Physical Education, Health and Social Sciences
https://journal-of-social-education.org
E-ISSN: 2958-5996
P-ISSN: 2958-5988

Predicting Heart Disease Risk Using Machine Learning Models and Feature Selection Techniques

Sohaib Latif[1], *, Raheel Khalid[2], Muhammad Raza Khan[2]

[1] Department of Computer Science and Software Engineering, Grand Asian University, Sialkot, Punjab, Pakistan. Corresponding Author, Email: sohaib.latif@gaus.edu.pk
[2] Department of Computer Science, The University of Chenab, Gujrat, Punjab, Pakistan 50700. raheelkhalid37@gmail.com, muhammadrazakhan28@gmail.com

DOI: https://doi.org/10.63163/jpehss.v3i2.229

Abstract
Heart disease is one of the leading causes of death worldwide, making early detection essential for improving patient outcomes. With advancements in machine learning (ML), predictive models now offer a powerful way to assist doctors in diagnosing heart disease more accurately and efficiently. This study explores various ML algorithms, including Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), Naïve Bayes (NB), and K-Nearest Neighbors (KNN), to identify the most effective approach for heart disease prediction. Using the Cleveland Heart Disease Dataset, which contains 1,025 patient records with 14 medical attributes, we preprocessed the data, selected key features, and optimized model parameters. After evaluating the models with 10-fold cross-validation, the Random Forest model achieved the highest accuracy (98%), followed by Decision Tree (97%). These results highlight the potential of ML-based tools in clinical decision-making, helping doctors detect heart disease at an earlier stage and make informed treatment plans.

Introduction
Heart disease, or cardiovascular disease (CVD), is still one of the biggest health issues we face today, claiming nearly one in three lives globally. If we don't act, the number of deaths from CVD could climb to 22 million by 2030. The problem often starts when plaque builds up in the arteries, narrowing them and making it harder for blood to flow, which can lead to heart attacks or strokes. A lot of it comes down to how we live—things like not moving enough, eating poorly, drinking too much, or smoking can all take a toll on our hearts. But the good news is that small changes can make a big difference. Eating less salt, filling our plates with more fruits and veggies, staying active, and avoiding harmful habits can help protect our hearts and keep us healthier in the long run. Technology has transformed healthcare, making it possible to collect and analyze massive amounts of patient data from hospitals and clinics. This information is now a cornerstone for diagnosing and managing diseases more effectively. Decision Support Systems (DSS) have become an invaluable tool for healthcare professionals, helping them sift through patient records, seek second opinions, and reduce unnecessary tests—ultimately saving time and resources. Machine learning (ML) has taken this a step further, enhancing our ability to predict and detect diseases earlier than ever before. A great example is using the Naïve Bayes (NB) algorithm in DSS for predicting heart disease. By leveraging historical data, like the Cleveland dataset, it identifies critical patterns and features,

improving diagnostic accuracy and helping doctors make better, more informed patient decisions. Heart disease has created a lot of serious concerned among research; one of the major challenges in heart disease is correct detection and finding presence of it inside a human. Early techniques have not been so much efficient in finding it even medical professor are not so much efficient enough in predicating the heart disease [1]. There are various medical instruments available in the market for predicting heart disease there are two major problems in them, the first one is that they are very much expensive and second one is that they are not efficiently able to calculate the chance of heart disease in human. According to latest survey conducted by WHO, the medical professional able to correctly be predicted only 67% of heart disease [2] so there is a vast scope of research in area of predicating heart disease in human. Several factors contribute to the occurrence of cardiovascular disease, which can be broadly categorized into genetic predispositions that include an extended family history of the disease, environmental factors such as smoking tobacco, abusing drugs, leading a sedentary lifestyle, and comorbidities such as uncontrolled diabetes, hypertension, dyslipidemia, associated lung diseases, mental illnesses, and other conditions that make a person more susceptible to MI [3]. It is an expensive endeavor. About 735,000 people have a heart attack in the United States alone each year, and 71.5% of those patients are first responders [4]. The prediction states that between 2015 and 2030, the incidence of coronary heart disease, the primary cause of MI, will increase by 18% [5]. The predicted cost of cardiovascular disease management by 2035 is expected to reach 1.1 trillion USD, up from 555 billion USD in 2015 [6]. Early detection of MI is essential to prevent cardiac failure, arrhythmia, or unexpected death. A variety of evaluation modalities, including electrocardiograms (ECGs) [5], magnetic resonance imaging (MRI) [7] and echocardiography [8], can be used to identify MI. The most often used technique for supporting cardiac functions and assessing the health of the myocardial and left ventricle is magnetic resonance imaging (MRI) [7]. Myocardial infarction is a pathological condition resulting from an anatomical issue with the left ventricle (LV). To tackle these challenges, Clinical Decision Support Systems (CDSS) powered by machine learning have been created to evaluate heart disease risk and suggest the best treatment options. Research shows that CDSS can improve decision-making, refine clinical evaluations, and play a key role in preventive care. Coronary artery disease (CAD), also referred to as ischemic heart disease (IHD), is one of the most widespread types of cardiovascular disease and a major cause of death, especially for adults over 35. In countries like Pakistan, the number of deaths linked to CAD has risen sharply, highlighting the critical need for reliable predictive tools. When arteries narrow and restrict blood flow to the heart, it can cause damage to the heart muscle, potentially leading to serious issues like irregular heartbeats or even sudden cardiac arrest. Coronary heart disease (CHD) is one of the top causes of death worldwide. It happens when fatty deposits, called atheroma, build up in the coronary arteries, narrowing them and reducing blood flow to the heart. Over time, this buildup—known as atherosclerosis—can lead to serious heart problems. Several factors increase the risk of CHD, including high cholesterol, high blood pressure, diabetes, smoking, and drinking too much alcohol. People with CHD often experience symptoms like chest pain (angina) and trouble breathing. While there's no complete cure for CHD, catching it early and making lifestyle changes—like eating healthier, exercising, and quitting smoking—can make a big difference. Combined with proper medical care, these steps can help patients live longer, healthier lives and lower the chances of complications. In section 2 review the recent related works, while section 3 provides the methodology detailed of our approach. Followed by the experiment result, the discussion, and the limitations of our study in section 4. Finally, section 5 concludes our contribution and provide some future works.

Literature review

In the growing field of data science and medical care, the need for automated diagnostic systems is increasing. Data scientists have developed several models, which have helped aid in the field of medical care. Previous studies have shown that neural networks, Naive Bayes classifiers, and associative classification are powerful methods for diagnosing coronary heart disease. This is because associative classification provides high data accuracy and data flexibility, which traditional classifiers lack [9]. In order to develop a heart disease classifier, a data mining algorithm was built for data gathering and for predictive modelling. Thousand CHD patient records were mined, and the authors used a Support Vector Machine (SVM), Artificial Neural Network (ANN), and a Decision Tree (DT) for the binary classification job. The models respectively produced accuracies of 92.1%, 91%, and 89.6%. Furthermore, K-folds validation and confusion matrices were used to evaluate the consistency, sensitivity, and specificity of the data [10]. Ensemble techniques have proved extremely powerful in predicting heart disease. A group of researchers [11] cross-compared three algorithms: c4.5, j4.8, and the bagging algorithm, and concluded that bagging was the most powerful, with an accuracy of 81.41%. This depicts the scope of ensemble techniques. Two researchers [12] combined various models and compared their respective strengths. The most powerful model was produced by combining a fuzzy Naive Bayes with a genetic algorithm. This had an accuracy of 97.14%. A group of researchers [13] helped develop a new cost function to address the limitations of the previous ensemble techniques: feature selection and low accuracy. Lastly, Baccouche et al. used an ensemble classifier with a BiLSTM or BiGRU model with a CNN model to achieve a F1 score of between 91 and 96% for prediction of heart disease [14]. The research highlighted that ensemble frameworks could overcome the problem of predicting upon an unbalanced dataset. Ashraf *et al.* [15] used both the individual learning algorithms and ensemble approaches like Bayes Net, J48, KNN, multilayer perceptron, Naïve Bayes, random tree, and random forest for prediction purposes. Of these, J48 had an accuracy of 70.77%. They subsequently employed new-fangled techniques of which KERAS obtained an 80% accuracy. A multi-task (MT) recurrent neural network was proposed to predict the onset of cardiovascular disease with the attention mechanism at work [16]. The proposed model benefits by an Area under Curve (AUC) increase between 2 and 6%.

Table 1: Research Studies on Heart Disease Detection and Analysis

Ref. No.	Authors	Title	Journal/ Conference	Year	Publisher
[17]	R. Tao, S. Zhang, X. Huang et al.	Magnetocardiography based ischemic heart disease detection and localization using machine learning methods	IEEE Transactions on Biomedical Engineering	2018	IEEE
[18]	N. L. Fitriyani, M. Syafrudin, G. Alfian, J. Rhee	HDPM: an effective heart disease prediction model for a clinical decision support system	IEEE Access	2020	IEEE
[19]	Q. Zhenya, Z. Zhang	A hybrid cost-sensitive ensemble for heart disease prediction	BMC Medical Informatics and Decision Making	2021	BMC

[20]	A. K. Biswal, D. Singh, B. K. Pattanayak, D. Samanta, S. A. Chaudhry, A. Irshad	Adaptive fault-tolerant system and optimal power allocation for smart vehicles in smart cities using controller area network	Security and Communication Networks	2021	-
[21]	F. I. Alarsan, M. Younes	Analysis and classification of heart diseases using heartbeat features and machine learning algorithms	Journal of Big Data	2019	-
[22]	V. Shorewala	Early detection of coronary heart disease using ensemble techniques	Informatics in Medicine Unlocked	2021	-
[23]	P. Sivakumar, R. Nagaraju, D. Samanta, M. Sivaram, M. N. Hindia, I. S. Amiri	A novel free space communication system using nonlinear InGaAsP microsystem resonators for enabling power-control toward smart cities	Wireless Networks	2020	-
[24]	S. U. Ghumbre, A. A. Ghatol	Heart disease diagnosis using machine learning algorithm	Advances in Intelligent and Soft Computing	2012	Springer
[25]	A. K. Gárate-Escamila, A. Hajjam El Hassani, E. Andrés	Classification models for heart disease prediction using feature selection and PCA	Informatics in Medicine Unlocked	2020	-
[26]	L. Verma, S. Srivastava, P. C. Negi	An intelligent noninvasive model for coronary artery disease detection	Complex & Intelligent Systems	2018	-
[27]	Chicco, D.; Jurman, G.	Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone	BMC Medical Informatics and Decision Making	2020	BMC
[28]	Karthick, D.; Priyadharshini, B.	Predicting the chances of occurrence of Cardio Vascular Disease (CVD) in people using classification techniques within fifty years of age	Proceedings of the 2nd International Conference on Inventive Systems and Control (ICISC)	2018	-
[29]	Sharma, H.; Rizvi, M.A.	Prediction of Heart Disease using Machine Learning Algorithms: A Survey	Int. J. Recent Innov. Trends Comput. Commun.	2017	-

Methodology

This section explains the proposed approach, covering the dataset used, data preprocessing steps, machine learning models, feature selection methods, and how performance was evaluated. Figure 1 provides a visual overview of the experimental workflow. The process starts with gathering the heart disease dataset in .csv format from the UCI Machine Learning Repository. Once the dataset was obtained, it was imported into Jupyter Notebook, a popular software tool, to analyze its attributes, data types, value ranges, and other statistical details. This step helps in understanding the dataset's structure and preparing it for further analysis.

Dataset Description and Statistics

The Heart disease dataset consists of 1025 instances with 14 attributes which are more suitable for research experimental purposes. The attribute descriptions for the Cleveland heart dataset are given in Table 1.

Table 2: Dataset Description

Attribute	Description	Type of Attribute	Attribute Value Range
Age	Description for age	int64	29 to 77
Sex	Description for sex	int64	0 is female and 1 is Male
Cp	Description for cp	int64	1 = typical angina, 2 = atypical angina, 3 = non-angina pain,
trestbps	Description for trestbps	int64	94 to 200
Chol	Description for chol	int64	126 to 564
Fbs	Description for fbs	int64	0 = false and 1= true
restecg	Description for restecg	int64	0 = normal, 1 = ST-T wave abnormality, 2 = definite left ventricular hypertrophy by Estes' criteria
thalach	Description for thalach	int64	0 = no 1 = yes
exang	Description for exang	int64	0 to 1
oldpeak	Description for oldpeak	float64	0.0 to 6.2
slope	Description for slope	int64	1 = upsloping, 2 = flat, 3 = downs loping
Ca	Description for ca	int64	0 to 4
Thal	Description for thal	int64	3 = normal, 6 = fixed defect, 7 = reversible defect
target	Description for target	int64	0 = no risk of heart disease, 1 to 4 = risk of heart disease

Attributes with fewer than ten unique categories are considered categorical variables. Here's a detailed look at some of the key attributes in the dataset:

- **Sex**: Represents gender, where 1 = male and 0 = female.
- **Chest Pain Type (cp)**: Divided into four categories:
  - 1: Typical angina
  - 2: Atypical angina
  - 3: Non-anginal pain
  - 4: Asymptomatic
- **Fasting Blood Sugar (fbs)**: Indicates if fasting blood sugar is above 120 mg/dL, with 1 = true and 0 = false.
- **Resting Electrocardiographic Results (restecg)**: Includes three outcomes:
  - 0: Normal
  - 1: ST-T wave abnormalities
  - 2: Left ventricular hypertrophy
- **Exercise-Induced Angina (exang)**: Shows whether the patient experiences angina during exercise, with 1 = yes and 0 = no.
- **Slope of the ST Segment (slope)**: Categorized based on the slope of the ST segment during peak exercise:
  - 1: Upsloping
  - 2: Flat
  - 3: Downsloping
- **Number of Major Vessels (ca)**: Represents the number of major blood vessels (ranging from 0 to 3) visible through fluoroscopy.
- **Thalassemia (thal)**: Describes heart status:
  - 3: Normal
  - 6: Fixed defect
  - 7: Reversible defect
- **Target (Heart Disease Risk)**: Originally included five classes:
  - 0: No risk of heart disease
  - 1 to 4: Different levels of heart disease risk

To simplify the study's goal of predicting heart disease risk, values 1 to 4 were combined into a single category (1 = at risk), making this attribute binary (0 or 1).

- **Numeric Attributes**: The following attributes are treated as numerical or continuous variables:
  - Age
  - Resting Blood Pressure (trestbps)
  - Cholesterol Level (chol)
  - Maximum Heart Rate Achieved (thalach)
  - ST Depression Induced by Exercise (oldpeak)

This breakdown helps in understanding how the dataset is structured and how each attribute contributes to predicting heart disease risk.

Data Preprocessing

Next, we moved on to data pre-processing, which involved checking for missing values and addressing them appropriately. Depending on the type of attribute, missing values were either filled with a user-defined constant or replaced with the mean value to ensure the machine learning models could perform at their best. Our study follows a clear, step-by-step approach to predict heart disease using key medical attributes. We used the HEART_DATA dataset, which contains 1,025 records and 14 attributes, for our analysis. After confirming that there were no missing values, we focused on important clinical factors such as age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, and other relevant measurements. Figure 2 provides a visual representation of all 14 attributes in the dataset. To improve the performance of our models, we removed unnecessary columns. The data was then standardized and normalized where needed. Categorical variables were processed using techniques like one-hot encoding and label encoding to make them suitable for analysis. Finally, the dataset was divided into two subsets: 80% for training the models and 20% for testing, ensuring a robust and reliable evaluation of the model's performance.

Model Selection

For phishing email detection, several machine learning models were selected based on various factors such as model diversity, ability to handle high-dimensional data, interpretability, and computational efficiency. The following models were chosen:

Random Forest (RF)

Random Forest is an ensemble method that builds multiple decision trees and merges their results to improve accuracy and prevent overfitting. Each tree is trained on a random subset of the features and data points, making it a robust and scalable model for classification tasks. Random Forest was selected due to its proven effectiveness in handling high-dimensional data and its ability to balance precision and recall, which is crucial for phishing detection.

Logistic Regression (LR)

Logistic Regression is a simple, yet effective linear model used for binary classification tasks. It estimates the probability that a given instance belongs to a particular class based on the input features. Logistic Regression was selected for its interpretability and computational efficiency. It provides a clear decision boundary and works well with preprocessed features like those used in this study.

Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a powerful algorithm used for classification and regression. It works by finding the best possible boundary (called a hyperplane) that separates different classes in the data. The goal is to maximize the distance between this boundary and the nearest data points from each class, ensuring a clear separation.

Naive Bayes (NB)

Naive Bayes is a probabilistic classifier based on Bayes' Theorem, which assumes that the features are independent. Despite this simplifying assumption, it can perform well in many classification tasks, including text classification. Naive Bayes was selected for its simplicity

and computational efficiency. It is particularly well-suited for text classification tasks and was expected to work well with the email content features in the dataset.

Decision Tree (DT)
A Decision Tree splits the data into subsets based on feature values, recursively creating decision rules. It is easy to interpret and visualize, making it a popular choice for classification tasks. Decision Trees are interpretable, simple to implement, and provide clear decision boundaries. While it may not always be as accurate as ensemble models like Random Forest, it serves as a useful benchmark for comparison.

Evaluation Metrics
The performance of the machine learning models was evaluated using several standard metrics, which are defined in terms of the following values: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These metrics are as follows:

Accuracy:
Accuracy measures the overall correctness of the model. It is the ratio of correctly classified instances (both true positives and true negatives) to the total number of instances.

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

Recall:
Recall measures the ability of the model to correctly identify positive instances (phishing emails). It is the ratio of true positives to the total actual positives (true positives + false negatives).

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

Precision:
Precision measures the accuracy of positive predictions. It is the ratio of true positives to the total predicted positives (true positives + false positives).

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

F1-Score:
The F1-score is the harmonic mean of precision and recall, providing a balance between the two. It is particularly useful when the class distribution is imbalanced.

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

These evaluation metrics allow for a comprehensive assessment of each model's performance, considering both the ability to correctly identify phishing emails and minimize false positives and false negatives. There are many reasons why data might be missing in a dataset. For example, respondents might forget to answer certain questions, choose not to respond, or simply skip them. Technical issues, like sensor failures, data loss during transfer, internet outages, or computational errors (such as division by zero), can also lead to gaps in the data.

Identifying missing values can be tricky because their impact isn't always obvious—sometimes they have a major effect on the results, while other times they don't. Even if a single variable has only a few missing entries, these gaps can add up across the dataset and become a significant problem. While it's possible to run an analysis with missing values, doing so can weaken the accuracy and reliability of the results. In our case, after carefully examining the dataset, we found that there were no missing values at all. Since every attribute contains complete data, we didn't need to use any techniques to handle missing values. This allowed us to move forward with our analysis confidently, knowing the data was fully intact.

Visualization of Attributes of Dataset

To gain a deeper understanding of the dataset, we started by generating descriptive statistics to analyze the distribution of numerical attributes. We also performed a correlation analysis to uncover relationships between different features, which helped us identify key patterns and connections. To visualize these trends, we used various techniques such as histograms, box plots, scatter plots, and pair plots. This allowed us to spot and address any anomalies in the dataset. We also performed feature importance analysis using mutual information and correlation heat maps to determine which variables had the most significant impact on predicting heart disease shown in Figures 1,2 and 3 respectively.

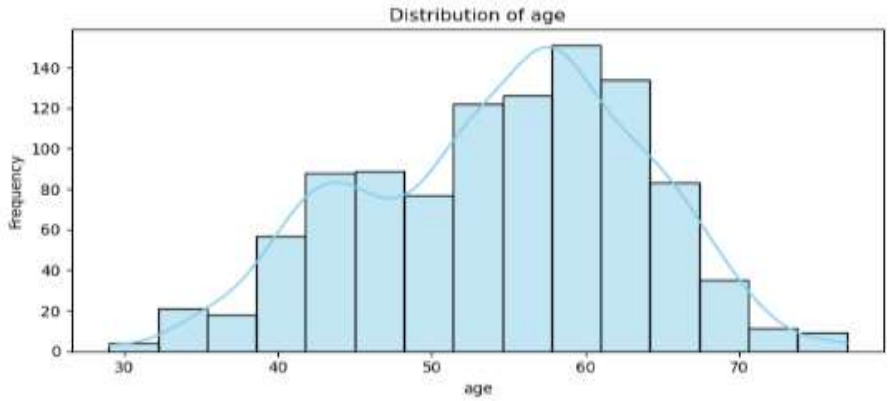
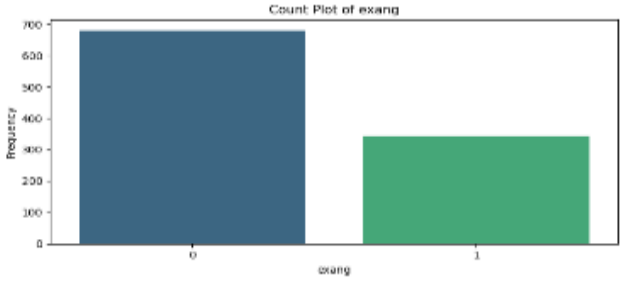
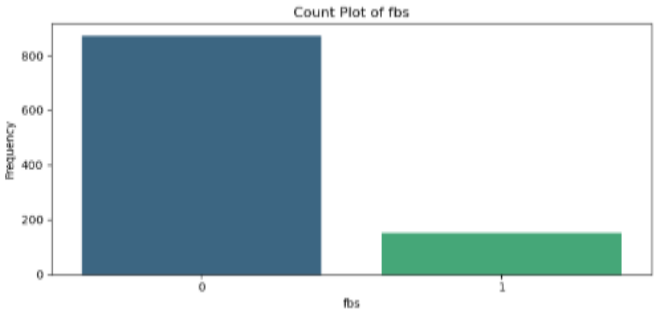
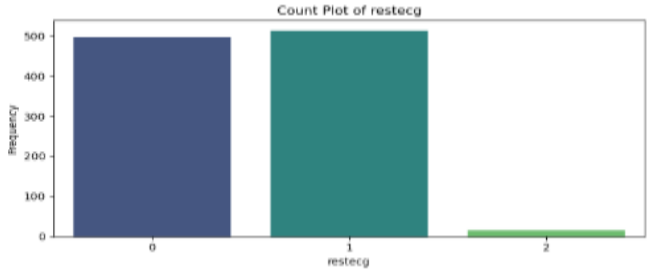


Figure 1: Histogram Plot





, Figure 2 (a)Count Plot for exang *Figure 2 (b) Count Plot for fbs*



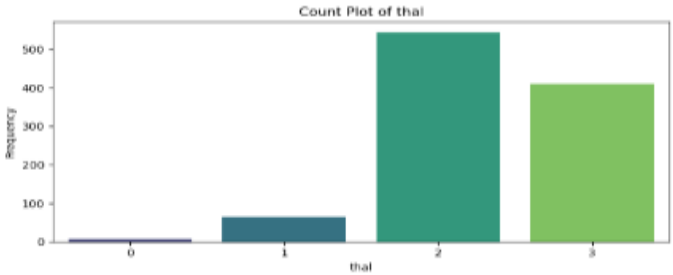
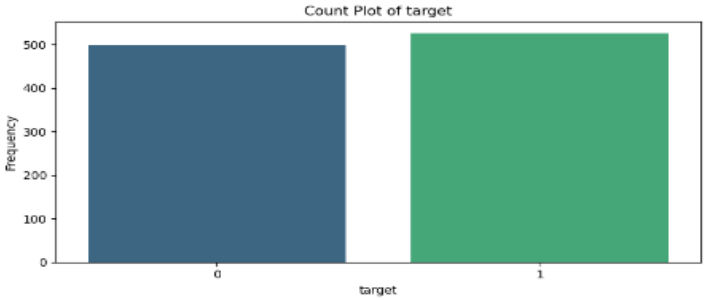


Figure 2 (c) Count Plot for restecg *Figure 2 (d) Count Plot for thal*



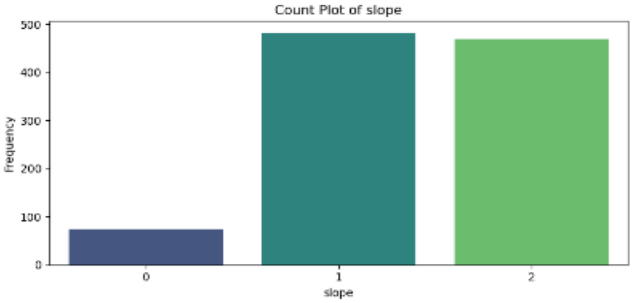


Figure 2 (e) Count Plot for target *Figure 2 (f) Count Plot for slope*

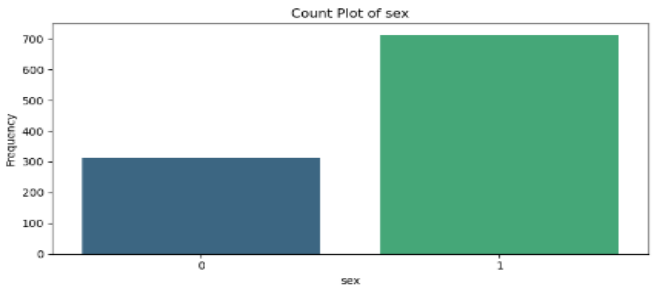


Figure 2 (g) Count Plot for sex

Figure 2: Count Plots for data visualization

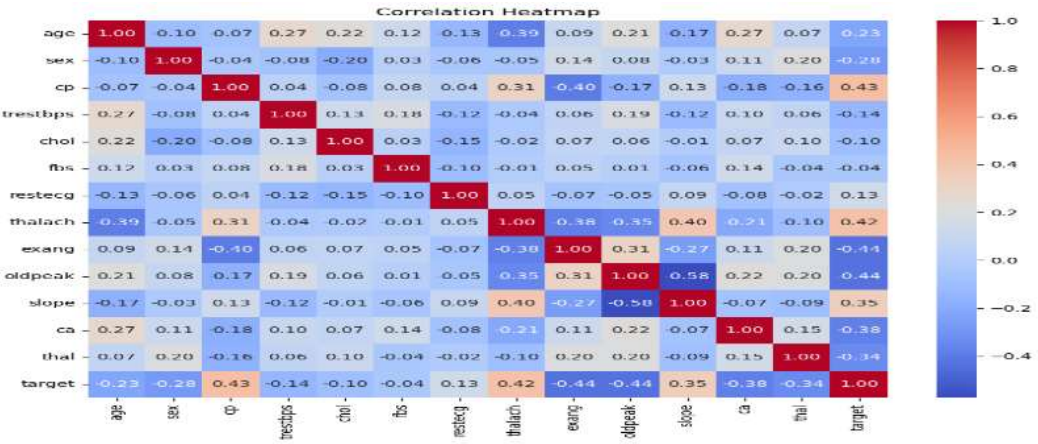


Figure 3: Confusion Matrix

For evaluating the models, we tested several machine learning algorithms, including Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), and Naïve Bayes.

Logistic Regression (LR)

Logistic Regression Analysis (LR) is a method used to determine the cause-effect relationship between the dependent variable and the independent variables, without being dependent on a certain distribution assumption, when the dependent variable is categorical, and the independent variables are mixed-scale. Using the maximum likelihood estimation method, LR estimates the unknown parameter values that maximize the probability obtained from the data set.

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n)}}$$

Decision Tree (DT)

A Decision Tree is a simple, yet powerful machine learning algorithm used for both classification and regression tasks. It works like a flowchart, where each decision is based on a question about the data, leading to different outcomes. The tree starts with a main question

(root node) and branches out based on the answers until it reaches a final decision (leaf node). However, if the tree grows too deep, it can memorize the training data instead of generalizing well, leading to overfitting. To prevent this, techniques like pruning (removing unnecessary branches) are used to simplify the model.

A decision tree uses a recursive splitting approach based on criteria like **Gini Impurity** or **Entropy**:

- **Gini Impurity**:

$$Gini = 1 - \sum_{i=1}^{c} p_i^2$$

- **Entropy**:

$$Entropy = -\sum_{i=1}^{c} p_i \log_2(p_i)$$

where p_i is the probability of class i.

Random Forest (RF)

A Random Forest is an advanced version of a decision tree that builds multiple trees instead of just one, making it more accurate and reliable. It works by creating many decision trees, each trained on a different random portion of the data. When making a prediction, the trees vote, and the most common result is chosen (for classification) or their average is taken (for regression). This approach reduces overfitting and makes the model more stable.

$$\hat{y} = \frac{1}{n} \sum_{i=1}^{n} T_i(X)$$

Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a powerful algorithm used for classification and regression. It works by finding the best possible boundary (called a hyperplane) that separates different classes in the data. The goal is to maximize the distance between this boundary and the nearest data points from each class, ensuring a clear separation.

For a linear SVM, the decision function is:

$$f(X) = w^T X + b$$

A new sample X is classified based on:

$$\hat{y} = \text{sign}(w^T X + b)$$

where w is the weight vector, and b is the bias.

For non-linear SVM, a kernel function $K(X, X')$ is used.

Naïve Bayes (NB)

Naïve Bayes is a simple and fast machine learning algorithm based on Bayes' theorem, which calculates probabilities to make predictions. It assumes that all features in the data are independent of each other, which isn't always true in real life but still works surprisingly well in many applications. Naïve Bayes is widely used for spam filtering, sentiment analysis, and medical diagnoses because it requires very little data to train and works efficiently with large datasets. However, its biggest limitation is the assumption of independence, which can lead to lower accuracy if features are related.

Based on Bayes' theorem, the probability of class C_k given input X is:

$$P(C_k|X) = \frac{P(X|C_k)P(C_k)}{P(X)}$$

Assuming independence of features, the likelihood is:

$$P(X|C_k) = P(x_1|C_k)P(x_2|C_k)...P(x_n|C_k)$$

where $P(C_k)$ is the prior probability of class C_k.

After applying these algorithms, we fine-tuned them to ensure they performed as effectively as possible. Finally, we conducted classification using the full set of attributes with the algorithms mentioned above and identified the model that delivered the best accuracy for predicting heart disease. The model performances of different algorithms are shown in Figure 4: -

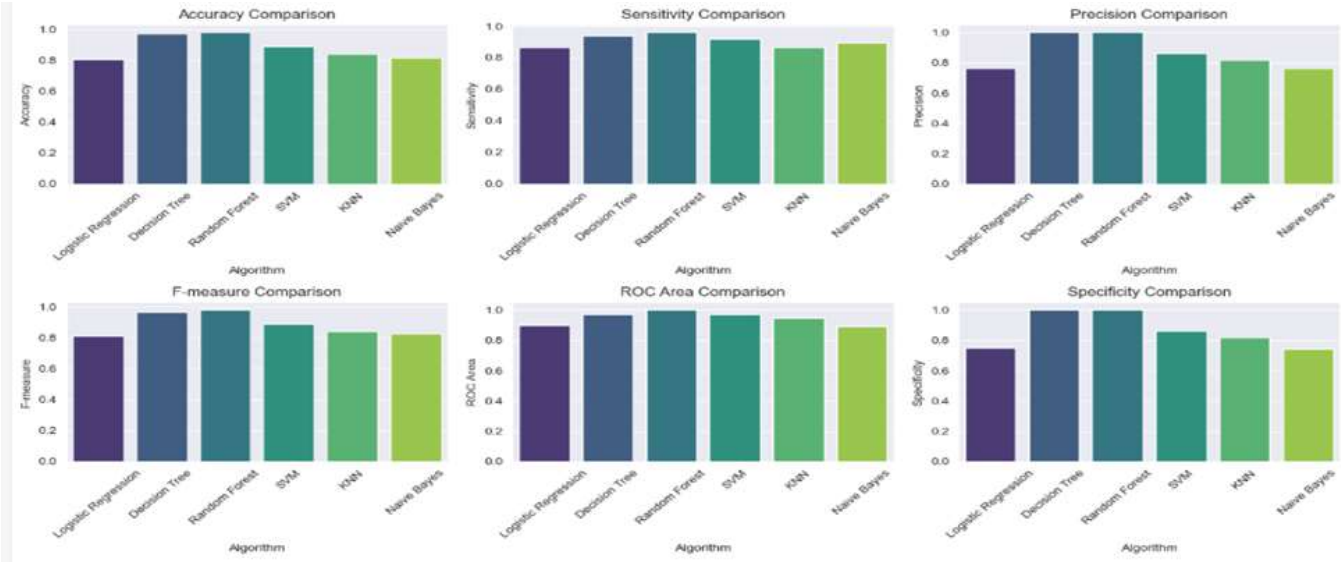


Figure 4: Comparison of Results

To assess the effectiveness of our models, we relied on key performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score. We also created confusion matrices to examine classification errors and used precision-recall curves and ROC curves to gain deeper insights into how well the models performed. To make the models more interpretable,

we applied SHAP (Shapley Additive Explanations) analysis, which helped us understand how each feature contributed to the predictions. The best-performing model was chosen based on a combination of high accuracy and clear interpretability. During exploratory data analysis, we discovered that certain features, like age and chest pain type, played a significant role in predicting heart disease. We analyzed key numerical attributes for their mean, standard deviation, and range. A correlation matrix revealed strong relationships between chest pain type, cholesterol levels, and the presence of heart disease. Additionally, we conducted statistical significance tests, including ANOVA and chi-square tests, to confirm the importance of these features in our analysis.

Results

Model Performance Evaluation

The performance of six machine learning models was evaluated on a balanced dataset of phishing emails. The accuracy, precision, recall, and F1-scores of each model were computed and analyzed to assess their effectiveness.

Logistic Regression

The Logistic Regression model achieved 80.51% accuracy, correctly classifying most instances. It has a high sensitivity (86.58%), meaning it effectively detects positive cases, but a moderate fallout (25.16%), indicating some false positives. The precision (76.33%) and F1-score (81.13%) show a good balance between precision and recall shown in Table 3. With an ROC area of 89.68%, the model demonstrates strong discrimination ability but may need adjustments to reduce false positives.

Table 3:

Accuracy	**80.51%**
MAE	19.48%
Sensitivity	86.58%
Fallout	25.16%
Precision	76.33%
F1-Score	81.13%
Roc Area	89.68%
Specificity	74.84%

Performance measure for Logistic Regression

Decision Tree

The Decision Tree model achieved a high accuracy of 97.07%, indicating strong classification performance. The Mean Absolute Error (MAE) of 2.92% suggests minimal prediction errors. With a sensitivity of 93.95%, the model effectively detects positive cases, while a fallout of 0% and specificity of 100% indicate that it makes no false positive errors. The precision of 100% means all predicted positives were correct, leading to a strong F1-score of 96.88%, balancing precision and recall. Additionally, the ROC area of 96.97% confirms excellent discrimination between positive and negative cases are shown in Table 4.

Table 4:

Accuracy	**97.07%**
MAE	02.92%
Sensitivity	93.95%
Fallout	0%
Precision	1.00%
F1-Score	96.88%
Roc Area	96.97%
Specificity	100%

Performance measure for Decision Tree

Random Forest

The Random Forest model achieved an accuracy of 98.05%, demonstrating excellent classification performance. The Mean Absolute Error (MAE) of 1.94% indicates very few prediction errors. With a sensitivity of 95.97%, the model effectively identifies positive cases, while a fallout of 0% and specificity of 100% confirm no false positives. The precision of 100% means all predicted positives were correct, leading to a high F1-score of 97.94% shown in Table 5, ensuring a strong balance between precision and recall. Additionally, the ROC area of 100% suggests perfect discrimination between classes, indicating an exceptionally robust model.

Table 5:

Accuracy	**98.05%**
MAE	01.94%
Sensitivity	95.97%
Fallout	0%
Precision	1.00%
F1-Score	97.94%
Roc Area	100%
Specificity	100%

Performance measure for Radom Forest

SVM

The Support Vector Machine (SVM) model achieved an accuracy of 88.96%, indicating strong classification performance. The Mean Absolute Error (MAE) of 11.03% suggests a moderate level of misclassification. With a sensitivity of 91.94%, the model effectively detects positive cases, though a fallout of 13.83% shows some false positives. The precision of 86.16%

indicates that most predicted positives are correct, leading to an F1-score of 88.96%, which balances precision and recall. The ROC area of 97.26% confirms excellent discrimination ability, while a specificity of 86.16% highlights the model's effectiveness in identifying negative cases shown in Table 6.

Table 6:

Accuracy	**88.96%**
MAE	11.03%
Sensitivity	91.94%
Fallout	13.83%
Precision	86.16%
F1-Score	88.96%
Roc Area	97.26%
Specificity	86.16%

Performance measures for SVM

KNN

The K-Nearest Neighbors (KNN) model achieved an accuracy of 84.09%, indicating decent classification performance. However, the Mean Absolute Error (MAE) of 15.90% suggests a relatively higher rate of misclassification compared to other models. The sensitivity of 86.57% shows that the model effectively detects positive cases, but a fallout of 18.23% indicates a notable false positive rate. The precision of 81.64% means that most predicted positives were correct, leading to an F1-score of 84.03%, which balances precision and recall shown in Table 7. The ROC area of 94.83% suggests strong overall classification ability, while the specificity of 81.76% indicates reasonable performance in identifying negative cases.

Table 7:

Accuracy	**84.09%**
MAE	15.90%
Sensitivity	86.57%
Fallout	18.23%
Precision	81.64%
F1-Score	84.03%
Roc Area	94.83%
Specificity	81.76%

Performance measures for KNN

Naive Bayes

The Naïve Bayes model achieved an accuracy of 84.09%, showing good classification performance. However, the Mean Absolute Error (MAE) of 18.50% indicates a relatively higher misclassification rate. The sensitivity of 89.26% suggests strong detection of positive

cases, but a fallout of 25.78% means a significant number of false positives. The precision of 76.43% indicates that some predicted positives were incorrect, leading to an F1-score of 82.35%, balancing precision and recall shown in Table 8. The ROC area of 89.46% shows decent discrimination ability, while the specificity of 74.21% suggests the model struggles somewhat with correctly identifying negative cases.

Table 8:

Accuracy	**84.09%**
MAE	18.50%
Sensitivity	89.26%
Fallout	25.78%
Precision	76.43%
F1-Score	82.35%
Roc Area	89.46%
Specificity	74.21%

Performance measures for Naïve Bayes

The performance of various machine learning classifiers was evaluated using both the complete set of attributes and an optimized subset selected through attribute evaluation techniques. As shown in Table 9, the Random Forest (RM) algorithm achieved the highest accuracy of 98%, followed closely by Decision Tree with 97% when using the full dataset. Furthermore, the SVM (Support vector Machine) algorithm also performed and achieved the accuracy of 88% other models in additional performance metrics, achieving a Mean Absolute Error (MAE) of 0.110, a sensitivity of 0.919, a fallout of 0.138, a precision of 0.861, an F-measure of 0.889, and a specificity of 0.97. Another algorithm, Logistic Regression (LR) algorithm also performed and achieved the accuracy of 80% other models in additional performance metrics, achieving a Mean Absolute Error (MAE) of 0.194, a sensitivity of 0.865, a fallout of 0.25, a precision of 0.763, an F- measure of 0.811, and a specificity of 0.74. Another algorithm, Naïve Bayes (NB) algorithm also performed and achieved the accuracy of 81% other models in additional performance metrics, achieving a Mean Absolute Error (MAE) of 0.159, a sensitivity of 0.892, a fallout of 0.257, a precision of 0.764, an F- measure of 0.823, and a specificity of 0.742. Another algorithm, KNN algorithm also performed and achieved the accuracy of 84% other models in additional performance metrics, achieving a Mean Absolute Error (MAE) of 0.159, a sensitivity of 0.865, a fallout of 0.182, a precision of 0.816, an F- measure of 0.840, and specificity of 0.817.

Table 9: Performance Comparison of Machine Learning Classifiers

Algorithm	Accuracy	Mae %	Sensitivity%	Fallout%	Precision%	F-Measure%	Roc Area%	Specificity%
Logistic Regression	80.51	19.48	86.57	25.15	76.33	81.13	89.67	74.84
Decision Tree	97.07	02.92	93.95	0	100	96.88	96.97	100
Random Forest	98.05	01.94	95.97	0	100	97.94	100	100

SVM	88.96	11.03	91.94	13.83	86.16	88.96	97.26	86.16
KNN	84.09	15.90	86.57	18.23	81.64	84.03	94.83	81.76
Naïve Bayes	81.49	18.50	89.26	25.78	76.43	82.35	89.46	74.21

Conclusion

This study confirms that machine learning can play a vital role in predicting heart disease risk, offering a reliable and efficient way to support early diagnosis. Among the models tested, Random Forest and Decision Tree outperformed others, achieving the highest accuracy rates. The research also shows that proper data preprocessing and feature selection can significantly improve model performance. While these results are promising, there is still room for improvement. Future research should explore deep learning techniques, larger and more diverse datasets, and real-time clinical applications to further enhance the accuracy and reliability of these models. Integrating ML-powered tools into healthcare systems could help doctors detect heart disease earlier, leading to faster interventions and better patient outcomes.

References

[1] Chandra, P. and Deekshatulu, B.L., 2012, November. Prediction of risk score for heart disease using associative classification and hybrid feature subset selection. In *2012 12th international conference on intelligent systems design and applications (ISDA)* (pp. 628-634). IEEE.

[2] Kirubha, V. and Priya, S.M., 2016. Survey on data mining algorithms in disease prediction. *International Journal of Computer Trends and Technology*, *38*(3), pp.124-128.

[3] Sing, C.F., Stengârd, J.H. and Kardia, S.L., 2003. Genes, environment, and cardiovascular disease. *Arteriosclerosis, thrombosis, and vascular biology*, *23*(7), pp.1190-1196.

[4] Acharya, U.R., Fujita, H., Sudarshan, V.K., Oh, S.L., Adam, M., Koh, J.E., Tan, J.H., Ghista, D.N., Martis, R.J., Chua, C.K. and Poo, C.K., 2016. Automated detection and localization of myocardial infarction using electrocardiogram: a comparative study of different leads. *Knowledge-Based Systems*, *99*, pp.146-156.

[5] Jafarian, K., Vahdat, V., Salehi, S. and Mobin, M., 2020. Automating detection and localization of myocardial infarction using shallow and end-to-end deep neural networks. *Applied Soft Computing*, *93*, p.106383.

[6] American Heart Association and American Stroke Association, 2017. Cardiovascular Disease: a costly burden for America—projections through 2035. *Cardiovascular Disease A Costly Burden [Internet]. American Heart Association*, *1*.

[7] Chen, Z., Lalande, A., Salomon, M., Decourselle, T., Pommier, T., Qayyum, A., Shi, J., Perrot, G. and Couturier, R., 2022. Automatic deep learning-based myocardial infarction segmentation from delayed enhancement MRI. *Computerized Medical Imaging and Graphics*, *95*, p.102014.

[8] Muraki, R., Teramoto, A., Sugimoto, K., Sugimoto, K., Yamada, A. and Watanabe, E., 2022. Automated detection scheme for acute myocardial infarction using convolutional neural network and long short-term memory. *PLoS One*, *17*(2), p.e0264002.

[9] Latha, C.B.C. and Jeeva, S.C., 2019. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, p.100203.

[10] Lv, J., Zhang, X., Han, X. and Fu, Y., 2007, November. A novel adaptively dynamic tuning of the contention window (CW) for distributed coordination function in IEEE 802.11 ad hoc networks. In *2007 International Conference on Convergence Information Technology (ICCIT 2007)* (pp. 290-294). IEEE.

[11] Shouman, M., Turner, T. and Stocker, R., 2011. Using Decision Tree for Diagnosing Heart Disease Patients. *AusDM*, *11*, pp.23-30.

[12] Singh, N., Firozpur, P. and Jindal, S., 2018. Heart disease prediction system using hybrid technique of data mining algorithms. *International Journal of Advance Research, Ideas and Innovations in Technology*, *4*(2), pp.982-987.

[13] Nourmohammadi-Khiarak, J., Feizi-Derakhshi, M.R., Behrouzi, K., Mazaheri, S., Zamani-Harghalani, Y. and Tayebi, R.M., 2020. New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection. *Health and Technology*, *10*(3), pp.667-678.

[14] Baccouche, A., Garcia-Zapirain, B., Castillo Olea, C. and Elmaghraby, A., 2020. Ensemble deep learning models for heart disease classification: A case study from Mexico. *Information*, *11*(4), p.207.

[15] Ashraf, M., Ahmad, S.M., Ganai, N.A., Shah, R.A., Zaman, M., Khan, S.A. and Shah, A.A., 2021. Prediction of cardiovascular disease through cutting-edge deep learning technologies: an empirical study based on TENSORFLOW, PYTORCH and KERAS. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2020, Volume 1* (pp. 239-255). Springer Singapore.

[16] Andreotti, F., Heldt, F.S., Abu-Jamous, B., Li, M., Javer, A., Carr, O., Jovanovic, S., Lipunova, N., Irving, B., Khan, R.T. and Dürichen, R., 2020. Prediction of the onset of cardiovascular diseases from electronic health records using multi-task gated recurrent units. *arXiv preprint arXiv:2007.08491*.

[17] Tao, R., Zhang, S., Huang, X., Tao, M., Ma, J., Ma, S., Zhang, C., Zhang, T., Tang, F., Lu, J. and Shen, C., 2018. Magnetocardiography-based ischemic heart disease detection and localization using machine learning methods. *IEEE Transactions on Biomedical Engineering*, *66*(6), pp.1658-1667.

[18] Fitriyani, N.L., Syafrudin, M., Alfian, G. and Rhee, J., 2020. HDPM: an effective heart disease prediction model for a clinical decision support system. *Ieee Access*, *8*, pp.133034-133050.

[19] Zhenya, Q. and Zhang, Z., 2021. A hybrid cost-sensitive ensemble for heart disease prediction. *BMC medical informatics and decision making*, *21*, pp.1-18.

[20] Biswal, A.K., Singh, D., Pattanayak, B.K., Samanta, D., Chaudhry, S.A. and Irshad, A., 2021. Adaptive Fault-Tolerant System and Optimal Power Allocation for Smart Vehicles in Smart Cities Using Controller Area Network. *Security and Communication Networks*, *2021*(1), p.2147958.

[21] Alarsan, F.I. and Younes, M., 2019. Analysis and classification of heart diseases using heartbeat features and machine learning algorithms. *Journal of big data*, *6*(1), pp.1-15.

[22] Shorewala, V., 2021. Early detection of coronary heart disease using ensemble techniques. *Informatics in Medicine Unlocked*, *26*, p.100655.

[23] Sivakumar, P., Nagaraju, R., Samanta, D., Sivaram, M., Hindia, M.N. and Amiri, I.S., 2020. A novel free space communication system using nonlinear InGaAsP microsystem resonators for enabling power-control toward smart cities. *Wireless Networks*, *26*(4), pp.2317-2328.

[24] Ghumbre, S.U. and Ghatol, A.A., 2012. Heart disease diagnosis using machine learning algorithm. In *Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012* (pp. 217-225). Springer Berlin Heidelberg.

[25] Gárate-Escamila, A.K., El Hassani, A.H. and Andrès, E., 2020. Classification models for heart disease prediction using feature selection and PCA. *Informatics in Medicine Unlocked*, *19*, p.100330.

[26] Verma, L., Srivastava, S. and Negi, P.C., 2018. An intelligent noninvasive model for coronary artery disease detection. *Complex & Intelligent Systems*, *4*, pp.11-18.

[27] Chicco, D. and Jurman, G., 2020. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC medical informatics and decision making*, *20*, pp.1-16.

[28] Karthick, D. and Priyadharshini, B., 2018, January. Predicting the chances of occurrence of Cardio Vascular Disease (CVD) in people using classification techniques within fifty years of age. In *2018 2nd international conference on inventive systems and control (ICISC)* (pp. 1182-1186). IEEE.

[29] Sharma, H. and Rizvi, M.A., 2017. Prediction of heart disease using machine learning algorithms: A survey. *International Journal on Recent and Innovation Trends in Computing and Communication*, *5*(8), pp.99-104.