

Early Heart Disease Prediction Using Data Mining Techniques

Dugguh Sylvester Aondonenge^{1*}, Ajayi Ore-Ofé², Kamorudeen Hassan Taiwo³, Abubakar Umar⁴, Isa Abdulrazaq Imam⁵, Dako Daniel Emmanuel⁶, Ibrahim Ibrahim⁷

^{1,2,4,5,6,7}Department of Computer Engineering, Faculty of Engineering, Ahmadu Bello University, Zaria, Nigeria

³Department of Family Medicine, Ahmadu Bello University Teaching Hospital, Zaria, Nigeria

Article Info

Article history:

Received December 13, 2024

Revised February 28, 2025

Accepted April 30, 2025

Keywords:

Data Mining

Heart Disease

Machine Learning Algorithms

Model Performance

Predictive Model

ABSTRACT

Heart disease is a leading cause of mortality worldwide, characterized by the buildup of plaque in the arteries, which can lead to severe cardiovascular complications. Predicting heart disease is complex due to the need to analyze multiple risk factors, such as age, cholesterol, and blood pressure. This study develops a predictive model for early heart disease detection using data mining techniques to enhance timely and accurate diagnosis. The model combines multiple machine learning algorithms, including Random Forest, Support Vector Machine, and a hybrid ensemble approach to improve prediction accuracy and reliability. The methodology follows five phases: data collection, data pre-processing, feature extraction, model construction, and model evaluation. Data was gathered from publicly available health repositories, preprocessed to remove missing values and irrelevant information, and subjected to feature extraction techniques to identify influential predictors. The hybrid model was trained and tested using an 80:20 data split and evaluated against various classification algorithms. It achieved an accuracy of 97.56%, precision of 98.04%, and recall of 97.09%, outperforming individual models. These results highlight the effectiveness of the hybrid approach in supporting early intervention for heart disease, particularly in healthcare settings with limited diagnostic resources. This study demonstrates that advanced data mining techniques provide a viable solution for improving patient outcomes through the early detection of heart disease.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



1. INTRODUCTION

Heart disease is a leading cause of death worldwide, and early detection is crucial for effective treatment. The rate at which people are suffering from heart disease is increasing geometrically due to the unhealthy lifestyle many people have adopted [1][2]. Heart failure often results from heart disease and may cause breathlessness when the heart becomes too weak to pump blood effectively [3][4]. In some cases, especially among older adults and individuals with diabetes, heart conditions can occur without noticeable symptoms. Congenital heart disease includes a variety of conditions, typically presenting symptoms such as excessive sweating, fatigue, rapid heartbeat, breathlessness, and chest pain [5]-[8]. These signs may not appear until adolescence, making diagnosis challenging and requiring significant expertise. Early detection of heart disease or the risk of a heart attack enables patients to adopt preventive measures and manage their condition more effectively. With advancements in healthcare, large volumes of patient data, including diagnostic reports, are being analyzed globally to predict heart attacks. Machine learning techniques are particularly valuable in processing and analyzing this extensive data to enhance the understanding and prevention of heart-related conditions. Technological advancements have significantly improved various aspects of life, including the early detection of heart diseases. Over the past decades, improved methods for data acquisition have been introduced, enabling precise sensing, collection, recording, and analysis of patients' physical conditions.

*Corresponding Author

Email: dugguhsylvester@gmail.com

Machine learning (ML) enables systems to acquire knowledge and improve based on experience autonomously, without requiring human intervention or reasoning [9].

These self-learning machines, called artificial intelligence, are categorized into supervised, unsupervised, semi-supervised, and reinforcement learning [10]. Supervised learning algorithms identify relationships and dependencies between input features and target outputs, allowing them to predict outcomes for new data based on the patterns learned from existing datasets [11]. In contrast, unsupervised learning involves training models with unlabeled data [12]. Semi-supervised learning combines labeled and unlabeled data, placing it between supervised and unsupervised approaches [13]. Reinforcement learning algorithms engage with their environment, performing actions and using the feedback received as rewards or errors to optimize behavior. This approach enables machines and software agents to autonomously determine the best course of action within a given context to enhance performance [10][14].

Predicting heart disease is a challenging and time-consuming process that often relies on the experience and expertise of doctors, along with medical tests. However, the large number of heart patients across hospitals worldwide provides valuable data that can be analyzed using data mining techniques. By identifying patterns and trends within this data, these techniques can help medical professionals better understand heart disease, improve diagnosis accuracy, and make more informed treatment decisions. This approach supports doctors' work and enhances their knowledge about heart disease, leading to better healthcare outcomes [15][16].

Data mining involves gathering necessary information from records to help with future predictions or making decisions. It is uncovering hidden patterns or valuable information from extensive data collections, often stored in databases or warehouses. It plays an essential role across many industries, such as finance, education, and healthcare, by helping organizations analyze vast amounts of data. This analysis supports better decision-making, improves efficiency, and leads to more effective outcomes. Many organizations rely on data mining to gain insights to shape strategies and achieve long-term goals [17].

In data science, Machine Learning is a key approach for deriving knowledge from prior research experiences and addressing challenges that traditional methods often struggle to resolve [18][19]. It involves creating models to analyze data patterns and produce reliable and accurate predictions. During the training phase, Machine Learning algorithms use data samples to learn foundational patterns, resulting in an automatic framework suitable for static and dynamic datasets. This framework typically splits data into training and testing sets, where the training data helps build the model, and the testing data evaluates its accuracy. Depending on the task, the model may employ classification or clustering techniques to process inputs like text or images and generate outcomes such as predictions or categorizations. Machine Learning has practical applications in various fields, including healthcare, where it is used to classify and predict cardiovascular diseases. For instance, it can automatically identify patients at high or low risk, enhancing early diagnosis and treatment planning. Machine Learning, while advantageous, encounters significant challenges, such as managing datasets with numerous variables. This issue, commonly known as the curse of dimensionality, arises when the vast quantity of data becomes too complex to analyze or interpret effectively [20]-[22].

Cardiovascular diseases (CVDs) were responsible for approximately 17.9 million deaths globally in 2019, accounting for 32% of all fatalities worldwide. Among these, 85% resulted from heart attacks and strokes. More than three-quarters of CVD-related deaths occurred in low and middle-income nations. Additionally, of the 17 million premature deaths (under 70 years) attributed to non-communicable diseases in 2019, CVDs accounted for 38% [23][24].

Cardiovascular diseases are hazardous and can lead to several health complications and even death, making early detection a critical challenge [25]-[27]. Timely identification of heart disease can significantly reduce mortality rates and prevent severe complications such as chest pain, shortness of breath, heart attacks, heart failure, and strokes. Given these risks, developing a reliable predictive model for early diagnosis is essential. This study leverages several classification techniques to create a high-accuracy heart disease detection model. Support Vector Machine (SVM) and Random Forest have proven particularly effective among these. SVM is advantageous for heart disease prediction because it can handle high-dimensional data and effectively separate classes using a hyperplane, making it robust even with limited datasets. Its use of kernel functions allows for flexible decision boundaries, improving classification accuracy [28]-[30]. On the other hand, Random Forest is highly effective due to its ensemble learning approach, which combines multiple decision trees to reduce overfitting and enhance predictive stability. Aggregating the outputs of numerous trees provides a more generalized model, making it well-suited for medical diagnosis where variability in patient data is common [31]-[33]. This research introduces a Machine Learning model designed to determine the likelihood of heart disease based on various medical factors. The model is trained using a patient information dataset, including attributes such as age, blood pressure, and cholesterol levels. Its goal is to provide accurate predictions to facilitate early diagnosis and treatment, proving valuable in the timely detection of heart conditions.

The structure of this article is as follows: Dataset loading and description, data preprocessing, Feature extraction, Model construction, Model evaluation, and Architecture of the developed system are explained in Section 2. Section 3 is comprehensive theoretical basis. In Section 4 there are discussions and simulations. The conclusion is presented in the last section.

2. METHOD

Developing and implementing an early prediction system for heart disease using data mining techniques was executed through five primary stages: dataset loading and description, data preprocessing, feature extraction, model construction, and model evaluation. The detailed methodology for creating and implementing the early heart disease prediction system is outlined below.

2.1. Dataset loading and description

Historical data is necessary for Machine Learning models to function. The Cleveland dataset from the UCI Machine Learning Repository was chosen due to its comprehensive medical attributes that are crucial for heart disease prediction, such as age, sex, chest pain type (cp), resting blood pressure (trestbps), serum cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), ST depression induced by exercise (oldpeak), the slope of the peak exercise ST segment (slope), number of major vessels colored by fluoroscopy (ca), thalassemia (thal), and the target variable indicating the presence or absence of heart disease. This study leveraged the Cleveland dataset to implement the proposed model. The dataset is accessible on “<https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>.” The dataset comprises 14 features, as shown in Figure 1, with details about the individual features described in Table 1.

| 1 | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|----|-----|-----|----|----------|------|-----|---------|---------|-------|---------|-------|----|------|--------|
| 2 | 63 | 1 | 3 | 145 | 233 | 1 | 0 | 150 | 0 | 2.3 | 0 | 0 | 1 | 1 |
| 3 | 37 | 1 | 2 | 130 | 250 | 0 | 1 | 187 | 0 | 3.5 | 0 | 0 | 2 | 1 |
| 4 | 41 | 0 | 1 | 130 | 204 | 0 | 0 | 172 | 0 | 1.4 | 2 | 0 | 2 | 1 |
| 5 | 56 | 1 | 1 | 120 | 236 | 0 | 1 | 178 | 0 | 0.8 | 2 | 0 | 2 | 1 |
| 6 | 57 | 0 | 0 | 120 | 354 | 0 | 1 | 163 | 1 | 0.6 | 2 | 0 | 2 | 1 |
| 7 | 57 | 1 | 0 | 140 | 192 | 0 | 1 | 148 | 0 | 0.4 | 1 | 0 | 1 | 1 |
| 8 | 56 | 0 | 1 | 140 | 294 | 0 | 0 | 153 | 0 | 1.3 | 1 | 0 | 2 | 1 |
| 9 | 44 | 1 | 1 | 120 | 263 | 0 | 1 | 173 | 0 | 0 | 2 | 0 | 3 | 1 |
| 10 | 52 | 1 | 2 | 172 | 199 | 1 | 1 | 162 | 0 | 0.5 | 2 | 0 | 3 | 1 |
| 11 | 57 | 1 | 2 | 150 | 168 | 0 | 1 | 174 | 0 | 1.6 | 2 | 0 | 2 | 1 |
| 12 | 54 | 1 | 0 | 140 | 239 | 0 | 1 | 160 | 0 | 1.2 | 2 | 0 | 2 | 1 |
| 13 | 48 | 0 | 2 | 130 | 275 | 0 | 1 | 139 | 0 | 0.2 | 2 | 0 | 2 | 1 |
| 14 | 49 | 1 | 1 | 130 | 266 | 0 | 1 | 171 | 0 | 0.6 | 2 | 0 | 2 | 1 |
| 15 | 64 | 1 | 3 | 110 | 211 | 0 | 0 | 144 | 1 | 1.8 | 1 | 0 | 2 | 1 |
| 16 | 58 | 0 | 3 | 150 | 283 | 1 | 0 | 162 | 0 | 1 | 2 | 0 | 2 | 1 |
| 17 | 50 | 0 | 2 | 120 | 219 | 0 | 1 | 158 | 0 | 1.6 | 1 | 0 | 2 | 1 |
| 18 | 58 | 0 | 2 | 120 | 340 | 0 | 1 | 172 | 0 | 0 | 2 | 0 | 2 | 1 |
| 19 | 66 | 0 | 3 | 150 | 226 | 0 | 1 | 114 | 0 | 2.6 | 0 | 0 | 2 | 1 |
| 20 | 43 | 1 | 0 | 150 | 247 | 0 | 1 | 171 | 0 | 1.5 | 2 | 0 | 2 | 1 |
| 21 | 69 | 0 | 3 | 140 | 239 | 0 | 1 | 151 | 0 | 1.8 | 2 | 2 | 2 | 1 |
| 22 | 59 | 1 | 0 | 135 | 234 | 0 | 1 | 161 | 0 | 0.5 | 1 | 0 | 3 | 1 |
| 23 | 44 | 1 | 2 | 130 | 233 | 0 | 1 | 179 | 1 | 0.4 | 2 | 0 | 2 | 1 |
| 24 | 42 | 1 | 0 | 140 | 226 | 0 | 1 | 178 | 0 | 0 | 2 | 0 | 2 | 1 |
| 25 | 61 | 1 | 2 | 150 | 243 | 1 | 1 | 137 | 1 | 1 | 1 | 0 | 2 | 1 |

Figure 1. Instantaneous sample of the dataset

2.2. Data preprocessing

The dataset underwent a series of preprocessing steps to ensure optimal model performance [34][35]. First, the dataset was split into features (independent variables) and the target (dependent variables). Gaussian noise was added to numerical features to improve the robustness of models against noise and variability. The dataset was then divided into training and testing sets using an 80-20 split to allow for model evaluation on unseen data. Next, missing values were handled using mean imputation for numerical features and mode imputation for categorical features. Categorical variables were encoded using one-hot encoding to convert them into a format suitable for machine learning models. To ensure consistency and enhance performance for algorithms sensitive to feature scaling, such as K-Nearest Neighbors (KNN) and Support Vector Machines (SVM), feature normalization was performed using a StandardScaler, which transforms features to have a mean of zero and a

standard deviation of one. These preprocessing steps helped prepare the dataset for practical model training and evaluation.

Table 1. Overview of dataset features

| S/N | Attribute | Description | Values |
|-----|-----------|---|--|
| 1 | Age | Represents the age of the individual in years. | - |
| 2 | Sex | Denotes the gender of the individual. | Female (0), Male (1) |
| 3 | CP | Type of chest pain experienced. | Typical angina (0), Atypical angina (1), Non-anginal pain (2), Asymptomatic (3) |
| 4 | Trestbps | Resting blood pressure in mm Hg. | 130-140 mm Hg |
| 5 | Chol | Serum cholesterol levels in mg/dl. | Ranges between 126 and 564 mg/dl |
| 6 | Fbs | Fasting blood sugar levels. | >120 mg/dl (True = 1, False = 0) |
| 7 | Restecg | Results of resting electrocardiogram. | Normal (0), ST-T wave abnormality (1), Left ventricular hypertrophy (2) |
| 8 | Thalach | Maximum heart rate achieved. | 71 to 202 |
| 9 | Exang | Indicates exercise-induced angina. | Yes (1), No (0) |
| 10 | Oldpeak | ST depression caused by exercise relative to rest. | Values range from 0 to 6.2 |
| 11 | Slope | Slope of the peak exercise ST segment. | Upsloping (1), Flat (2), Downsloping (3) |
| 12 | Ca | Number of major vessels obstructed as observed in fluoroscopy. | 0 to 3 |
| 13 | Thal | Type of defect observed. | Normal (3), Fixed defect (6), Reversible defect (7) |
| 14 | Target | Target variable indicating the presence of disease. | Yes (1), No (0) |

2.3. Feature extraction

Feature extraction was crucial in simplifying the dataset while retaining the most meaningful information for model training. Feature extraction helped improve model performance, enhance interpretability, and decrease computational complexity by reducing dimensionality [36][37]. This research utilized Recursive Feature Elimination (RFE) with a Random Forest Classifier to identify the most significant features. RFE operates iteratively, eliminating the least relevant features while constructing models to assess and rank feature importance. This study applied RFE by first training a Random Forest model on the full dataset, evaluating feature importance scores, and recursively removing the least essential features until the optimal subset was identified. Random Forest was chosen as the base estimator due to its ability to handle high-dimensional data, robustness to overfitting, and built-in feature importance ranking. Through this process, 13 critical features were selected, significantly influencing the prediction task. These identified features were compiled into a new DataFrame for streamlined management and application in later analytical stages. By Focusing on the most essential features, the model achieved better accuracy and efficiency while reducing computational demands.

2.4. Model construction

In prior research, multiple machine learning algorithms were evaluated to identify the most suitable model, as no single classifier could consistently deliver optimal performance without experimentation. This limitation arises from variations in the learning mechanisms of different ML algorithms. Consequently, several machine learning models were designed and tuned to analyze data efficiently. The models utilized included Random Forest, Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), XGBoost (XGB), Logistic Regression, and Gradient Boosting. These models were selected based on their unique strengths. Random Forest and Decision Tree are robust for handling non-linear relationships and missing data, while KNN is effective for classification tasks with well-separated data. SVM is known for its high accuracy in high-dimensional spaces. XGBoost and Gradient Boosting offer strong predictive performance through boosting techniques, and Logistic Regression provides interpretability in binary classification problems. To further enhance predictive accuracy, a Voting Classifier was implemented as an ensemble method, integrating the outputs of Random Forest, SVM, and XGB. The rationale for this approach lies in leveraging the complementary strengths of these models—Random Forest's ability to reduce overfitting, SVM's effectiveness in high-dimensional spaces, and XGB's powerful feature selection and handling of complex patterns. By

combining these models, the ensemble aimed to improve overall performance and reduce individual model biases. Before applying the algorithms, the dataset was partitioned into training and testing sets in an 80:20 ratio. This split was chosen to balance training the models effectively and reserving sufficient data for evaluation. An 80% training set ensures that the models learn from a substantial portion of the data, improving generalization. The 20% test set provides an adequate sample size for assessing the models' performance without excessive data leakage.

2.5. Model evaluation

After training, the performance of each model was thoroughly evaluated to determine how well it predicted outcomes. Metrics such as accuracy, precision, recall, and F1-score were calculated, as each plays a crucial role in assessing the model's effectiveness [38]-[40]. Accuracy provides an overall measure of correct predictions. At the same time, precision and recall highlight the balance between false positives and false negatives, which is critical in applications where misclassification can have significant consequences. The F1-score, a harmonic mean of precision and recall, ensures a balanced evaluation, especially for imbalanced datasets. The ROC-AUC score was also used to assess the models' ability to distinguish between classes by measuring how well they rank positive instances higher than negative ones. To gain deeper insights into the models' performance, confusion matrices were generated to visualize the distribution of correct and incorrect predictions, making it easier to identify common misclassification patterns. A soft voting approach was employed for the hybrid model, which combined the probability outputs of individual models in the ensemble to make final predictions. This method improved the system's overall reliability by integrating multiple perspectives from different models, reducing bias, and enhancing predictive stability. These evaluation tools are essential in determining how well a model performs and understanding its limitations, guiding improvements, and ensuring its suitability for real-world applications.

2.6. Architecture of the developed system

The developed system architecture, shown in Figure 2, consists of five major phases: data collection, data pre-processing, feature extraction or selection, model construction or formulation, and model evaluation. Each phase plays a crucial role in ensuring the accuracy and efficiency of the final predictive model. The process begins with the data collection phase, where essential medical parameters such as cholesterol, blood pressure, maximum heart rate achieved, and chest pain are gathered. These variables form the foundation for training and testing the model. Next, the data pre-processing phase refines the dataset to ensure consistency and reliability. This involves techniques such as filtration, tokenization, and stop-word removal to eliminate anomalies and standardize the data, making it suitable for analysis. Following this, the feature extraction and selection phase identifies the most relevant features from the dataset. The system employs Recursive Feature Elimination (RFE) with a Random Forest Classifier, systematically removing less significant features to retain the 13 most relevant ones. This step enhances model performance by reducing complexity while preserving essential predictive information. Different machine learning algorithms are trained and tested in the model construction phase to determine the most effective model for predicting heart disease risk. This phase is critical, as selecting the best-performing algorithm ensures optimal accuracy and generalization to unseen data. Finally, the model evaluation phase assesses the model's performance using various evaluations. The system ensures a structured and systematic approach to heart disease prediction, where each phase contributes to the overall accuracy and reliability of the final model.

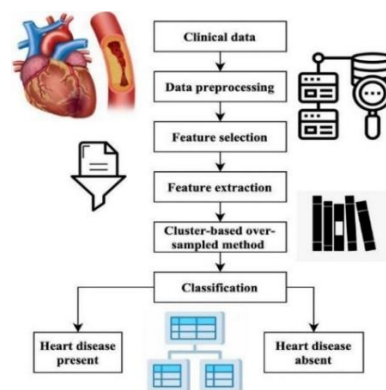


Figure 2. System architectu

3. COMPREHENSIVE THEORETICAL BASIS

3.1. Heart diseases

Heart diseases rank among the top causes of death globally, mainly due to factors like smoking, unhealthy eating habits, obesity, lack of physical activity, high blood pressure, diabetes, and abnormal cholesterol levels. Managing these factors is crucial for lowering the chances of heart-related issues. Recent patterns show that heart disease and stroke are leading causes of death, especially in low- and middle-income nations such as Nigeria. Furthermore, the death rate from these illnesses is significantly higher in women than in men [41].

Cardiovascular disease (CVD) continues to be a significant public health concern and is one of the leading causes of death globally. It includes a range of conditions such as atherosclerosis, stroke, heart failure, irregular heartbeats, heart valve problems, chest pain, and heart muscle disorders. Developing countries have experienced a noticeable increase in CVD cases and related risk factors, with younger people being more affected compared to developed nations. This trend is often linked to limited awareness and insufficient preventive measures, which are further worsened by widespread poverty in these regions [42]. CVD poses a significant health threat in sub-Saharan Africa, accounting for over one-fifth of deaths and 7% of disability-adjusted life years in the region. In Nigeria, the prevalence of CVD has been increasing, driven by rapid urbanization and the adoption of westernized lifestyles, which now contribute to 11% of deaths nationwide. This rising challenge is associated with factors such as poor dietary habits, insufficient physical activity, tobacco consumption, and socioeconomic issues, including financial difficulties and stress at both individual and national levels. According to the World Health Organization, the impact of CVD is projected to grow further. It is anticipated to overtake infectious diseases as the leading health concern in developing nations by 2030 [43]. Congenital heart disease (CHD) refers to a range of heart defects that occur from birth, affecting the heart's structure and function. These defects vary in severity, from minor conditions with little impact on health to complex abnormalities needing urgent medical care. Diagnosing CHD can be difficult due to the wide variation in symptoms, which depend on the type and severity of the defect. Key diagnostic methods include assessing family history, conducting physical examinations, and using diagnostic and genetic testing. One crucial indicator is the presence of abnormal heart sounds caused by irregular blood flow within the heart, often signaling structural issues like a hole or a narrowed valve. Healthcare providers listen for these sounds during physical exams and may recommend further tests to identify the underlying problem [44].

Coronary heart disease (CHD) occurs when the major blood vessels of the heart, known as coronary arteries, develop plaque buildup, leading to narrowing and reduced blood flow to the heart. Different tests are used to diagnose CHD, including electrocardiography (ECG), echocardiography (ECHO), and single-photon emission computed tomography (SPECT) scans. Among these methods, ECG stands out as the most accessible, cost-effective, and easy to perform, making it especially suitable for rural and resource-limited areas common in sub-Saharan Africa [45].

3.2. Data mining

Data mining is an essential part of the knowledge discovery in databases (KDD) process, focusing on identifying hidden patterns within data through repeated use of specific methods [46]. The purpose of KDD is to simplify these patterns for easier understanding and better interpretation of the data [47][48]. Over time, the study and application of data mining techniques have grown, particularly in handling real-world databases that are often large and require multiple scans. Standard techniques include classification, which relies on methods like decision trees and neural networks, association rule mining, clustering, and analyzing sequential patterns, all of which aim to uncover valuable insights from data [49]. Broadly, data mining techniques can be classified into predictive and descriptive categories. Predictive techniques include classification, regression, and time series analysis, while descriptive techniques encompass clustering, summarization, and association rule analysis, each designed to explore and interpret data from different perspectives [50].

3.3. Machine learning

Different machine learning methods have been suggested for detecting heart diseases. This section reviews the current machine learning techniques used for heart disease detection. Machine learning aims to allow computers to learn and improve from data without requiring explicit programming [51][52]. It teaches machines how to handle and analyze data more efficiently. The objective is to enable systems to adapt and enhance their performance by learning from the data they encounter. Machine learning methods are commonly grouped based on how they learn: supervised, unsupervised, and reinforcement learning.

Machine learning can perform tasks like regression, classification, and clustering, each having distinct roles. Regression is a supervised learning technique used to predict results by examining the relationship between dependent and independent variables. It includes methods such as Simple Linear Regression, Multiple Linear Regression, and Non-Linear Regression. For example, Simple Linear Regression establishes a direct

line relationship between variables, as demonstrated in Figure 3, where the model adjusts input values (X) to generate corresponding outputs (Y). Classification, another supervised learning method, involves using labeled data to match input data with its appropriate category, with techniques like K-Nearest Neighbors and Support Vector Machines (SVM) being typical examples. On the other hand, clustering is an unsupervised learning method that deals with unlabeled data to identify patterns, group similar data, and decide which group new data should belong to. K-MEANS is a clustering algorithm automatically detecting patterns and organizing data into groups. A classification algorithm graph shown in Figures 4 and 5.

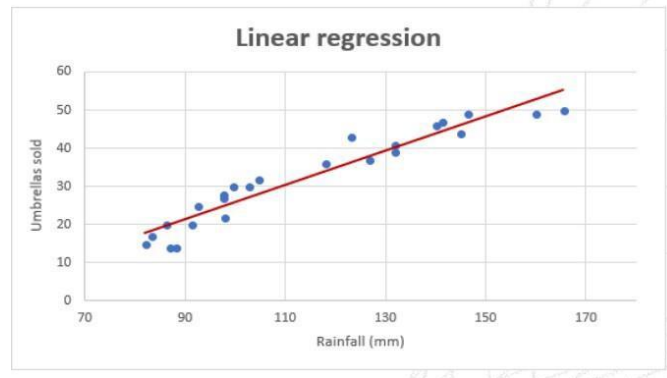


Figure 3. A simple linear regression graph [22]

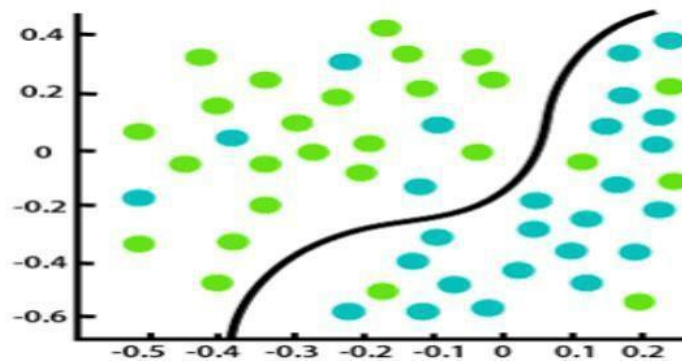


Figure 4. A classification algorithm graph [21]

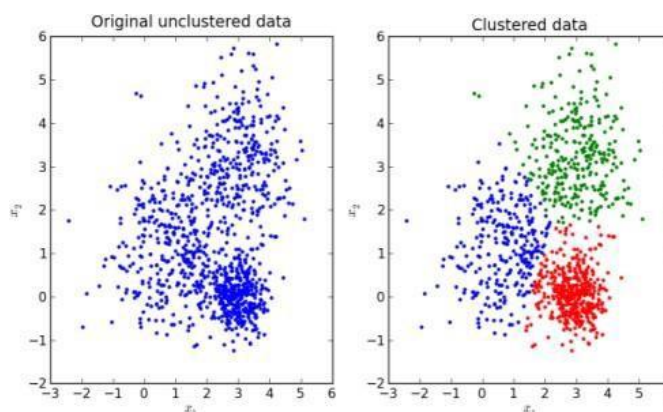


Figure 5. A clustering algorithm graph [21]

3.3.1. Supervised learning

Supervised learning is a method that creates a connection between input variables (X) and an output variable (Y), allowing predictions for new data that hasn't been seen before. This method is commonly used in machine learning and is crucial for managing multimedia data. Supervised learning algorithms used to identify heart diseases. They explored heart disease prediction by applying different supervised learning techniques to analyze the condition while evaluating their effectiveness and precision. The study compared three

classification methods: K-Nearest Neighbor (KNN), Support Vector Classifier, and Multi-Layer Perceptron (neural network), concluding that all these methods were effective at classifying heart disease [53].

3.3.2. Unsupervised learning

Unsupervised learning involves various techniques to uncover patterns or insights from data, particularly focusing on features like X_1, X_2, \dots, X_p . This type of learning is usually more difficult, as the goal is less clear-cut compared to methods like prediction, making the process more subjective. It is often used as part of an exploratory data analysis to help understand the data better [54]. Unsupervised learning models come in different forms, such as K-means, Principal Component Analysis (PCA), and hierarchical clustering. A prediction model for detecting heart diseases using Principal Component Analysis (PCA) was proposed by [55]. In an unsupervised learning method, classification is done without prior information about the sorted data. The process occurs without guidance or supervision, making it part of the unsupervised machine learning approach [56].

3.3.3. Ensemble learning

Ensemble learning involves creating and combining multiple models, like classifiers or experts, to address specific computational problems [57]. Its primary purpose is to enhance the performance of tasks such as classification, prediction, and function approximation while minimizing the risk of relying on an underperforming model. Additionally, ensemble learning is applied to tasks like improving decision confidence, choosing the best features, merging data, supporting incremental and adaptive learning, and correcting errors. Ensemble learning techniques combine several machine learning algorithms to generate predictions that, on their own, may not be strong. By analyzing data from various angles and merging the predictions using voting methods, these techniques improve accuracy compared to relying on a single algorithm. These techniques enhance the accuracy of predictions by using multiple models that are trained separately and then merging their outcomes.

4. RESULTS AND DISCUSSION

4.1. Performance metrics

This section assesses the machine learning algorithms applied in this study, using necessary measures like accuracy, precision, recall, and F1 score. Bar charts display the results based on these evaluation metrics to help compare the performance of different algorithms.

4.1.1. Model accuracy

Figure 6 shows the accuracy comparison of different models in heart disease prediction. The Hybrid Model achieved the highest accuracy at 97.56%, demonstrating its superior ability to identify patterns associated with early heart disease. XGBoost followed with an accuracy of 93.17%, highlighting its effectiveness in handling complex relationships within the dataset. The Logistic Regression model, with the lowest accuracy of 78.05%, indicates its limitations in capturing non-linear patterns, making it less suitable for this predictive task than the other models. The accuracy differences underscore the importance of selecting models that balance interpretability and predictive power for reliable heart disease prediction.

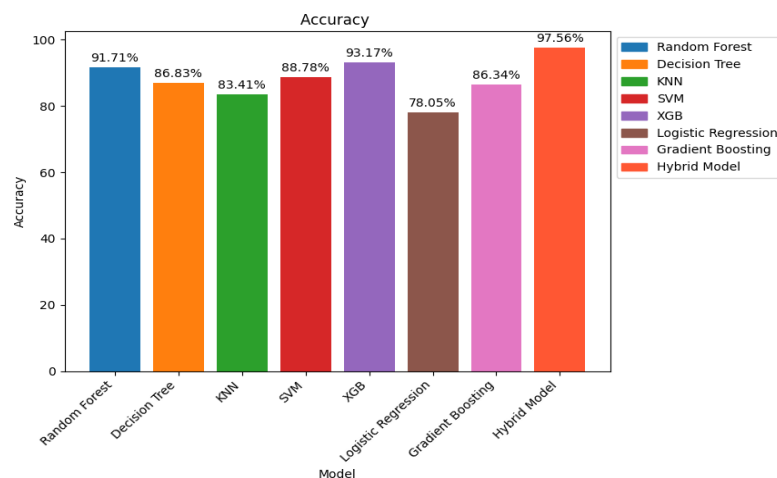


Figure 6. Model accuracy comparison

4.1.2. Model precision

Figure 7 presents the plot of precision comparison, the Hybrid Model leads with the highest precision at 98.04%, followed closely by XGBoost with a precision of 94.95%, suggesting their superior ability to reduce false positives. Logistic Regression, however, has the lowest precision at 73.77%, showing that it may misclassify a significant number of non-risk cases as at-risk.

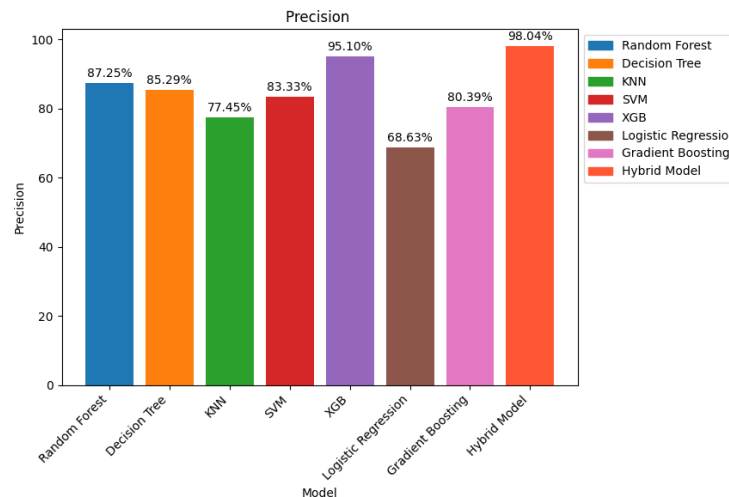


Figure 7. Model precision comparison

4.1.3. Model recall

Figure 8 presents the model recall comparison. Random Forest achieved the highest recall at 96.12%, closely followed by the Hybrid Model with a recall of 97.09%, highlighting their effectiveness in identifying positive cases of heart disease. The Logistic Regression model shows the lowest recall at 87.38%, which may lead to more missed cases in heart disease detection.

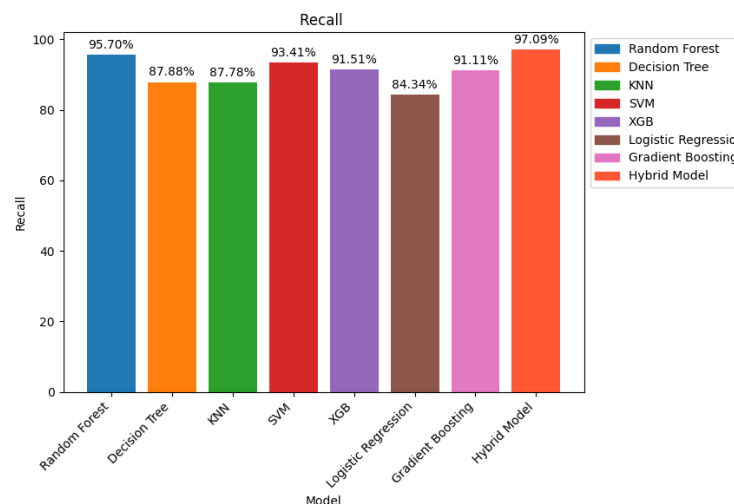


Figure 8. Model recall comparison

4.1.4. Model F1 score

Figure 9 presents the model F1 score comparison; the Hybrid Model demonstrates the highest F1 score at 97.56%, effectively balancing precision and recall. XGBoost also performs well with an F1 score of 93.07%. At the lower end, Logistic Regression has an F1 score of 80.0%, indicating it is less reliable than other models for this predictive task.

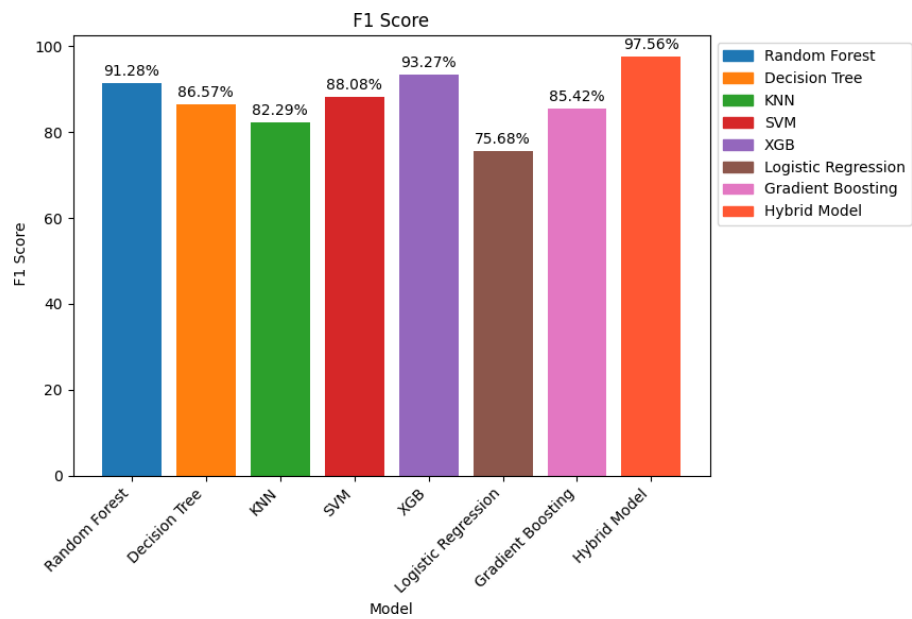


Figure 9. Model F1 score comparison

Table 2. Summary of performance metrics of the models

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1 Score (%) |
|---------------------|--------------|---------------|------------|--------------|
| Random Forest | 91.71 | 88.39 | 96.12 | 92.09 |
| Decision Tree | 86.83 | 85.85 | 88.35 | 87.08 |
| KNN | 83.41 | 80.0 | 89.32 | 84.4 |
| SVM | 88.78 | 85.09 | 94.17 | 89.4 |
| XGB | 93.17 | 94.95 | 91.26 | 93.07 |
| Logistic Regression | 78.05 | 73.77 | 87.38 | 80.0 |
| Gradient Boosting | 86.34 | 82.61 | 92.23 | 87.16 |
| Hybrid Model | 97.56 | 98.04 | 97.09 | 97.56 |

4.1.5. Summary of performance metrics of the models

Table 2 summarizes each model's performance metrics, and Figure 10 shows the confusion matrix. The Hybrid Model outperformed the other models, achieving the lowest False Positive Rate (FPR) at 1.46% and False Negative Rate (FNR) at 0.98%. In comparison, XGBoost exhibited a higher FPR of 2.44% and an FNR of 4.39%, indicating a relatively higher misclassification rate. Logistic Regression performed the worst, with an FPR of 15.61% and an FNR of 6.34%, suggesting significant difficulties distinguishing between classes. This comparison highlights the Hybrid Model's superiority in minimizing false positives and negatives, making it the most reliable approach among the evaluated models.

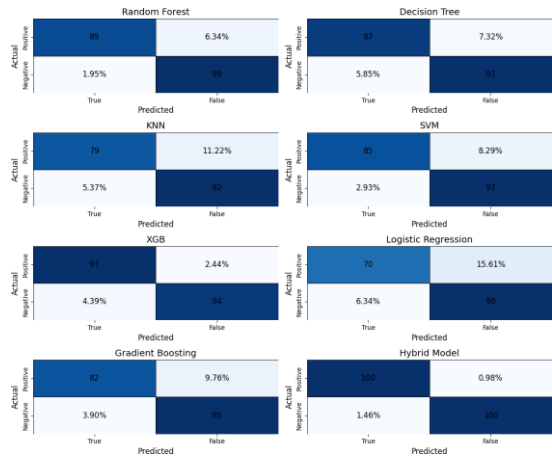


Figure 10. Confusion matrix of all the models

4.2. Receiver Operating Characteristic (ROC) curve

This section presents the Receiver Operating Characteristic (ROC) curve, an essential tool for evaluating the efficacy of binary classification models in predicting early heart disease risk. The ROC curve plots the True Positive Rate (TPR), or sensitivity, against the False Positive Rate (FPR) across a range of threshold values, allowing us to visualize the trade-off between sensitivity and specificity.

Figure 11 shows the Area under the ROC Curve (AUC), which provides a quantitative measure of each model's performance, with a higher AUC generally indicating better predictive accuracy. In early heart disease prediction, a high AUC is particularly critical, as it reflects the model's ability to differentiate between individuals at risk correctly and those not at risk. Given the severe consequences of misclassification in medical applications, a high AUC ensures fewer false negatives, reducing the likelihood of undiagnosed at-risk patients. This is essential for timely medical intervention, improving patient outcomes, and enhancing the reliability of the predictive system in real-world healthcare settings.

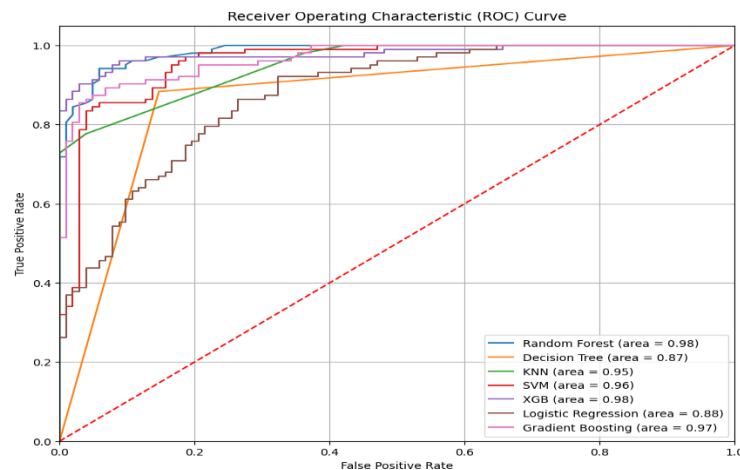


Figure 11. Area under the curve of all models

The ROC curve presented in Figure 12 achieves an AUC score of 99%, reflecting a very high level of accuracy for the heart disease detection system. This score demonstrates the hybrid model's ability to differentiate between individuals with and without heart disease. An AUC of 0.99 indicates that the model ranks heart disease cases above non-cases 99% of the time. This accuracy underscores the model's reliability in balancing sensitivity (True Positive Rate) and specificity (True Negative Rate). Additionally, the curve closely approaches the top-left corner of the graph, highlighting the model's strong performance. Such precision is vital in clinical settings, where reducing false positives and negatives is critical for accurate and safe diagnoses.

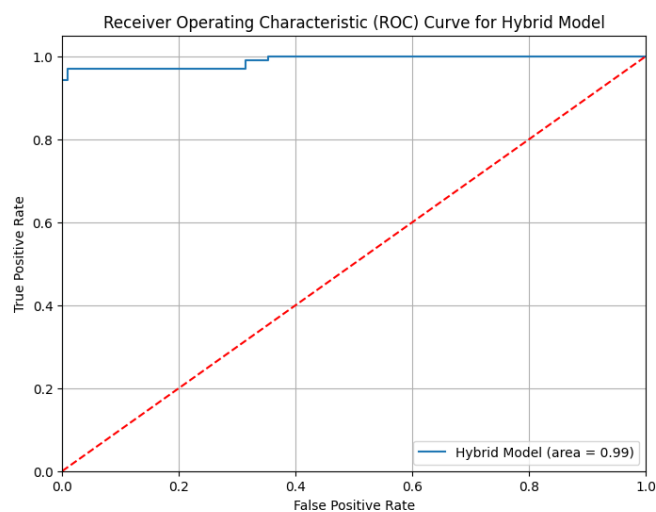


Figure 12. Area under the curve of the hybrid mode

4.3. Graphical User Interface (GUI) of the developed model

The Graphical User Interface (GUI) for the developed heart disease prediction model, presented in [Figure 13](#), is an interactive platform for users to input essential patient information related to heart disease risk factors. These factors include age, sex, chest pain type, resting blood pressure, cholesterol levels, and fasting blood sugar status, which are critical in determining the likelihood of heart disease. The GUI streamlines the prediction process by enabling users to enter relevant data, upon which the model generates an output prediction and a probability score. The prediction classifies whether the patient is likely to have heart disease, while the probability score quantifies the model's confidence in its assessment. A score between 0 and 0.49 suggests little to no risk, whereas a score from 0.5 to 1 indicates a high risk, assisting in the early identification of at-risk individuals who may require further medical evaluation. Figure 13 visually illustrates the model's decision-making process, demonstrating how user inputs are processed to generate predictions. The interface also includes a "Flag" button to highlight high-risk cases for further review. This functionality is particularly beneficial for healthcare professionals, as it aids in prioritizing urgent cases and streamlining patient management. By integrating these features, the GUI enhances usability, facilitates quick data entry, and provides a transparent and interpretable risk assessment, ultimately supporting healthcare providers in making informed clinical decisions. The visual representation in [Figure 13](#) underscores the model's effectiveness in translating patient data into actionable insights, reinforcing its value as a decision-support tool in heart disease prediction.

Heart Disease Prediction

Enter patient data to predict heart disease risk.

| | | |
|----------|----|---|
| age | 50 | output Heart Disease: Yes, Probability: 0.72 Flag |
| sex | 0 | |
| cp | 2 | |
| trestbps | 66 | |
| chol | 5 | |
| fbs | 0 | |

Activate Windows
Go to Settings to activate Windows.

Figure 13. Graphical User Interface (GUI) of the developed model

5. CONCLUSION AND LIMITATION

This research evaluated different machine learning algorithms for predicting early heart disease risk, focusing on the effectiveness of a Hybrid Model. The Hybrid Model outperformed Random Forest, Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), XGBoost, Logistic Regression, and Gradient Boosting. It achieved high accuracy at 97.56%, precision of 98.04%, recall of 97.09%, and an F1 score of 97.56%. The model's reliability was further confirmed by a low false positive rate (1.46%) and false negative rate (0.98%), alongside an AUC score of 99, which indicated its strong ability to differentiate high-risk patients from low-risk ones. These findings demonstrate the model's potential in helping healthcare providers identify at-risk individuals for timely interventions, offering key insights into using predictive analytics in healthcare. However, the study faced several limitations. The analysis was based on a specific dataset, which may limit the model's generalizability to broader populations. Differences could also influence real-world accuracy in patient demographics, healthcare systems, and variations in data collection methods. Additionally, while the model demonstrated strong predictive performance, it does not account for real-time clinical variables or dynamic health conditions, which may impact its practical applicability. The study primarily focused on algorithmic evaluation, with limited consideration of the interpretability of model predictions, integration with existing medical systems, and ethical concerns related to AI-driven diagnostics.

Future research should focus on validating the model with more diverse and representative datasets to enhance its generalizability. Longitudinal studies are also necessary to assess the model's adaptability and long-term reliability in real-world clinical applications. Moreover, exploring explainable AI techniques could improve model transparency and trust among healthcare professionals. Addressing these aspects will ensure the robustness, usability, and ethical deployment of AI-based predictive models in healthcare environments.

REFERENCES

- [1] K. Karthick, S. Aruna, R. Samikannu, R. Kuppusamy, Y. Teekaraman, & A. Thelkar, "Implementation of a Heart Disease Risk Prediction Model Using Machine Learning," *Computational and Mathematical Methods in Medicine*, vol. 2022, pp. 1-14, 2022. <https://doi.org/10.1155/2022/6517716>
- [2] H. Jindal, S. Agrawal, R. Khera, R. Jain, & P. Nagrath, "Heart Disease Prediction using Machine Learning Algorithms," *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1, pp. 12072, 2021. <https://doi.org/10.1088/1757-899x/1022/1/012072>
- [3] J. Mehta, G. Kaur, H. Buttar, H. Bagabir, R. Bagabir, & S. Bagabir, "Role of the renin-angiotensin system in the pathophysiology of coronary heart disease and heart failure: diagnostic biomarkers and therapy with drugs and natural products," *Frontiers in Physiology*, vol. 14, 2023. <https://doi.org/10.3389/fphys.2023.1034170>
- [4] C. Razo, C. Welgan, C. Johnson, S. McLaughlin, V. Iannucci, & A. Rodgers, "Effects of elevated systolic blood pressure on ischemic heart disease: a burden of proof study," *Nature Medicine*, vol. 28, no. 10, p. 2056-2065, 2022. <https://doi.org/10.1038/s41591-022-01974-1>
- [5] A. Armoundas, S. Narayan, D. Arnett, K. Spector-Bagdady, D. Bennett, & L. Celi, "Use of Artificial Intelligence in Improving Outcomes in Heart Disease: A Scientific Statement From the American Heart Association," *Circulation*, vol. 149, no. 14, 2024. <https://doi.org/10.1161/cir.0000000000001201>
- [6] C. Jurgens, C. Lee, D. Aycock, R. Creber, Q. Denfeld, & H. DeVon, "State of the Science: The Relevance of Symptoms in Cardiovascular Disease and Research: A Scientific Statement From the American Heart Association," *Circulation*, vol. 146, no. 12, 2022. <https://doi.org/10.1161/cir.0000000000001089>
- [7] N. Bansal, L. Zelnick, R. Scherzer, M. Estrella, & M. Shlipak, "Risk Factors and Outcomes Associated With Heart Failure With Preserved and Reduced Ejection Fraction in People With Chronic Kidney Disease," *Circulation: Heart Failure*, vol. 17, no. 5, 2024. <https://doi.org/10.1161/circheartfailure.123.011173>
- [8] K. Yang and M. Song, "New Insights into the Pathogenesis of Metabolic-Associated Fatty Liver Disease (MAFLD): Gut-Liver-Heart Crosstalk," *Nutrients*, vol. 15, no. 18, p. 3970, 2023. <https://doi.org/10.3390/nu15183970>
- [9] R. Sarra, A. Dinar, M. Mohammed, M. Ghani, & M. Albahar, "A Robust Framework for Data Generative and Heart Disease Prediction Based on Efficient Deep Learning Models," *Diagnostics*, vol. 12, no. 12, p. 2899, 2022. <https://doi.org/10.3390/diagnostics12122899>
- [10] A. Bhowmick, K. Mahato, C. Azad, & U. Kumar, "Heart Disease Prediction Using Different Machine Learning Algorithms," 2022 *IEEE World Conference on Applied Intelligence and Computing (AIC)*, 2022. <https://doi.org/10.1109/aic55036.2022.9848885>
- [11] C. Navarro, J. Damen, T. Takada, S. Nijman, P. Dhiman, & J. Ma, "Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review," *BMJ*, p. n2281, 2021. <https://doi.org/10.1136/bmj.n2281>
- [12] R. Sarra, A. Dinar, M. Mohammed, & K. Abdulkareem, "Enhanced Heart Disease Prediction Based on Machine Learning and χ^2 Statistical Optimal Feature Selection Model," *Designs*, vol. 6, no. 5, p. 87, 2022. <https://doi.org/10.3390/designs6050087>
- [13] X. Liu, D. Lü, A. Zhang, Q. Liu, & G. Jiang, "Data-Driven Machine Learning in Environmental Pollution: Gains and Problems," *Environmental Science & Technology*, vol. 56, no. 4, p. 2124-2133, 2022. <https://doi.org/10.1021/acs.est.1c06157>
- [14] S. Kutiname, R. Millham, A. Adekoya, M. Tetley, B. Weyori, & P. Appiahene, "Application of Machine Learning Algorithms in Coronary Heart Disease: A Systematic Literature Review and Meta-Analysis," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, 2022. <https://doi.org/10.14569/ijacsa.2022.0130620>
- [15] B. Kaur and G. Kaur, "Heart Disease Prediction Using Modified Machine Learning Algorithm," *Lecture Notes in Networks and Systems*, p. 189-201, 2022. https://doi.org/10.1007/978-981-19-2821-5_16
- [16] R. Rastogi and M. Bansal, "Diabetes prediction model using data mining techniques," *Measurement: Sensors*, vol. 25, p. 100605, 2023. <https://doi.org/10.1016/j.measen.2022.100605>
- [17] T.R. Ramesh, U. Lilhore, M. Poongodi, S. Simaiya, A. Kaur, & M. Hamdi, "Predictive Analysis of Heart Diseases With Machine Learning Approaches," *Malaysian Journal of Computer Science*, pp. 132-148, 2022. <https://doi.org/10.22452/mjcs.sp2022no1.10>
- [18] E. Onyema, O. Khalaf, C. Tavera, S. Tayeb, S. Ghouali, & G. Abdulsahib, "A Classification Algorithm-Based Hybrid Diabetes Prediction Model," *Frontiers in Public Health*, vol. 10, 2022. <https://doi.org/10.3389/fpubh.2022.829519>
- [19] Z. Zhou, "Open-environment machine learning," *National Science Review*, vol. 9, no. 8, 2022. <https://doi.org/10.1093/nsr/nwac123>

- A. Kwekha-Rashid, H. Abduljabbar, & B. Alhayani, "Coronavirus disease (covid-19) cases analysis using machine-learning applications," *Applied Nanoscience*, vol. 13, no. 3, pp. 2013-2025, 2021. <https://doi.org/10.1007/s13204-021-01868-7>
- [20] W. Li, Y. Chai, F. Khan, S. Jan, S. Verma, & V. Menon, "A Comprehensive Survey on Machine Learning-Based Big Data Analytics for IoT-Enabled Smart Healthcare System," *Mobile Networks and Applications*, vol. 26, no. 1, pp. 234-252, 2021. <https://doi.org/10.1007/s11036-020-01700-6>
- [21] M. Pichler and F. Härtig, "Machine learning and deep learning—a review for ecologists," *Methods in Ecology and Evolution*, vol. 14, no. 4, pp. 994-1016, 2023. <https://doi.org/10.1111/2041-210x.14061>
- [22] S. Safiri, N. Karamzad, K. Singh, K. Carson-Chahhoud, C. Adams, & S. Nejadghaderi, "Burden of ischemic heart disease and its attributable risk factors in 204 countries and territories, 1990–2019," *European Journal of Preventive Cardiology*, vol. 29, no. 2, pp. 420-431, 2021. <https://doi.org/10.1093/eurjpc/zwab213>
- [23] S. Khan, J. Coresh, M. Pencina, C. Ndumele, J. Rangaswami, & S. Chow, "Novel Prediction Equations for Absolute Risk Assessment of Total Cardiovascular Disease Incorporating Cardiovascular-Kidney-Metabolic Health: A Scientific Statement From the American Heart Association," *Circulation*, vol. 148, no. 24, pp. 1982-2004, 2023. <https://doi.org/10.1161/cir.0000000000001191>
- [24] N. Wenger, D. Lloyd-Jones, M. Elkind, G. Fonarow, J. Warner, & H. Alger, "Call to Action for Cardiovascular Disease in Women: Epidemiology, Awareness, Access, and Delivery of Equitable Health Care: A Presidential Advisory From the American Heart Association," *Circulation*, vol. 145, no. 23, 2022. <https://doi.org/10.1161/cir.0000000000001071>
- [25] Y. Zhuang, Y. Wang, P. Sun, J. Ke, & F. Chen, "Association between triglyceride glucose-waist to height ratio and coronary heart disease: a population-based study," *Lipids in Health and Disease*, vol. 23, no. 1, 2024. <https://doi.org/10.1186/s12944-024-02155-4>
- [26] M. Cushman, C. Shay, V. Howard, M. Jiménez, J. Lewey, & J. McSweeney, "Ten-Year Differences in Women's Awareness Related to Coronary Heart Disease: Results of the 2019 American Heart Association National Survey: A Special Report From the American Heart Association," *Circulation*, vol. 143, no. 7, 2021. <https://doi.org/10.1161/cir.0000000000000907>
- [27] N. Absar, E. Das, S. Shoma, M. Khandaker, M. Miraz, & M. Faruque, "The Efficacy of Machine-Learning-Supported Smart System for Heart Disease Prediction," *Healthcare*, vol. 10, no. 6, p. 1137, 2022. <https://doi.org/10.3390/healthcare10061137>
- [28] H. El-Sofany, B. Bouallègue, & Y. El-Latif, "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," *Scientific Reports*, vol. 14, no. 1, 2024. <https://doi.org/10.1038/s41598-024-74656-2>
- [29] A. Ogunpola, F. Saeed, S. Basurra, A. Albarrak, & S. Qasem, "Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases," *Diagnostics*, vol. 14, no. 2, p. 144, 2024. <https://doi.org/10.3390/diagnostics14020144>
- [30] S. Mondal, R. Maity, Y. Omo, S. Ghosh, & A. Nag, "An Efficient Computational Risk Prediction Model of Heart Diseases Based on Dual-Stage Stacked Machine Learning Approaches," *IEEE Access*, vol. 12, pp. 7255-7270, 2024. <https://doi.org/10.1109/access.2024.3350996>
- [31] P. Kokol, M. Kokol, & S. Zagoranski, "Machine learning on small size samples: a synthetic knowledge synthesis," *Science Progress*, vol. 105, no. 1, 2022. <https://doi.org/10.1177/00368504211029777>
- [32] G. Ahmad, H. Fatima, S. Ullah, A. Saidi, & A. Imdadullah, "Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV," *IEEE Access*, vol. 10, pp. 80151-80173, 2022. <https://doi.org/10.1109/access.2022.3165792>
- [33] A. Ahmad and H. Polat, "Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm," *Diagnostics*, vol. 13, no. 14, p. 2392, 2023. <https://doi.org/10.3390/diagnostics13142392>
- [34] E. ShafieiBavani, B. Goudey, I. Kiral-Kornek, P. Zhong, A. Yepes, & A. Swan, "Predictive models for cochlear implant outcomes: performance, generalizability, and the impact of cohort size," *Trends in Hearing*, vol. 25, 2021. <https://doi.org/10.1177/23312165211066174>
- [35] A. Ulloa, L. Jing, J. Pfeifer, S. Raghunath, J. Ruhl, & D. Rocha, "rECHOmmend: An ECG-Based Machine Learning Approach for Identifying Patients at Increased Risk of Undiagnosed Structural Heart Disease Detectable by Echocardiography," *Circulation*, vol. 146, no. 1, pp. 36-47, 2022. <https://doi.org/10.1161/circulationaha.121.057869>
- [36] I. Mienye and N. Jere, "Optimized Ensemble Learning Approach with Explainable AI for Improved Heart Disease Prediction," *Information*, vol. 15, no. 7, p. 394, 2024. <https://doi.org/10.3390/info15070394>
- [37] S. Babu, P. Ramya, & J. Gracewell, "Revolutionizing heart disease prediction with quantum-enhanced machine learning," *Scientific Reports*, vol. 14, no. 1, 2024. <https://doi.org/10.1038/s41598-024-55991-w>
- [38] H. Yang, Z. Chen, Y. Huajian, & M. Tian, "Predicting Coronary Heart Disease Using an Improved LightGBM Model: Performance Analysis and Comparison," *IEEE Access*, vol. 11, pp. 23366-23380, 2023. <https://doi.org/10.1109/access.2023.3253885>
- [39] M. Shams, A. Elshewey, E. El-kenawy, A. Ibrahim, F. Talaat, & Z. Tarek, "Water quality prediction using machine learning models based on grid search method," *Multimedia Tools and Applications*, vol. 83, no. 12, pp. 35307-35334, 2023. <https://doi.org/10.1007/s11042-023-16737-4>
- [40]

- [41] K. Okorie-Ufere, P. Regidor, & S. Adeniyi, "Assessment of the Knowledge of Risk Factors Associated with Heart Diseases among Women of Reproductive Age in Nigeria," *International Journal of Nursing, Midwife and Health Related Cases*, vol. 10, no. 2, pp. 36-56, 2024. <https://doi.org/10.37745/ijnmh.15/vol10n23656>
- [42] N. Odunaiya, T. Adesanya, E. Okoye, & O. Oguntibeju, "Towards cardiovascular disease prevention in nigeria: a mixed method study of how adolescents and young adults in a university setting perceive cardiovascular disease and risk factors," *African Journal of Primary Health Care & Family Medicine*, vol. 13, no. 1, 2021. <https://doi.org/10.4102/phcfm.v13i1.2200>
- [43] N. Odunaiya, O. Adegoke, A. Adeoye, & O. Oguntibeju, "Preliminary study of perceived cardiovascular disease risk and risk status of adults in small rural and urban locations in ibadan, nigeria," *AIMS Public Health*, vol. 10, no. 1, pp. 190-208, 2023. <https://doi.org/10.3934/publichealth.2023015>
- [44] G. Wang, B. Wang, & P. Yang, "Epigenetics in Congenital Heart Disease," *Journal of the American Heart Association*, vol. 11, no. 7, 2022. <https://doi.org/10.1161/jaha.121.025163>
- [45] C. Ezeude, A. Ezeude, M. Abonyi, M. Nkpozi, C. Ugwueze, & K. Akhidue, "Associations of asymptomatic coronary heart disease in a cohort of stable type 2 diabetic subjects in a tertiary health center in south eastern Nigeria: A cross -sectional study," *International Journal of Scholarly Research in Multidisciplinary Studies*, vol. 4, no. 1, pp. 001-012, 2024. <https://doi.org/10.56781/ijrms.2024.4.1.0090>
- [46] T. Kwan, S. Wong, Y. Hong, A. Kanaya, S. Khan, & L. Hayman, "Epidemiology of Diabetes and Atherosclerotic Cardiovascular Disease Among Asian American Adults: Implications, Management, and Future Directions: A Scientific Statement From the American Heart Association," *Circulation*, vol. 148, no. 1, pp. 74-94, 2023. <https://doi.org/10.1161/cir.0000000000001145>
- [47] G. Isola, A. Polizzi, A. Alibrandi, R. Williams, & A. Giudice, "Analysis of galectin-3 levels as a source of coronary heart disease risk during periodontitis," *Journal of Periodontal Research*, vol. 56, no. 3, pp. 597-605, 2021. <https://doi.org/10.1111/jre.12860>
- [48] T. Yang, Y. Liu, L. Li, Y. Zheng, Y. Wang, & J. Su, "Correlation between the triglyceride-to-high-density lipoprotein cholesterol ratio and other unconventional lipid parameters with the risk of prediabetes and type 2 diabetes in patients with coronary heart disease: a rcsd-tcm study in china," *Cardiovascular Diabetology*, vol. 21, no. 1, 2022. <https://doi.org/10.1186/s12933-022-01531-7>
- [49] M.S. Sousa, M.L.Q. Mattoso & N.F.F. Ebecken, "Data Mining: A Database Perspective," *WIT Transactions on Information and Communication Technologies*, vol. 22, p.19, 2024. <http://doi.org/10.2495/DATA980301>
- [50] D. Shankar, A. Azhakath, N. Khalil, J. Sajeev, T. Mahalakshmi, & K. Sheeba, "Data mining for cyber biosecurity risk management – a comprehensive review," *Computers & Security*, vol. 137, p. 103627, 2024. <https://doi.org/10.1016/j.cose.2023.103627>
- [51] A. Almulihi, H. Saleh, A. Hussien, S. Mostafa, S. El-Sappagh, & K. Alnowaiser, "Ensemble Learning Based on Hybrid Deep Learning Model for Heart Disease Early Prediction," *Diagnostics*, vol. 12, no. 12, p. 3215, 2022. <https://doi.org/10.3390/diagnostics12123215>
- [52] W. Ng, G. Goh, G. Goh, J. Ten, & W. Yeong, "Progress and Opportunities for Machine Learning in Materials and Processes of Additive Manufacturing," *Advanced Materials*, vol. 36, no. 34, 2024. <https://doi.org/10.1002/adma.202310006>
- [53] K. Kumar, V. Rohini, J. Yadla, & J. VNRaju, "A Comparison of Supervised Learning Algorithms to Prediction Heart Disease," 2023 *International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*, 2023. <https://doi.org/10.1109/iceconf57129.2023.10084035>
- [54] G. James, D. Witten, T. Hastie, R. Tibshirani, & J. Taylor, "Unsupervised Learning," *Springer Texts in Statistics*, pp. 503-556, 2023. https://doi.org/10.1007/978-3-031-38747-0_12
- [55] S. Sharma, M. Kaur, & S. Gupta, "A Comparison of Machine Learning Approaches for Forecasting Heart Disease with PCA Dimensionality Reduction," *Smart Innovation, Systems and Technologies*, pp. 333-347, 2023. https://doi.org/10.1007/978-981-99-3982-4_29
- [56] R. Kumar, S. Polepaka, & D. Krishna, "An Insight on Machine Learning Algorithms and its Applications," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 11S2, pp. 432-436, 2019. <https://doi.org/10.35940/ijitee.k1069.09811s219>
- [57] I. Mienye and N. Jere, "Optimized Ensemble Learning Approach with Explainable AI for Improved Heart Disease Prediction," *Information*, vol. 15, no. 7, p. 394, 2024. <https://doi.org/10.3390/info15070394>

BIOGRAPHIES OF AUTHORS





Dugguh Sylvester Aondonenge is a data analyst at Federal Inland Revenue Service, in 2017, he obtains a bachelor degree in Computer Science from Federal University Kashere, Gombe State. He further advanced his studies in 2024 where he obtain a Masters degree in Information Technology (MIT) from Ahmadu Bello University, Zaria. His area of interest is Data Analysis and Machine Learning. He can be contacted via email at dugguhsylvester@gmail.com.



Ajayi Ore-Ofe is a lecturer at the Department of Computer Engineering, Ahmadu Bello University, Zaria, Nigeria. He received his MSc and Ph.D from Computer Engineering in Control Engineering, in 2017 and 2022 respectively. He received his MSc and Ph.D from the department of Computer Engineering in Ahmadu Bello University, Zaria, Nigeria. He is mainly research in control engineering. He can be contacted at email: ajayi.oreofe17@gmail.com.



Abubakar Umar   is a lecturer in the Department of Computer Engineering at Ahmadu Bello University, Zaria, Nigeria. He earned his BEng Degree from Electrical Engineering Department Ahmadu Bello University, Zaria, Nigeria, in 2011, MSc, and Ph.D. degrees from Computer Engineering Department, Ahmadu Bello University, Zaria, Nigeria, in 2017 and 2024. He specializes in various aspects of computer engineering. His primary research focus is in Control Engineering, where he explores the development and optimization of control systems for different applications. He is dedicated to advancing his research and contributing to academic knowledge in this field. He can be contacted via email at abuumar@abu.edu.ng, abubakaru061010@gmail.com



Kamorudeen Hassan Taiwo is a Senior Registrar in the department of Family Medicine at Ahmadu Bello University Teaching Hospital, Zaria, Nigeria. He holds a Bachelor of Medicine, Bachelor of Surgery (M.B.B.S) as well as Master's degree in Disaster Risk Management and Development Studies (MDRMDS) from the Department of Medicine and Department of Geography in Ahmadu Bello University, Zaria, Nigeria, respectively. Dr. Taiwo is an Associate Fellow of National Postgraduate Medical College of Nigeria (NPMCN) and a Fellow of the Institute of Disaster Management and Safety Science, Nigeria (FDMSS). His primary area of research is in clinical medicine with focus on Artificial Intelligence (AI) applications in healthcare. He can be reached at via email at drtaiwokamar@gmail.com.