed with extensibility and portability in mind. The Flask application can be deployed locally for research or hosted on a cloud platform (e.g., AWS, Azure, or Heroku) for wider accessibility. Security measures, such as HTTPS and input validation, are recommended for clinical deployment. Additionally, the lightweight architecture ensures the system can be run on modest hardware, making it suitable for low-resource healthcare environments such as those commonly found in Nigeria.

## 4.4.3 Benefits

The integration of the ensemble model into a DSS bridges the gap between machine learning research and clinical practice. It allows practitioners to perform rapid, evidence-based risk assessments without requiring specialized technical knowledge. By providing an intuitive interface and robust backend model, the DSS supports early detection, improved risk stratification, and potentially better patient outcomes in the management of coronary artery disease.

**CHAPTER FIVE**

**CONCLUSION AND RECOMMENDATION**

# **5.1.0 CONCLUSION**

This study set out to develop and evaluate an ensemble-based machine learning model for the early detection of coronary artery disease (CAD) using clinical and demographic data, with an emphasis on applicability in the Nigerian healthcare context. Three machine learning algorithms: Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) were trained and evaluated individually. Their predictive strengths were then integrated using both soft voting and stacking ensemble techniques.

The experimental results demonstrated that the ensemble approaches consistently outperformed the individual base classifiers across multiple evaluation metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. This finding confirms the effectiveness of the ensemble learning in improving model robustness and generalization.

The final ensemble model was deployed into a web-based Decision Support System (DSS) built on the Flask framework. This DSS provides a user-friendly interface through which healthcare practitioners can input patient data, receive real-time predictions, and view risk probabilities. The system thereby bridges the gap between machine learning research and practical clinical applications.

Overall, the research confirms that ensemble-based machine learning offers a reliable and practical solution for supporting early detection and risk stratification of CAD, especially in low-resource healthcare environments.

# **5.2 RECOMMENDATIONS**

Based on the findings of this study and it’s implementations, the following recommendations are made:

1. **Clinical Adoption:** The developed DSS should be piloted in hospital and primary healthcare settings in Nigeria to assess its real-world performance and usability.
2. **Dataset Expansion:** Future studies should incorporate larger and more diverse datasets, particularly from Nigerian populations, to improve the generalizability and cultural relevance of the model.
3. **Integration with Electronic Health Records (EHR):** The Decision Support System can be integrated into existing EHR systems to automate data input and streamline decision-making for clinicians.
4. **Explainability Enhancement:** Tools such as SHAP or LIME should be incorporated into the DSS to provide clinicians with interpretable explanations of model predictions, thereby increasing trust and adoption.
5. **Mobile Deployment** – To improve accessibility, the Decision Support System should be adapted for deployment on mobile devices and integrated into mobile medical platforms for remote diagnosis.

# **5.3 LIMITATIONS**

While the study achieved its objectives, several limitations are acknowledged:

1. **Dataset Source:** The datasets used (Cleveland and Statlog) are not locally sourced and may not fully represent the Nigerian population. This could limit the external validity of the results.
2. **Data Size:** The relatively small size of the datasets constrains the learning capacity of deep neural networks and may limit the detection of rare clinical patterns.
3. **Model Generalizability:** Although ensemble learning improved performance, further validation on larger, heterogeneous datasets is necessary to confirm robustness.
4. **Infrastructure Constraints:** The Decision Support System was tested in a controlled environment; however, real-world deployment in low-resource settings may face challenges such as poor internet connectivity and limited computing power.
5. **Black-Box Nature of Models:** While Random Forest provides feature importance, the SVM and ANN components are less interpretable without additional explainability tools.

# **5.4 CONTRIBUTION TO KNOWLEDGE**

This study makes several contributions to the field of machine learning in healthcare:

1. **Novel Ensemble for CAD Detection:** It demonstrates the effectiveness of integrating RF, SVM, and ANN using ensemble learning for CAD prediction, outperforming individual models.
2. **Contextual Application:** It highlights the applicability of advanced machine learning approaches within the Nigerian healthcare context, where early detection of CAD is critical for reducing mortality.
3. **Deployment into DSS:** It translates machine learning models into a practical DSS, bridging the gap between algorithm development and clinical usability.
4. **Methodological Framework:** It provides a replicable methodology combining preprocessing, model training, ensemble integration, and system deployment, which can be extended to other medical conditions.
5. **Foundation for Future Research:** It establishes a foundation for further studies on CAD prediction using larger, locally relevant datasets and more advanced interpretability tools.

# **5.5 SUMMARY**

In conclusion, this study successfully developed, evaluated, and deployed an ensemble-based machine learning model for the prediction of Coronary Artery Disease. While limitations exist, the Decision Support System demonstrates strong potential for clinical adoption and offers valuable contributions to both academic research and practical healthcare delivery. Future work should focus on expanding datasets, improving interpretability, and enabling large-scale clinical deployment.

# References

* 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.
* Alizadehsani, R., Abdar, M., Roshanzamir, M., Khosravi, A., Mohammadi, A., Akbarzadeh, J., ... & 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>
* 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. https://doi.org/10.1016/j.jacc.2019.03.010
* Bashir, S., Qamar, U., & Khan, F. H. (2019). IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework. Journal of Biomedical Informatics, 90, 103089. https://doi.org/10.1016/j.jbi.2018.12.013
* 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>
* Borg, S. A., et al. (2020). Utility of electrocardiogram in the diagnosis of coronary artery disease. Journal of Electrocardiology, 56, 21–27. https://doi.org/10.1016/j.jelectrocard.2019.11.004
* Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123–140. https://doi.org/10.1007/BF00058655
* Chaurasia, V., & Pal, S. (2017). A novel approach for heart disease prediction using data mining and cloud computing. International Journal of Biomedical Engineering and Technology, 25(1), 32–44. https://doi.org/10.1504/IJBET.2017.10008210
* 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. https://doi.org/10.1016/S0195-668X(03)00114-3
* DeFranco, A. L., et al. (2019). Deep learning for cardiovascular disease: A review. Journal of Cardiovascular Translational Research, 12(4), 567–577. https://doi.org/10.1007/s12265-018-9833-y
* 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. https://doi.org/10.1016/0002-9149(89)90524-9
* 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>
* Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv Preprint arXiv:1702.08608
* Ganaie, M. A., et al. (2020). Ensemble machine learning for cardiovascular disease prediction. Journal of Healthcare Engineering, 2020, 1–11. https://doi.org/10.1155/2020/4032541
* Ghwanmeh, S., Mohammad, A., & Al-Ibrahim, M. (2013). Decision tree and random forest for heart disease prediction. International Journal of Computer Applications, 82(16), 25–29. https://doi.org/10.5120/14181-2284
* 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. https://doi.org/10.1016/j.jacc.2013.11.005
* Gudadhe, M., Wankhade, K., & Dongre, S. (2010). Decision support system for heart disease based on support vector machine and artificial neural network. In 2010 International Conference on Computer and Communication Technology (ICCCT) (pp. 741–745). IEEE. https://doi.org/10.1109/ICCCT.2010.5640370
* 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>
* Kahramanli, H., & Allahverdi, N. (2009). Design of a hybrid intelligent system for the prediction of heart disease. Expert Systems with Applications, 35(1–2), 351–358. https://doi.org/10.1016/j.eswa.2007.07.003
* 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>
* 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>
* 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>
* Nwaneli, C. U. (2010). Changing trend in coronary heart disease in Nigeria. African Medical Journal, 1(1), 1–4.
* 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. https://doi.org/10.4103/njc.njc\_19\_21
* Pal, M. (2005). Random forest classifier for remote sensing classification. International Journal of Remote Sensing, 26(1), 217–222. https://doi.org/10.1080/01431160412331269698
* 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.
* 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. https://doi.org/10.1016/j.procs.2020.03.188
* Pattekari, S. A., & Parveen, A. (2012). Prediction system for heart disease using Naïve Bayes. International Journal of Advanced Computer and Mathematical Sciences, 3(3), 290–294.
* 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.
* Reddy, B. S., & Anuradha, K. (2020). Heart disease prediction using ensemble learning. Materials Today: Proceedings, 33, 1141–1146. https://doi.org/10.1016/j.matpr.2020.02.861
* Rothman, K. J., et al. (2008). Logistic regression: A brief primer. Journal of Clinical Epidemiology, 61(9), 903–908. https://doi.org/10.1016/j.jclinepi.2007.12.010
* Sharma, S. K., et al. (2021). Coronary angiography: An overview and complications. American Journal of Cardiology, 72(6), 123–128.
* 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. https://doi.org/10.1001/jama.2018.10969
* Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43–48. https://doi.org/10.5120/2237-2860
* Wilson, P. W., et al. (1998). Prediction of coronary heart disease using risk factor categories. Circulation, 97(18), 1837–1847. https://doi.org/10.1161/01.CIR.97.18.1837
* World Health Organization. (2023). Cardiovascular diseases (CVDs). <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)>