Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 6, 2025

HEARTSMART: IMPROVED CVD RISK PREDICTION VIA RECURSIVE FEATURE ELIMINATION: VALIDATION ON EXTENDED DATASET

Waqas Tariq Paracha[*1], **Haleema Inam**[2], **Maliha Manzoor**[3]

[*1,3]*Gomal research institute of computing (GRIC), Faculty of Computing, Gomal University, DIKhan (KP), Pakistan*
[2]*Department of Computer Science & Information Technology Virtual University Islamabad*

[*1]waqasparacha125@gmail.com, [2]haleemainam786@gmail.com, [3]maliha.manzoor6@gmail.com

DOI: https://doi.org/10.5281/zenodo.15772033

Keywords

Article History
Received on 21 May 2025
Accepted on 21 June 2025
Published on 30 June 2025

Corresponding Author: *
Waqas Tariq Paracha

Abstract

Heart disease continues to rank among the leading causes of mortality worldwide, posing serious challenges for global healthcare systems. Early detection and precise risk prediction are vital for reducing death rates and ensuring timely medical interventions. This study investigates a machine learning–based framework to enhance heart disease prediction by employing Recursive Feature Elimination (RFE), a robust feature selection technique that systematically removes less significant features to boost model performance and minimize computational costs.

Initially, the research utilized a real-world dataset comprising 70,000 patient records sourced from Kaggle. [1] To further strengthen the analysis, the dataset was expanded to 100,000 samples using the Synthetic Minority Oversampling Technique (SMOTE) in Python, enabling a more balanced and enriched data representation. Multiple machine learning algorithms were then applied, including Random Forest, Decision Tree, Naïve Bayes, K-Nearest Neighbor (KNN), and XGBoost, to evaluate their predictive capabilities. Among these, the Random Forest classifier continued to demonstrate superior results, achieving a high accuracy of 99.55% and an AUC of 1.00 on the augmented dataset, showing a minor yet promising improvement over the original performance.

The findings confirm the effectiveness of RFE in isolating the most relevant features, thereby improving interpretability, enhancing model efficiency, and reducing unnecessary computational burden. By removing redundant or irrelevant features, RFE ensures that the model focuses on the most critical indicators of heart disease risk. This research contributes to the advancement of a predictive framework capable of assisting healthcare professionals in making more informed clinical decisions. With an accurate and efficient model, early detection and proactive treatment planning become more feasible, ultimately improving patient outcomes and reducing the global burden of heart disease through the integration of machine learning in medical diagnostics.

INTRODUCTION

1.1 Heart Disease Overview

CVD remains one of the leading causes of death worldwide, with the World Health Organization (WHO) identifying it as a major contributor to global mortality. The heart is a vital organ responsible for pumping and circulating blood throughout the entire body, including supplying the brain. If the heart fails to deliver adequate blood flow to the central nervous system, it can result in the failure of the nervous system, leading to the dysfunction of nerves and tissues across the body and ultimately causing death. Therefore, the heart is a critical organ essential for sustaining human life. Therefore, for an individual to live a healthy life, their heart must work in a proper function. According to [3], decreasing the death rate requires early disease detection and timely administration of the proper treatment. In addition, the most significant challenge facing the community now is the diagnosis of heart disease. In 2016, Nearly 17.9 million individuals [2] lost their lives because of heart disease. Heart disease can be split down into several subtypes, each of which needs to be diagnosed at an earlier stage. This is among the emerging illnesses that are rapidly disseminating across the globe and it is responsible for a rise in the high mortality rate everywhere.

1.1.1 Heart Disease Risk Factors

Risk factors are various for the causes of the increasing blockage. Modifiable and non-modifiable risk factors are the classifications used to organize these risk factors. Gender, age, and genetics are non-modifiable risk factors in a person's health. No matter what, these risk factors will continue to be the main reason why people get heart disease. Potential dangers that can be changed are called changeable risk factors. Changing risk factors like those related to Habitat, stress, food, and biochemical risk factors. Heart disease exists in different forms: coronary, rheumatic, congenital, myocarditis, arrhythmia, angina, and atherosclerosis.

The risk of heart disease is higher in those who engage in certain lifestyle habits or circumstances that are considered risk factors. Moreover, the presence of these risk factors carries the possibility of the disease being present and becoming even more severe. **Figure 1.1**is a graphical representation of the risk factors that impact the heart system.

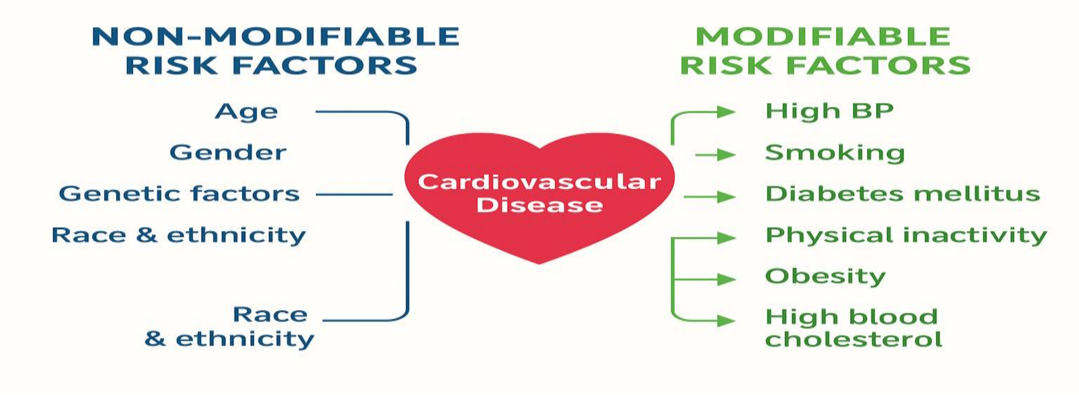


Figure 1.1: Heart Disease Risk Factors

The diagnosis of heart disease may involve a variety of blood tests, imaging procedures (such as MRIs and CT scans), and electrocardiograms (ECGs).

1.2 Need For Heart Disease Prediction

An estimated 17.5 million people die from heart disease each year in the world. Seventy-five percent of people die from heart disease, which primarily affects those with middle-class and lower-class incomes. Furthermore, heart attacks and strokes account for 80% of deaths caused by heart disease. India is among the nations where the number of people with heart disease is increasing each year, according to a WHO report. Every year, two lakh open heart surgeries are performed due to the rise in the number of patients affected by heart disease. The biggest concern in recent years has been the 20%–30% growth in the patient count.

1.3 Various Methods for Heart Disease Prediction

Predicting heart disease using machine learning involves leveraging various algorithms to analyze medical data and identify patterns indicative of cardiovascular conditions. Traditional methods like Logistic Regression are often used for their simplicity and interpretability, making them suitable for binary classification tasks such as predicting the presence or absence of heart disease. Decision Trees and Random Forests are also popular due to their ability to handle non-linear relationships and feature importance ranking, which helps in understanding the contribution of factors like cholesterol levels, blood pressure, and age. Support Vector Machines (SVM) are effective for high-dimensional data and can classify patients by finding the optimal hyperplane that separates healthy individuals from those at risk. K-Nearest Neighbors (KNN) is another approach that classifies patients based on similarity measures, though it can be computationally expensive for large datasets. Advanced techniques like Gradient Boosting Machines (GBM) and XGBoost have gained traction for their high accuracy and ability to handle imbalanced datasets, which are common in medical data.

1.4 Feature Selection Method.

Feature selection is used in machine learning to get new factors from raw data, which is needed for the processes of machine learning. You can get rid of noisy, repetitive, and useless data with the help of a feature selection method. This process also makes the classification more accurate. The main goal of the feature selection method is to get the smallest set of features from the problem so that the description of the original feature is accurate. An important part of identifying heart disease is picking out the right characteristics. The results that are collected during the research phase of feature selection for heart disease forecast rely on the quality groups of patients and the traits that were picked ahead of time. There are three ways to choose features: Extra Tree Classifier [5], Recursive Feature Elimination [6], and Information Gain [7]. Recursive Feature Elimination will be used in this study to reduce the number of features to overcome the complexity of time.

1.5 Motivation

his study is motivated by the observation that individuals often attempt to diagnose themselves based on their prior knowledge. While this may occasionally lead to correct decisions, in most cases, people lack a sufficient understanding of the nature and severity of their condition. As a result, self-treatment or delayed medical consultation can increase both the risk and severity of disease, potentially leading to more advanced stages of illness, higher healthcare costs, and greater treatment challenges. To address this, the present study extends the dataset from our previous work to improve the accuracy and reliability of heart disease prediction models, supporting earlier and more effective diagnosis.

1.6 Aims of the Study

The study aims to enhance the accuracy of heart disease diagnosis, ensuring early intervention and reducing mortality rates. Additionally, it investigates the impact of Recursive Feature Elimination (RFE) on both the accuracy and interpretability of predictive models. Through comparative analysis with established benchmark models, the study seeks to demonstrate the effectiveness of the proposed approach in improving heart disease prediction [8].

1.7 Research Objectives

The main objectives of this research are

1. How Machine Learning algorithms can be used in the diagnosis of heart disease by building an optimized model that can be used to predict heart diseases.
2. To predict heart disease from the given heart disease dataset by way of the use of a machine learning models and extended this data set.
3. To check the impact of Recursive Feature Elimination (RFE) on the accuracy and interpretability of machine learning models for predicting heart disease
4. Evaluating the advised model's performance to that of other benchmark works.

1.8 Research Questions

The questions addressed by this study are:

Research Question No 1: How can machine learning algorithms be used in the diagnosis of heart disease by building an optimized model that can be used to predict heart diseases?

Research Question No. 2: How can machine learning models predict better heart disease from the extended heart disease dataset?

Research Question No. 3: How does Recursive Feature Elimination (RFE) impact the accuracy and interpretability of machine learning models for predicting heart disease?

Research Question No. 4: What is the accuracy of the proposed models regarding similar research studies?

1.9 Thesis Outline

This thesis is structured into five main chapters. The **Introduction** provides an overview of heart disease, highlighting its types, risk factors, and the necessity for predictive modeling using machine learning. It outlines various methods employed for heart disease prediction, the motivation behind this research, and the objectives it aims to achieve. The **Related Work** chapter reviews previous studies on heart disease prediction, analyzing existing methodologies, datasets, and machine learning models. It highlights gaps in current research and positions the proposed approach within the broader academic landscape. The **Research Methodology** chapter details the dataset used, preprocessing techniques, feature selection methods, and model development strategies. It discusses evaluation metrics and tools employed while addressing potential risks and ethical concerns. The **Experiments and Results** chapter presents the implementation and performance evaluation of various machine learning models, comparing their effectiveness in heart disease prediction. Finally, the **Conclusion and Future Work** chapter summarizes the key findings, discusses limitations, and suggests potential improvements, including real-time applications and wearable health monitoring technologies. This structured approach ensures a comprehensive exploration of heart disease prediction using machine learning.

2. Related Work

Review of the Research on Machine Learning in the Prediction of Heart Disease in Table 2.1.

Table2.1 Comparison of Existing Studies

Year	Study	Dataset	Methods and Accuracy	Research Gap
2020	[66]	299 heart failure patients collected from Faisalabad Institute of Cardiology and Allied Hospital Faisalabad	Logistic Regression: 83%	Limited to specific biomarkers; may not generalize across diverse populations.
2021	[67]	University College Dublin (UCD), Ireland (410 subjects), University Hospital of Ioannina (77 subjects).	Rotation Forest: 91.23%	Class Imbalance and Undersampling
2021	[63]	Kaggle heart disease dataset consisting of 303 patient	Random Forest (RF) 0.88 Support Vector Machine (SVM) 0.83 Naive Bayes (NB) 0.85 Logistic Regression (LR) 0.87	Limited Dataset Size and Diversity
2023	[62]	Statlog Heart, Cleveland, Hungarian, Switzerland, and Long Beach	A Meta Model (87 % Accuracy) combining Random Forest, Gaussian Naive Bayes, Decision Tree, and k-nearest Neighbor algorithms.	The complexity of Heart systems and the need for diverse data integration pose challenges.
2023	[71]	Kaggle Dataset contains 70000 records	RF: 95%, Decision Tree:94%, Multilayer Perceptron: 95%	The study does not validate the models on an external dataset or in a real-world clinical setting.

Year	Study	Dataset	Methods and Accuracy	Research Gap
			XGBoost: 95%	
2024	[64]	Cleveland Heart Disease dataset	Ensemble Mode (Logistic Regression, Decision Trees SVM, Random Forests, Neural networks) 65.55%	Imbalanced Data Distribution
2024	[65]	Kaggle heart disease dataset consisting of 303 patient	Convolution Neural Network (CNN) 91%	Lack of Real-Time Prediction Capabilities
2024	[68]	Cleveland Heart Disease dataset	SVM: 82% NB: 81.32% LSTM: 79.51	Lack of Models Diversity and Comparison
2024	[69]	Kaggle Dataset contains 1100 records with 14 different features	LR: 84.07% SVM: 65.19% KNN: 62.69% NB:85.19% DNN:76.05%	The study didn't discuss the computational complexity and time efficiency of the dimensionality reduction techniques.
2024	[70]	Kaggle Dataset contains 303 records with 14 features	(Logistic Regression, Decision Trees, XGBoost, Gradient boosting, Random Forests, Support Vector Machines, and Artificial Neural Networks) Accuracy Not Available	Machine Learning models did not achieve a significant advantage; clinical feasibility concerns.
2024	[73]	Dataset from JUH in Amman, Jordan	SVM: 94.3%	The study does not explicitly address the issue of imbalanced datasets
2024	[74]	Kaggle Public and Private Dataset with 505 records and 14 different Features	XGBoost: 97.57%	Limited Dataset Size and Diversity

Table 2.1 shows an entire analysis of machine learning research studies that predict heart disease. Different research efforts from 2020 to 2024 are summarized by showing their chosen datasets and applied ML models with accurate results and identified study limitations. Multiple studies involving Cleveland Heart Diseases, Kaggle, Statlog, and University Hospital datasets have utilized machine learning models including Logistic Regression, Support Vector Machines (SVM), Random Forest, XGBoost, Gradient Boosting, Neural Networks, and CNNs. The predictive achievement reached maximum values by XGBoost at 97.57% along with Random Forest which reached 95%.

3 Research Methodology

The primary objective of this study is to empower patients and healthcare providers with a digital heart disease prediction model capable of accurately assessing the risk of developing cardiac conditions. The comprehensive framework for heart disease prediction is illustrated in **Figure 3.1**.

The dataset utilized in this research was sourced from Kaggle, specifically from the following resource: Heart Disease Dataset.This dataset comprises a vast collection of 70,000 records, providing valuable insights into various heart-related conditions.

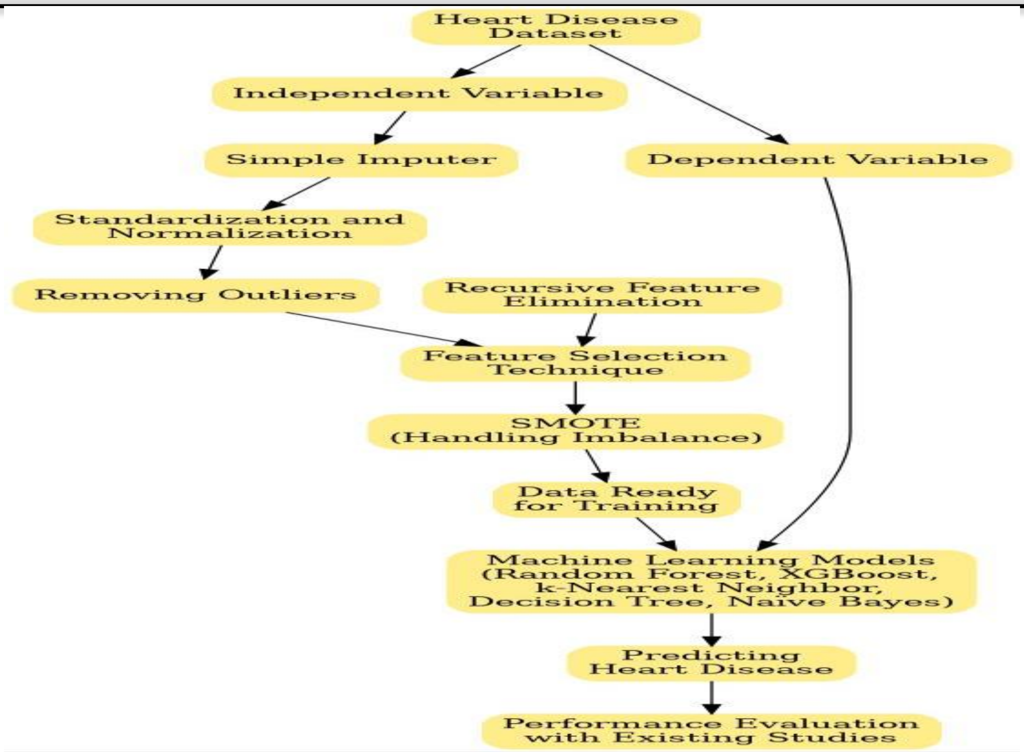

Heart Disease Dataset
Independent Variable
Simple Imputer
Dependent Variable
Standardization and Normalization
Removing Outliers
Recursive Feature Elimination
Feature Selection Technique
SMOTE (Handling Imbalance)
Data Ready for Training
Machine Learning Models (Random Forest, XGBoost, k-Nearest Neighbor, Decision Tree, Naive Bayes)
Predicting Heart Disease
Performance Evaluation with Existing Studies

Figure 3-1: A Proposed Model for Predicting Heart Disease

3.1 Heart Disease Dataset

The Heart Disease dataset typically contains information about people and different health-related variables. Here's a breakdown of the usual features and the target variable you might find in such a dataset:

Age: This numeric characteristic refers to the age of the person age.

Sex: This categorical feature describes the person's gender and is usually represented as 0 and 1 (for instance, 0 for female and 1 for male).

BMI (Body Mass Index): BMI is an estimated value obtained from the person's height and weight. It measures whether the individual falls under the category of underweight, normal weight, overweight, or obese. A high BMI may increase the risk of heart disease.

Blood Pressure: Blood pressure includes two readings: systolic and diastolic.

Cholesterol Levels: This categorical aspect presents high, normal, or low cholesterol levels. High cholesterol levels, especially high LDL (bad) cholesterol levels, increase the risk of heart disease.

Blood Sugar/Glucose Levels: This reads the blood glucose levels of an individual. Hyperglycemia or high glucose amounts are associated with diabetes, and people affected suffer the risk of having heart disease.

Smoking: 1 for a smoker, 0 for a non-smoker is a binary feature indicating if the individual smokes. Smoking is a contributing factor to heart disease.

Alcohol Intake: Another binary identifying whether the person drinks alcohol (1 for yes, 0 for no). Alcohol overuse may affect heart health.

Physical Activity: Represents how much physical activity or exercise the individual does. Physically active individuals are generally associated with lower risks of heart disease.

Family History: 1 for a positive history of heart disease and 0 for no family history of heart disease. A family history of heart ailment can make individuals more prone to high risk.

Target: Heart Disease: 0 for absence, 1 for presence of heart disease in the patient.
By using machine learning techniques, researchers and data scientists could create models that use these factors to figure out the risk of heart disease. That study separates the heart disease information into dependent variables and independent variables. Type of person, height, weight, blood pressure, cholesterol, smoking, glucose levels, alcohol intake, physical exercise, and family background are all examples of independent variables. The Target variable is the only thing that makes up the dependent variable. So that an exact model can be made to identify heart disease, the goal is to find patterns and connections between the features of the independent variable and the features of the dependent variable.

3.2 Data Preprocessing

A usual procedure of preprocessing a heart dataset includes several stages for analyzing machine learning models. In data preprocessing, different steps are included in the research model to process the data, such as data inspection and understanding, handling missing values through the imputation process, handling duplicate records, and performing a Normalization through the Min-Max process to enhance the model's accuracy. Every step included in the research model is defined below.

3.2.1 Handling Missing Values

Data are prepared for analysis or modeling by processing missing values to ensure the robustness and reliability of data. First, missing values usually refer to null or NaN, identified through functions such as isnull() in Python’s Pandas. Recognizing the pattern and nature of missingness across columns or rows can guide handling it best.
Imputation can be a common technique in which missing values are replaced by estimates based on statistical measurements. To replace the qualitative data, these can be completed by inserting the mean median or mode of the corresponding column without losing the integrity of a dataset. Even missing values may have to be dropped if missingness is large, systematic, and significantly affects the analysis.

3.2.2 Handling Duplicate Data

Finding and managing instances of identical or similar records within a dataset constitutes handling duplicate data. Finding duplicate data is essential because it can affect the quality of models based on the data, overstate patterns or trends, and change the outcomes of analyses. The identification procedure usually entails comparing rows or columns across the collection to find records with matching values. Once found, duplicates can be eliminated from the dataset using functions like duplicated in Python's Pandas package. Eliminating duplicates guarantees that
every observation in the dataset is distinct, avoiding biases in analysis or modeling and redundancy.

3.3 Standardization

Standardization, also called z-score normalization or zero-mean normalization, is a preprocessing strategy often used in machine learning and data analysis to scale the features of a dataset.

Calculate the mean (average): Compute the average of the feature for all data points.

Calculate the standard deviation: Determine the standard deviation of the feature. The distribution or variance of a data collection could be shown by calculating its standard deviation.

Subtract the mean: For each feature data point, deduct the mean. This makes zero the center of the data.

Divide by the standard deviation: For each centered point, divide by the standard deviation. This normalizes the data so that the values are measured with a consistent unit.
Benefits of standardization include:

Equalizes scales: Standardization provides equal scales for all features, which is crucial to algorithms based on distance metrics like k-nearest neighbors or clustering.

Improves convergence: The gradient descent and most other optimization algorithms converge faster if the features are at a similar scale.

Facilitates interpretation: Values are standardized to the same units, which makes it easier to compare the importance or contribution of one feature in a model.

3.4 Normalization

Normalization is a method of data preprocessing that employs scaling and standardization of the characteristics of a dataset. The purpose of normalization is to ensure that the algorithm or model can effectively learn from the data by bringing all of the features under consideration to a comparable scale.
Here are some of the more frequent approaches to normalization:

Min-Max Normalization: This approach adjusts the attributes to conform to a specific interval, frequently extending from 0 to 1.

Z-score Standardization: This method standardized the characteristics by aligning them with a mean and standard deviation.

Robust Scaling: This approach is related to the median and interquartile distance (IQR), making it less sensitive to outliers.

3.5 Removing Outliers

A dataset contains outliers if there are data points that significantly differ from most of the other data points.

Measurement errors: Outliers may appear when errors occur during data collection or
recording. For instance, if the wrong recording of a person’s height or weight is made, this may cause an outlier.

Natural variation: Sometimes, outliers may arise from the inherent variability in the data. But they might still be important and give a lot of information.

Data processing errors: Outliers can also be introduced if errors occur during data entry, transformation, or processing. It is essential to meticulously clean and pre-process data to decrease the impact of such errors.

Extreme values: The outliers can also be considered a legitimate data point representing the distribution's extreme values. The context of the data and the domain play a vital role in distinguishing between meaningful outliers and errors.

Some of the different ways to find outliers are:
Visual inspection: The data can be plotted using histograms, box plots, or scatterplots to visualize the distribution and identify any outliers.

Statistical methods: It could be useful to compute summary statistics like standard deviation, median, and mean. Two major statistical approaches for detecting outliers are Z-scores and IQR.

Machine learning techniques: Some machine learning algorithms, for instance, clustering or anomaly detection models, can detect outliers automatically.

3.6 Feature Selection

An important part of machine learning is feature selection to identify the most relevant and informative features for creating predictive models. Different feature selection techniques exist, but a Recursive Feature Elimination approach is used to select the most significant features in this research. Here are various techniques used for feature selection:

Filter Methods: Filter Methods use statistical properties like correlation, variance, or information gain to determine the effectiveness of features. Here are some commonly used filter method techniques.

Pearson's Correlation Coefficient: Determines whether a feature is correlated linearly with the target variable. High correlating features are taken as more relevant.

Feature-Feature Correlation: Identifies pairs of highly correlated features and helps to remove redundant features.

Low Variance Filter: Removes characteristics with low variance, assuming they are less informative because their values do not vary commonly within the dataset.

Mutual Information: It measures the information obtained about one variable by monitoring another. The features that have high mutual information with the target are seen to be more informative.

Chi-Square Test: It evaluates the independence of categorical variables by comparing observed and expected frequencies. It identifies characteristics most likely to be independent of the target variable and hence, less relevant.

Wrapper Methods: Some of the wrapper methods for feature selection involve leveraging a given machine learning algorithm to assess different subsets of features, choosing the subset that offers the best model performance. In these methods, the machine learning model directly contributes to selecting features, which can be computationally expensive. Here are some standard wrapper method techniques:

Recursive Feature Elimination (RFE): According to the success of a machine learning model, Recursive Feature Elimination (RFE) is a feature selection wrapper method that gets rid of less important features repeatedly. To get the target number of features or optimal subset, the method involves applying the model repeatedly on subsets of characteristics, identifying and removing the least significant ones.

Process:

1. **Model Training:** The whole set of data is used to train a certain machine learning model, like linear regression, support vector machines, and so on.
2. **Feature Importance:** Feature weight coefficients or the importance of each feature are determined after training, using a measure specific to the model, such as feature importance in tree-based models.
3. **Feature Elimination:** Low-ranked or low-importance features are removed from the present set of attributes.
4. **Iteration:** 1-3 are performed recursively on the reduced feature set until either an arbitrarily small number of features remain or a stopping criterion (such as a plateau in model fitting) is reached.

Forward Selection
Process: It begins with no attributes and incrementally adds one feature at a time, monitoring how each feature affects the model's performance until the required number of features is met.

Backward Elimination
Process: All features are added first, and then the least important feature is taken away one at a time until the best group or desired number of features is reached.

Embedded Methods
Lasso (L1 Regularization): Sparsity-enforcing penalties in linear models that automatically pick which features are relevant.

Tree-Based Feature Importance: Models that use decision trees, like Random Forest or Gradient Boosting provide the feature importances that help select the most essential features.

Mutual Information: Measuring the amount of knowledge about one variable obtained through other variable observation.

Sequential Forward Selection (SFS): One feature at a time should be changed while its effectiveness is checked until a certain level is reached.
Once the most important feature from the dataset has been chosen, the data is ready to be used to train the model. Several types of machine-learning methods such as Random Forest, Decision Tree,

Naïve Bayes, K-Nearest Neighbor, and XGBoostare used in this study.

3.7 Modelling

The dataset is partitioned into two subsets, with 80% allocated for training and 20% reserved for testing. The training dataset serves as the foundation for model learning, while the testing dataset is utilized to evaluate the model's predictive performance. Various machine learning classifiers, including Decision Tree, XGBoost, Naïve Bayes, Random Forest, and K-Nearest Neighbor, are applied to the structured dataset. To measure the effectiveness of each model, key evaluation metrics such as accuracy, precision, recall, and F-measure are employed, ensuring a comprehensive assessment of their performance.

The specific models—Decision Tree, XGBoost, Naïve Bayes, Random Forest, and K-Nearest Neighbor—were chosen for this study based on their effectiveness in classification tasks and their ability to handle complex datasets. The selection process considered various factors, including accuracy, interpretability, and computational efficiency.

Decision Tree

A Decision Tree is a fundamental machine-learning model that mirrors human decision-making by structuring data into a tree-like hierarchy. At its core, the Root Node serves as the initial input, setting the foundation for the decision process. As the tree branches out, Internal Nodes represent feature attributes, each forming a decision point based on specific criteria. These branches extend further until they reach Leaf Nodes, which signify the final classification or prediction outcome.

Gradient Boosting

Gradient Boosting algorithm begins by training an initial base model (often a simple one, like a shallow decision tree) to make predictions on the dataset. The errors (residuals) from this model are then calculated, and a new model is trained to predict these residuals, effectively learning from the mistakes of the first model. This new model is added to the ensemble, and the process is repeated iteratively. At each step, the algorithm minimizes a loss function (e.g., mean squared error for regression or log loss for classification) by adjusting the model parameters in the direction of the negative gradient of the loss function. This gradient-based optimization ensures that each

subsequent model focuses on the most challenging aspects of the data, gradually improving overall performance.

XGBoost

XGBoost works by iteratively adding decision trees to an ensemble, with each new tree designed to correct the errors (residuals) made by the previous ensemble. This iterative process continues until the model's predictions converge to an optimal solution, minimizing a predefined loss function.

Naive Bayes

At its core, Naive Bayes relies on Bayes' theorem to estimate the probability of a class based on the observed features. The algorithm consists of two main phases: training and prediction. During the training phase, it calculates the probabilities of each feature occurring within each class, essentially learning the likelihood of specific features given a class label. In the prediction phase, for a new instance, Naive Bayes computes the posterior probability of each class using the observed features and selects the class with the highest probability as the predicted output. This process is both simple and efficient, making Naive Bayes a popular choice for many applications.

K-Nearest Neighbors (KNN)

At the heart of KNN lies the idea that similar data points should have similar labels or values. The algorithm assumes that if a new data point is close to a set of neighboring points in the feature space, its label or value should resemble those of its neighbors. To quantify this similarity, KNN relies on a distance metric, with the Euclidean distance being the most commonly used. This metric calculates the straight-line distance between two points in an n-dimensional space, providing a measure of how close or far apart they are. Other distance metrics, such as Manhattan or Minkowski distances, can also be used depending on the nature of the data.

A critical parameter in KNN is the choice of K, which represents the number of nearest neighbors considered when making a prediction. The value of

K significantly influences the algorithm's behavior. A larger K tends to smooth out decision boundaries, making the model more robust to noise, while a smaller K can make the model more sensitive to local variations in the data. Selecting an appropriate K is essential for balancing bias and variance, and it often requires experimentation and tuning based on the dataset.

3.8 Evaluation Metrics in Machine Learning

The research uses different evaluation metrics to compare the different models. Evaluation metrics are essential to machine learning as they help judge the model's performance and its ability on new, unseen data. However, the specific choice of metrics for evaluation depends on the problem you are trying to solve (our case is based on classification).

Classification Metrics in Machine Learning

A machine learning model that classifies input data based on several categories or classes is assessed using classification metrics. In the case of a binary classification, where there are two classes – positive and negative –, some standard measures include Accuracy, Precision, Recall, F1 Score, and AUC-ROC. Let's explore these metrics in detail:

Accuracy:

Formula: TP + TN / (TP + TN + FP + FN)

Accuracy is the number of correctly predicted instances divided by the total number. It gives a general indication of the accuracy of the model.

Interpretation: Although accuracy is the most commonly used metric, it can be misleading when dealing with imbalanced datasets where one class vastly dominates another. For instance, in a system where 95% of instances are in class A and 5% in class B, a model that predicts all the cases in class A can score an accuracy of almost 95%, which may be poor.

Precision (Positive Predictive Value)

Formula: TP / (TP + FP)

Precision refers to the accuracy of positive predictions made by the model. It answers the question: From all instances that were predicted to be positive, how many of them turned out to be positive?

Interpretation: In situations where false positives are expensive, precision is significant. For instance, when making medical diagnoses, you want to reduce false positives to avoid putting patients through unnecessary treatments.

Recall (Sensitivity, True Positive Rate)

Formula: TP / (TP + FN)

Explanation: Recall determines the capability of the model to reflect all instances of a class. It computes the rate of true positive predictions over total actual positives (true positives + false negatives).

F1 Score

Formula: P * R / (P + R)

Explanation: F1 Scoring is a harmonic mean of precision and recall. It offers a compromise between precision and recall, making this metric helpful in unbalanced class distribution or when false positives and negatives matter.

Area Under the ROC Curve (AUC-ROC)

Explanation: The ROC curve graphically depicts the relationship between true positive rate (sensitivity) and false positive rate(1-specificity), with different thresholds. The AUC-ROC measures also indicate the area under this curve. The higher AUC-ROC signifies better discriminatory capabilities of the model.

Area Under the Precision-Recall Curve (AUC-PR):

Explanation: Unlike AUC-ROC, the AUC-PR metric measures the area under the precision-recall curve. This is particularly useful in unbalanced sets when the number of negative instances greatly exceeds positive ones.

Confusion Matrix

Explanation: A confusion matrix can summarize the model's performance in a tabular form, whereby predictions are binned into TP, TN, FP, and FN. It is helpful in understanding where the model goes wrong and evaluating the various areas of performance.

These classification-based metrics allow practitioners to understand several important features of a model's operation, enabling them to make critical decisions regarding the performance of the given classifier. After performing evaluation metrics of different machine-learning models. The study selects the best model from all machine learning models discussed in the proposed study.

3.9 Risk Involved

In the proposed model, some of the risks include the following:

- A trained model based on machine learning is extremely dependent on the effectiveness of the algorithms used to train it, and there are occasions when these algorithms may not produce optimal results.
- The implementation phase may take longer than expected due to professional and personal responsibilities.
- Furthermore, given that the amount of time it takes for a program to run is partially determined by the speed of the central processing unit (CPU) and the processor, if the program takes a long time to execute, it is likely that the dataset will need to be reduced.

3.10 Tools used

This study makes use of a wide variety of tools. Every one of them is open source and offers no cost.

1. Python version 3.6.5
2. Pgmpy
3. Pandas version 0.23.
4. NumPy version 1.14.3
5. Matplotlib version 2.2.2
6. SciPy and Scikit-learn version 0.19.1 both
7. Seaborn version 0.8.1

3.10 Computational Constraints Addressed

The study tackles multiple computational constraints that typically arise in machine learning-based heart disease prediction. One of the major challenges in handling large-scale medical datasets, such as the 70,000-record heart disease dataset from Kaggle, is computational efficiency. To address this, the dataset from the previous work was extended from 70,000 to 100,000 samples using the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance and enrich the training data. This larger, balanced dataset supports improved model performance and greater generalizability for heart disease prediction.

3.11 Ethical Concern

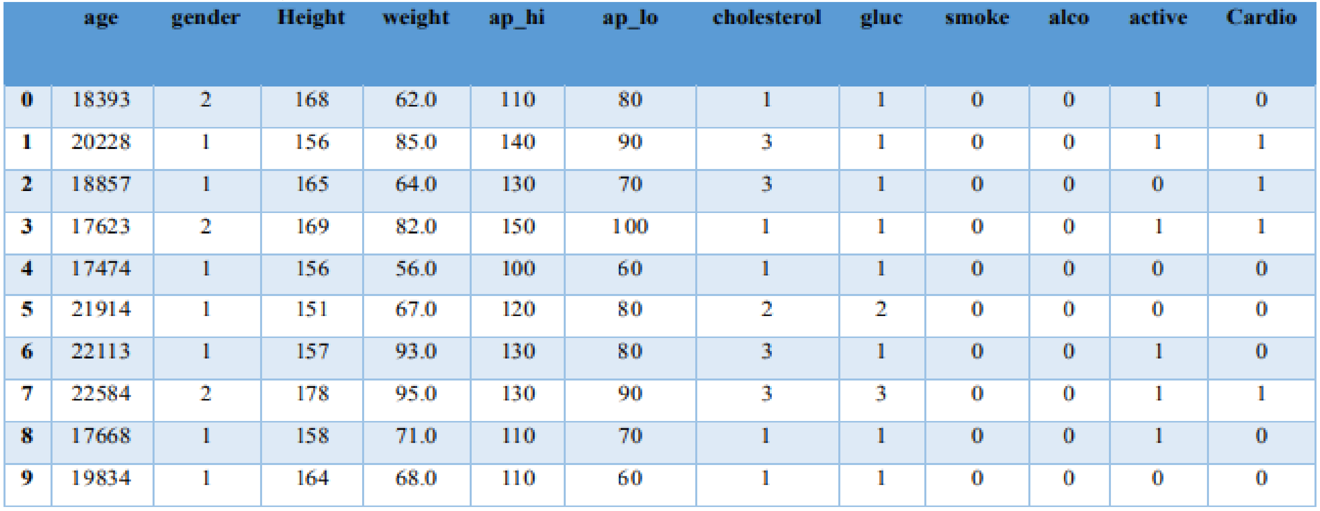
The ethical considerations include:

1. **Fair Representation**: Ensuring that the dataset includes diverse patient demographics to prevent biases that could lead to inaccurate predictions for underrepresented groups.
2. **Transparency in Feature Selection**: Using techniques like Recursive Feature Elimination (RFE) to systematically select features that are most relevant while avoiding those that could introduce bias.
3. **Minimizing Algorithmic Discrimination**: Ensuring that machine learning models do not disproportionately misclassify certain groups by evaluating fairness metrics.
4. **Ethical AI Practices**: Following responsible AI guidelines to avoid unintended consequences in healthcare applications.

4. Experiment and Results

The following section delves into the experimental results and performance evaluation metrics of the study. To develop and implement the machine learning models, the research utilized Jupyter Notebook and the Python programming language, alongside a suite of powerful machine learning libraries, including NumPy, Pandas, Pyplot, and Scikit-Learn.The computational setup for this study comprised a system equipped with a 320 GB hard drive and 8 GB of RAM, running on Windows 10 Professional. This infrastructure provided the necessary resources to handle and process the extensive dataset efficiently.As previously mentioned, the dataset was sourced from the Kaggle machine learning repository and consists of 70,000 records and 12 distinct features related to heart disease. A detailed breakdown of the dataset's attributes and their significance is provided in **Table 4.1**.

Table 4.1: Dataset Sample



	age	gender	Height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	Cardio
0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	17474	1	156	56.0	100	60	1	1	0	0	0	0
5	21914	1	151	67.0	120	80	2	2	0	0	0	0
6	22113	1	157	93.0	130	80	3	1	0	0	1	0
7	22584	2	178	95.0	130	90	3	3	0	0	1	1
8	17668	1	158	71.0	110	70	1	1	0	0	1	0
9	19834	1	164	68.0	110	60	1	1	0	0	0	0

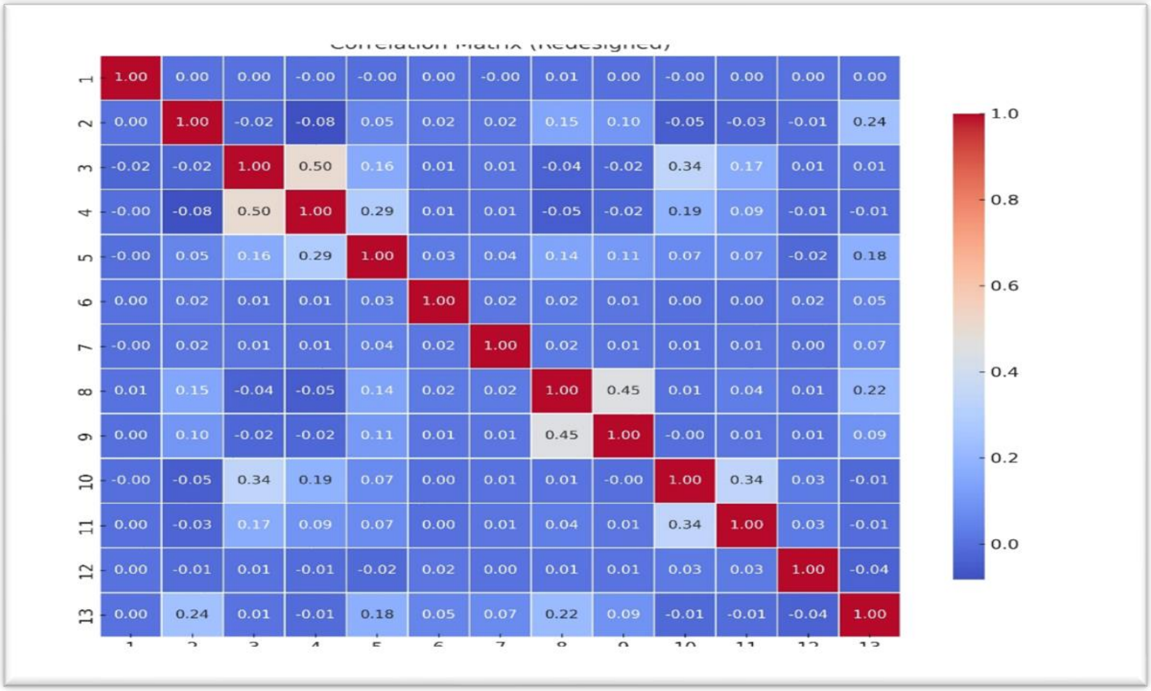
Data preparation occurs after data collection. After reviewing the data set, the imputation process replaces all missing values with the best value. The process of replacing missing or incomplete data with estimated values is imputation.

The dataset after standardization and normalization is shown in **Table 4.2**.

Table 4.2: Dataset after standardization and Normalization

	Age	Gender	Height	weight	Ap_hi	Ap_lo	Cholesterol	gluc	smoke	alco	active
0	-0.66	-0.17	0.52	0.17	-0.32	-0.05	-0.03	-0.21	-0.15	-0.12	-0.09
1	-0.51	0.09	-0.22	-0.30	0.22	0.02	-0.01	0.71	-0.12	-0.09	-0.07
2	-0.46	-0.07	-0.19	0.02	-0.19	0.00	-0.04	0.63	-0.10	-0.08	-0.06
3	-0.66	-0.28	0.52	0.22	0.21	0.05	0.01	-0.21	-0.15	-0.12	-0.09
4	-0.51	-0.24	-0.22	-0.30	-0.37	-0.05	-0.06	-0.16	-0.12	-0.09	-0.07
....											
99996	0.37	0.27	-0.16	-0.17	0.77	0.02	-0.01	0.20	0.29	-0.07	-0.05
99997	0.27	-0.03	0.21	0.35	0.33	0.05	-0.01	0.37	-0.06	-0.05	0.64
99998	0.51	0.36	-0.22	-0.05	-0.05	0.01	-0.03	-0.16	0.40	-0.09	-0.07
99999	0.73	0.18	-0.31	0.29	-0.06	-0.02	-0.04	0.39	-0.17	-0.13	-0.10

After completing the standardization and normalization procedure, design a correlation matrix that displays the correlation coefficients between the variables.**Figure 4.1** is a graph that illustrates the correlation matrix by showing the correlation coefficients between the various characteristics of the heart disease Kaggle dataset.



	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1.00	0.00	0.00	-0.00	-0.00	0.00	-0.00	0.01	0.00	-0.00	0.00	0.00	0.00
2	0.00	1.00	-0.02	-0.08	0.05	0.02	0.02	0.15	0.10	-0.05	-0.03	-0.01	0.24
3	-0.02	-0.02	1.00	0.50	0.16	0.01	0.01	-0.04	-0.02	0.34	0.17	0.01	0.01
4	-0.00	-0.08	0.50	1.00	0.29	0.01	0.01	-0.05	-0.02	0.19	0.09	-0.01	-0.01
5	-0.00	0.05	0.16	0.29	1.00	0.03	0.04	0.14	0.11	0.07	0.07	-0.02	0.18
6	0.00	0.02	0.01	0.01	0.03	1.00	0.02	0.02	0.01	0.00	0.00	0.02	0.05
7	-0.00	0.02	0.01	0.01	0.04	0.02	1.00	0.02	0.01	0.01	0.01	0.00	0.07
8	0.01	0.15	-0.04	-0.05	0.14	0.02	0.02	1.00	0.45	0.01	0.04	0.01	0.22
9	0.00	0.10	-0.02	-0.02	0.11	0.01	0.01	0.45	1.00	-0.00	0.01	0.01	0.09
10	-0.00	-0.05	0.34	0.19	0.07	0.00	0.01	0.01	-0.00	1.00	0.34	0.03	-0.01
11	0.00	-0.03	0.17	0.09	0.07	0.00	0.01	0.04	0.01	0.34	1.00	0.03	-0.01
12	0.00	-0.01	0.01	-0.01	-0.02	0.02	0.00	0.01	0.01	0.03	0.03	1.00	-0.04
13	0.00	0.24	0.01	-0.01	0.18	0.05	0.07	0.22	0.09	-0.01	-0.01	-0.04	1.00

Figure 4.1: Correlation Matrix for Kaggle Heart Disease dataset feature

Looking at the matrix shown in **Figure 4.1**, the correlation coefficient between the cardio and api_hi feature is 0.066. The value indicates that this feature has a considerable impact on diagnosis when taking into consideration this dataset. When looking at the matrix, it is possible to observe that all of the coefficients in the diagonal are 1. This is because the correlation between a variable and itself is always one.

An outlier process is performed on all the dataset attributes to find a correlation between the different features. There are outliers in the dataset, as shown in **Figure 4.2**. Potential errors in entering the data led to these outliers. Eliminating these outliers might make our prediction model work better. To fix this, we eliminated all the ap_hi, ap_lo, weight, and height values that were not between 2.5% and 97.5%. Finding and getting rid of outliers had to be done by hand. The number of rows dropped from 70,000 to 62745 after cleaning the data.

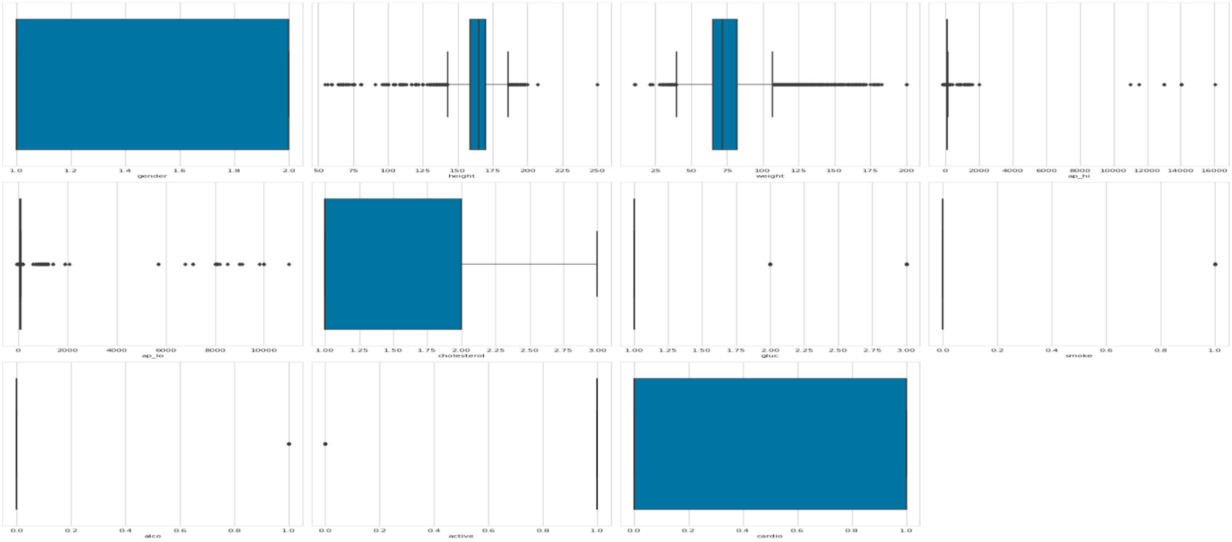


Figure 4.2: Boxplot of all attributes for removing outliers

The data is then analyzed using feature selection methods of Recursive feature elimination.Feature selection means features relevant to the problem and appropriate for the model. On the other hand, irrelevant features would result in a “Garbage in, garbage out” situation in data analysis and machine learning. The feature selection method is chosen to increase the model's accuracy. The total no of attributes in the selected dataset was 12. The Recursive feature elimination method only selects six different attributes with the most highly relevant. These attributes are given to the model to overcome its complexity.

Table 4.3 shows the features selected from the recursive feature elimination method.

Table 4.3: Selected attributes taken from RFE

Sr.No	Selected Attributes Through RFE	High Relevance Attribute
1	height	(0.205464)
2	cholesterol	(0.168511)
3	gender	(0.159489)
4	ap_lo	(0.148742)
5	ap_hi	(0.136044)
6	gluc	(0.104750)

This study used many techniques, such as XGBoost classifier, Naïve Bayes, K-Nearest Neighbor, decision tree, and random forest. Area under the ROC curve, precision, recall, accuracy, and F1 score were some of the performance metrics used in this research. Splitting the dataset in half allowed us to train the model using 80% of the data and test it with 20%.

Naïve Bayes Model Implementation

A summary of the binary classification results is provided in terms of recall, accuracy, precision, and f1-score in **Table 4.4**. The table displays the findings from various evaluation matrices required to verify the model's performance. The model performs well in terms of 98.89% accuracy.

Table 4-4: Classification report of the Naïve Bayes model

Classes	Precision	Recall	F1-score	Support
0	0.99	0.98	0.99	6232
1	0.98	1.00	0.99	7768
Weighted Avg	0.99	0.99	0.99	14000
Accuracy	98.89%			

Confusion matrices were used to validate the results further to classify heart disease according to binary class, shown in **Figure 4.3**.

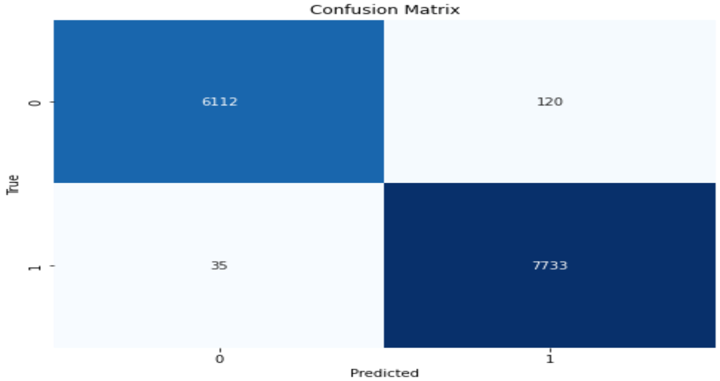


Figure 4.3: Confusion matrix of Naïve Bayes model

A total of 14000 test data points were chosen to evaluate the model. The model correctly predicts that 6112 corresponds to Class 0, while 7733 corresponds to Class 1. The confusion matrix result shows that the Naïve Bayes model correctly classifies most diseases.

Plotting the ROC curves of the binary classification model in **Figure 4.4** allowed for a better understanding of the class differentiation. Utilizing a range of threshold values derived from the likely outcomes of the Naïve Bayes classifier, a receiver operating characteristic (ROC) curve is utilized to display the true class rate compared to the false class rate. **Figure 4.4** provides evidence that the proposed model successfully classified the binary category.

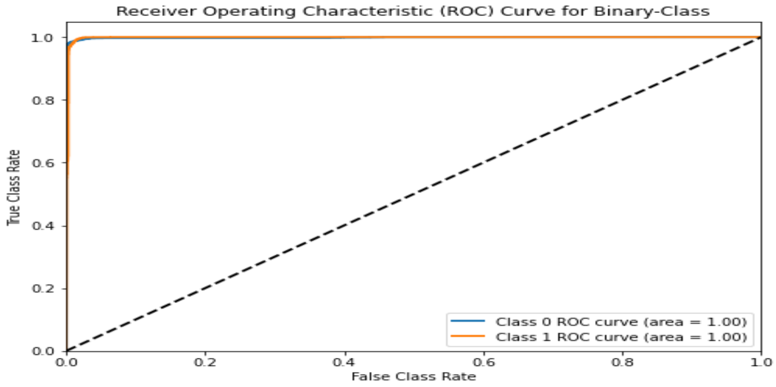


Figure 4.4: ROC curve for Naïve Bayes Classifier

Decision Tree Model Implementation

Table 4.5 presents a summary of the model's performance in binary classification, evaluating accuracy, precision, memory, and F1-score. The results highlight the model's effectiveness, achieving an impressive success rate of 98.59%.

Table 4-5: Classification report of the Decision Tree model

Classes	Precision	Recall	F1-score	Support
0	0.99	0.98	0.99	6232
1	0.98	1.00	0.99	7768
Avarage	0.99	0.99	0.99	14000
Accuracy:	98.89 %			

Confusion matrices were used to validate the results further to classify heart disease according to binary class, shown in **Figure 4.5**.

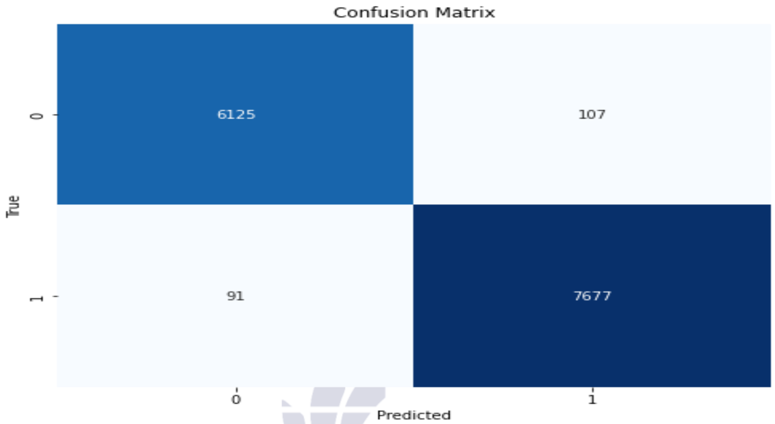


Figure 4.5: Confusion Matrix of Decision Tree Classifier

A total of 14000 test data points were chosen to evaluate the model. The model correctly predicts that 6125 corresponds to Class 0, while 7677 corresponds to Class 1. The confusion matrix result shows that the Decision Tree model correctly classifies most diseases.

Plotting the ROC curves of the binary classification model in **Figure 4.6** allowed for a better understanding of the class differentiation. Utilizing a range of threshold values that are derived from the likely outcomes of the Decision tree classifier, a receiver operating characteristic (ROC) curve is utilized to display the true class rate in comparison to the false class rate. **Figure 4.6** provides evidence that the proposed model successfully classified the binary category.

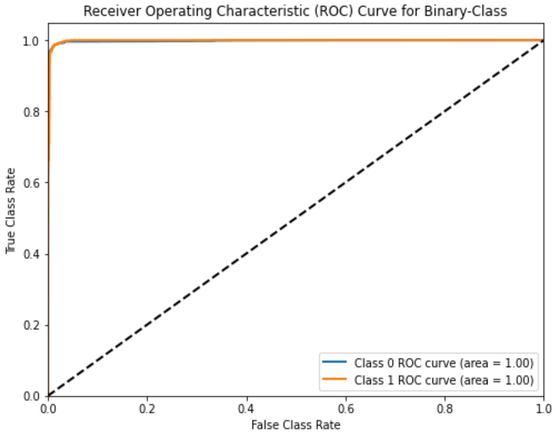


Figure 4.6: ROC Curve for Decision Tree Classifier

K-nearest neighbors Model Implementation

The results of the KNN classifier in terms of accuracy, precision, recall, and f1-score are summarized in **Table 4.6**. The table displays the findings from various evaluation matrices required to verify the model's performance. The model performs well in terms of 98.89% accuracy.

Table 4-6: Classification report of the KNN model

Classes	Precision	Recall	F1-score	Support
0	1.00	0.99	0.99	6232
1	0.99	1.00	0.99	7768
Avarage	0.99	0.99	0.99	14000
Accuracy	99.13%			

Confusion matrices were used to validate the results further to classify heart disease according to binary class, shown in **Figure 4.7.**

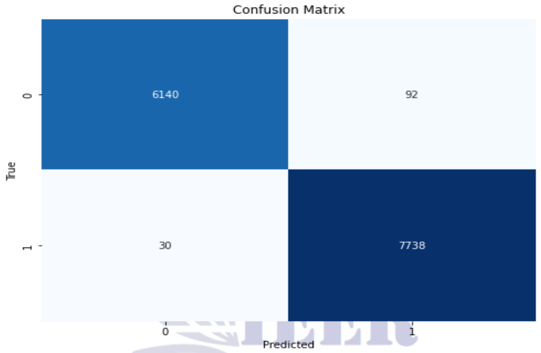


Figure 4.7: Confusion Matrix of KNN Classifier

A total of 14000 test data points were chosen to evaluate the model. The model correctly predicts that 6140 corresponds to Class 0, while 7738 corresponds to Class 1. The confusion matrix result shows that the Decision Tree model correctly classifies most diseases.

Plotting the ROC curves of the binary classification model in **Figure 4.8** allowed for a better understanding of the class differentiation. Utilizing a range of threshold values that are derived from the likely outcomes of the KNN classifier, a receiver operating characteristic (ROC) curve is utilized to display the true class rate in comparison to the false class rate. **Figure 4.8** provides evidence that the proposed model successfully classified the binary category.

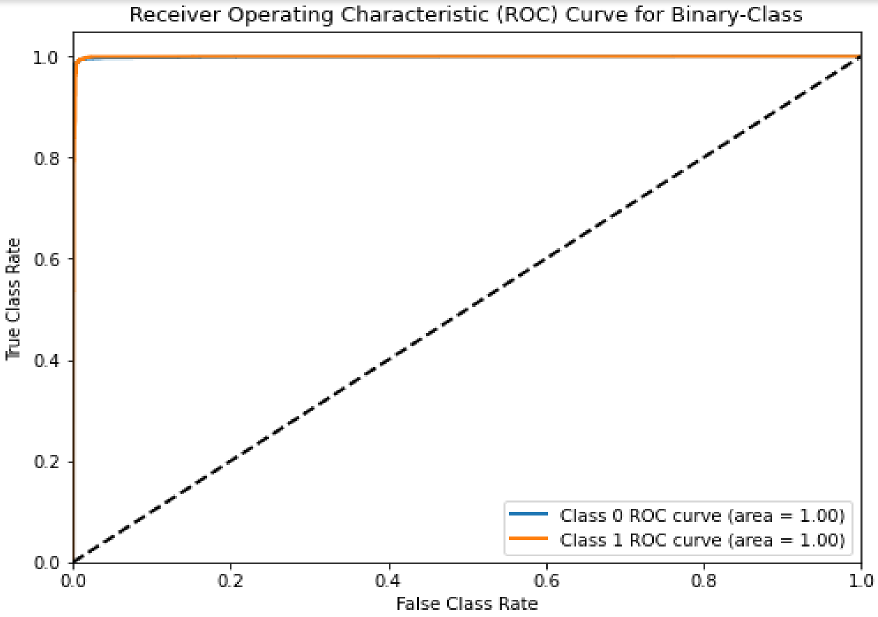


Figure 4.8: ROC Curve of KNN Classifier

XGBoost Classifier Implementation

The outcomes of the XGBoost classifier include factors such as accuracy, precision, recall, and f1-score, as shown in **Table 4.7**. The table presents the results obtained from the numerous evaluation matrices necessary to validate the model's performance. The model's accuracy is 98.09%, which is an outstanding performance.

Table 4-7: Classification report of the XGBoost model

Classes	Precision	Recall	F1-score	Support
0	0.96	1.00	0.98	6232
1	1.00	0.97	0.98	7768
Weighted Avg	0.98	0.98	0.98	14000
Accuracy	98.80			

Confusion matrices were utilized to validate the data further and classify heart disease according to binary class, as depicted in **Figure 4.9**.

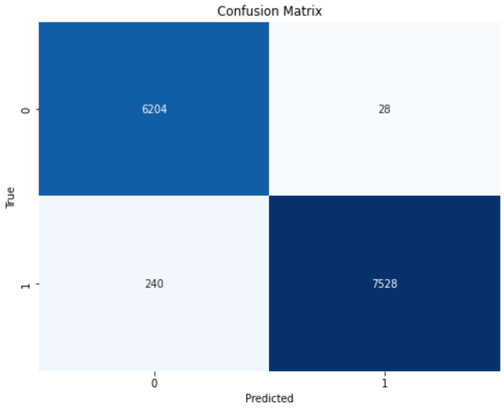


Figure 4.9: Confusion Matrix of XGBoost Classifier

The evaluation of the model uses a total of 14000 test data points. Following the model's predictions, the value 6204 is associated with Class 0, while the value 7528 is associated with Class 1. The outcome of the confusion matrix demonstrates that the XGBoost model makes accurate classifications of most diseases.

By plotting the ROC curves of the binary classification model with the help of **Figure 4.10**, it was possible to grasp the differentiation between the classes better. A receiver operating characteristic (ROC) curve is applied to represent the true class rate concerning the false class rate. This curve is created by utilizing a range of threshold values determined from the likely outcomes of the XGBoost classifier. **Figure 4.10** shows that the suggested model successfully classified the binary category using the binary classification system.

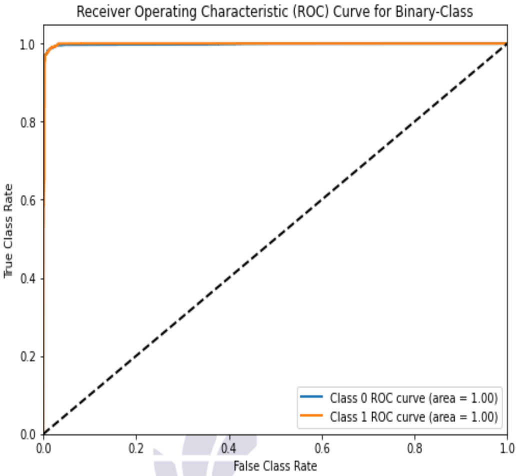


Figure 4.10: ROC Curve of XGBoost Classifier

Random Forest Classifier Implementation

The results of the Random Forest classifier in terms of accuracy, precision, recall, and f1-score are summarized in **Table 4.8**. The table displays the findings from various evaluation matrices required to verify the model's performance. The model performs well in terms of 98.89% accuracy.

Table 4-8: Classification report of the Random Forest model

	Precision	Recall	F1-score	Support
0	0.99	1.00	0.99	6232
1	1.00	1.00	1.00	7768
Weighted Avg	1.00	1.00	1.00	14000
Accuracy	99.55			

Confusion matrices were utilized to validate the data further and classify heart disease according to binary class, as depicted in **Figure 4.11**.

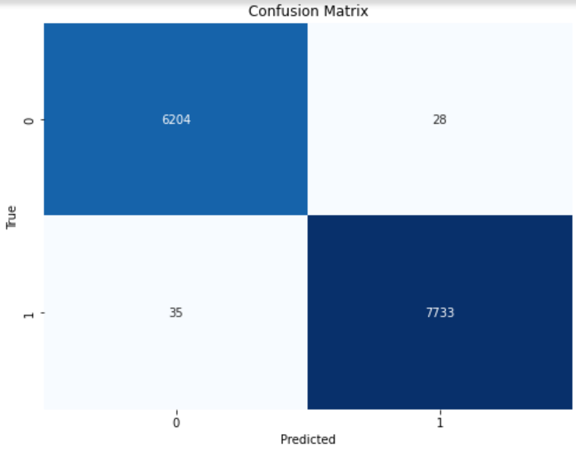


Figure 4.11: Confusion Matrix of Random Forest Classifier

It was feasible to understand better the class separation by visualizing the binary classification model's ROC curves with the aid of Figure 4.12. The true class rate with the false class rate is represented by a receiver operating characteristic (ROC) curve. The Random Forest classifiers' expected outcomes are used to determine a range of threshold values that are then used to create this curve. Figure 4.12 proves that the proposed model successfully used the binary classification technique to categorize the binary category.

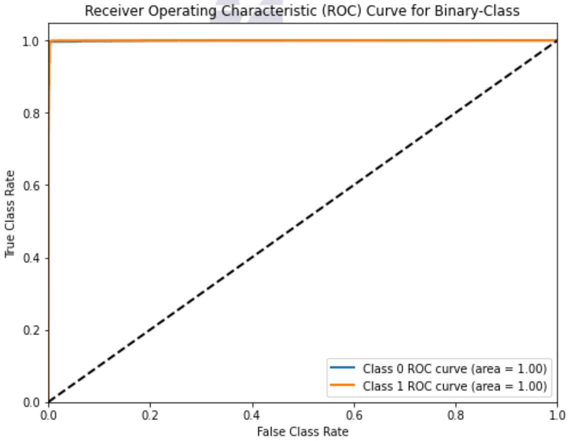


Figure 4.12: ROC curve of Random Forest Classifier

Discussion

Comparison with Similar Studies

Comparing the findings of one study with those of other studies is the most significant aspect of research since they validate and clarify the study's findings. It is essential to draw comparisons and contrasts between the findings of related studies to recognize patterns, highlight areas where there are gaps in knowledge, and identify areas that require more exploration. An extensive amount of studies have been carried out on intrusion detection systems using a variety of measurements, and the proposed work is the most appropriate concerning these studies.

Comparison with Benchmarks

The effectiveness of the provided model was analyzed compared to an existing academic study. It is possible to directly compare the suggested model and other studies comparable to the performance evaluation findings, which are displayed in **Table 4.9**. The following section provides a comprehensive summary of the analysis of the comparison.

Table 4.9: Comparison of the proposed model with existing studies

Works	Model	Accuracy(%)	Precision(%)	Recall(%)	F1 Score(%)
Existing Studies (Kaggleheart disease dataset (70,000 patients, 12 attributes))					
[8]	RF	87.05	89.42	83.43	86.32
	DT	86.37	89.58	81.61	85.42
	XGBoost	86.87	88.93	83.57	86.16
[58]	KNN	70.00	70.00	73.00	71.00
	NB	70.00	76.00	80.00	71.00
[75]	KNN	73.6	66.6	73.5	69.9
	DT	74.8	67.4	76.2	71.5
	RF	74.4	67.3	76.6	71.2
	XGB Boost	73.6	73.56	75.95	71.7
[71]	RF	95	89.42	85.91	86.32
	Decision	94	90.10	81.17	85.42
	Multilayer Perceptron	95	89.03	84.85	86.71
	XGBoost	95	89.62	82.11	86.30
Proposed Model	KNN	99.13	99	99	99
	DT	98.89	99	99	99
	RF	99.55	100	100	100
	XGB Boost	98.09	98	98	98
	NB	98.99	99	99	99

The strength of the presented model was quite attractive, and it was able to predict evidence of having heart disease in a specific individual by utilizing KNN, RF, NB, XGBoost, and DT, which showed good accuracy in contrast to the classifiers that were previously employed and other similar methods. The utilization of the provided model in determining the probability of the classifier correctly and accurately identifying the heart condition has resulted in releasing considerable pressure, which has been a significant relief. By completing this experiment, substantial knowledge has been gained that can assist us in predicting people who are suffering from heart disease.From the study, the evaluation of various models, including Decision Tree, Random Forest, Naïve Bayes, KNN, and XGBoost, highlights their high accuracy rates, with Random Forest achieving 99.55% accuracy.One possible reason for Random Forest's superior performance is its ensemble nature, which reduces overfitting by aggregating multiple decision trees. Additionally, Random Forest assigns equal weight to each tree, leading to more stable predictions. However, solely relying on accuracy without delving into misclassification patterns can be misleading. Metrics like precision, recall, and F1-score, along with confusion matrix analysis, were utilized to see the overall pictures of the models.

5. Conclusion and Future Work

This study employed five machine-learning classification algorithms — Random Forest, Naïve Bayes, Decision Tree, XGBoost, and K-Nearest Neighbors — to develop a predictive model for heart disease detection. By utilizing patients' clinical histories, including features such as blood sugar, blood pressure, and other relevant indicators, the proposed models effectively predict the likelihood of heart disease. This approach supports both patients and medical professionals in making more informed, data-driven decisions and could contribute to reducing heart disease-related fatalities.

To further improve data balance, the dataset from previous work was extended using the Synthetic Minority Oversampling Technique (SMOTE), increasing the sample size from 70,000 to 100,000. This adjustment yielded a minor but meaningful improvement in performance, particularly observed in the XGBoost classifier compared to earlier studies. The models achieved high predictive accuracy: 99.13% for KNN, 98.80% for XGBoost, 98.89% for Decision Tree, 98.99% for Naïve Bayes, and 99.55% for Random Forest, demonstrating their superiority over traditional approaches. Among these, Random Forest showed the highest accuracy of 99.55%, making it the most promising model.

For future work, this research can be extended by incorporating larger and more diverse datasets from multiple sources, testing on real-time data streams, and exploring advanced ensemble or deep learning methods to further enhance performance and generalizability.

5.1 Limitation

Despite the encouraging results, this study has several limitations that must be acknowledged. First, the analysis was performed on a single dataset, which may limit the generalizability of the findings to other populations or patient groups. Second, the study considered only a limited range of clinical and demographic variables, potentially overlooking other important factors. Finally, it did not account for additional contributors to heart disease risk, such as lifestyle changes or genetic predispositions, which could further impact prediction outcomes.

5.2 Future Work

In the future, interdisciplinary collaborations with medical professionals could be explored to refine the model's real-world usability. By integrating insights from healthcare practitioners, the predictive framework can be tailored to meet clinical needs more effectively. This collaboration would enable the development of a decision-support system that aligns with established medical protocols and enhances interpretability for clinicians. Additionally, integrating machine learning models into wearable health-monitoring devices, such as smartwatches or fitness trackers, could provide continuous, real-time monitoring of patients at risk of heart disease. These devices could detect early warning signs and alert both patients and physicians, ensuring timely medical intervention. Future research could also focus on testing the model across diverse datasets from various healthcare institutions to validate its generalizability and reliability. Moreover, incorporating explainable AI techniques would improve transparency, allowing medical professionals to understand and trust the model's decision-making process. These advancements would bridge the gap between machine learning and clinical practice, ultimately improving patient outcomes and reducing the burden on healthcare systems.

Reference

[1] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," *IEEE Access*, vol. 8, no. August, pp. 133034–133050, 2020, doi: 10.1109/ACCESS.2020.3010511.

[2] WHO, "Cardiovascular Diseases (Cvds)," 2020.

[3] Paracha, W. T., Nasir, J. A., Farhan, M., Paracha, M. F. K., & Iqbal, M. J. (2022). Optimizing heart disease prediction with machine learning and feature selection techniques. *International Journal of Computational Intelligence in Control*, *14*(1), 538. ISSN 0974-8571..

[4] A. Kumar, "The impact of obesity on cardiovascular disease risk factor," *Asian J. Med. Sci.*, vol. 10, no. 1, pp. 1–12, 2018, doi: 10.3126/ajms.v10i1.21294.

[5] M. Farhan *et al.*, "Network Intrusion Detection by using a Sequential Deep Neural Network with an Extra Tree Classifier," *Int. J. Comput. Intell. Control Copyrights @Muk Publ.*, vol. 14, no. 1, pp. 434–446, 2022.

[6] M. Awad and S. Fraihat, "Recursive Feature Elimination with Cross-Validation with Decision Tree: Feature Selection Method for Machine Learning-Based Intrusion Detection Systems," *J. Sens. Actuator Networks*, vol. 12, no. 5, 2023, doi: 10.3390/jsan12050067.

[7] A. Hameed and N. Z. Bawany, "Network intrusion detection using oversampling technique and machine learning algorithms," *PeerJ Comput. Sci. 8e820*, pp. 1–19, 2022, doi: 10.7717/peerj-cs.820.

[8] P. Singh and I. S. Virk, "Heart Disease Prediction Using Machine Learning Techniques," *2023 Int. Conf. Artif. Intell. Smart Commun. AISC 2023*, pp. 999–1005, 2023, doi: 10.1109/AISC56616.2023.10085584.

[9] A. Narin, Y. Isler, and M. Ozer, "Early prediction of Paroxysmal Atrial Fibrillation using frequency domain measures of heart rate variability," pp. 1–4, 2017, doi: 10.1109/tiptekno.2016.7863110.

[10] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," *SN Comput. Sci.*, vol. 1, no. 6, pp. 1–6, 2020, doi: 10.1007/s42979-020-00365-y.

[11] K. Drożdż *et al.*, "Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: a machine learning approach," *Cardiovasc. Diabetol.*, vol. 21, no. 1, pp. 1–12, 2022, doi: 10.1186/s12933-022-01672-9.

[12] R. Sali, H. Shavandi, and M. Sadeghi, "A clinical decision support system based on support vector machine and binary particle swarm optimisation for cardiovascular disease diagnosis," *Int. J. Data Min. Bioinform.*, vol. 15, no. 4, pp. 312–327, 2016, doi: 10.1504/IJDMB.2016.078150.

[13] P. B. K. Janki Jayvant Dalvi, Sayali Mukund Khole, "Heart disease prediction using machine learning techniques: A survey," *Int. J. Eng. Technol.*, vol. 7, no. 2.8 Special Issue 8, pp. 684–687, 2018, doi: 10.14419/ijet.v7i2.8.10557.

[14] A. Garg, B. Sharma, and R. Khan, "Heart disease prediction using machine learning techniques," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, no. 1, 2021, doi: 10.1088/1757-899X/1022/1/012046.

[15] J. Bohacik and M. Zabovsky, "Nearest neighbor method using non-nested generalized exemplars in breast cancer diagnosis," *2017 IEEE 14th Int. Sci. Conf. Informatics, INFORMATICS 2017 - Proc.*, vol. 2018-Janua, no. February, pp. 40–44, 2017, doi: 10.1109/INFORMATICS.2017.8327219.

[16] F. S. Alotaibi, "Implementation of machine learning model to predict heart failure disease," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 6, pp. 261–268, 2019, doi: 10.14569/ijacsa.2019.0100637.

[17] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction," *Expert Syst. Appl.*, vol. 68, pp. 163–172, 2017,

doi: 10.1016/j.eswa.2016.10.020.

[18] M. & A. Hassoon, M, Kouhi, MS, Zomorodi-Moghadam, "No TitleRule optimization of boosted c5. 0 classification using genetic algorithm for liver disease prediction'," in *in 2017 International Conference on Computer and Applications (ICCA)*, 2017, pp. 299–305.

[19] M. Abdar, W. Książek, U. R. Acharya, R. S. Tan, V. Makarenkov, and P. Pławiak, "A new machine learning technique for an accurate diagnosis of coronary artery disease," *Comput. Methods Programs Biomed.*, vol. 179, no. July, 2019, doi: 10.1016/j.cmpb.2019.104992.

[20] K. H., J. H., and G. J., "Diagnosing Coronary Heart Disease using Ensemble Machine Learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 10, pp. 30–39, 2016, doi: 10.14569/ijacsa.2016.071004.

[21] A. Alarifi, A. Tolba, Z. Al-Makhadmeh, and W. Said, "A big data approach to sentiment analysis using greedy feature selection with cat swarm optimization-based long short-term memory neural networks," *J. Supercomput.*, vol. 76, no. 6, pp. 4414–4429, 2020, doi: 10.1007/s11227-018-2398-2.

[22] B. Jin, C. Che, Z. Liu, S. Zhang, X. Yin, and X. Wei, "Predicting the Risk of Heart Failure with EHR Sequential Data Modeling," *IEEE Access*, vol. 6, no. c, pp. 9256–9261, 2018, doi: 10.1109/ACCESS.2017.2789324.

[23] I. D. Mienye, Y. Sun, and Z. Wang, "Improved sparse autoencoder based artificial neural network approach for prediction of heart disease," *Informatics Med. Unlocked*, vol. 18, p. 100307, 2020, doi: 10.1016/j.imu.2020.100307.

[24] F. Babic, J. Olejar, Z. Vantova, and J. Paralic, "Predictive and descriptive analysis for heart disease diagnosis," *Proc. 2017 Fed. Conf. Comput. Sci. Inf. Syst. FedCSIS 2017*, vol. 11, pp. 155–163, 2017, doi: 10.15439/2017F219.

[25] C. Bemando, E. Miranda, and M. Aryuni, "Machine-Learning-Based Prediction Models of Coronary Heart Disease Using Naïve Bayes and Random Forest Algorithms," *Proc. - 2021 Int. Conf. Softw. Eng. Comput. Syst. 4th Int. Conf. Comput. Sci. Inf. Manag. ICSECS-ICOCSIM 2021*, no. August 2021, pp. 232–237, 2021, doi: 10.1109/ICSECS52883.2021.00049.

[26] I. Yekkala, D. S. Institutions, S. Dixit, D. S. Institutions, and J. Akhil, "Prediction of heart disease using Ensemble Learning and Particle Swarm Optimization," no. November 2018, 2017, doi: 10.1109/SmartTechCon.2017.8358460.

[27] N. Hasan and Y. Bao, "Comparing different feature selection algorithms for cardiovascular disease prediction," *Health Technol. (Berl).*, vol. 11, no. 1, pp. 49–62, 2021, doi: 10.1007/s12553-020-00499-2.

[28] A. C. Karthik, S., Santhosh, M., Kavitha, M. S., & Paul, "Automated Deep Learning Based Cardiovascular Disease Diagnosis Using ECG Signals," *Comput. Syst. Sci. Eng.*, vol. 42, no. 1, 2022.

[29] R. S. Patil and M. Gangwar, "Heart Disease Prediction Using Machine Learning and Data Analytics Approach," *Lect. Notes Networks Syst.*, vol. 435, no. 3, pp. 351–361, 2022, doi: 10.1007/978-981-19-0976-4_29.

[30] A. M. Diaa Salama AbdElminaam, Nada Mohamed , Hady Wael , Abdelrahman Khaled, "MLHeartDisPrediction: Heart Disease Prediction using Machine Learning," *J. Comput. Commun.*, vol. 2, no. 1, pp. 50–65.

[31] K. S. Archana, B. Sivakumar, R. Kuppusamy, Y. Teekaraman, and A. Radhakrishnan, "Automated Cardioailment Identification and Prevention by Hybrid Machine Learning Models," *Comput. Math. Methods Med.*, vol. 2022, 2022, doi: 10.1155/2022/9797844.

[32] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, no. 1, 2021, doi: 10.1088/1757-899X/1022/1/012072.

[33] F. I. Alarsan and M. Younes, "Analysis and classification of heart diseases using heartbeat features and machine learning algorithms," *J. Big Data*, 2019, doi: 10.1186/s40537-019-0244-x.

[34] M. Trigka, "Efficient Data-Driven Machine Learning Models for Cardiovascular Diseases Risk Prediction," no. January, 2023, doi: 10.3390/s23031161.

[35] K. M. Almustafa, "Prediction of heart disease and classifiers' sensitivity analysis," *BMC Bioinformatics*, vol. 21, no. 1, pp. 1–19, 2020, doi: 10.1186/s12859-020-03626-y.

[36] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[37] B. F. D. H, H. B. F. David, and S. A. Belcy, "HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES," no. November, 2018, doi: 10.21917/ijsc.2018.0254.

[38] M. B. Yildiz, E. T. Yasin, and M. Koklu, "A Detailed Analysis of Detecting Heart Diseases Using Artificial Intelligence Methods INTELLIGENT METHODS A Detailed Analysis of Detecting Heart Diseases Using Artificial Intelligence Methods," no. December, 2023, doi: 10.58190/imiens.2023.71.

[39] V. S. K. Reddy, P. Meghana, N. V. S. Reddy, and B. A. Rao, "Prediction on Cardiovascular disease using Decision tree and Naïve Bayes classifiers," *J. Phys. Conf. Ser.*, vol. 2161, no. 1, 2022, doi: 10.1088/1742-6596/2161/1/012015.

[40] C. and M. Methods in Medicine, "Retracted: Implementation of a Heart Disease Risk Prediction Model Using Machine Learning," *Comput. Math. Methods Med.*, vol. 2023, pp. 1–1, 2023, doi: 10.1155/2023/9764021.

[41] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Informatics Med. Unlocked*, vol. 20, no. July, p. 100402, 2020, doi: 10.1016/j.imu.2020.100402.

[42] M. D. A. Hossen, T. Tazin, S. Khan, and E. Alam, "Supervised Machine Learning-Based Cardiovascular Disease Analysis and Prediction," vol. 2021, no. Ml, 2021.

[43] H. Sun and J. Pan, "Heart Disease Prediction Using Machine Learning Algorithms with Self-Measurable Physical Condition Indicators," pp. 1–10, 2023, doi: 10.4236/jdaip.2023.111001.

[44] M. M. Nishat *et al.*, "A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset," vol. 2022, no. Cvd, 2022.

[45] N. Biswas *et al.*, "Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques," vol. 2023, 2023.

[46] I. Hossain *et al.*, "Heart disease prediction using distinct artificial intelligence techniques : performance analysis and comparison," *Iran J. Comput. Sci.*, vol. 6, no. 4, pp. 397–417, 2023, doi: 10.1007/s42044-023-00148-7.

C. Anwar, J. Iqbal, R. Irfan, S. Hussain, and A. D. Algarni, "Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers," 2022.

[49] S. K. D. Khandaker Mohammad Mohi Uddin , Rokaiya Ripa , Nilufar Yeasmin , Nitish Biswas, "Machine learning-based approach to the diagnosis of cardiovascular vascular disease using a combined dataset," *Intell. Med.*, vol. 7, 2023.

[50] C. S. Wu, M. Badshah, and V. Bhagwat, "Heart disease prediction using data mining techniques," *ACM Int. Conf. Proceeding Ser.*, pp. 7–11, 2019, doi: 10.1145/3352411.3352413.

[51] B. Bahrami and M. H. Shirvani, "Prediction and Diagnosis of Heart Disease by Data Mining Techniques," *J. Multidiscip. Eng. Sci. Technol.*, vol. 2, no. 2, pp. 3159–3199, 2015, [Online]. Available: www.jmest.org

[52] S. A. Hannan, A. V. Mane, R. R. Manza, and R. J. Ramteke, "Prediction of heart disease medical prescription using radial basis function," *2010 IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC 2010*, pp. 735–740, 2010, doi: 10.1109/ICCIC.2010.5705900.

[53] U. Nagavelli, D. Samanta, and P. Chakraborty, “Machine Learning Technology-Based Heart Disease Detection Models,” *J. Healthc. Eng.*, vol. 2022, 2022, doi: 10.1155/2022/7351061.

[54] S. Nashif, M. R. Raihan, M. R. Islam, and M. H. Imam, “Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System,” *World J. Eng. Technol.*, vol. 06, no. 04, pp. 854–873, 2018, doi: 10.4236/wjet.2018.64057.

[55] A. M. A. Barhoom, A. Almasri, B. S. Abu-nasser, and S. S. Abu-naser, “Prediction of Heart Disease Using a Collection of Machine and Deep Learning Algorithms,” vol. 6, no. 4, pp. 1–13, 2022.

[56] A. Alqahtani, S. Alsubai, M. Sha, L. Vilcekova, and T. Javed, “Cardiovascular Disease Detection using Ensemble Learning,” *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/5267498.

[57] F. Kanwal, M. K. Abid, M. S. Maqbool, N. Aslam, and M. Fuzail, “Optimized Classification of Cardiovascular Disease Using Machine Learning Paradigms,” *VFAST Trans. Softw. Eng.*, vol. 11, no. 2, pp. 140–148, 2023.

[58] J. Maiga, G. G. Hungilo, and Pranowo, “Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data,” *Proc. - 1st Int. Conf. Informatics, Multimedia, Cyber Inf. Syst. ICIMCIS 2019*, pp. 45–48, 2019, doi: 10.1109/ICIMCIS48181.2019.8985205.

[59] Purushottam, K. Saxena, and R. Sharma, “Efficient Heart Disease Prediction System,” *Procedia Comput. Sci.*, vol. 85, pp. 962–969, 2016, doi: 10.1016/j.procs.2016.05.288.

[60] C. S.Dangare and S. S. Apte, “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques,” *Int. J. Comput. Appl.*, vol. 47, no. 10, pp. 44–48, 2012, doi: 10.5120/7228-0076.

[61] M. Shouman and T. Leonard Turner, “Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients,” *Appl. k-Nearest Neighb. Diagnosing Hear. Dis. Patients*, vol. 2, no. 3, pp. 220–223, 2012.

[62] I. Mahmud, M. M. Kabir, M. F. Mridha, S. Alfarhood, M. Safran, and D. Che, “Cardiac Failure Forecasting Based on Clinical Data Using a Lightweight Machine Learning Metamodel,” *Diagnostics*, vol. 13, no. 15, pp. 1–18, 2023, doi: 10.3390/diagnostics13152540.

[63] N. S. Mansur Huang, Z. Ibrahim, and N. Mat Diah, “Machine Learning Techniques for Heart Failure Prediction,” *Malaysian J. Comput.*, vol. 6, no. 2, p. 872, 2021, doi: 10.24191/mjoc.v6i2.13708.

[64] K. Ahirwar, V. Yadav, A. Tiwari, M. S. Pandey, M. K. Ahirwar, and R. K. Ranjan, “A Hybrid Optimized Model for Predicting and Analyzing Heart Attacks Using Machine Learning,” vol. XXV, pp. 5–12, 2024.

[65] I. Journal and O. F. Science, “Heart Disease Prediction Using CNN with Various Feature Selection,” vol. 2030, no. 1, pp. 3960–3968, 2024.

[66] D. Chicco and G. Jurman, “Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone,” vol. 5, pp. 1–16, 2020.

[67] D. K. Plati *et al.*, “A machine learning approach for chronic heart failure diagnosis,” *Diagnostics*, vol. 11, no. 10, pp. 1–15, 2021, doi: 10.3390/diagnostics11101863.

[68] S. Saikia *et al.*, “Analysis of Heart Disease Prediction using Novel Machine Learning and Deep Learning Techniques,” pp. 2114–2125, 2024.

[69] J. Kamwele and A. Kipkorir, “Exploring the Role of Dimensionality Reduction in Enhancing Machine Learning Algorithm Performance,” vol. 17, no. 5, pp. 157–166, 2024, doi: 10.9734/AJRCOS/2024/v17i5445.

[70] G. Logabiraman, D. Ganesh, M. S. Kumar, and A. V. Kumar, “Heart disease prediction learning algorithms using machine,” vol. 01122, 2024.

[71] C. M. Bhatt and T. G. and P. L. M. , Parth Patel, "Effective Heart Disease Prediction Using Machine Learning Techniques," *Algorithms*, 2023.

[72] E. Kokori *et al.*, "Machine learning in predicting heart failure survival: a review of current models and future prospects," *Heart Fail. Rev.*, 2024, doi: 10.1007/s10741-024-10474-y.

[73] M. Alshraideh, N. Alshraideh, A. Alshraideh, Y. Alkayed, Y. Al Trabsheh, and B. Alshraideh, "Enhancing Heart Attack Prediction with Machine Learning : A Study at Jordan University Hospital," vol. 2024, 2024, doi: 10.1155/2024/5080332.

[74] H. El-sofany, B. Bouallegue, and Y. M. A. El-latif, "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," pp. 1–18, 2024.

[75] V. Shorewala, "Early detection of coronary heart disease using ensemble techniques," *Informatics Med. Unlocked*, vol. 26, p. 100655, 2021, doi: 10.1016/j.imu.2021.100655.