



Improved Heart Disease Diagnosis: Classification and Feature Selection Through Leveraging Machine Learning Algorithms

Kalaivani B¹ Ranichitra A²

¹Research Scholar, Department of Computer Science, Sri S.Ramasamy Naidu Memorial College (Affiliated to Madurai Kamaraj University, Madurai), Sattur- 626 203, Tamilnadu, India

²Assistant Professor, Department of Computer Science, Sri S.Ramasamy Naidu Memorial College, Sattur-626 203, Tamilnadu, India

Abstract: Heart disease remains a leading cause of death globally, underscoring the urgent demand for precise and affordable diagnostic solutions. Given the complexity of predicting cardiac disease, advanced data mining techniques are essential for enhancing prediction accuracy. This work introduces a novel classification approach that identifies key predictive features, helping to reduce diagnostic costs without compromising accuracy. The methodology emphasizes robust data preprocessing, incorporating outlier detection through z-score analysis and effective feature selection, both of which significantly boost data quality and model reliability. Exploratory Data Analysis (EDA) is also instrumental in uncovering patterns and relationships between patient attributes and heart disease, supporting well-informed decisions during model development. The Random Forest classifier outperformed the other machine learning methods assessed, attaining an accuracy of 90.21%, which increased to 96.51% following feature selection. This outcome highlights the model's capacity to handle complex datasets and underscores the importance of eliminating low-correlation features. The primary objective of this work is to enhance heart disease detection by integrating rigorous preprocessing techniques with powerful machine learning algorithms.

Keywords: Classification, EDA, Feature Ranking (FR), Machine Learning, Z –Score.

I. Introduction

The World Health Organization (WHO) identifies heart disease as the leading causes of death globally, responsible for approximately 17.7 million deaths (31%) in 2015, with numbers projected to rise beyond 20 million deaths annually [1]. A substantial portion of the world's population succumbs to heart disease each year [2]. Cardiovascular disease (CVD) is a broad term that encompasses conditions affecting the heart and blood vessels, including coronary heart disease (CHD), also known as coronary artery disease (CAD), which is influenced by lifestyle choices and age. CHD occurs when fatty deposits, such as cholesterol, calcium, and cellular waste, build up in the coronary arteries, causing atherosclerosis and narrowing the arteries. This reduces blood flow to the heart, potentially leading to angina (chest pain) or a heart attack if an artery becomes completely blocked. Key risk factors include obesity, diabetes, smoking, high blood pressure, high cholesterol, and inactivity. Heart attacks, or myocardial infarctions, occur when a blood clot obstructs a coronary artery, reducing blood flow to the heart muscle and causing chest pain [3]. Angina, a symptom of reduced blood flow, often manifests as chest pain radiating to the left arm and shoulder. Modifiable risk factors include smoking, high blood pressure, high cholesterol, obesity, poor diet, diabetes, depression, and stress, while non-modifiable risks include age, gender, genetics, race, and ethnicity [4]. Heart failure, or congestive heart failure, often arises from conditions like diabetes, high blood pressure, and heart disease, and occurs when the heart is unable to pump blood effectively.

As heart disease becomes increasingly prevalent and often presents without clear symptoms, early diagnosis is vital. Regular monitoring of key health indicators such as body temperature, respiratory rate, pulse, blood pressure, glucose levels, ECG, EEG and EMG, is essential for effective disease management [5, 6]. Healthcare facilities generate vast amounts of data, yet much of it remains underutilized [7]. Integrating clinical decision support

systems with electronic health records has the potential to improve patient outcomes, reduce inconsistencies in care and also reduce the possibility of medical errors [8, 9]. This work aims to determine which specific features are critical for predicting cardiovascular disease and to develop a highly accurate predictive model using the Cleveland and Statlog heart disease datasets from the UCI repository [10].

Machine learning classifiers are essential for improving prediction accuracy in complex datasets. The Random Forest (RF) classifier was employed due to its demonstrated ability to enhance prediction accuracy. Combining effective models with feature extraction techniques to enhance performance demonstrates its efficacy in healthcare applications.

This work attempts to achieve the following main objectives:

1. **Collect the Data:** Improve diagnosis accuracy by utilizing the Cleveland and Statlog heart disease databases.
2. **Data Preprocessing:** Exploratory Data Analysis (EDA) and Z-score analysis were performed to identify significant features and address outliers, respectively. This enhanced data quality and mitigated the impact of skewed values.
3. **Balanced Dataset:** The dataset was balanced to ensure equal representation of patients with and without heart disease. This minimized class bias and improved the accuracy of model predictions.
4. **Feature Importance:** Random Forest facilitated the identification of the most influential features for heart disease prediction through feature importance analysis. This knowledge can be valuable for developing targeted prevention and intervention strategies.
5. **Improved Prediction:** The Random Forest model, combined with effective feature selection and data preparation techniques, demonstrated reliable performance in predicting heart disease.

The primary objective is to develop a heart disease prediction model that is both efficient and cost-effective, while ensuring high accuracy through optimized data processing and feature selection methods. Figure 1 illustrates the framework of the proposed model.

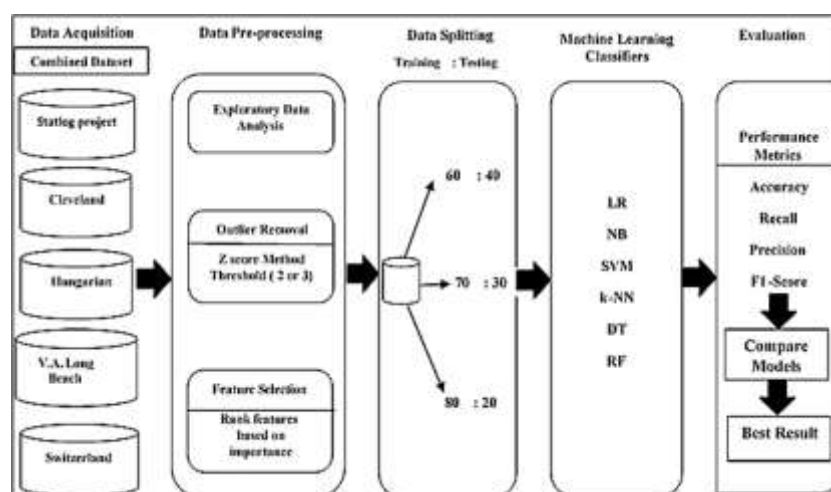


Figure 1 : The framework of the Proposed Model

This work is organized as follows: The literature review in the field of these research applications is summarized in Section II. Section III describes the methodology used in the proposed work. In Section IV, the comparative assessment of many classifiers is thoroughly examined, along with the performance details, results, and proposed



research related to heart disease. In Section V, the work results are analysed and potential directions for future research are discussed.

II. Literature Review

Cardiovascular disease (CVD) encompasses a range of conditions that result from disruptions in the heart's ability to pump blood and supply essential oxygen and nutrients to the body. These conditions include coronary artery disease (CAD), arrhythmias, and various structural heart abnormalities. Accurate diagnosis and prediction are crucial for effective treatment and management of these disorders [11]. This section reviews studies on feature selection and classification techniques that are used to predict heart disease, highlighting their importance in improving diagnostic accuracy and patient outcomes.

Sentilkumar et al. [12] suggest a methodology called the Hybrid Random Forest Linear Model (HRFLM), which uses machine learning techniques to improve the accuracy of CVD prediction.

This approach emphasizing the integration of multiple algorithms and feature selection to enhance model accuracy. It demonstrates high performance in evaluating classification metrics such as error rates, precision, F1 score, sensitivity, and specificity. The authors conclude that the HRFLM technique improves the accuracy of cardiac disease predictions; however, the primary limitation is the lack of a real-world dataset for validation.

Cai et al. [13] conducted to study cardiac arrhythmias and 12-lead ECGs. A one-dimensional deep densely connected neural network was utilized in order to identify atrial fibrillation. An accuracy of 99.35%, a sensitivity of 99.19%, and a specificity of 99.44% were demonstrated by the test dataset findings.

Magesh and Swarnalatha et al. [14] conducted an analysis of cardiovascular disease focusing on the rural population. They identified several risk factors and illnesses associated with smoking-related coronary disease. To examine the goal level distribution from samples and find pertinent characteristics, the authors employed entropy. In predicting cardiac illness, they discovered that utilizing Random Forest in conjunction with cluster-based decision tree learning produced an accuracy of 89.30%, as opposed to 76.70% without cluster-based learning.

Miled et al. [15] examined electronic medical records (EMRs), including prescriptions, notes, and diagnoses. They identified and predicted instances of dementia and non-dementia using machine learning techniques. They obtained a 77.43% accuracy rate by using Random Forest algorithms on three EMR datasets. This section examines a number of research on feature selection and classification techniques used to forecast heart disease, emphasizing how important a precise diagnosis is to maintaining cardiovascular health. Although there are still issues with dataset availability and practical application, the findings show that machine learning models have the potential to enhance heart disease prediction.

III. Methodology

A. Data Preprocessing -EDA Analysis

Exploratory Data Analysis (EDA), a vital initial stage in data analysis, including using statistical and visual techniques to investigate a dataset's fundamental properties [16]. The machine learning model's primary goals are to identify patterns, discover errors, validate assumptions, and make connections between data. In many fields, including healthcare, this preliminary study is crucial for uncovering hidden patterns, predicting trends, and obtaining insights [17].

Data is graphically represented using graphical approaches, which facilitates the detection of underlying structures and abnormalities. Multivariate strategies look at correlations between several variables at once, whereas univariate approaches concentrate on analysing the characteristics of a single variable. A typical component of this all-encompassing method is bivariate analysis, which examines the connection between two variables;

however, analysis involving three or more variables may also be carried out as required [18]. Before moving on to more intricate interactions, univariate study of individual components is frequently advised.

The key objectives of this phase include identifying errors, validating assumptions, selecting appropriate models for further analysis, and understanding the relationships between explanatory variables and outcomes. In conclusion, this thorough analysis ensures that the data is properly prepared for advanced analytical techniques, making it suitable for developing accurate and reliable prediction models.

B. Z score Analysis

Outlier identification plays a crucial role in minimizing noise and detecting data points that deviate significantly from the rest of the dataset. This technique is essential for uncovering valuable insights and hidden patterns in various fields such as healthcare, finance, and engineering. Detecting and removing outlier's leads to more accurate and reliable analyses. One of the most common methods for outlier detection is the Z-score test [19], which calculates the Z-score for each data point to determine its deviation from the mean in terms of standard deviations. Values that fall outside a predefined threshold—typically two or three standard deviations from the mean—are considered outliers. Data points exceeding a threshold of 3 or -3 standard deviations are flagged as outliers and removed from the dataset to prevent distortion of the analysis [20].

Outliers can be detected and removed using the formula in equation (1),

$$Z \text{ score} = \frac{(x-\bar{x})}{\sigma} \quad \text{----- (1)}$$

Where, x represents Standardized random variable, $\bar{x}$ denotes the Mean, σ indicates the Standard deviation

Removing outliers based on z-score analysis helps in refining data quality, which can lead to more accurate and reliable results in subsequent analyses. This method is widely employed across various fields to uncover valuable insights and improve the robustness of data-driven models.

C. Algorithms for Data Mining and Machine Learning

The practice of analysing data to uncover new and useful information is known as data mining. The judgment of physicians might be greatly enhanced by this process, which could also become a crucial part of advanced healthcare systems. Among the techniques used in data mining include regression, association analysis, classification, clustering, and outlier detection. Machine learning is one of the main areas of expertise in the interdisciplinary field of data mining [21]. The following section provides an explanation of the machine learning methods used to classify and predict heart disease:

1) Logistic Regression

Logistic regression is a technique for calculating the probability of a categorical outcome based on independent inputs. It creates probabilities between 0 and 1 by altering the output of a linear model using the logistic (sigmoid) function. As a result, it may be applied to binary classification tasks, such calculating the likelihood that an event will occur or identifying a patient's illness based on a number of risk factors [22].

2) Naive Bayes

The Naive Bayes (NB) classifier, often known as the "independent feature model," is a probabilistic classification technique that relies on the Bayes theorem. The underlying assumption of this approach is that the features of categorization are conditionally independent of one other. Because the method evaluates each characteristic separately to determine if a data point is likely to be in a certain class, this "naïve" assumption simplifies the computation. The Naive Bayes algorithm is widely used in supervised learning and performs particularly well in high-dimensional data contexts because of its computational efficiency and ability to handle a large number of

characteristics. Because it uses historical knowledge and observed data to calculate the chance of a particular outcome, it is a useful tool for predictive modelling [23].

3) Support Vector Machine

Support Vector Machine (SVM), a supervised machine learning method, is used for both classification and regression. The SVM approach aims to find the best line or decision boundary that may separate n-dimensional space into classes, simplifying the process of classifying new data points in the future. One term for this ideal decision boundary is a hyperplane [24]. The block diagram of the SVM is explained in Figure 2.

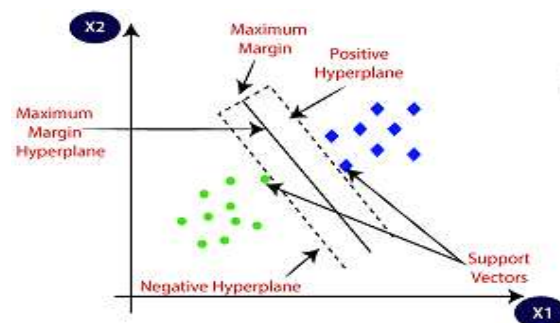


Figure 2 : A block Diagram of Support Vector Machine

4) K-Nearest Neighbors

The K-Nearest Neighbors (K-NN) technique is a straightforward memory-based classifier that compares incoming data points to the nearest previous samples to identify their category. Using Euclidean distance to determine the k closest neighbors and majority voting for classification, it makes predictions using the complete dataset. While K-NN is simple to use and converges rapidly, it is susceptible to noise and irrelevant information, necessitating attribute normalization and careful parameter k adjustment for best results [25].

5) Decision Tree

A decision tree classifier organizes data items into a tree structure to build a model. In this model, the leaves represent class labels, while the branches indicate the attributes that differentiate these labels. The model is trained using the X-train and Y-train parameters, and its accuracy is evaluated with the X-test and Y-test parameters through the score() function. The decision tree shows the decision-making process visually, with the root node at the top and the leaves showing the ultimate results [26]. The block diagram for DT is shown in Figure 3.

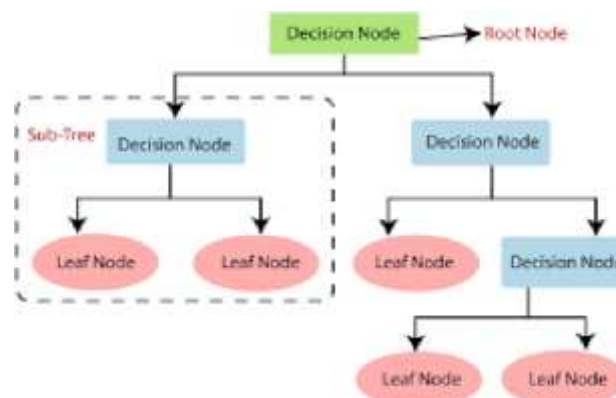


Figure 3: A Block Diagram of Decision Tree

6) Random Forest

A supervised classification method called the Random Forest algorithm builds a "forest" out of many decision trees. Each tree in the forest offers a prediction, and the final classification is decided by the majority vote of all the trees. By using different parts of the training data for each tree, this ensemble method increases the accuracy of predictions [27]. This method improves model flexibility and reduces overfitting by selecting random subsets of data and constructing decision trees from them. A specified number of trees (N) are built to create a forest, with repeated selection and tree creation to enhance diversity and stability. Finally, predictions for new data are made through majority voting, which minimizes individual tree bias and results in more accurate and reliable predictions.

The process can be outlined in the following steps:

- Step 1: From the training set, choose K data points at random.

Step 2: Build decision trees using the chosen subsets of data.

Step 3: Establish how many N should be used to build the decision trees.

Step 4: Go through Steps 1 and 2 again.

Step 5: Using the majority vote, categorize the new data points after obtaining predictions from each decision tree.

Data mining plays a vital role in refining clinical decision-making and extracting meaningful patterns from vast databases. It enables healthcare professionals to identify trends, make informed decisions, and improve patient outcomes by leveraging hidden information within the data. In healthcare, this process not only aids in identifying useful information but also supports advanced medical systems by enabling more informed and effective decision-making.

D. Feature Selection Approaches

The large volumes of data generated from multiple sources are now crucial for collecting, storing, retrieving, and sharing information, but they remain challenging to interpret and analyze effectively [28]. The vast amount of data and the increasing cost of diagnosis necessitate feature selection, which enhances the model's accuracy and improves illness prediction outcomes. Algorithms for feature ranking aim to evaluate each feature's importance independently of supervised or unsupervised learning methods. Each feature is given a score by these algorithms, and only these scores are used to select the features [29]. The original dataset consists of numerous features, some of which may not be relevant and therefore must be eliminated during data preparation [30]. Consequently, feature selection is an essential part of preprocessing data as it eliminates unnecessary features and improves the classification model's performance. Numerous feature selection techniques have been applied in heart disease prediction studies, including wrapper approaches, filters, embedded and ensemble methods, and hybrid algorithms.

E. Tools for Measuring Performance Evaluation

Performance metrics play a crucial role in evaluating the effectiveness and accuracy of machine learning models. In the context of heart disease prediction, positive classification refers to the correct identification of a person as having heart disease (HD), while negative classification refers to incorrectly labelling someone as not having HD. Several key metrics are used to assess model performance, which are calculated using equations (1) to (4) as outlined in references [31] and [32].

These metrics include True Positive (TP), which measures the number of correctly identified heart disease patients, and True Negative (TN), which represents the number of correctly identified non-heart disease patients. On the other hand, False Negative (FN) occurs when the model wrongly classifies a heart disease patient as healthy, and False Positive (FP) occurs when a non-heart disease patient is incorrectly classified as having heart disease.

$$Accuracy (Acc) = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad Accuracy (Acc) = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad \text{----- (1)}$$

$$Precision = \frac{(TP)}{(TP + FP)} \quad Precision = \frac{(TP)}{(TP + FP)} \quad \text{----- (2)}$$

$$Recall \text{ or Sensitivity } (Sen) = \frac{(TP)}{(TP + FN)} \quad Recall \text{ or Sensitivity } (Sen) = \frac{(TP)}{(TP + FN)} \quad \text{----- (3)}$$

$$F1 - score = \frac{2(Precision \times Recall)}{(Precision + Recall)} \quad F1 - score = \frac{2(Precision \times Recall)}{(Precision + Recall)} \quad \text{----- (4)}$$

IV. Result And Discussion

A. Data source

The five datasets included in this study are those from Statlog, V.A. Medical, Hungarian, Switzerland, and Cleveland that deal with cardiac disease. The UCI Machine Learning Repository is open to the public [10]. These datasets were merged into a single dataset in order to enhance model performance. 1190 instances that show the presence or absence of heart disease across 14 parameters are included in the merged dataset. Table 1 displays the characteristics of the individual and combined datasets for cardiac disease [33].

Table 1. Features of the five and combined cardiac disease datasets

Collection of Datasets	No of Instances	No of Attributes	Missing Values
Statlog project	270	12	No
Cleveland	303	12	No
Hungarian	294	12	Yes
V.A. Long Beach	200	12	Yes
Switzerland	123	12	Yes
Combined Dataset	1190	12	Missing values replaced with Mode Value

B. Description of the Attributes

The heart disease dataset consists of twelve attributes, each providing unique information related to the condition [34]. The main result that our research aims to forecast is the 'target' variable, which is one of these. It allows one to determine whether a patient has cardiac disease. The remaining 11 characteristics: These input features are what our prediction model is built and trained with. They offer the information required to forecast the "target" characteristic. These attributes comprise a variety of health-related measurements and indications that may

influence a patient's diagnosis of heart disease. For a comprehensive understanding of each attribute and its role in the dataset, refer to Table 2, which provides detailed descriptions of all 12 attributes.

Table 2. Description of attributes is derived from five distinct datasets [33]

S.no	Attribute	Description	Values/Range
1	Age	Describes the patient's age.	Between 29 and 77
2	Sex	Describes the gender of the patient.	1 = male; 0 = female
3	Chest Pain type (CP)	Describes the level of chest pain (CP).	Level of cp value 1: typical angina Level of cp Value 2: atypical angina Level of cp Value 3: non typical angina Level of cp Value 4: asymptomatic
4	Resting bps	Describes the blood pressure (BP) of the patient	Between 94 and 200
5	cholesterol	Describes the cholesterol level recorded while admitting the patient in the hospital.	Between 126 and 564
6	Fasting Blood Sugar	Describes the fasting blood sugar level of the patient. It has binary classified values.	1 if yes, 0 if not
7	Resting ECG	Describes the ECG from 0 to 2. Where each value is showing the severity of the pain.	0, 1, 2
8	Max Heart Rate	Describes the maximum value of heartbeat.	Between 71 and 202
9	Exercise Angina	Describes whether there is exercise-induced angina or not.	1 for yes, 0 for not
10	Old Peak	Describes the patient's depression status.	Between 0 and 6.2
11	ST Slope	Describes the condition of the patient during an exercise stress test.	Value 1: up-sloping Value 2: flat Value 3: down-sloping
12	Target	Describes the number of categories (classes) indicating patients with heart disease or not.	1 for yes, 0 for not

C. Data Preprocessing

A key component of getting the dataset ready for modelling is exploratory data analysis (EDA), which finds patterns, identifies outliers, and comprehends the connections between variables. Features like cholesterol and resting blood pressure in the heart disease dataset showed notable outliers, with cholesterol levels ranging from 0

to 603. It was also verified that the dataset was balanced, with 47% of instances representing normal people (561 cases) and 53% representing patients with heart disease (628 cases). A prediction model that is more accurate and dependable is made possible by this balanced distribution, which lowers the possibility of class bias. The figure 4 displays the overview of the heart disease patients in a dataset. Resolving these outliers is crucial to enhancing model accuracy and avoiding distorted findings.

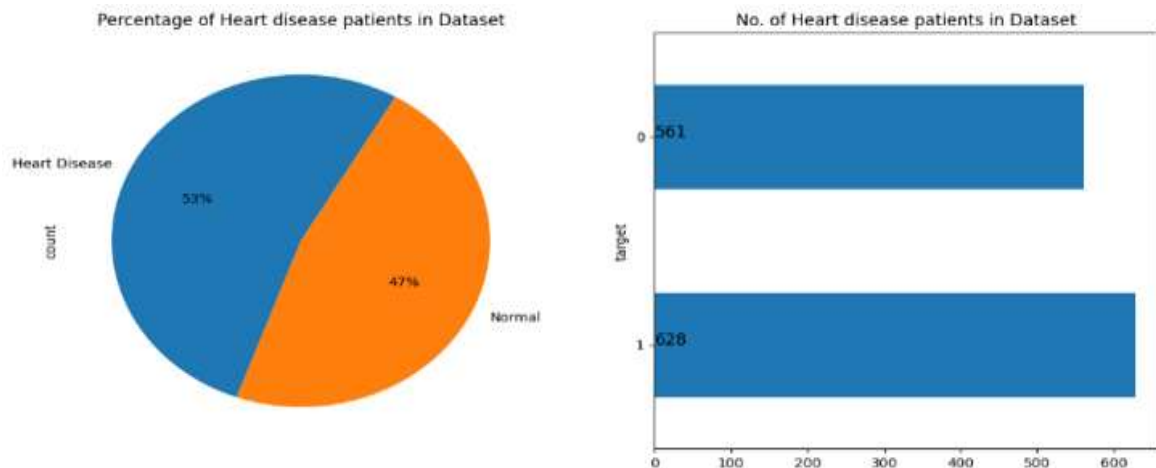


Figure 4. Heart Disease Overview

The significance of blood pressure, cholesterol, and age as major risk factors was brought to light by demographic insights, which also emphasised the necessity of careful preprocessing procedures including feature selection and outlier treatment. The almost equal representation of classes ensures that the model can discriminate between individuals with and without cardiac disease. Figure 5 displays the distribution of healthy patients' ages and genders (those without heart disease).

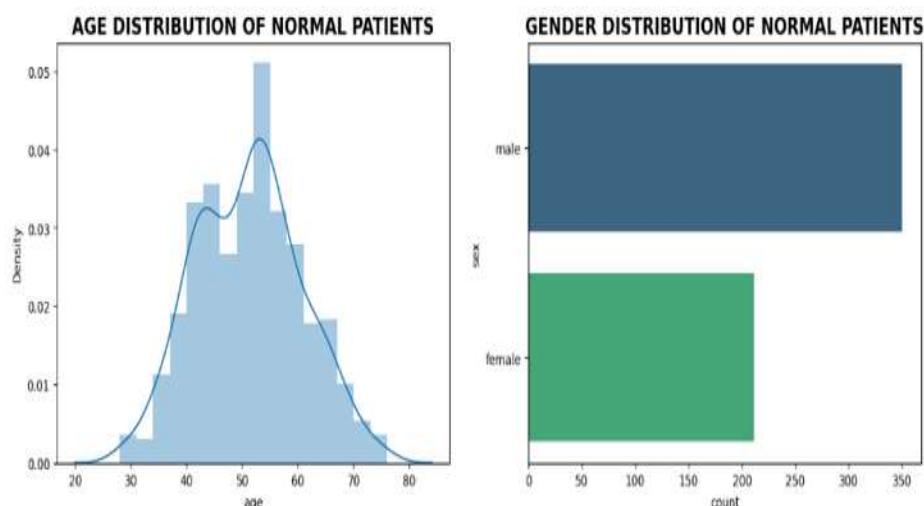


Figure 5. Age and Gender Distribution of Normal Patients

According to the left histogram, the majority of healthy patients are between the ages of 40 and 60, with a balanced distribution between 20 and 80. The majority of patients are middle-aged, as the density curve shows. In the typical patient group, there is more male (350) than female (200), as seen in the right bar chart. The dataset's gender distribution indicates that men, labelled as '0,' are more represented than women, labelled as '1'.

This demographic study is crucial because it aids model building by providing insights into the makeup of the data, guarantees a fair representation of various subgroups, and helps in identifying potential biases in the dataset.

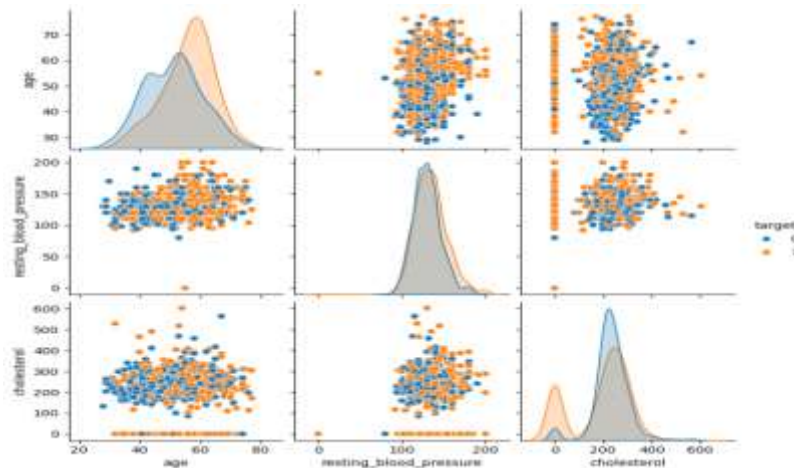


Figure 6. Classification of Target value based on the 3 attributes

Figure 6 shows that age, resting blood pressure, and cholesterol do not significantly impact heart rate. The associations between these characteristics and the target variable, likely heart disease are graphically depicted by the pair plot. With normal blood pressure and cholesterol readings, the variables are distributed around middle age. Age, resting blood pressure, and cholesterol all show positive associations, which may indicate a connection to heart disease. These connections provide crucial information on how these variables affect the risk of heart disease. The pair plot offers a thorough synopsis of the data, which helps identify factors linked to heart disease and directs future studies and preventative initiatives.

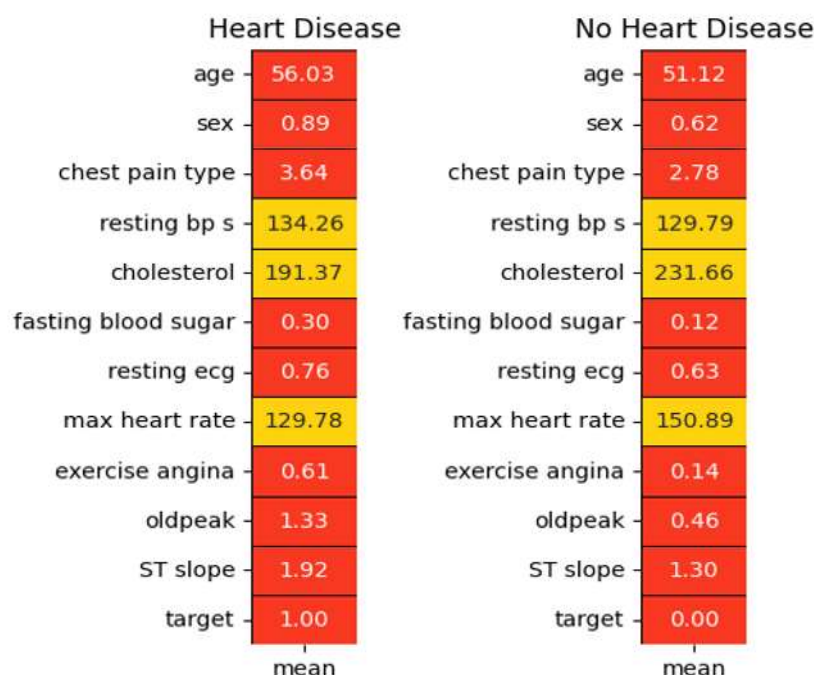


Figure 7. Heart Disease Risk Factors: A Comparative Analysis



The figure 7 presents a comparison of heart disease patients and those without heart disease based on twelfth attributes. Patients with cardiac disease had a somewhat higher average age (56.03) than people without the condition (51.12). Chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, oldpeak, and ST slope are also greater in heart disease patients, indicating a relationship with these variables [35]. However, they have lower values for sex, max heart rate, and exercise angina, indicating potential protective effects. Confirming the anticipated categorization, the target variable, which is likely a measure of heart disease presence or absence, is 1.00 for individuals with heart disease and 0.00 for those without. Conversely, they have lower values for sex, max heart rate, and exercise angina, indicating potential protective effects. The target variable confirms the correct classification of heart disease status, providing a solid foundation for further research and prevention strategies.

Finally, it is helpful to determine the key heart disease risk factors by examining characteristics such as age, cholesterol, and resting blood pressure. These results have a direct impact on model development. As seen by the almost equal representation of patients with and without heart disease, the balanced sample reduces class bias. This results in more accurate and dependable model predictions. Demographic parameters like age and gender distribution can be analysed to gain a deeper knowledge of how these factors may affect the risk of disease. Selecting the most significant predictors is further aided by finding connections between characteristics such as blood pressure, cholesterol, and age. This comprehensive analysis ensures that the model built from this dataset is robust, as it uses well understood data that has undergone careful outlier treatment and feature selection, ultimately improving predictive performance and enhancing the validity of the results.

D. Z score Analysis of Dataset

There are 1190 entries in the collection. A record with a st_slope of 0 is removed during preprocessing in order to confirm the distribution of categorical values. One target variable is left for analysis out of 1189 records and 11 characteristics. The dataset, which had 12 features and 1189 cases from the heart disease dataset, was subjected to a similar process. In the dataset, 937 cases, or around 80% of the instances, were utilised as training sets, while 235 cases, or 20% of the total, were used as test sets. After removing the 17 outlier instances, a total of 1172 occurrences could be further analysed. First, Table 3 displays the Z-score analysis comparison for the Cleveland datasets.

Table 3. The total Instances calculate the Z-Score Outlier Detection Instances

Dataset			Z score Detection			
	Original Instances	Thresh old	Outlier Removal	Z score Analysis	Training (80%)	Testing (20%)
Combined (Cleveland Dataset)	1189	3	17	1172	937	235
		2	290	899	719	180

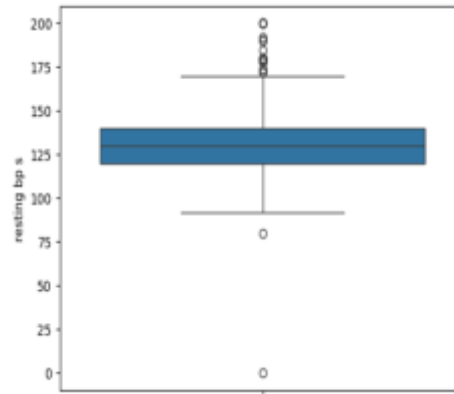


Figure 8. Before Z-Score Outlier Removal of resting BPS

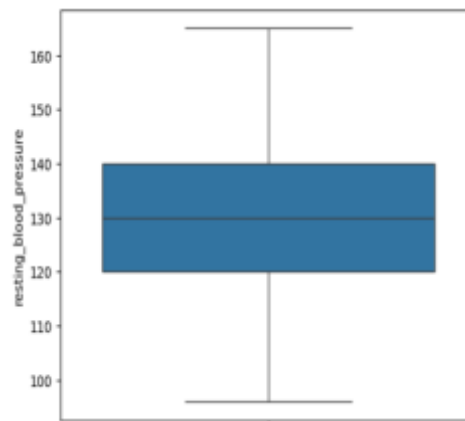


Figure 9. After Z-Score Outlier Removal of resting BPS

Figures 8 and 9 illustrate the effects of applying the Z-score removal method for outlier detection in blood pressure data. In Figure 8, representing "resting bps", several outliers were initially observed, indicating extreme values. Through Z-score analysis, these outliers identified by their high positive or negative Z-scores were removed, leaving only values within the normal range. In contrast, Figure 9, representing "resting blood pressure," showed no outliers even before applying the Z-score method, confirming that the data distribution was already normal and did not require further adjustments.

Table 4. Distribution of target Variables

Dataset	Distribution of Target Variables		
	Class	Training Instances	Testing Instances
Combined (Cleveland Dataset)	0	446	112
	1	491	123
Total		937	235



Table 4 presents the distribution of target variables following the 80:20 data split. To reduce bias in model training, the training set is balanced, containing 491 instances of class '1' and 446 instances of class '0'. Likewise, the test set ensures that the model is assessed on a proportionate distribution by including 123 cases of class '1' and 112 instances of class '0'. In order to ensure feature space consistency, the test set has 235 samples with the same 12 features as the training set, which consists of 937 samples with 12 features. This congruence between the training and test sets reduces the likelihood of overfitting or underfitting, which facilitates model performance evaluation. The model's capacity to generalise well and increase its accuracy in forecasting unknown data is supported by the balanced distribution between the two sets.

E. Performance on Combined Dataset

On a composite dataset that incorporates data from five heart disease datasets, this work applied different classification techniques. To effectively assess the performance of each classification model, the Numpy and Pandas libraries were utilized for data manipulation and evaluation. The combined dataset, consisting of 12 key features, was used to ensure comprehensive analysis across all models. In Table 5 and figure 10, the results of the performance evaluation are presented, with each classification method's accuracy compared. The Random Forest classifier showed the greatest predicted accuracy of the five algorithms used, with an astounding 90.21%. The usefulness of Random Forest in managing the intricacies of heart disease data is demonstrated by this exceptional result.

Table 5. Comparison of Classification Algorithms Based on Performance Metrics with 12 features

Algorithm ms	Performance Measurements						
	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	mathew_correlation
LR	0.80851	0.78261	0.87805	0.73214	0.82759	0.8051	0.61803
NB	0.81702	0.80769	0.85366	0.77679	0.83004	0.81522	0.61803
SVM	0.82553	0.80147	0.88618	0.75893	0.8417	0.82255	0.61803
k-NN	0.80851	0.78677	0.86992	0.74107	0.82626	0.8055	0.61803
DT	0.82979	0.81679	0.86992	0.78571	0.84252	0.82782	0.61803
RF	0.90213	0.87313	0.95122	0.84821	0.91051	0.89972	0.80655

LR: Logistic Regression; NB: Naive Bayes; SVM: Support vector machine; k-NN: k- nearest neighbors; DT: Decision Tree; RF: Random Forest

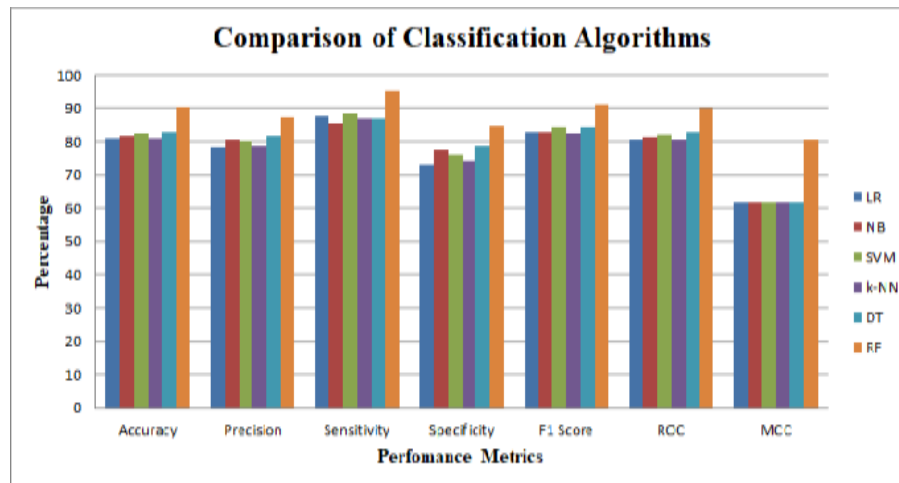


Figure 10. Comparison of Classification Algorithms based on Performance Metrics

F. Performance on Feature Selection Attributes

The correlation matrix in figure 11 visually represents the relationships between different factors in a heart disease dataset. Each cell shows how closely two factors are connected. A red cell means the factors increase or decrease together, while a blue cell means they move in opposite directions. Interestingly, the types of chest pain and exercise-induced angina are strongly linked to heart disease. Remarkably, a higher maximum heart rate appears linked to a reduced risk of heart disease. Other factors, such as resting blood pressure and age also show some correlation with the likelihood of developing heart-related disorders. These findings suggest that factors with a strong positive or negative relationship to heart disease are crucial for predicting it. Factors with a weak or no connection might not be as important and could be removed to simplify the model without losing accuracy. The analysis also revealed significant correlations between age, resting blood pressure, chest pain, and exercise-induced angina.

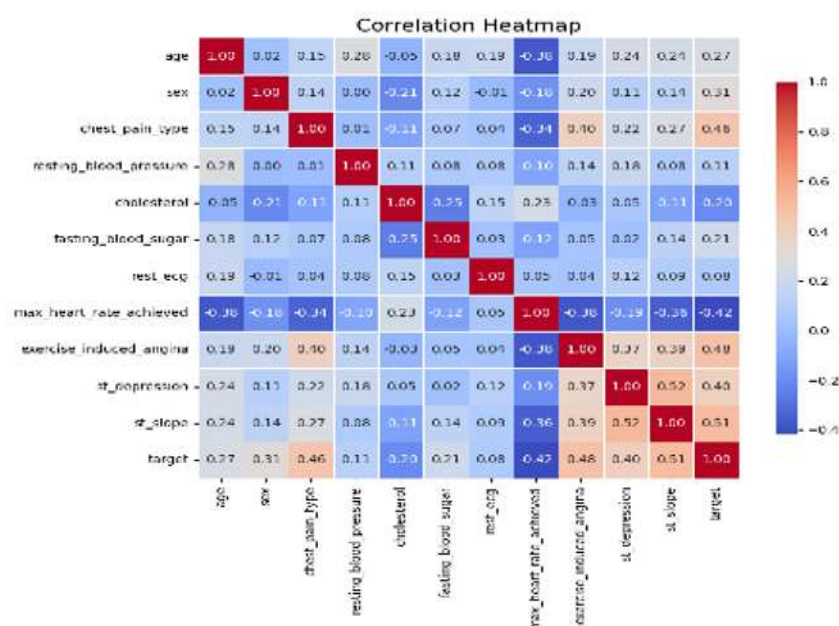
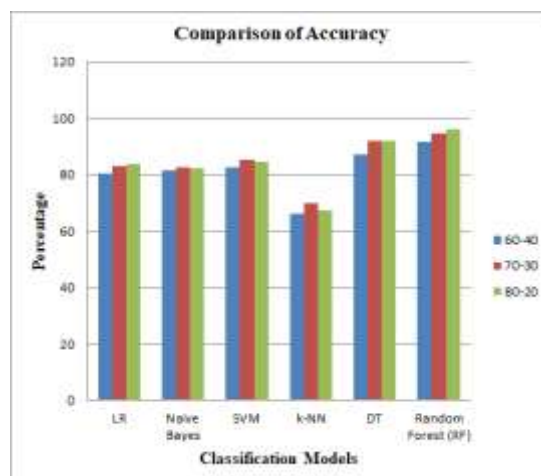


Figure 11. Correlation relationships visualized for the combined dataset

Table 6. Performance comparisons between various percentage splits of the combined dataset using threshold-based outlier removal (T=2)

(T=2)	60-40	70-30	80-20									
Algorithms	Accuracy	Recall	Precision	F1-Score	Accuracy	Recall	Precision	F1-Score	Accuracy	Recall	Precision	F1-Score
LR	80.33	81.71	77.46	79.53	82.96	82.82	81.54	82.17	83.89	86.25	79.31	82.63
NB	81.39	81.39	79.77	80.47	82.59	82.59	83.08	82.13	82.22	82.22	79.31	81.18
SVM	82.78	81.36	83.24	82.29	85.19	84.09	85.38	84.73	84.44	83.91	83.91	83.91
k-NN	66.11	67.35	57.23	61.88	70.0	72.07	61.54	66.39	67.22	67.5	62.07	64.67
DT	87.22	86.29	87.28	86.78	91.85	92.86	90.0	91.41	92.22	95.06	88.51	91.67
RF	91.67	89.5	93.64	91.53	94.81	93.28	96.15	94.7	96.11	96.51	95.4	95.95

T: Threshold; LR: Logistic Regression; NB: Naive Bayes; SVM: Support vector machine; k-NN: k- nearest neighbors; DT: Decision Tree; RF: Random Forest



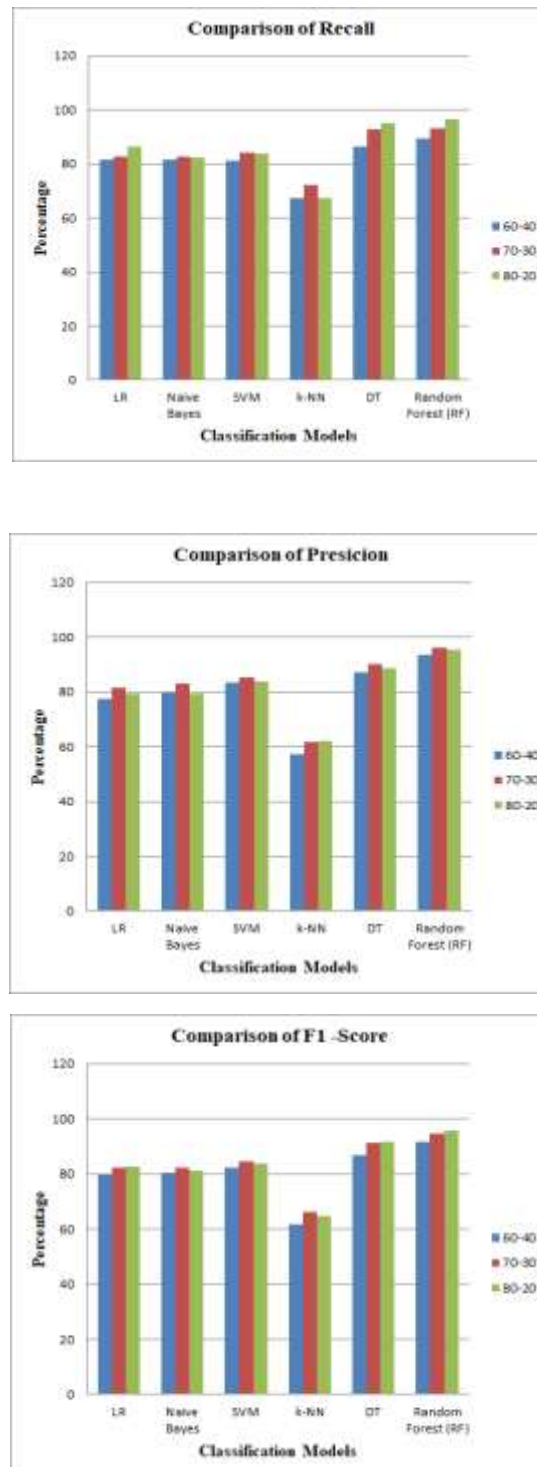
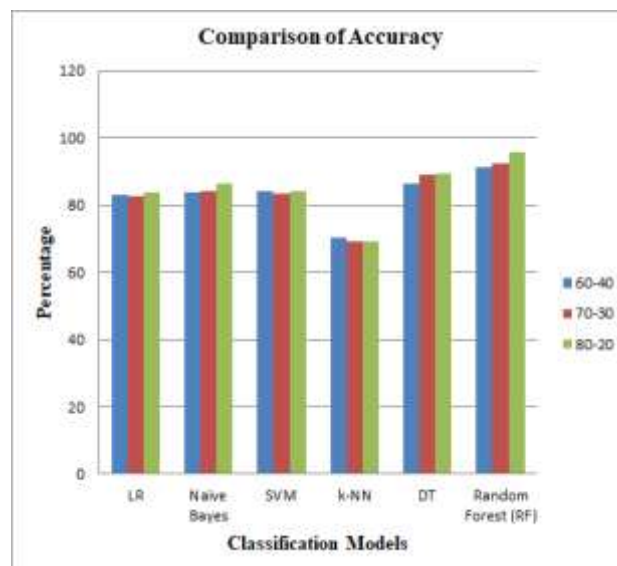


Figure 12. Performance Comparison of Classification Models across Different Train-Test Splits with Threshold (T=2)

Table 7. The performance comparisons of various percentage splits of combined dataset using threshold-based outlier removal (T=3)

(T=3) Algorithms	60-40				70-30				80-20			
	Accuracy	Recall	Precision	F1-Score	Accuracy	Recall	Precision	F1-Score	Accuracy	Recall	Precision	F1-Score
LR	82.98	82.3	84.39	83.33	82.67	82.58	83.05	82.82	83.83	85.22	82.35	83.76
NB	83.83	83.83	83.54	83.99	84.09	84.09	83.62	84.09	86.38	86.38	85.71	86.44
SVM	84.04	83.47	85.23	84.34	83.24	83.15	83.62	83.38	84.26	85.34	83.19	84.26
k-NN	70.43	69.76	73	71.34	69.03	68.68	70.62	69.64	69.36	67.67	75.63	71.43
DT	86.38	84.74	89.03	86.83	88.92	87.1	91.53	89.26	89.36	90.52	88.24	89.36
RF	91.28	88.28	95.36	91.68	92.33	89.47	96.05	92.64	95.74	94.31	97.48	95.87

T: Threshold; LR: Logistic Regression; NB: Naive Bayes; SVM: Support vector machine; k-NN: k- nearest neighbors; DT: Decision Tree; RF: Random Forest



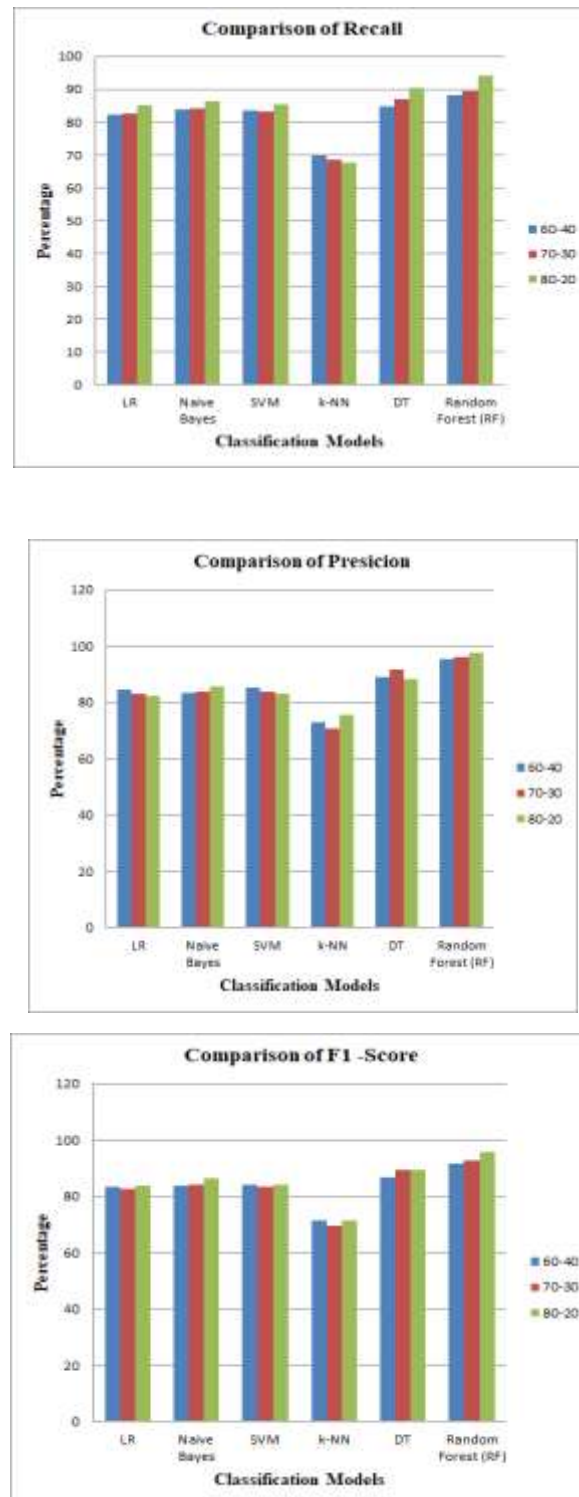


Figure 13. Performance Comparison of Classification Models across Different Train-Test Splits with Threshold (T=3)



Finding outliers in a dataset is mostly dependent on the Z-score threshold, and the selection of threshold (2 or 3) has a big impact on which data points are marked as outliers. According to the studies in Tables 6, Figure 12, and Table 7, Figure 13, a Z-score cut-off of 2 is more effective than 3 because it finds and removes more outliers, creating a more efficient dataset that may enhance model performance. Consequently, the combined dataset had been cleared of characteristics like "age," "resting blood pressure," "cholesterol," and "max heart rate reached."

In models such as Random Forest or Gradient Boosting, feature importance measures the contribution of each feature to the predictive power of the model. It is determined by calculating the improvement in prediction performance when a feature is used for decision-making. In Random Forest, this is often quantified by metrics such as the mean decrease in accuracy or reduction in Gini impurity when a feature is excluded. For instance, in a heart disease dataset, a Random Forest classifier can be applied to rank features based on their importance, identifying which features most significantly reduce uncertainty (impurity) across decision trees. This ranking helps in focusing on the most relevant features, improving model efficiency, interpretability, and performance. By selecting and evaluating the top features, dimensionality is reduced, and the model becomes more accurate by concentrating on the features that have the highest predictive power.

Table 8. The ranking of heart disease risk factors according to the Rank Feature Selection Algorithm [36]

Feature	Importance	Rank
ST slope	0.194539	1
Chest pain type	0.134759	2
Oldpeak	0.116515	3
Max heart rate	0.116115	4
Cholesterol	0.111434	5
Age	0.094971	6
Resting BP	0.078512	7
Exercise angina	0.068288	8
Sex	0.03824	9
Resting ECG	0.027517	10
Fasting blood sugar	0.01911	11

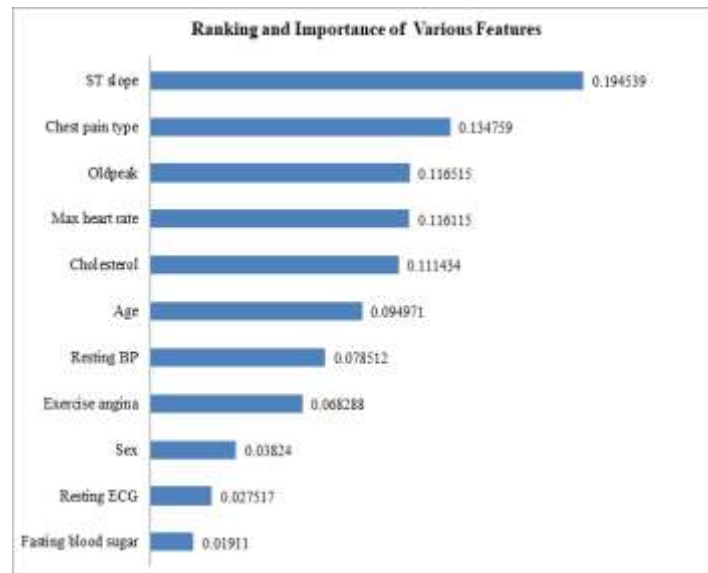


Figure 14. Ranking and Importance of Various Features

Feature ranking is a systematic process that evaluates and prioritizes the importance of individual variables in predicting outcomes, such as heart disease in this case. Each feature's contribution to the predictive capability of the model is indicated by the rankings that are produced from the feature significance scores. The order of importance is crucial for further work because it guides future analyses and model refinement. After applying the several feature selection techniques, it was discovered that 11 features were consistently chosen using the Feature Ranking (FR) method. 'ST slope,' 'chest pain type,' 'oldpeak' and 'max heart rate' were determined to be the most important traits in predicting heart disease, and they were ranked in order of importance [37]. Table 8 demonstrates that the characteristics chosen using the FR technique had a comparable ranking order, indicating their importance in regression analysis. The figure 14 shows the ranking and importance of various features in predicting heart disease. ST slope, chest pain type and oldpeak were identified as the most significant factors, followed by maximum heart rate, cholesterol, and age. [38] These results emphasize the importance of these features in evaluating an individual's heart disease risk.

Additionally, understanding the importance of these attributes can lead to improved feature selection techniques in subsequent modelling phases, optimizing predictive accuracy and interpretability. [39] In the end, improved decision-making and resource allocation in cardiovascular health research are fostered by a clear knowledge of feature value.

V. Conclusion and Future Work

This work explored the effectiveness of machine learning techniques for heart disease prediction, using a combined dataset from five publicly available sources with 1,190 cases and 12 attributes representing various patient characteristics. The research underscores the crucial need for accurate heart disease detection, stressing that reliable prediction is key to addressing this significant global health concern.

To ensure data quality, rigorous preprocessing steps, including outlier removal and feature selection, were employed. Exploratory Data Analysis (EDA) offered key insights into feature distributions and important relationships among risk factors like cholesterol levels and resting blood pressure, strengthening model reliability. Performance assessments across multiple classification algorithms highlighted the superiority of the Random Forest classifier, which reached an outstanding predictive accuracy of 90.21%, outperforming other evaluated models. This success underscores the model's capability to navigate the complexities of heart disease data and



emphasizes the significance of advanced data mining techniques in clinical decision-making processes. [40] Notably, the application of feature selection further improved the model's performance, yielding a maximum accuracy of 96.51% after eliminating four features with low correlation to the target variable. Additionally, the balanced representation of patients with and without heart disease minimized class bias, contributing to the robustness of the model. Overall, these findings illustrate how combining thorough preprocessing with effective machine learning algorithms can significantly enhance diagnostic accuracy in heart disease prediction, paving the way for more reliable clinical applications. However, it is important to acknowledge limitations, such as the potential for biases within the original datasets and the need for further validation on independent datasets.

Future research directions could involve incorporating additional features, exploring more sophisticated models, and applying these methods in a clinical setting.

References

- [1] Bhatnagar, P., Wickramasinghe, K., Williams, J., Rayner, M., and Townsend, N. (2015). The Epidemiology of Cardiovascular Disease in the UK 2014. *Heart*, 101, 1182-1189.
- [2] Yadav, D.C. and Pal, S.A.U.R.A.B.H., 2020. Prediction of heart disease using feature selection and random forest ensemble method. *International Journal of Pharmaceutical Research*, 12(4), pp.56-66.
- [3] Thiagaraj, M., and Suseendran, G. (2017). Survey on Heart Disease Prediction System Based on Data Mining Techniques. *Indian Journal of Innovations and Developments*, 6(1), 1-9.
- [4] <https://www.heartfoundation.org.au/your-heart/heart-conditions>.
- [5] Raghupathi, W., and Raghupathi, V. (2014). Big Data Analytics in Healthcare: Promise and Potential. *Health Information Science and Systems*, 2, 3.
- [6] Palanisamy, V. and Thirunavukarasu, R. (2017). Implications of Big Data Analytics in Developing Healthcare Frameworks – A Review. *Journal of King Saud University–Computer and Information Sciences*, 1-11, in press.
- [7] Ghasemi, M. and Amyot, D. (2016). Process Mining In Healthcare: A Systematised Literature Review. *Int. J. of Electronic Healthcare*, 9(1), 60-88.
- [8] Dominic, Vinitha., Deepa, Gupta. and Sangita, Khare. (2015.) An Effective Performance Analysis of Machine Learning Techniques for Cardiovascular Disease. *Applied Medical Informatics*, 36(1), 23-32.
- [9] Chandralekha, M. and Shenbagavadivu, N. (2018). Performance Analysis of Various Machine Learning Techniques to Predict Cardiovascular Disease: An Empirical Study. *Appl. Math. Inf. Sci.*, 12(1), 217-226.
- [10] UCI Machine Learning Repository: <http://archive.ics.uci.edu/ml/datasets/Heart+Disease> , <http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29>.
- [11] Pushpavathi, T.P., Kumari, S. and Kubra, N.K., 2021, January. Heart Failure Prediction by Feature Ranking Analysis in Machine Learning. In 2021 6th International Conference on Inventive Computation Technologies (ICICT) (pp. 915-923). IEEE.
- [12] Mohan, S., Thirumalai, C. and Srivastava, G., 2019. Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, 7, pp.81542-81554.
- [13] Cai, W., Chen, Y., Guo, J., Han, B., Shi, Y., Ji, L., ... & Luo, J. (2020). Accurate detection of atrial fibrillation from 12-lead ECG using deep neural network. *Computers in biology and medicine*, 116, 103378.



-
- [14] Magesh, G., & Swarnalatha, P. (2020). Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction. *Evolutionary Intelligence*, 1-11.
- [15] Miled, Z. B., Haas, K., Black, C. M., Khandker, R. K., Chandrasekaran, V., Lipton, R., & Boustani, M. A. (2020). Predicting dementia with routine care EMR data. *Artificial Intelligence in Medicine*, 102, 101771.
- [16] Komorowski M, Marshall D. C , J, Saliciccioli J D and Crutain Y, (2016). Chapter 15- Exploratory Data Analysis - Secondary Analysis of Electronic Health Records. DOI: 10.1007/978-3-319-43742-2_15.
- [17] Valdiviezo-Diaz, P., Reátegui, R., Barba-Guaman, L., Ortega, M. (2022). Exploratory Data Analysis on Cervical Cancer Diseases. In: Botto-Tobar, M., Montes León, S., Torres-Carrión, P., Zambrano Vizuetes, M., Durakovic, B. (eds) *Applied Technologies. ICAT 2021. Communications in Computer and Information Science*, vol 1535. Springer, Cham. https://doi.org/10.1007/978-3-031-03884-6_32.
- [18] Huang, CW., Lu, R., Iqbal, U. et al. (2015). A richly interactive exploratory data analysis and visualization tool using electronic medical records. *BMC Med Inform Decis Mak* 15, 92. <https://doi.org/10.1186/s12911-015-0218-7>.
- [19] Rahmayanti, N., Pradani, H., Pahlawan, M. and Vinarti, R.: Comparison of machine learning algorithms to classify fetal health using cardiotocogram data. *Procedia Computer Science*, 197, pp.162-171 (2022).
- [20] Rishika, R., A. and Suresh, K. P. (2016). Predictive Big Data Analytics in Healthcare. Second International Conference on Computational Intelligence & Communication Technology, IEEE.
- [21] Neesha, Jothi., Nur, Aini., Abdul, Rashid. And Wahidah, Husain. (2015). Data Mining in Healthcare—A Review. *Procedia Computer Science*, 72, 306-313.
- [22] Khanna, D., Sahu, R., Baths, V. and Deshpande, B., 2015. Comparative study of classification techniques (SVM, logistic regression and neural networks) to predict the prevalence of heart disease. *International Journal of Machine Learning and Computing*, 5(5), p.414.
- [23] Repaka, A.N., Ravikanti, S.D. and Franklin, R.G., 2019, April. Design and implementing heart disease prediction using naives Bayesian. In 2019 3rd International conference on trends in electronics and informatics (ICOEI) (pp. 292-297). IEEE.
- [24] Dhawan, H., Rakshit, S., Gupta, T., Narula, R. and Kumar, V., 2023, October. Analysing the Best Algorithm and Hyperparameter for Brain Tumour Detection. In *International Conference on Smart Systems: Innovations in Computing* (pp. 439-454). Singapore: Springer Nature Singapore.
- [25] Jabbar, M.A., Deekshatulu, B.L. and Chandra, P., 2013. Heart disease classification using nearest neighbor classifier with feature subset selection. *Anale. Seria Informatica*, 11, pp.47-54.
- [26] Mia, M., Masruriyah, A.F.N. and Pratama, A.R., 2022. The Utilization of Decision Tree Algorithm In Order to Predict Heart Disease. *JURNAL SISFOTEK GLOBAL*, 12(2), p.138.
- [27] ThanhNoi, P. and Kappas, M., 2017. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. *Sensors*, 18(1), p.18. Aggarwal, V., Gupta, V., Singh, P., Sharma, K. and Sharma, N.: Detection of spatial outlier by using improved Z-score test. In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 788-790). IEEE (2019)
- [28] Suganya, R., Rajaram, S., Sheik, Abdullah, A. and Rajendran, J. (2016). A Novel Feature Selection Method for Predicting Heart Disease with Data Mining Techniques. *Asian J of Info Tech*, 15, 1314-1321.



- [29] H. Wang, T. M. Khoshgoftaar, K. Gao, A comparative study of filter-based feature ranking techniques, in: 2010 IEEE Int. Conf. Inf. Reuse Integr., IEEE, 2010, 43–48.
- [30] Jain,D.,Singh,V.(2018).FeatureSelectionandClassificationSystemsforChronicDiseasePrediction: A Review. Egyptian Informatics J. 1-11,inpress.<https://doi.org/10.1016/j.eij.2018.03.002>.
- [31] F. M. J. M. Shamrat, P. Ghosh, M. H. Sadek, A. Kazi, and S. Shultana, “Implementation of machine learning algorithms to detect the prognosis rate of kidney disease,” in Proc. IEEE Int. Conf. Innov. Technol.,Nov. 2020, pp. 1–7.
- [32] S. Shultana, M. S. Moharram, and N. Neehal, “Olympic sports events classification using convolutional neural networks,” in Proc. Int. Joint Conf. Comput. Intell. (IJCCI), Dhaka, Bangladesh, 2018, pp. 507–518.
- [33] Sumwiza, K., Twizere, C., Rushingabigwi, G., Bakunzibake, P. and Bamurigire, P., 2023. Enhanced cardiovascular disease prediction model using random forest algorithm. Informatics in Medicine Unlocked, 41, p.101316.
- [34] Reddy, N.S.C., Nee, S.S., Min, L.Z. and Ying, C.X., 2019. Classification and feature selection approaches by machine learning techniques: Heart disease prediction. International Journal of Innovative Computing, 9(1).
- [35] Srilakshmi, G., Avinash, G.S., Gayathri, V., Abishek, K.B., Swarna, A.D. and Srinivas, P.V.V.S., 2023, July. Forecasting Congestive Heart Failure using Deep Learning. In 2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 1128-1136). IEEE.
- [36] Karna, V.V.R., Karna, V.R., Janamala, V., Devana, V.K.R., Ch, V.R.S. and Tummala, A.B., 2024. A Comprehensive Review on Heart Disease Risk Prediction using Machine Learning and Deep Learning Algorithms. Archives of Computational Methods in Engineering, pp.1-33.
- [37] Bandyopadhyay, S., Sarma, M. and Samanta, D., 2023, December. Cardiac Risk Factors Identification with Hybrid Statistical Approach and Prediction of Cardiovascular Disease with Machine Learning. In 2023 OITS International Conference on Information Technology (OCIT) (pp. 41-46). IEEE.
- [38] Abdulaal, A. H., Yassin, R. A., Valizadeh, M., Abdulwahhab, A. H., Jasim, A. M., Mohammed, A. J., Jabir, H. J., Albaker, B. M., Dheyaa, N. H., & Amirani, M. C. (2024). Cutting-edge CNN approaches for breast histopathological classification: The impact of spatial attention mechanisms. ShodhAI: Journal of Artificial Intelligence, 1(1), 109–130. <https://doi.org/10.29121/shodhai.v1.i1.2024>
- [39] Matumueni, H. C. (2025). Knowledge-based expert system to assist physician for differential diagnosis of cholera tropical diseases. International Journal of Engineering Science Technologies, 9(2), 62–71. <https://doi.org/10.29121/ijoest.v9.i2.2025>
- [40] Huang, C.-H. (2024). Real-time motion image pose decomposition, classification and analysis. International Journal of Engineering Technologies and Management Research, 11(6), 8–13. <https://doi.org/10.29121/ijetmr.v11.i6.2024.602>