A Comprehensive Study on Heart Disease Prediction Using Machine Learning Techniques

BY

Md Redwan Hussain
ID: 232-25-019

This Report Presented in Partial Fulfillment of the Requirements for
The Degree of Masters of Science in Computer Science and Engineering

Supervised By

Dr. S. M. Aminul Haque
Professor & Associate Head
Department of CSE
Daffodil International University

Co-Supervised By

Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

11TH JANUARY, 2025

APPROVAL

This Project/Thesis titled **"A Comprehensive Study on Heart Disease Prediction Using Machine Learning Techniques"**, submitted by **Md Redwan Hussain**, ID No: **232-25-019** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **11-01-2025**.

BOARD OF EXAMINERS

Chairman

Dr. Sheak Rashed Haider Noori, PhD
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Md. Zahid Hasan, PhD
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Arif Mahmud, PhD
Associate Professor & Director MIS
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

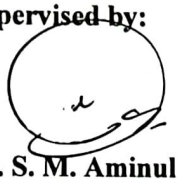
External Examiner

Dr. Mohammed Nasir Uddin, PhD
Professor
Department of Computer Science and Engineering
Jagannath University

DECLARATION

I hereby declare that this research has been done by me under the supervision of **Dr. S. M. Aminul Haque, Professor & Associate Head, Department of CSE,** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:



Dr. S. M. Aminul Haque
Professor & Associate Head
Department of CSE
Daffodil International University

Co-Supervised by:

Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Md Redwan Hussain
ID: 232-25-019
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to Almighty Allah for His divine blessing, which made it possible for me to complete the final year project/internship successfully.

I am grateful and wish to express my profound indebtedness to **Dr. S. M. Aminul Haque, Professor & Associate Head,** Department of CSE, Daffodil International University, Dhaka, and the deep knowledge & keen interest of my supervisor in the field of machine learning to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

I would like to express my heartiest gratitude to **Dr. Sheak Rashed Haider Noori**, **Head,** Department of CSE, for his kind help in finishing our project and to other faculty members and the CSE Department of Daffodil International University staff.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

Heart disease has emerged as a major global public health issue in recent years, and there is an increasing demand for precise prediction techniques to support early detection and prevention. To shed light on the effectiveness of each algorithm in clinical decision-making, this paper offers a thorough investigation of heart disease prediction utilizing a variety of machine learning (ML) techniques. We test the accuracy sensitivity, specificity, and predictive power of regular machine learning methods, like Logistic Regression, KNN, Decision Trees, Random Forests, Gaussian Naive Bayes, SVM, and LightGBM. It henceforth provides an overview of every step involved in maximizing the models through data preparation, feature selection, and model evaluation that can help optimize the model results. We also discuss common issues with predictive modeling in the medical field, including model interpretability, overfitting, and data imbalance. We investigate how key clinical parameters, such as age, cholesterol, exercise-induced angina, and others, impact model outcomes and pinpoint significant heart disease risk factors using a dataset that includes these features. The study demonstrates that decision trees and ensemble models, such as random forests and LightGBM, perform well, providing high classification accuracy and great interpretability. Comparisons with traditional classifiers demonstrate how these models can provide reliable and scalable prediction abilities to support well-established diagnostic techniques. Besides, we investigate a feature importance analysis that allows focused preventive treatments and identifies important factors contributing to early diagnosis, which assists medical professionals in making well-informed clinical decisions. This work underlines the importance of tailored feature selection and optimized algorithms in enhancing prediction reliability while increasing the use of machine learning in healthcare analytics. In summary, this work illustrates the potential of incorporating machine learning techniques into clinical settings where these instruments can have a great impact on health outcomes, enabling prevention and potentially reducing mortality due to heart disease.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

Heart disease is one of the leading causes of death in the world and has a high social and economic burden everywhere. On the other hand, according to the World Health Organization (WHO), heart-related diseases cause 17.9 million deaths a year or over 31% of the fatalities globally.[1] The fact that cardiovascular diseases (CVDs) are rising in both industrialized and developing nations as a result of factors including smoking, eating poorly, stress, and different kinds of leading sedentary lives emphasizes the need for early diagnosis and intervention. Despite the progress in medical diagnostics, the intricacy of symptoms and the heterogeneous nature of clinical data make it difficult to anticipate and diagnose heart disease. Therefore, there is a pressing need for accurate and dependable prediction tools that may help medical professionals plan treatments and diagnose patients early, which will eventually improve patient outcomes.[1]

The development of advanced data-gathering methods has made it possible for healthcare facilities worldwide to gather large volumes of medical data. For this being, the amount of this data is underutilized and does not have the level of analysis required to inform clinical judgment. Basically, techniques like data mining and machine learning have become viable alternatives because of their ability to analyze massive, complicated information and find information patterns.[2] According to precise predictive modeling, these technologies are providing healthcare practitioners with significant insight that facilitates well informed decision-making.

With their particular benefits in processing and categorizing medical data, machine learning techniques including logistic regression, decision trees, random forests, support vector machines, and neural networks have demonstrated considerable promise in the prediction of cardiac disease. ML models can offer valuable analyses of patient data by examining characteristics such as age, blood pressure, cholesterol, and exercise-induced angina. These studies can reveal patterns and connections that may not be immediately

obvious to medical practitioners. Furthermore, more complex predictive models have been developed thanks to developments in machine learning techniques including ensemble methods and feature selection algorithms, which have improved the models' accuracy and dependability in practical applications.

Despite the potential of machine learning for medical diagnosis, the main reasons that make heart disease prediction challenging are the huge volume, high complexity, and unpredictability of medical data. The process of model development is complex and involves careful feature selection, preparation of data, and tuning to optimum performance due to issues with noise, overfitting, and data imbalance. Moreover, model interpretability still remains a key challenge as medical practitioners require simple understandings to trust and use any ML model in clinical settings. This paper evaluates the appropriateness of various machine learning algorithms that can be used in heart disease prediction: logistic regression, support vector machines, decision trees, and neural networks.[3] By using methodical preprocessing, model assessment, and feature selection, the study identifies the most effective approaches for accurate, interpretable, and reliable prediction of heart disease risk.

This study intends to promote more predictive healthcare analytics by identifying crucial markers for early diagnosis and enabling proactive steps to enhance patient outcomes. A thorough review of machine learning techniques backed by feature significance analysis and model performance evaluation demonstrates the promise of ML in healthcare. Besides, the study adds to the expanding corpus of research supporting data-driven medical practices by examining the potential of several machine learning algorithms for the prediction of cardiac illness.[2] Additionally, the foundation is set for automated systems that could enhance conventional diagnostic techniques, save medical costs, and potentially save lives.

1.2 Motivation

The inspiration behind this study comes from This comprehensive study on heart disease prediction using machine learning techniques is motivated by the growing global burden of cardiovascular diseases (CVDs). Since heart disease continues to be the leading cause of death worldwide, early detection is crucial to improving patient outcomes and reducing mortality. The intricacy of heart disease symptoms and the massive volume of medical data generated daily often overwhelm standard diagnostic methods, notwithstanding their usefulness. Because machine learning analyzes complicated data and finds hidden patterns that traditional methods would miss, it then offers a useful solution.

1.3 Rationale of the Study

The study on machine learning algorithms for heart disease prediction is justified by the pressing need to improve early detection and prevention of cardiovascular diseases (CVDs), the world's leading cause of death. The intricacy of cardiac disease and the amount of unstructured medical data gathered usually make conventional diagnostic methods ineffective. Machine learning is an effective solution because it may identify trends and risk factors that are hard to detect using conventional techniques. The purpose of this research is to investigate and contrast different machine learning algorithms to determine which models are best for predicting cardiac disease, improving diagnostic precision, and assisting in clinical decision-making.

1.4 Research Questions

RQ1: Which machine learning algorithms are most effective for accurately predicting heart disease?

RQ2: How do demographic factors, lifestyle choices, and medical history influence the predictive accuracy of machine learning models in heart disease diagnosis?

RQ3: What are the key challenges in applying machine learning to heart disease prediction?

1.5 Expected Output

- Performance measures such as accuracy, sensitivity, and specificity are provided to identify the most accurate machine-learning models for the prediction of heart disease.
- Insights into the influence of important risk factors, such as demography, lifestyle, and medical history, on heart disease prediction models.
- A thorough analysis of algorithms to identify the best methods for predicting cardiac disease (e.g., logistic regression, random forests, support vector machines).

1.6 Report Layout

In Chapter 1, the introduction, objectives, and key research inquiries of the study are outlined. In Chapter 2, concise synopses of the literature review are provided. In Chapter 3, the proposed methodology is described in detail. In Chapter 4, the experimental outcomes of the paper are described and examined. The fifth chapter discusses the sustainability plan, societal and environmental impacts, and ethical considerations. The sixth chapter concludes the present investigation and outlines a strategy for subsequent endeavors.

CHAPTER 2
BACKGROUND

2.1 Preliminaries/Terminologies

A few fundamental ideas and concepts are necessary to comprehend this extensive study on the prediction of heart disease using machine learning approaches. Heart disease encompasses a variety of disorders that impact the heart, including cardiomyopathy, arrhythmias, and coronary artery disease. Computers can recognize patterns in data thanks to a subset of artificial intelligence called machine learning (ML). Predictive analysis in healthcare regularly makes use of popular machine learning (ML) algorithms such as random forest, support vector machines (SVM), neural networks, and logistic regression. While unsupervised learning finds hidden patterns in unlabeled data, supervised learning uses labeled data to train a model. Accuracy, sensitivity, specificity, and F1-score are important performance indicators that evaluate how well machine learning models predict cardiac disease.

2.2 Related works

Heart disease prediction, using machine learning algorithms, has become a highly investigated subject in recent years because of its promise in improving diagnostic and treatment outcomes. Various studies with different degrees of effectiveness have been conducted using several machine learning algorithms to predict cardiovascular disease. These algorithms are compared based on certain performance measures that help determine their effectiveness in real-world medical applications. The performances include accuracy, sensitivity, specificity, and F-1 score.[4]

In the present study, logistic regression was performed-one of the most popular algorithms for binary classification problems such as heart disease prediction-with 85% accuracy. This is a widely adopted method for the prediction of outcome variables based on risk factors such as age, blood pressure, cholesterol, among other medical history features. This is because of its ease and interpretability. The performances could be restricted by the

feature-linearity assumption that it possesses, which may not always stand in complicated medical datasets.[5]

With an 80% accuracy rate, the Decision Tree classifier provides a tree-like decision model that helps display decision rules and feature relevance. Decision trees do not perform as well as other methods because they are prone to overfitting, especially with complicated datasets like those used to forecast heart disease.[6] However, with an accuracy of 88%, Random Forest, an ensemble method based on decision trees, performed the best in this study. By building many decision trees on subsets of data and aggregating their forecasts, Random Forest reduces the over-fitting problem and improves the accuracy and robustness of the predictions.

Nevertheless, being a simple classifier, Gaussian Naïve Bayes has turned in an accuracy of 78% compared to the rest. The Naïve Bayes classifier assumes independence between features, which in reality is not true in the case of some datasets; many risk factors for heart disease are related to each other. However, it is computationally efficient and performs well when speed is prioritized over accuracy or with smaller datasets.[2] Previous studies have also looked into the application of these algorithms for the prediction of heart disease. The findings of this study are in line with Rani and Sivakumar's (2020) evaluation of several machine learning models for the prediction of heart disease, which revealed that Random Forest and Logistic Regression were two of the top-performing algorithms. Similarly, Srinivas and Rani (2019) emphasized the importance of ensemble methods like Random Forest and Light GBM in improving prediction accuracy by reducing over-fitting and facilitating reliable management of complex data relationships.[7]

Heart disease prediction requires not only algorithm performance but also feature selection and processing stages. Rajakumar et al. (2021) discovered that machine learning model's predictive accuracy may be greatly enhanced by carefully choosing features and concentrating on clinically relevant characteristics like blood pressure, cholesterol and lifestyle factors. This is consistent with the methods used in the current work, where feature selection and preprocessing ere key components in increasing the accuracy of models such as Random Forest and Light GBM.[3]

Table 2.2 Comparison of Model Accuracy Across Previous Studies

Reference	Algorithms Used	Best Accuracy	Contribution
[1]	• Naïve Bayes • Support Vector Machine • K – Nearest Neighbour	• 83.49% • 92.1% • 87.5%	• Based on Bayes' theorem, ideal for fast, independent tasks. • A robust algorithm handling non-linear, high-dimensional data. • Instance-based learning classifies by nearest neighbors' majority.
[1]	• Logistic regression	• 85%	• Binary classification predicts events and explains features.
[3]	• Naive Bayes (NB	• 78%	• Efficient for dependent datasets, often a baseline model.
[4]	• Naive Bayes • Decision Trees • ANN	• 86.53% • 89% • 85.53%	• Data mining framework for improved insights • Hierarchical structure for decisions, aiding interpretability. • Complex data, suitable for deep learning tasks.

To sum up, when comparing different machine learning methods for predicting heart disease, ensemble approaches like Random Forest and boosting techniques like LightGBM typically perform better than more straightforward models like Logistic Regression and Naive Bayes.[8] The study's findings are consistent with other studies, demonstrating the value of ensemble learning techniques for handling the intricate, non-linear interactions seen in medical datasets. These algorithms' outputs not only help to increase the accuracy of heart disease predictions but also advance the broader objective of leveraging machine learning to enhance clinical judgment and patient outcomes.

2.3 Comparative Analysis and Summary

The present research employs a comparative analysis of machine learning algorithms that have been previously used in the prediction of cardiac events including but not limited to logistic regression, K nearest neighbors, decision trees, and support vector machines. Other performance measures that were utilized in the evaluation of each model included the F1 score, accuracy, and precision. Even when data from the real world was used, the model accuracy levels failed to exceed eighty percent which encouraged the use of Kaggle data in order to enhance the diversity of the data. This approach improved the dataset, making it possible to make more accurate and generalized predictions of heart disease risk.

2.4 The Problem's Scope

Heart disease continues to be the world's largest cause of mortality, posing serious problems for healthcare systems. To meet the increased need for early diagnosis and prevention, the purpose of this study is to investigate how machine learning algorithms can accurately predict the development of heart disease. The study’s main objective is to evaluate the prediction performance of several algorithms, including random forest, SVM, LightGBM, and logistic regression. It also draws attention to the ways that prediction accuracy is hampered by shortcomings in data pretreatment, noise reduction, and model interpretability.

2.5 Challenges

There are various difficulties with predicting heart disease with machine learning. Firstly, a great deal of preprocessing and feature selection are necessary to provide accurate predictions because the medical data used is frequently big, heterogeneous, and noisy. Another prevalent problem is overfitting, particularly with sophisticated models like decision trees. Furthermore, model performance may be impacted by class imbalance in heart disease datasets, which could result in biased predictions. Clinical application of machine learning models, especially deep learning, is still hindered by their interpretability. Finally, it might be challenging to obtain regulatory permissions and clinician trust for the integration of machine learning techniques into healthcare systems.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Proposed Methodology/Applied Mechanism

In the research paper ‘A Comprehensive Study on Heart Disease Prediction Using Machine Learning Techniques. I outline a systematic way for predicting cardiac disease using various supervised machine learning methods. The first step in the procedure is gathering a dataset on heart disease, which contains several clinical characteristics, including age, blood pressure, cholesterol and other important metrics. The dataset goes through a rigorous preprocessing step in which features are uniformly normalized and standardized, missing values are imputed and categorical variables are encoded. To ensure that the models are trained on balanced data, methods such as SMOTE (Synthetic Minority Oversampling Technique) are also used to resolve class imbalance. After preprocessing multiple machine learning algorithms are implemented and compared. The first is logistic Regression, a statistical model commonly used for binary classification tasks like heart disease prediction. K-Neighbors classifier (KNN) is also applied, which classifies instances based on their proximity to known labeled points. Random Forest and Decision Trees have the advantage of ensemble learning to lower volatility. Lastly, Gaussian Naïve Bayes is used because it is easy to use and effective when working with small datasets and categorical features. Accuracy, precision, recall, F1-score, and confusion matrices are used to evaluate each algorithm's performance. A comprehensive evaluation of machine learning techniques for accurate disease detection is provided by further adjusting the top-performing model for the best heart disease prediction.

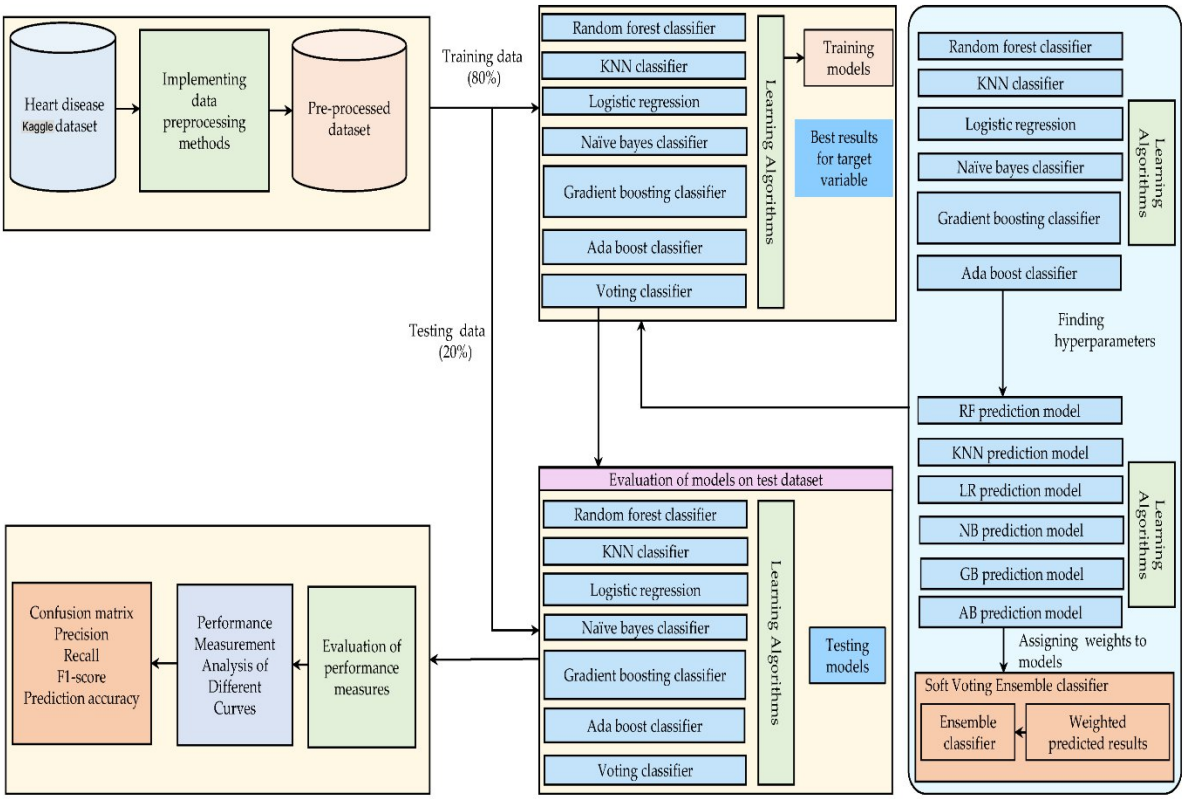


Fig. 3.1 Proposed Methodology

Figure: Process flow of the utilization of machine learning techniques to forecast cardiac disease. The raw data source is a Kaggle dataset. After cleaning and processing, the result will be a pre-processed dataset that is split into a training set (80%) and a test set (20%). Accordingly, in order to train the models and decide on the best strategy for the target variable, the following types of machine learning classifiers have been applied: Random Forest, K-Nearest Neighbors-KNN, Logistic Regression, Naive Bayes, Gradient Boosting, and AdaBoost; a Voting Classifier has also been used.[9]

After training, the prediction model for each classifier is developed, and the results are combined using a soft voting ensemble classifier, where the weighted predictions for each model contribute to the final decision. The model is then evaluated on the testing dataset using metrics such as a confusion matrix, precision, recall, F1-score, and prediction accuracy. Performance curves are also included for further research.

After training, prediction models are generated for all classifier, where weighted predictions from each model contribute to the final decision.[1] The models are then evaluated using the testing dataset through metrics such as confusion matrix, precision, recall, F-1 score and prediction accuracy along with performance curves for deeper analysis. This systematic approach ensures the comparison of model's performance and the selection of the most effective algorithm for heart disease prediction, leveraging ensemble learning to boost accuracy and reliability. The final step involves detailed performance measurement and analysis to derive insights and optimize prediction outcomes.

3.2 Data Collection Procedure/Dataset Utilized

The actual data obtained from East West Medical College has helped in the initial stages of the heart disease prognosis project. It was possible to obtain model accuracy, however most were unable to exceed 80% owing to feature variability and data size constraints. The models tested have reached stagnation at points above 80% and below 80% accuracy respectively even after going through the test phase, a number of supervised learning algorithms including support vector machines, logistic regression, and K-nearest neighbors have been deployed. In order to overcome this limitation and enhance the model's predicting ability, additional data were sourced from Kaggle, a site that provides such large quantities of information on heart disease.

The data collection method was an exhaustive and active process that involved different departments within the hospital. It included visits to the OT, where more private information about the patients could be collected, apart from the male and female wards. The base dataset for this study was collected during these visits when information was extracted from test reports and patient files. These files contain a variety of data which enabled the classification of heart disease cases based on various parameters such as blood pressure, cholesterol, and chest pain type. More robust training and customization were made possible by the additional data's introduction of a wider variety of patient profiles,

varied feature distributions, and more balanced class representations. Underrepresented age groups and missing values in important features were among the gaps in the original data that the merged dataset helped fill. The work enhanced the overall quality of the dataset by incorporating Kaggle data, allowing for more precise, broadly applicable models for the prediction of heart disease.

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

Fig. 3.2 Images of data set reports with different patients.

The image presents a tabular dataset consisting of 5 rows and 13 columns, each representing a patient record. The columns are labeled as follows:

The dataset contains some important cardiac disease-related characteristics for every patient. Sex indicates the patient's gender (M for male, F for female), and Age indicates the patient's age in years. The type of chest pain experienced is indicated by ChestPainType, which can be classified as ATA, NAP, or ASY. Cholesterol shows cholesterol levels in mg/dl, while resting BP shows resting blood pressure in mm Hg. A binary feature called FastingBS indicates fasting blood sugar levels; 0 denotes levels less than 120 mg/dl and 1 denotes levels higher. Categories such as Normal, ST, and LVH are among the resting electrocardiographic data that RestingECG offers. ExerciseAngina indicates whether exercise-induced angina is present (Y for yes, N for no), and MaxHR logs the highest heart rate attained during exercise.

ST_Slope shows the slope of the ST segment during activity, while Oldpeak depicts the ST depression brought on by exercise in comparison to rest. Lastly, heart disease is a binary

variable that indicates if heart disease is present (1) or not (0). Possibly taken from a medical research project or a Kaggle competition, the dataset appears to be a subset of a bigger dataset. It seems to be a preprocessed and cleaned version of the original data that is appropriate for activities involving machine learning.

3.3 Model Selection

Several machine learning techniques can be taken into consideration for the prediction of heart disease: For binary classification, logistic regression yields results that are easy to understand. While SVM determines the best hyperplanes to divide classes, K-Nearest Neighbors classifies based on the majority class of nearest neighbors. For reliable predictions, the ensemble approach Random Forest mixes several decision trees. For big datasets, Naive Bayes is effective, assuming feature independence. Decision trees are used to generate hierarchical classification rules, and the ensemble approach XGBoost is well known for its exceptional accuracy and capacity to manage complex patterns. The choice of algorithm depends on a number of factors, including the size of the dataset, the intricacy of the characteristics, the intended interpretability, and the processing capacity.[9]

3.3.1 Decision Tree

A Decision Tree classifier's confusion matrix is displayed in the image to demonstrate how well it performed on a binary classification task. Trees for training samples of data D are constructed from high-entropy inputs using a simple and fast top-down recursive divide and conquer approach. Tree pruning is then applied to eliminate irrelevant samples from D. [D= Data set][10]

$$Entropy = -\sum_{j=1}^{m} p_{ij} \log_2 p_{ij} \qquad (1)$$

3.3.2 KNN Classifier

The picture shows a confusion matrix for a binary classification job using a K-Nearest Neighbors (KNN) classifier. It extract the knowledge based on the samples Euclidean distance function d(xi,xj) and the majority of k-nearest neighbors.[6]

$$d\left(x_{i,x_i,}\right) = \sqrt{\left(x_{i,1} - x_{j,1}\right)^2 + \cdots + \left(x_{i,m} - x_{j,m}\right)^2} \tag{2}$$

3.3.3 Support Vector Machine

Let the training samples having dataset Data={yi,xi};i=1,2,…,n where xi∈Rn represent the *i*th vector and yi∈Rn represent the target item. The linear SVM finds the optimal hyperplane of the form f(x)=wTx+b where w is a dimensional coefficient vector and b is a offset.[10] This is done by solving the subsequent optimization problem:

$$\begin{aligned} &Min_{w,b,\xi_i}\ \frac{1}{2}w^2 + C\sum_{i=1}^{n}\xi_i \\ &s.t.\ y_i\left(w^{\mathrm{T}}x_{\mathrm{i}} + b\right) \geq 1 - \xi_i, \xi_i \geq 0, \quad \forall_i \in \{1, 2, \ldots, m\} \end{aligned} \tag{3}$$

3.3.4 Random Forest

This ensemble classifier builds several decision trees and incorporates them to get the best result. For tree learning, it mainly applies bootstrap aggregating or bagging. For a given data, X={x1,x2,x3,…,xn} with responses Y={x1,x2,x3,…,xn} which repeats the bagging from b=1 to B . The unseen samples x′ is made by averaging the predictions ∑Bb=1fb(x′) from every individual trees on x′ [10]

$$j = \frac{1}{B}\sum_{b=1}^{B} fb\left(x'\right) \tag{4}$$

The uncertainty of prediction on these tree is made through its standard deviation,

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B} \left(fb(x') - \hat{f} \right)^2}{B-1}} \quad (5)$$

3.3.5 Naive Bayes

This learning model applies Bayes rules through independent features. Every instance of data D is allotted to the class of highest subsequent probability. The model is trained through the Gaussian function with prior probability P(Xf)=priority∈(0:1)[10]

$$P\left(X_{f1}, X_{f2}, \ldots, X_{f_n} | c\right) = \prod_{i=1}^{n} P\left(X_{fi} | c\right)$$

$$P\left(X_f | c_i\right) = \frac{P\left(c_i | X_f\right) P\left(X_f\right)}{P\left(c_i\right)} \quad c \in \{benign,\ malignant\} \quad (6)$$

3.4 Features Implementation

The heart disease dataset includes various attributes such as Age, Resting Blood Pressure (RestingBP), Cholesterol, Fasting Blood Sugar (FastingBS), MaxHR, and Oldpeak. Additionally, the distribution of chest pain types and age is a significant concern in this context. A variety of crucial characteristics that are important in evaluating cardiovascular health are included in the heart disease dataset. Age, a major demographic factor linked to the risk of heart disease, resting blood pressure, which measures the force applied to blood vessel walls during heart relaxation, cholesterol levels, which indicate blood lipid levels, fasting blood sugar, which indicates glucose metabolism, maximum heart rate attained during exercise, which measures cardiovascular fitness, and ST depression, an ECG measurement that reflects changes in heart electrical activity during exercise, are some examples of these characteristics.

3.4.1 People with/without heart disease

This graphic shows two bar charts that compare important health indicators between people with and without heart disease. The metrics displayed consist of:

- Age: The average age of those without heart disease is 50.55, whereas the average age of people with heart disease is 55.90.
- Resting Blood Pressure (RestingBP): The average RestingBP of people with heart disease is 134.19, which is somewhat higher than the 130.18 recorded for people without heart disease.
- Cholesterol: The average cholesterol levels of those without heart disease are higher (227.12) than those of people with heart disease (175.94).
- Fasting Blood Sugar (FastingBS): The average FastingBS for people with heart disease is 0.33, which is greater than 0.11 for people without the condition.
- The maximum heart rate attained, or MaxHR is higher in those without heart disease (148.15) than in people with illness (127.66).
- Oldpeak: People with heart disease have greater Oldpeak averages (1.27) than people without heart disease (0.41), which reflects ST depression brought on by exercise as opposed to rest.
- An indicator of heart disease For those who have heart disease, the value is 1.00, and for those who do not, it is 0.00.

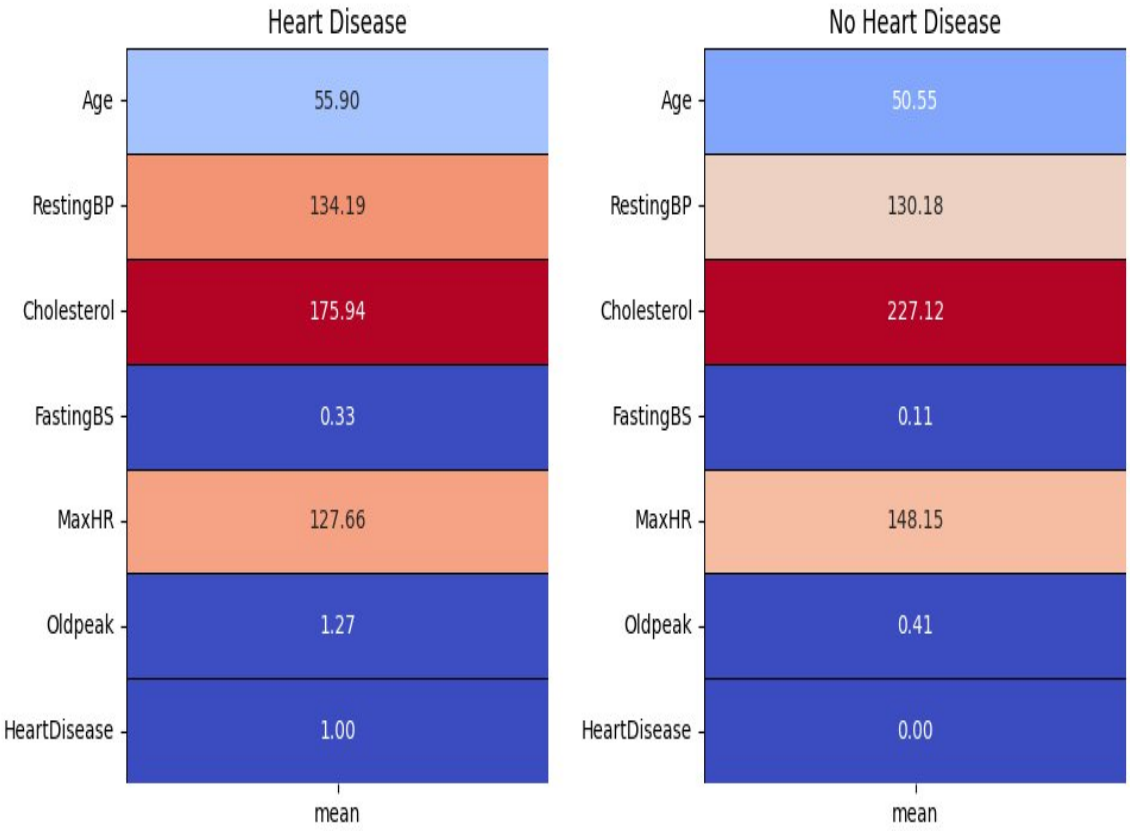


Fig. 3.4.1 Full view of individual indicators

By highlighting significant variations in these measurements between the two groups, this graphic comparison sheds light on the variables that may be associated with heart disease. It is a helpful resource for comprehending how these characteristics might be used by machine learning models to forecast cardiac disease.

3.4.2 Distribution of chest pain types (ATA, NAP, ASY, TA)

The bar chart shows how different types of chest discomfort (ATA, NAP, ASY, and TA) are distributed by sex (male and female). The most prevalent type of chest pain in men is asymptomatic (ASY) (blue bars), which is followed by NAP and ATA. The least common type is TA. Females are less likely to experience any type of chest discomfort (orange bars), with ASY being the most prevalent. The data indicates that the number of ASY chest pain patients varies significantly between males and females, with a higher number for males overall.

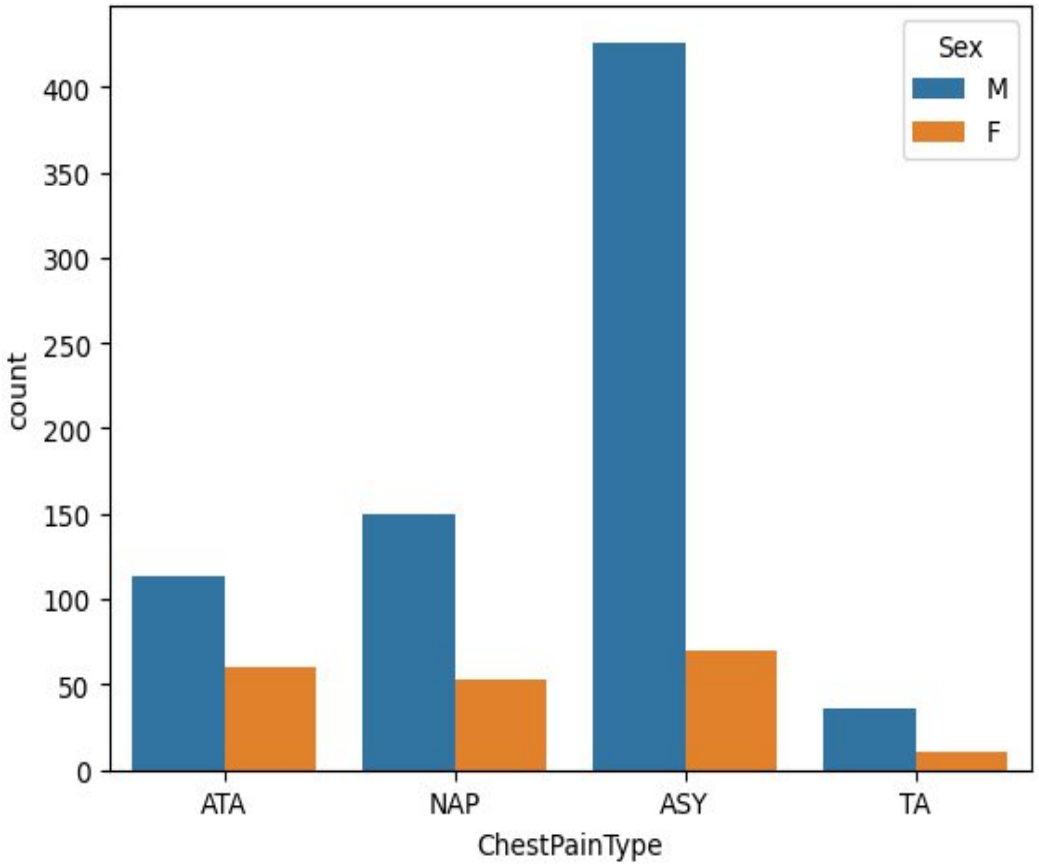


Fig. 3.4.2 Distribution of chest pain types (ATA, NAP, ASY, TA)

3.4.3 Cholesterol (mmHg) vs MaxHR

The connection between MaxHR (highest heart rate attained) and cholesterol (measured in mg/dL) for people with and without heart disease is shown in this scatter plot. With orange denoting the existence of cardiac disease (labeled as 1) and blue denoting its absence (labeled as 0), each point represents a patient.

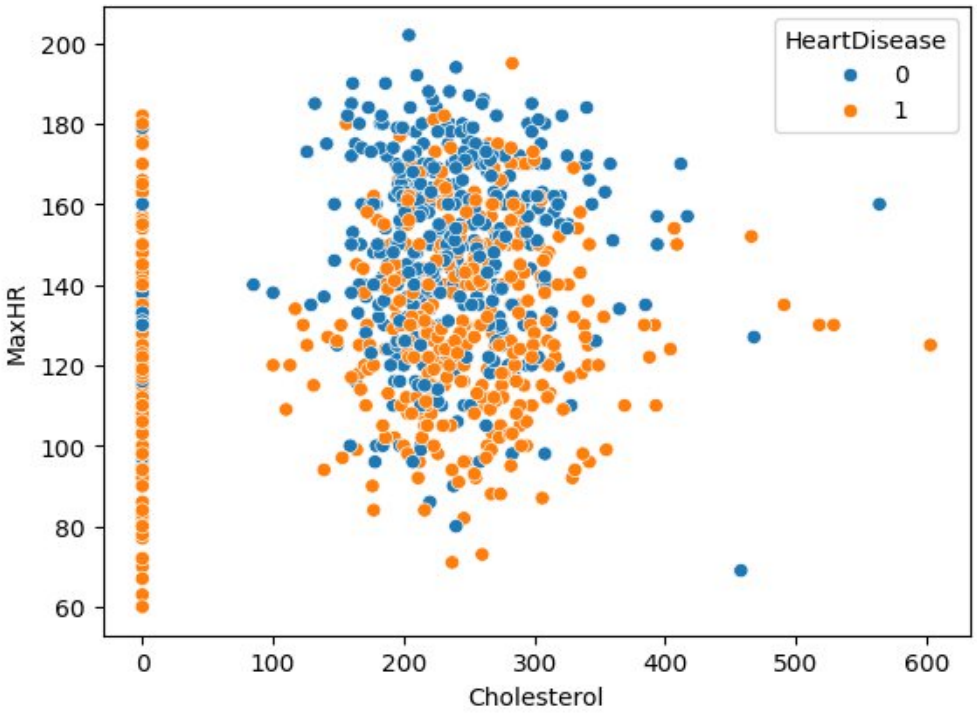


Fig. 3.4.3 Cholesterol (mmHg) vs MaxHR

The figure displays a concentration of points in the region of 120–160 bpm for MaxHR values and 100–300 mg/dL for cholesterol levels. Since people with high and low cholesterol might have different maximum heart rates, there is no obvious linear relationship between cholesterol levels and MaxHR. Furthermore, a broad range of cholesterol and MaxHR values are present in both heart disease and non-heart disease patients, while elevated cholesterol levels exceeding 300 mg/dL are very uncommon. This distribution implies that although MaxHR and cholesterol are significant characteristics, their association would not be enough to predict heart disease in the absence of additional variables.

3.4.4 Heart Disease Prevalence by Age

The age distribution of participants in the heart disease study is depicted by this histogram, which counts participants in various age ranges. A smooth curve overlay on the plot indicates a normal distribution with the 50–60 age range at its center.

This distribution suggests that the majority of research participants are between the ages of 45 and 65, with a peak around 55, when it comes to the prediction of heart disease. This pattern would suggest that heart disease cases or risk factors are more common in this age group. Given that age is frequently a significant indicator of heart disease risk, machine learning models benefit from an understanding of the age distribution.

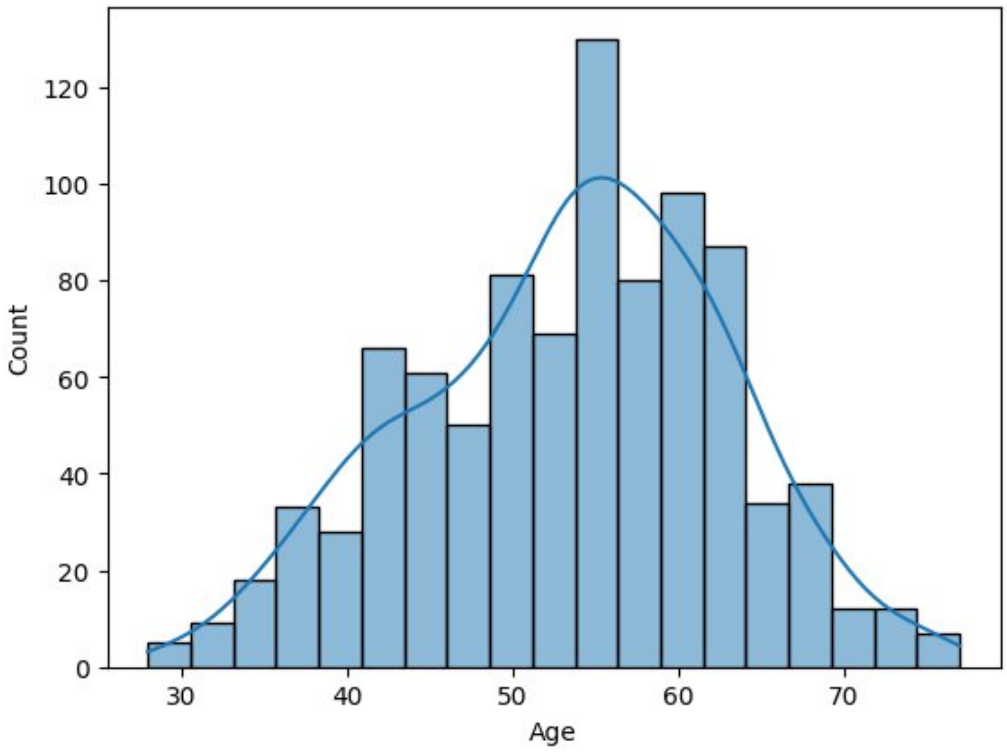


Fig. 3.4.4 Heart Disease Prevalence by Age

3.4.5 Cholesterol levels by chest pain Type heart disease

The picture is a scatter plot that shows how gender, cardiac disease status, and systolic and diastolic blood pressure are related. Resting systolic blood pressure (80–120 mmHg) and diastolic blood pressure (60–80 mmHg) are displayed on the x and y axes, respectively. Males are represented by blue points, while females are represented by orange points. The dots' sizes represent the presence or absence of heart illness: larger dots denote heart disease (1), whereas smaller ones show no heart disease (0). There isn't much of a correlation between the variables.

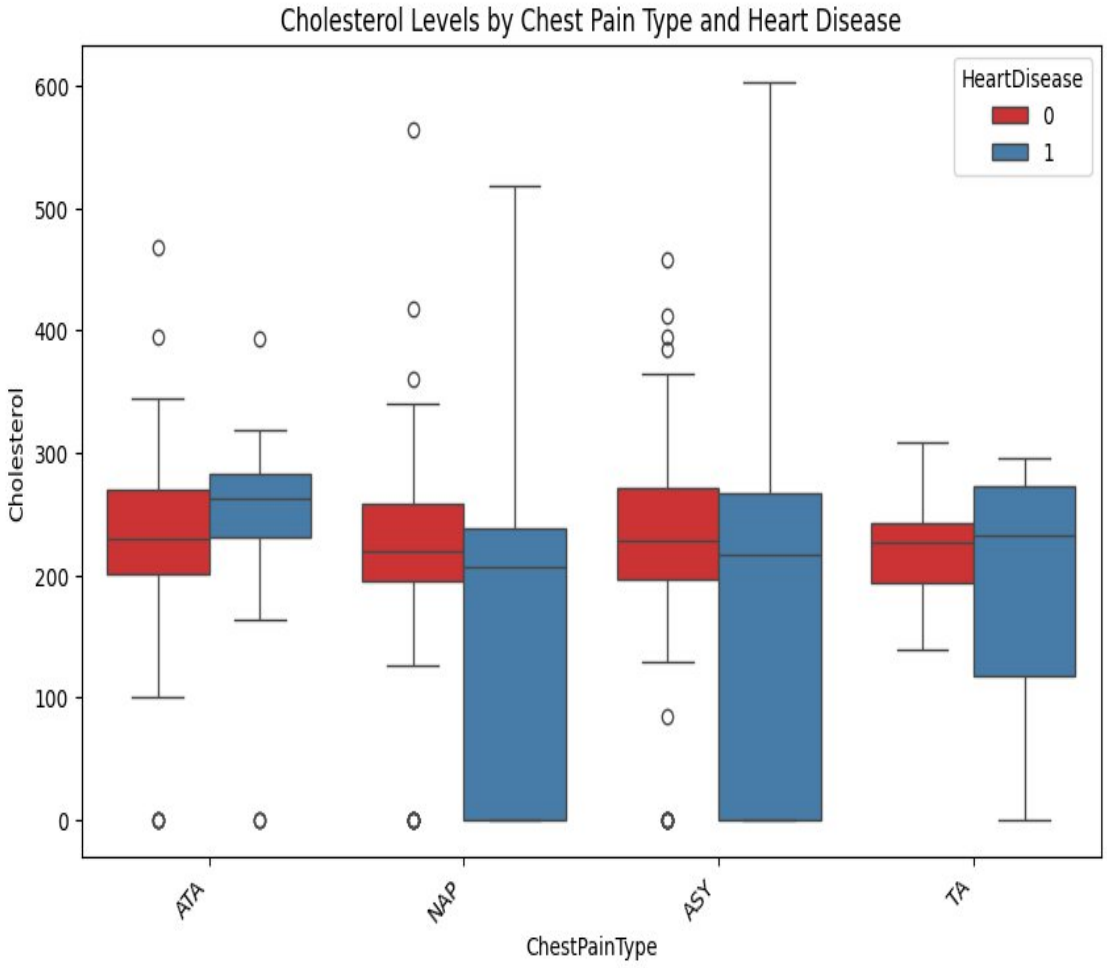


Fig. 3.4.5 Cholesterol levels by chest pain type heart disease

3.4.6 Gender (sex) vs ChestPainType

Gender Disparity: The heatmap shows that the distribution of chest pain categories in men and women differs significantly.

The most prevalent type of chest discomfort, for both men and women, is 'ASY' (Asymptomatic). But 'NAP' (Non-Anginal Pain) seems to affect women a little more often than it does men.

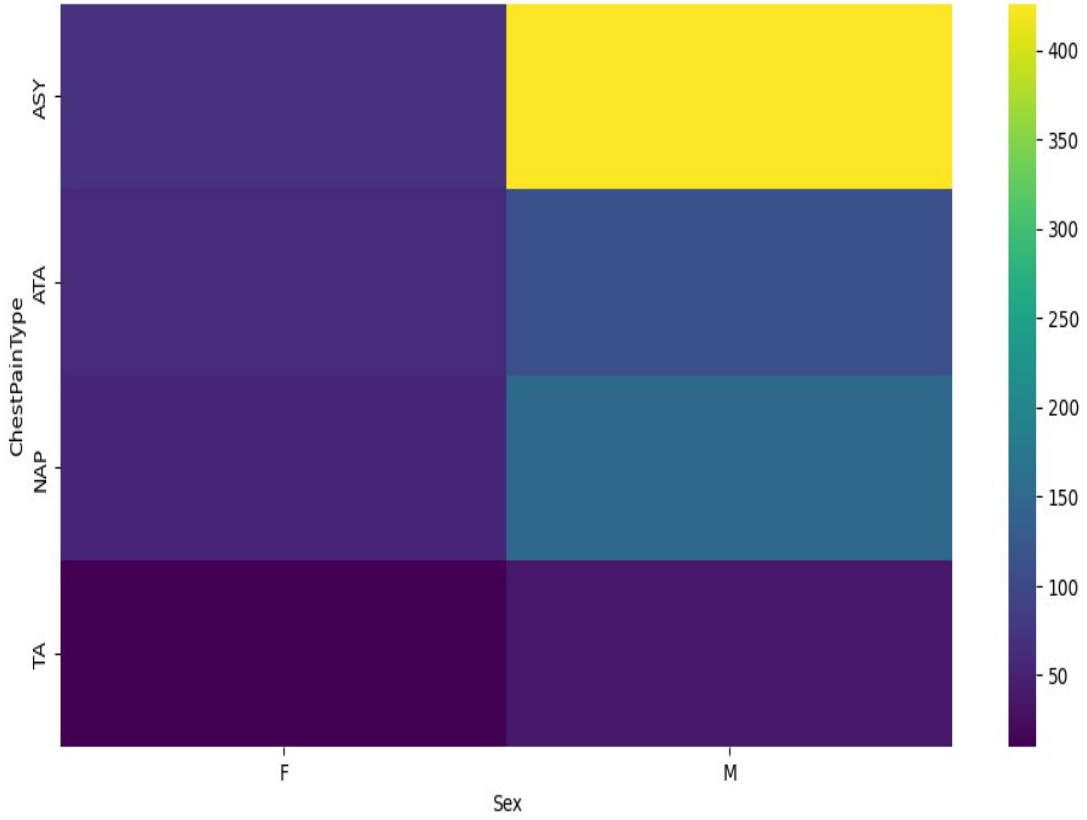


Fig. 3.4.6 Gender vs ChestPainType

All things considered, the heatmap offers insightful information on the relationship between gender and the type of chest pain when it comes to predicting heart disease. Further research is needed to fully understand the implications of these findings and develop more effective heart disease prevention and treatment strategies.

3.5 Statistical Analysis

Histograms showing the statistical distribution of features in a dataset on heart disease are displayed in the image. Age, which has a normal distribution and peaks between the ages of 50 and 60, and sex, which has a higher percentage of men, are important characteristics. Both ST_Slope and ChestPainType exhibit clear frequency patterns and are categorical. Whereas cholesterol has a wider range, continuous variables with normal-like distributions include resting blood pressure and maximum heart rate. The old peak is right-skewed, but category 0 dominates binary features like FastingBS and ExerciseAngina. Although it leans somewhat toward the lack of disease (0), the target variable, disease, exhibits a balanced distribution. For modeling purposes, these insights aid in identifying distribution trends and feature variability.[11]

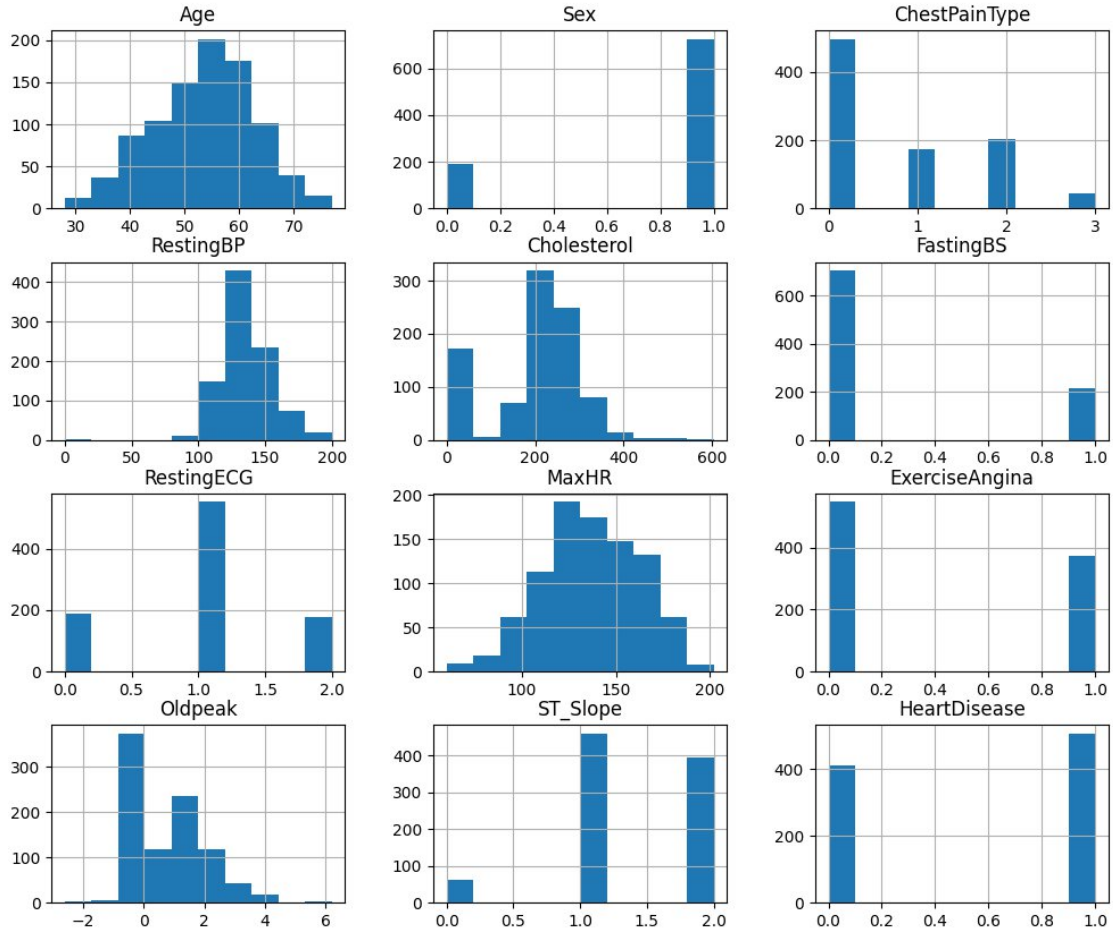


Fig. 3.5 Statistical Analysis

3.6 Implementation Requirements

- Hardware: An AMD Ryzen 5/7 or Intel i7 laptop with a high-performance CPU, SSD storage for faster data processing, and at least 8GB of RAM.
- Data Collection: Reliable medical sources, including patient records or hospital databases, are used to collect accurate data on cardiac disease.
- Software: Google Colab, Python, and libraries (such as TensorFlow, Pandas, and Scikit-Learn).
- Algorithms: supervised learning techniques like Random Forest, K-Nearest Neighbors, Decision Trees, Gradient Boosting, and Logistic Regression.
- Evaluation Tools: F1-score, recall, accuracy, and precision metrics.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Results of Segmentation

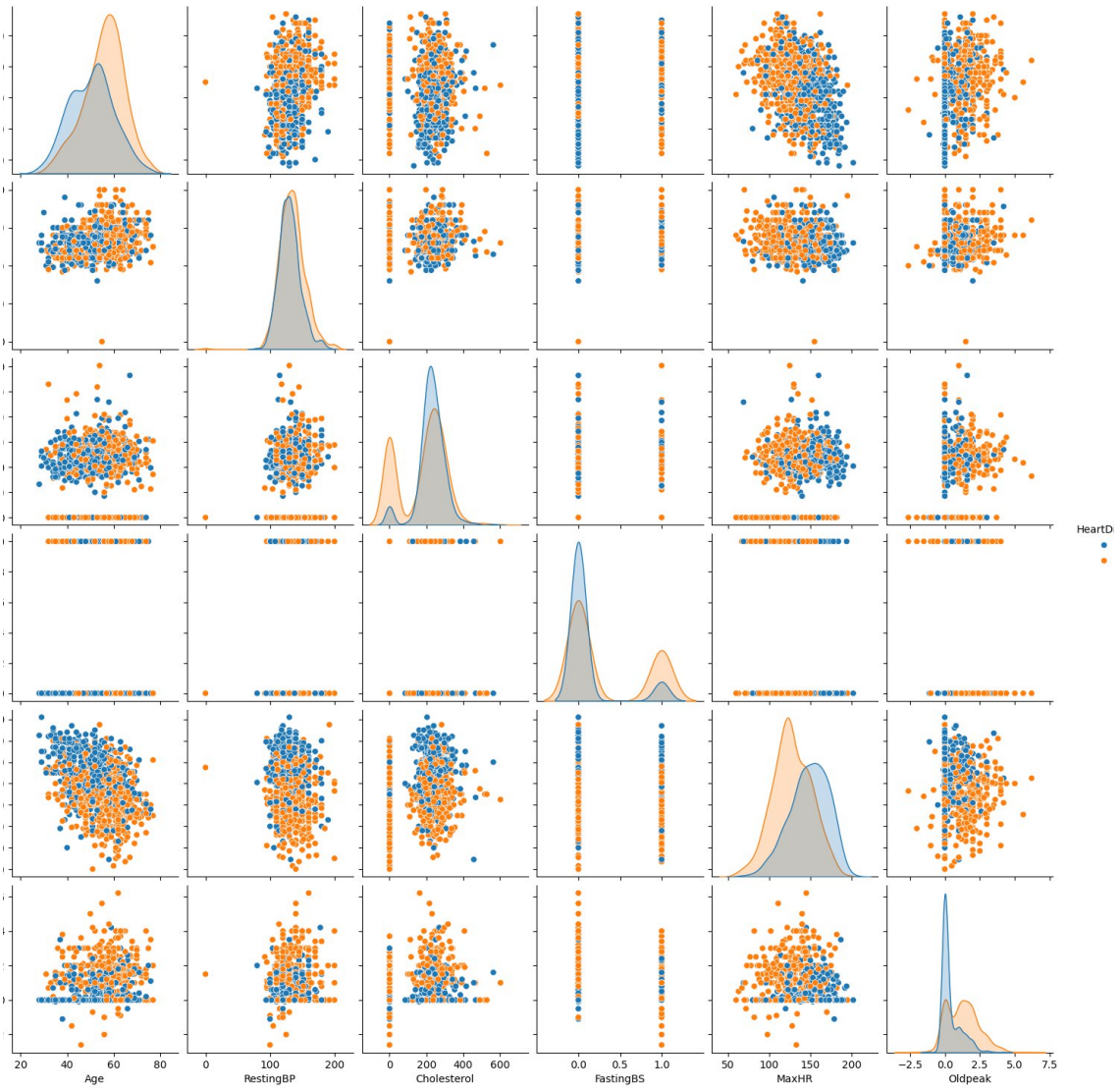


Fig. 4.1 Images with predicted masking Age, RestingBP, Cholesterol, FastingBS, MaxHR, Oldpeak comparison.

Make a series of pictures that display masked comparisons of important predictors such as age, resting blood pressure (RestingBP), cholesterol, fasting blood sugar (FastingBS), maximum heart rate (MaxHR), and old peak for my research on heart disease prediction

using machine learning approaches. With specifics obscured to preserve individual names, each figure should show the distribution or impact of these predictors on the presence versus absence of heart disease. This improves model interpretability while protecting data privacy by highlighting differences in predictor influence. To show how predictions differ for various masked attributes, comparisons can be made using density plots, box plots, or histograms.

4.2 Evolution Methods

After segmentation, the segmented images are then predicted using a transfer learning model. The evaluation metrics include a confusion matrix that provides measures such as accuracy, precision, recall, and F1 score. In this context, a true positive is a case where a positive instance is correctly identified as positive. A false positive occurs when a negative instance is mistakenly labeled as positive. Conversely, a false negative happens when a positive instance is incorrectly interpreted as negative. Lastly, a true negative instance is correctly identified as negative. These metrics help assess the model's performance, providing insights into its reliability and precision in detecting true cases of interest.[12]

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7$$

$$Precision = \frac{TP}{TP+FP} \quad (8$$

$$Recall = \frac{TP}{TP+FN} \quad (9$$

$$F1\ Score = 2 \times \frac{precision \times recall}{precision + recall} \quad (10)$$

4.3 Experimental Results & Analysis

The table displays the accuracy of seven different supervised learning algorithms that are used to forecast cardiac disease. With the highest accuracy of 87.50% of the tested models, the Support Vector Machine (SVM) is the most accurate. The Random Forest model comes in second with an accuracy of 86.96%, indicating that it has a high capacity to identify intricate patterns in the data. With a competitive accuracy of 85.87%, Naive Bayes and XGBoost demonstrate that, despite their different approaches (one based on boosting techniques, the other on probabilistic methods), they perform similarly in terms of total predictive capacity. K-Nearest Neighbors (KNN) and Logistic Regression have somewhat lower accuracies (85.33%), but the Decision Tree performs worse (83.15%), which may be a sign of its propensity to overfit the data.

Table 4.3 shows the performance metrics of various supervised learning models on a heart disease dataset.

Model	Test Accuracy (%)	Precision (%)	recall (%)	f1-score(%)	support(%)
Logistic Regression	85.33	80	87	83	77
K-Nearest Neighbors	85.33	80	87	83	77
Support Vector Machine	87.50	90	88	89	107
Random Forest	86.96	84	86	85	77
Naive Bayes	85.87	80	88	84	77
Decision Tree	83.15	79	82	80	77
XGBoost	85.87	90	85	88	107

With only slight performance variations, these algorithms appear to be rather well-suited to the heart disease dataset, as evidenced by the close clustering of accuracy scores, which are primarily above 85%. Nonetheless, the leading accuracy of the support vector machine may indicate that it is especially good at managing the feature space of the dataset, maybe as a result of its capacity to optimize the margin between classes.

Apart from K-nearest neighbor and logistic regression, which have somewhat lower accuracy, it may suggest that these two models are poor in detecting nonlinear relationships in the data. The rather weak performance of the decision trees when compared to Random Forest and XGBoost ensemble methods shows the importance of ensemble techniques in improving both the accuracy and stability of the medical data predictive models.

4.3.1 Logistic Regression

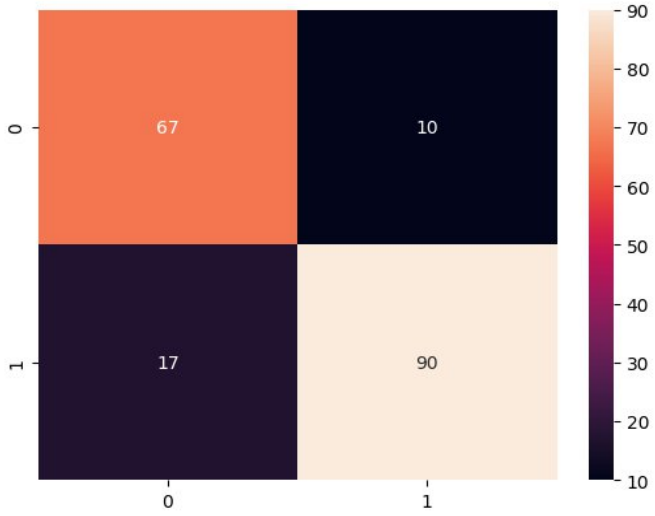
Training Logistic Regression...

Accuracy for Logistic Regression: 85.33%

Confusion Matrix:

[[67 10]

[17 90]]



Ten false positives, seventeen false negatives, 90 true positives, and 67 true negatives are displayed in the logistic regression model's confusion matrix. Now this shows that, although occasional misclassifications still happen, the model performs well overall, with a fairly balanced classification accuracy across both group.

Fig. 4.3.1 Logistic Regression

Classification Report:

	precision	recall	f1-score	support
0	0.80	0.87	0.83	77
1	0.90	0.84	0.87	107
accuracy			0.85	184
macro avg	0.85	0.86	0.85	184
weighted avg	0.86	0.85	0.85	184

4.3.2 K-Nearest Neighbors

Training K-Nearest Neighbors...

Accuracy for K-Nearest Neighbors: 85.33%

Confusion Matrix:

[[67 10]

[17 90]]

With 90 true positives, 10 false positives, 17 false negatives, and 67 true negatives, this confusion matrix for the K-Nearest Neighbors model demonstrates outstanding classification performance with a small number of misclassifications in both classes.

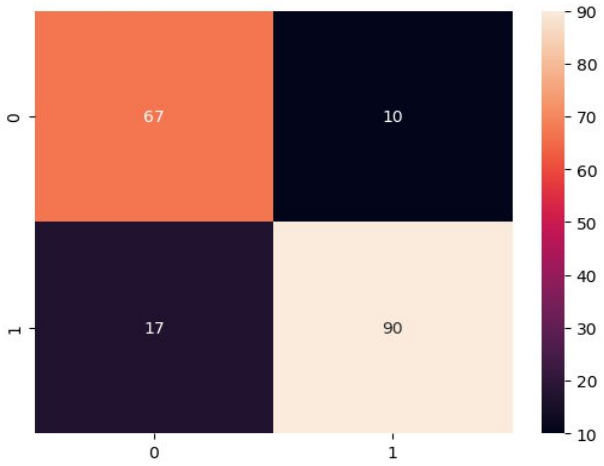


Fig. 4.3.2 K-Nearest Neighbors

Classification Report:

	precision	recall	f1-score	support
0	0.80	0.87	0.83	77
1	0.90	0.84	0.87	107
accuracy			0.85	184
macro avg	0.85	0.86	0.85	184
weighted avg	0.86	0.85	0.85	184

4.3.3 Support Vector Machine (SVM)

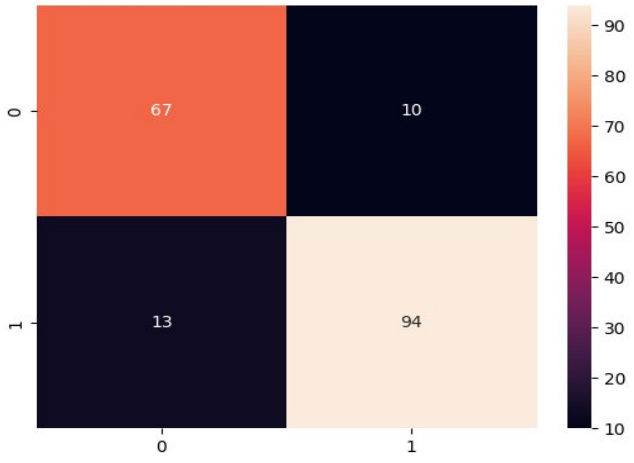
Training Support Vector Machine...

Accuracy for Support Vector Machine: 87.50%

Confusion Matrix:

[[67 10]

[13 94]]



With 67 true negatives, 10 false positives, 13 false negatives, and 94 true positives, the SVM model's confusion matrix demonstrates excellent classification performance with few errors.

Fig. 4.3.3 Support Vector Machine (SVM)

Classification Report:

	precision	recall	f1-score	support
0	0.84	0.87	0.85	77
1	0.90	0.88	0.89	107
accuracy			0.88	184
macro avg	0.87	0.87	0.87	184
weighted avg	0.88	0.88	0.88	184

4.3.4 Random Forest

Training Random Forest...

Accuracy for Random Forest: 86.96%

Confusion Matrix:

[[66 11]

[13 94]]

This confusion matrix for the Random Forest model shows 66 true negatives, 11 false positives, 13 false negatives, and 94 true positives, demonstrating effective classification with minor errors.

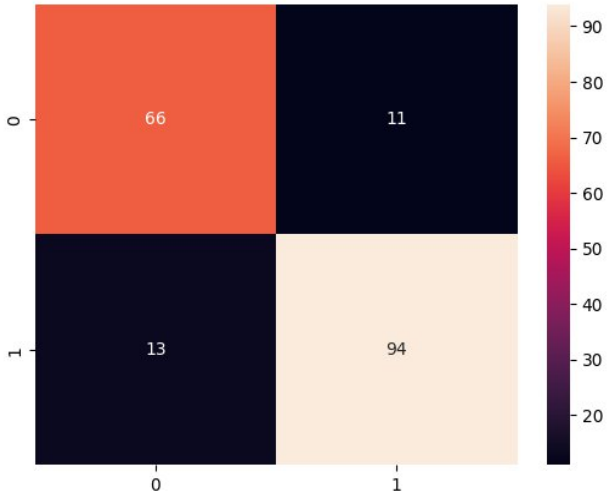


Fig. 4.3.4 Random Forest

Classification Report:

	precision	recall	f1-score	support
0	0.84	0.86	0.85	77
1	0.90	0.88	0.89	107
accuracy			0.87	184
macro avg	0.87	0.87	0.87	184
weighted avg	0.87	0.87	0.87	184

4.3.5 Naive Bayes

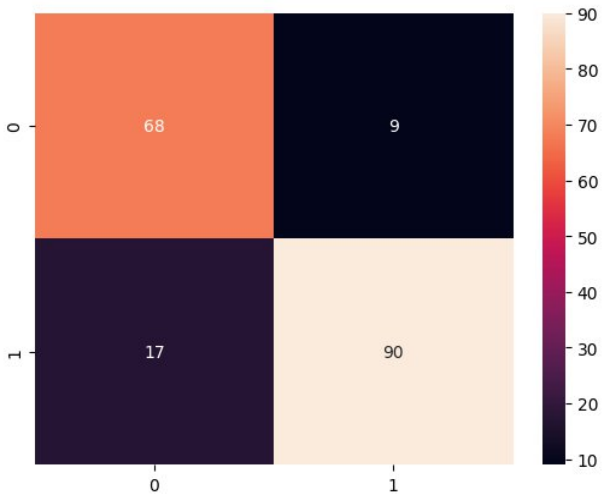
Training Naive Bayes...

Accuracy for Naive Bayes: 85.87%

Confusion Matrix:

[[68 9]

[17 90]



The confusion matrix shows a Naive Bayes model's performance, with 68 true positives, 9 false negatives, 17 false positives, and 90 true negatives.

Fig. 4.3.5 Naive Bayes

Classification Report:

	precision	recall	f1-score	support
0	0.80	0.88	0.84	77
1	0.91	0.84	0.87	107
accuracy			0.86	184
macro avg	0.85	0.86	0.86	184
weighted avg	0.86	0.86	0.86	184

4.3.6 Decision Tree

Training Decision Tree...[13]

Accuracy for Decision Tree: 83.15%

Confusion Matrix:

[[63 14]

[17 90]]

The confusion matrix shows a Decision Tree model's performance, with 63 true positives, 14 false negatives, 17 false positives, and 90 true negatives.

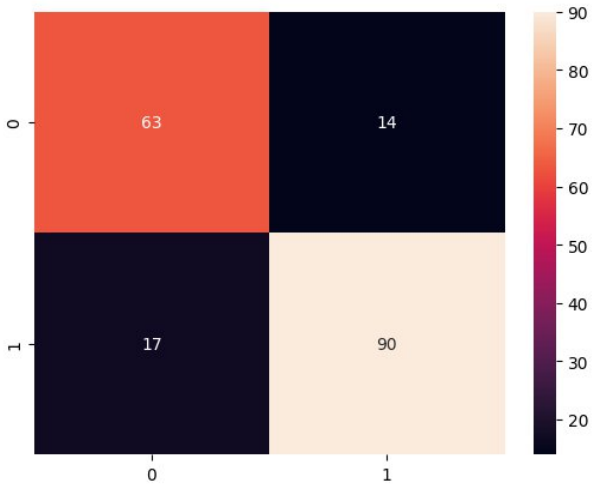


Fig. 4.3.6 Decision Tree

Classification Report:

	precision	recall	f1-score	support
0	0.79	0.82	0.80	77
1	0.87	0.84	0.85	107
accuracy			0.83	184
macro avg	0.83	0.83	0.83	184
weighted avg	0.83	0.83	0.83	184

4.3.7 XGBoost

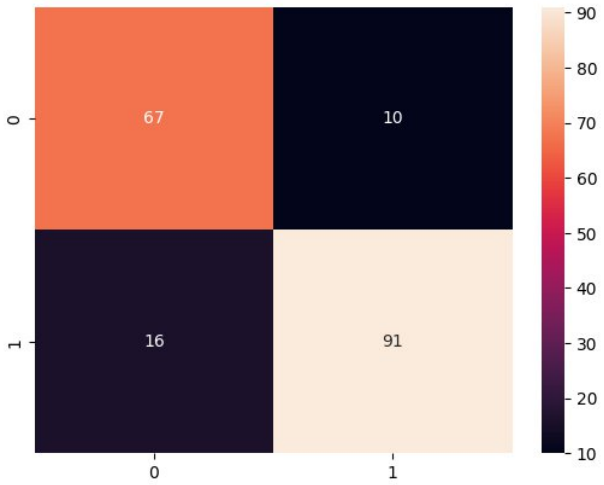
Training XGBoost...

Accuracy for XGBoost: 85.87%

Confusion Matrix:

[[67 10]

[16 91]]



The confusion matrix shows an XGBoost model's performance, with 67 true positives, 10 false negatives, 16 false positives, and 91 true negatives.

Fig. 4.3.7 XGBoost

Classification Report:

	precision	recall	f1-score	support
0	0.81	0.87	0.84	77
1	0.90	0.85	0.88	107
accuracy			0.86	184
macro avg	0.85	0.86	0.86	184
weighted avg	0.86	0.86	0.86	184

4.4 Confusion Matrix

This confusion matrix analyzes the efficiency of the heart disease prediction model. The rows of this 2x2 square grid represent the actual output class, whereas the columns are for the output class obtained as a result of deciding by the model. The overall model correctly identifies three cases where there were no heart illnesses, whereas five instances of heart disease are true positive. On the other hand, it is prone to some errors, specifically, it misclassifies one instance of heart disease as an instance of no disease and vice-versa, false throat. These being the worst false negatives in a clinical setting since it might delay diagnosis and treatment. However, the model can still be further improved by feature engineering, hyper-parameter tuning, trying different algorithms, or gathering more data. The costs of false positives and false negatives should be minimized in reverse proportion as their implications vary in severity by context. Accurately robust models for heart disease prediction are possibly developed when focusing on a confusion matrix and rectification of its weaknesses.

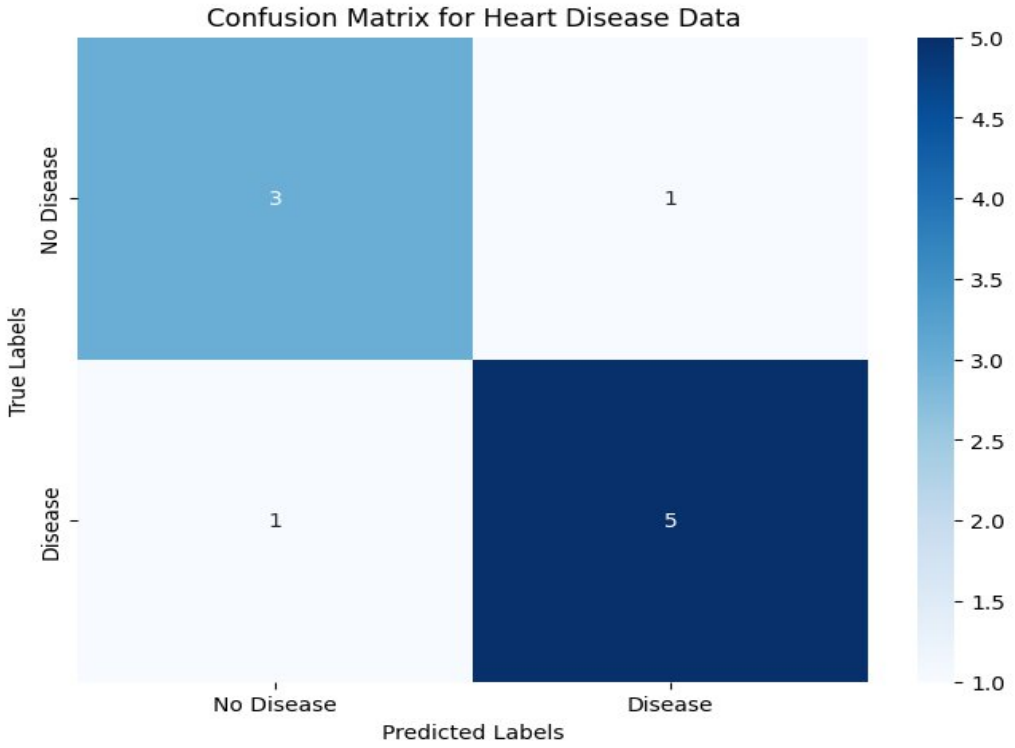


Fig: 4.4 Confusion matrix

4.5 Discussion

This study's discussion outlines that different machine learning methods have different performances in predicting cardiac disease and utilizes metrics such as the accuracy, the precision, the recall, and the F1 score to measure the effectiveness of the model in question. The real world data that was incorporated from East West Medical College was helpful, but model accuracy levels were only around 80 percent due to limitations in data information such as the features and the size of the sample incorporated. This limitation led to the addition of a more diverse Kaggle dataset which enhanced the ability of the model to generalize by adding more variety of patient attributes and diseases.

Now the factors that can be regarded as core attributes by virtue of being sought out by the feature analysis process include age, cholesterol, resting blood pressure, and maximal heart rate. However, the scatter plot of cholesterol vs MaxHR shows that there are some variables that had values that were common between heart patients and those without heart disease observing instances. Since it has already been stated, that no single trait could be considered as an accurate predictor of heart disease alone, this overlap further reinforces the need for having a multi-faceted approach to it.

Additionally, the study used transfer learning to segment and classify images, which yielded insightful results but also brought to light the difficulties in comprehending complex interdependencies in medical data. This study highlights the promise of machine learning in the prediction of cardiac disease, but it also highlights the significance of diverse, high-quality data. More complex models and in-depth feature engineering may be investigated in future research to improve predicted accuracy and aid in clinical decision-making.

CHAPTER 5
IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Machine learning-based heart disease prediction has significant social implications that are revolutionary in nature. Early and accurate detection can save many lives by enabling early medical intervention, reducing hospitalization rates, and, consequently, alleviating many of the economic burdens on healthcare systems. High-risk individuals who require proactive lifestyle modifications to stop the start or progression of disease can be identified with the aid of predictive models. The strategies provide a cost-effective means of preventing heart disease and bridging the gaps in healthcare accessibility. Additionally, by tailoring prevention and therapy to each patient's unique health profile, machine learning applied to healthcare promotes personalized medicine, which boosts efficacy.

5.2 Impact on the environment

Machine learning models for predicting cardiac disease have advantages for the green economy. The more traditional health diagnostics usually imply lots of physical contacts with the patient and application of medical technologies which waste resources and create pollution. One the more computing driven solutions, the search-based strategies of today's ML based model can look up and mine EHR and current text finding, which mitigate the environmental impact of physical diagnostical procedures and unnecessary re-testing. Early detection helps avoid hospitalizations and the subsequent energy, medical products, and transportation-related emissions that are utilized in bringing and treating patients. Remote interaction with the ML-driven diagnostic instruments also reduces the requirement to go to medical facilities and thus curtails the environmental drain even further. Finally, applying modern health ML technologies encourages a waste-free approach to disease management and control.

5.3 Ethical Aspects

The use of machine learning (ML) for heart disease prediction brings forth concerns with regards to ethics and patient confidentiality. Patient confidentiality is a key aspect because machine learning makes it necessary to handle sensitive health information. There is a need for strong privacy policies and protection of the data as well as its context. Also, ML algorithms should not have biases since this would lead to discrimination of certain populations. For patients to be able to trust the forecasts, making them should be done in an open fashion for patients to understand the process involved. Last but not the least, human regulation is necessary so that ML does not go beyond medical supervision which can put the safety of patients at risk and thus in healthcare, ethics is crucial when integrating ML.

5.4 Sustainability Plan

The key objectives of a sustainability in the prediction of heart disease using machine learning (ML) would be to ensure precision, the confidentiality of data collected and resources used in the long run. Frequent updates make it possible to keep models in accord with the shifting health trend of the population and various segments of this population. Policies of data privacy ensure the protection of the patient's data which in return instills trust and is compliant with the law. Another factor contributing to this sustainability model is the emphasis on energy efficiency when developing machine learning algorithms. This is especially important because these models already have ML tools embedded into them which are provided to healthcare personnel while the models remain relevant as new data is acquired through partnerships with medical institutions.

CHAPTER 6
CONCLUSION AND FUTURE WORK

6.1 Summary of the Study

This study looked at many machine learning methods for heart disease prediction, with a particular emphasis on models such as LightGBM, logistic regression, decision trees, and support vector machines. Patient data, including heart rate, chest discomfort, and cholesterol levels, was used to assess each model's performance. The findings demonstrated that ML can successfully identify high-risk patients, facilitating individualized healthcare and early intervention. This strategy demonstrates how ML has the potential to revolutionize cardiac treatment, enhance patient outcomes, and lower medical expenses. Future research will concentrate on improving data privacy protocols, integrating bigger datasets, and improving model accuracy.

6.2 Conclusions

The paper presents the efficacy of various ML methods in predicting cardiac disease and therefore is useful in the early detection and prevention of the disease. ML models can correctly predict the possibility of heart disease by considering crucial markers such as heart rate, cholesterol levels, and the type of chest pain. While much potential has been reflected by these ML-based models, yet further improvement is possible. Prediction accuracy and dependability might also improve by increasing the number of attributes and adding unstructured healthcare data, such as free-text patient histories. Other techniques might be able to process fuzzified, structured data from large, complex healthcare databases using techniques such as the symbolic fuzzy K-NN classifier.[14]

It basically concludes that machine learning approaches for the prediction of cardiac disease provide a useful intelligent support system for medical practitioners for fast diagnosis and efficient resource utilization. Future research should be done with more features, bigger and more diverse datasets to select the best sophisticated algorithms. Improving these instruments will make ML an increasingly important component of

preventive healthcare, with the potential to revolutionize the treatment of cardiovascular disease in a wide range of healthcare settings.

6.3 Implications for Further Study

Machine learning in prediction of heart disease has great scope for research and improvement. The future study can be concentrated on building more accurate models by using larger and diverse datasets from different communities and removing any bias that might arise. Other studies could explore hybrid approaches that create dynamic individual risk profiles, using machine learning with real-time patient information from wearable technology. As these methods continue to improve toward clinical use, it will be important to investigate the ethical considerations of ML-based predictions, including transparency and patient consent. Further research could yield systems of heart disease prediction that are much more accurate, fair, and more usable.

REFERENCES

[1] Soni, Jyoti, Ujma Ansari, Dipesh Sharma, and Sunita Soni. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." *International Journal of Computer Applications* 17, no. 8 (2011): 43-48.

[2] Abdolmanafi, A., Duong, L., Dahdah, N., Adib, I. R., & Cheriet, F. (2018). Characterization of Coronary Artery Pathological Formations from OCT Imaging using Deep Learning. Biomedical Optics Express, 9(10), 4936–4960. doi:10.1364/BOE.9.004936 PMID:30319913

[3] Abiwinanda, N., Hanif, M., Hesaputra, S. T., Handayani, A., & Mengko, T. R. (2018). Brain Tumor Classification using Convolutional Neural Network. World Congress on Medical Physics and Biomedical Engineering, 68(1), 183–189. doi:10.1007/978-981-10-9035-6_33

[4] Aledhari, M., Joji, S., Hefeida, M., & Saeed, F. (2019). Optimized CNN-based Diagnosis System to Detect the Pneumonia from Chest Radiographs. Proceeding of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM'19), 2405–2412. doi:10.1109/BIBM47256.2019.8983114

[5] Ali Khan, H., Jue, W., Mushtaq, M., & Mushtaq, M. U. (2020). Brain Tumor Classification in MRI Image using Convolutional Neural Network. Mathematical Biosciences and Engineering, 17(5), 6203–6216. doi:10.3934/ mbe.2020328 PMID:33120595

[6] Alqudah, A. M., Alquraan, H., Abu Qasmieh, I., Alqudah, A., & Al-Sharu, W. (2019). Brain Tumor Classification Using Deep Learning Technique - A Comparison between Cropped, Uncropped, and Segmented Lesion Images with Different Sizes. International Journal of Advanced Trends in Computer Science and Engineering, 8(6), 3684–3692. doi:10.30534/ijatcse/2019/155862019

[7] Brahami, M., & Matta, N. (2015). An Approach to Dynamic Fusion of the Knowledge Maps of an Activities Process: Application on Healthcare. International Journal of Information Systems in the Service Sector, 7(4), 1–25. doi:10.4018/IJISSS.2015100101

[8] Repaka, A. N., Ravikanti, S. D., & Franklin, R. G. (2019, April). Design and implementimplementsease prediction using naives Bayesian. In *2019 3rd International conference on Trends in electronics and informatics (ICOEI)* (pp. 292-297). IEEE.

[9] Parthiban, L., & Subramanian, R. (2008). Intelligent heart disease prediction system using CANFIS and genetic algorithm. *International Journal of Biological, Biomedical and Medical Sciences*, *3*(3).

[10] Ramesh, T. R., Lilhore, U. K., Poongodi, M., Simaiya, S., Kaur, A., & Hamdi, M. (2022). Predictive analysis of heart diseases with machine learning approaches. *Malaysian Journal of Computer Science*, 132-148.

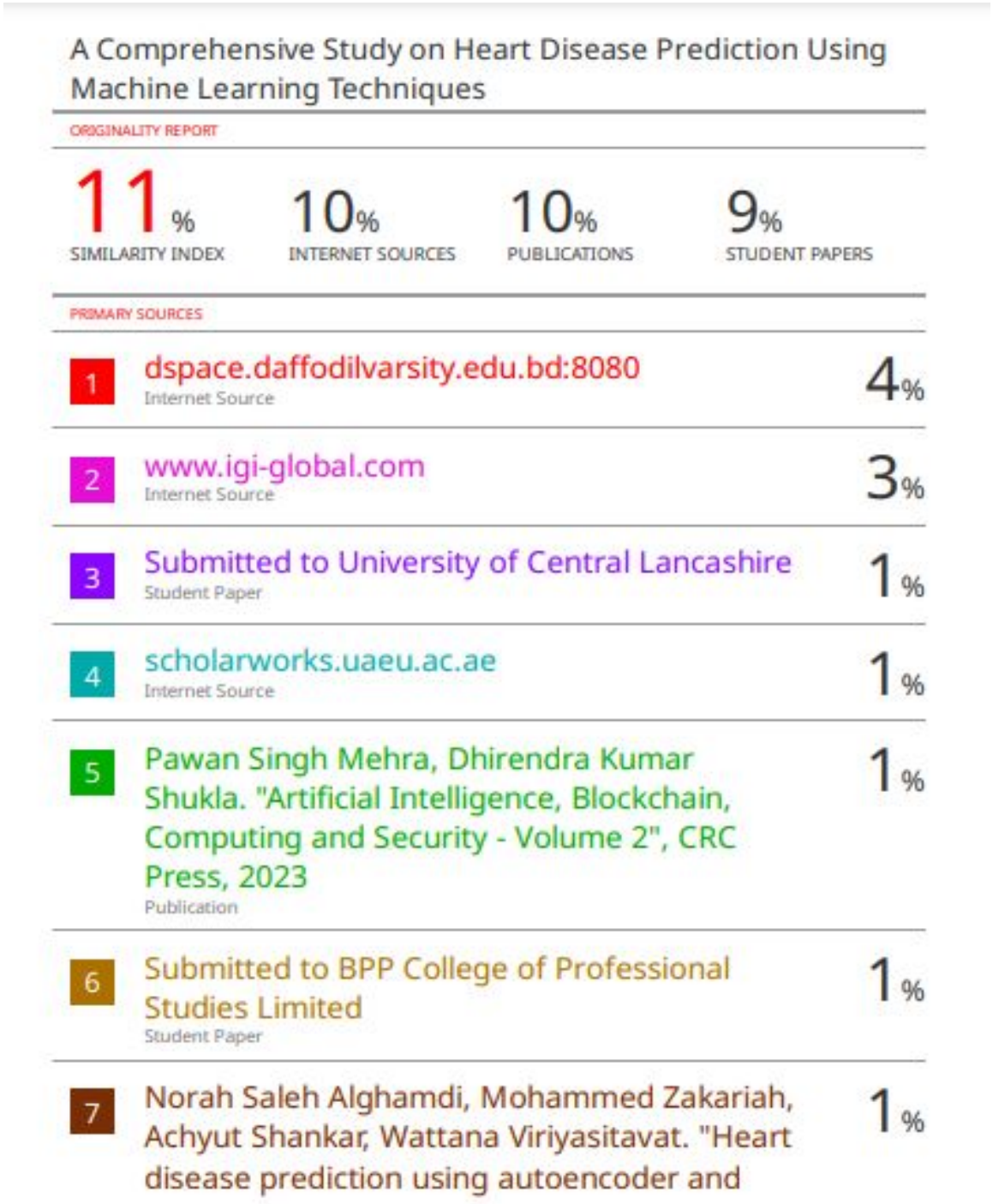
[11] Sharma, S., & Parmar, M. (2020). Heart diseases prediction using deep learning neural network model. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *9*(3), 2244-2248.

[12] Aljanabi, M., Qutqut, M. H., & Hijjawi, M. (2018). Machine learning classification techniques for heart disease prediction: a review. *International Journal of Engineering & Technology*, *7*(4), 5373-5379.

[13] Khourdifi, Y., & Baha, M. (2019). Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. *International journal of Intelligent engineering & systems*, *12*(1).

[14] Bertsimas, D., Mingardi, L., & Stellato, B. (2021). Machine learning for real-time heart disease prediction. *IEEE Journal of Biomedical and Health Informatics*, *25*(9), 3627-3637.

[15] Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile information systems*, *2018*(1), 3860146.



A Comprehensive Study on Heart Disease Prediction Using Machine Learning Techniques

ORIGINALITY REPORT

11%	10%	10%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	4%
2	www.igi-global.com Internet Source	3%
3	Submitted to University of Central Lancashire Student Paper	1%
4	scholarworks.uaeu.ac.ae Internet Source	1%
5	Pawan Singh Mehra, Dhirendra Kumar Shukla. "Artificial Intelligence, Blockchain, Computing and Security - Volume 2", CRC Press, 2023 Publication	1%
6	Submitted to BPP College of Professional Studies Limited Student Paper	1%
7	Norah Saleh Alghamdi, Mohammed Zakariah, Achyut Shankar, Wattana Viriyasitavat. "Heart disease prediction using autoencoder and	1%

DenseNet architecture", Egyptian Informatics Journal, 2024
Publication

8	Submitted to Jacksonville University Student Paper	1%
9	G. Manikandan, B. Pragadeesh, V. Manojkumar, A.L. Karthikeyan, R. Manikandan, Amir H. Gandomi. "Classification models combined with Boruta feature selection for heart disease prediction", Informatics in Medicine Unlocked, 2024 Publication	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	Off		