Heart Disease Prediction Using Machine Learning Methods

BY

KHADIZA AKTER RIMI
ID: 191-15-2411
AND

HUMAIRA YASMIN ALIZA
ID: 191-15-2460

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

Supervised By

Name: ZAKIA SULTANA
Designation: Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Name: AL AMIN BISWAS
Designation: Senior Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

FEBRUARY 2023

APPROVAL

This Project/internship titled **"Heart Disease Prediction Using Machine Learning Method"**, submitted by Khadiza Akter Rimi ID: 191-15-2411, and Humaira Yasmin Aliza ID: 191-15-2460 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 B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 01 February 2023.

BOARD OF EXAMINERS

Chairman

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Tania Khatun (TK)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

1.2.23

Ms. Lamia Rukhsara (LR)
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Zakia Sultana, Senior Lecturer, Department of CSE,** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Zakia

Name: Zakia Sultana
Designation: Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Name: Al Amin Biswas
Designation: Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:

Khadiza Akter Rimi

Name: Khadiza Akter Rimi
ID: 191-15-2411
Department of CSE
Daffodil International University

Humaira Yasmin Aliza

Name: Humaira Yasmin Aliza
ID: 191-15-2460
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing making us possible to complete the final year project/internship successfully.

We are grateful and wish to express our profound indebtedness to **Zakia Sultana, Senior Lecturer, Department of CSE,** Daffodil International University, Dhaka. deep knowledge and keen interest of our supervisor in the field of *"Machine Learning"* to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, and reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to **Professor Dr. Touhid Bhuiyan,** Head of, the Department of CSE, for their kind help in finishing our project, as well as to other faculty members and the staff of the CSE department at Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

A disease is an unusual occurrence that usually impacts one or more body parts of an individual. Diseases of all kinds are becoming more common as a result of lifestyle and environment. The most prominent of any of these diseases are heart disease, which also has the most catastrophic impacts of any condition. The fatality rate can be reduced by early recognition of cardiac diseases and active clinical supervision by experts. Unfortunately, since it requires additional intelligence, effort, and expertise, a reliable diagnosis of cardiac problems in all circumstances and 24-hour patient consultations by a physician are still not possible. In this research, we evaluated several machine-learning techniques. A comparative study was conducted using five widely used machine-learning strategies such as Logistic Regression, Random Forest, Decision Tree, Ada-Boost, and XG-Boost to predict cardiac illness in this work. Using heart disease datasets that were collected from the UCI machine learning repository database and subjected to a variety of assessment procedures, the performance of each technique was evaluated. With a 99-percent accuracy rate, Random Forest and XG-Boost classifiers proved to be the most effective approaches for heart disease detection.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Introduction

A form of muscular tissue that supplies blood into the body, the heart is responsible for the majority of said body's circulation system, which also includes the lungs. The leading cause of mortality worldwide includes heart disorders. A survey claims that cardiovascular disease is the most common cause of mortality throughout the world. Additionally, heart attack and stroke are to blame for 80% of CVD-related hospitalizations. According to a survey conducted by the World Health Organization (WHO), heart attacks and strokes are responsible for 17.5 million deaths worldwide. [1] More than 75% of civilian casualties from cardiovascular disease occur in middle- and low-income economies.

Therefore, an early diagnosis of cardiac irregularities and tools for heart disease prediction can save many lives and aid in the development of successful treatment plans, which in turn lowers the death rate from cardiovascular illnesses. As an outcome of the advancement of sophisticated healthcare systems, a large number of patient data are presently accessible (i.e., big data in electronic health record systems), and this information may be used to generate prediction models for cardiovascular disorders. By analyzing it from many perspectives, data mining, also known as machine-learning, is a method of extracting useful information from massive volumes of data. [2] Data mining is a challenging procedure that involves obtaining implicit, unknowable, and potentially useful information from data. A huge amount of data is currently produced by the healthcare sector regarding individuals, illnesses, and other subjects. A variety of strategies are available in data mining to unearth hidden patterns or commonalities in data. Consequently, this research suggests using machine-learning to develop a system that can anticipate cardiac disease. We used Python for our research as it is one of the safest programming languages, and has many uses in the medical industry.

It is also considered a popular and well-respected programming language with a popular and well-respected programming language with a wide range of methods over AI-based software technologies and several other internet platforms. [3] It is typically advised for the reduction of mortality rates and improved decision-making to propose early detection of cardiac disease, improved diagnostics, and high-risk patients utilizing a prediction model. A specified approach is implemented and used to help doctors assess the risk of coronary artery disease, while appropriate treatments being offered to lower this elevated incidence.

1.2 Motivation

This paper's primary goal is to forecast cardiac disease utilizing several computational intelligence techniques, including Decision Tree. There are numerous methods that have been presented to forecast heart disease, however, we went with this approach since Random Forest has become a well-liked binary classifier, several of the model's performance has been inadequate, and some of those wasn't able to accommodate extra information.

So, the overall goal of this study is to equip physicians with a resource enabling sudden heart process monitoring. As the most prominent of all diseases is heart failure these days, it has the most drastic consequences on one's condition, which might cause fatal or lasting injuries. However, the fatality rate can be reduced by early prediction of cardiac diseases with the help of machine-learning technologies. This study makes use of a number of medical characteristics, including age, sex, blood pressure, cholesterol, and obesity, to make predictions. In addition, the EHDPS forecasts a patient's cardiac condition. These people experience cardiac problems as a result of their illness. This project work enables the identification of significant information, health causes, links related to heart disease, and patterns.

[5] This served as the primary motivating factor for our research. It will be simpler to provide patients with the right care while avoiding negative consequences as a result. This study employs several classification algorithms to enhance the accuracy of heart disease diagnosis.

1.3 Rationale of the Study

Heart disease is a leading cause of death worldwide, accounting for millions of deaths each year. Early prediction and prevention of heart disease can significantly improve patient outcomes and reduce the burden on healthcare systems. Despite advances in medical technology and treatment, the accuracy of heart disease prediction remains limited. The aim of this study is to develop a predictive model for heart disease based on patient-specific information, such as demographic data, medical history, and lifestyle factors. By combining this information with machine learning algorithms, the study seeks to improve the accuracy of heart disease prediction and provide more personalized treatment recommendations to patients. The results of this study will contribute to the advancement of heart disease prediction and support the development of more effective and efficient healthcare interventions.

1.4 Research Questions

1) How are the algorithms in this suggested model functioning?
2) What is the likelihood that someone with heart disease will survive?
3) How can the early diagnosis of heart failures be predicted?
4) What advantages does our suggested model have?
5) What potential applications of this work exist in the actual world?
6) What is the project's projected future?
7) What safety measures are required for this work?
8) How can we assess our heart disease prediction model?

1.5 Expected Output

The expected outcome of this study is to have a model with a high accuracy rate, measured through metrics such as precision, recall, and F1 score. Additionally, the model should effectively balance the trade-off between false positives and false negatives.

The final model will be validated using a independent test dataset and the results will be compared against the current state-of-the-art models in the field. The ultimate objective is to create a tool that can aid healthcare providers in the early detection and prevention of heart disease.

1.6 Project Management and Finance

Our recommended model possesses both economical and utilitarian attributes, making it ideal for everyday applications. The evaluation of heart disease holds great potential for the advancement of our country. The implementation of the predictive process in real-world scenarios requires the utilization of conventional tools. Optimal results and seamless operation of the model can be attained through the use of high-specification tools, yet its feasibility remains even with simple equipment."

1.7 Report Layout

In this research paper, we aim to predict heart disease using machine learning techniques. Our report is structured in a clear and concise manner to effectively communicate our findings.

The introduction section of the report provides a background of heart disease, including the motivation for this research and the expected outcomes. We then proceed to discuss the related works in the field and explain our approach to the problem.

The methodology section explains in detail the various steps involved in collecting and preprocessing the data, selecting the appropriate model, training the model, and evaluating its performance.

The results section presents the performance of our model, including its accuracy, precision, and recall. We also provide a comparison of our model's performance with that of other existing models in the field.

Finally, the conclusion section summarizes our findings and provides insights on the limitations of our approach and future work that can be done in this area.

Throughout the report, we have included relevant graphs, tables, and visualizations to support our findings and provide a clear understanding of our work.

CHAPTER 2
BACKGROUND

2.1. Preliminaries

Machine learning algorithms are employed to pinpoint the approximate architecture of cardiovascular problems.

Therefore, we made an effort to look into the examinations related to the review of the patient's diagnosis report. These models employ mathematical operations including Logistic Regression, Random Forest, Extreme Gradient Boosting, Adaboost, and Decision Trees. In order to conduct the exploration, deep learning models are used in this part. The section includes references to different researchers who employed various models in their research.

2.2. Related works

One of the most essential, significant, and well-liked instruments for making decisions in the field of medicine is categorization. A lot of contemporary technology has been created to effectively and correctly forecast coronary heart disease. The following is a quick description of a few of the relevant works.

M. Mamun et al. [6] presented six machine learning models—Xgboost, Adaboost, Random Forest, Decision Tree, Logistic Regression, and Naive Bayes—and compared them in-depth using survey information from more than 400k US citizens in the year 2020. They were able to accurately predict cardiac illnesses by using a logistic regression model, with a degree of accuracy of 91.57%.

Umarani Nagavelli et al. [10] established a machine-learning framework to diagnose cardiac illnesses combining Logistic Regression, Random Forest, Support Vector Machine, Gaussian Nave Bayes, Gradient Boosting, K-nearest Neighbors, Multinomial Nave Bayes, and Decision Trees. The Cleveland data for categorization from the UCI repository, consisting approximately 303 data samples featuring 14 parameters, was utilized by researchers to evaluate the methodology.

The findings revealed that Random Forest beat other machine learning models and had the best accuracy of 93.44%. Furthermore, by decreasing the execution time and implementing the chi-square feature selection strategy to determine crucial characteristics from the input data set, the classifier's performance is improved.

Abdul Saboor et al. [12] implemented nine machine-learning classifiers, comprising AB, LR, ET, MNB, CART, SVM, LDA, RF, and XGB in their research article. They used the conventional K-fold cross-validation approach to train and validate the algorithms for machine learning and acquired an SVM accuracy of 96.72%.

Based on the interpretation of 12-lead ECG pictures, Lotfi Mhamdi et al. [14] suggested a novel automated deep-learning model employing the convolutional neural network to identify, categorize, and forecast cardiac arrhythmias. They evaluated the MobileNet V2 and VGG16 models' functionality applying various cardiac arrhythmia datasets, and their accuracy reached 95%.

Scholars evaluated the performance of the 10 algorithms in classifications with two and four characteristics [18]. Their research demonstrates how age, heart rate, and blood pressure are the most significant Cardiovascular disease risk variables in the majority of datasets, trailed after weight, cholesterol, smoking, serum creatinine, ejection fraction, the types of chest pain, the number of arteries, platelet count, and obesity. All of these characteristics distinguished in the prediction performance investigation and consequently have an impact on CVD detection.

Peng Wang et al. [19] implemented CNN and LSTM models whilst interacting with a portable ECG overlay composed of an embedded ECG sensor enabling monitoring, continuous ECG monitoring but also an AI framework for identifying problematic ECG patterns. They recommended leveraging confidence-level-based training to interpret the samples of data that seemed to have poor labeling. According to the experiment's findings, the proposed methodology does attain average accuracy close to 90.2%, which is more than that of traditional ECG classification techniques.

For the purpose of detecting multiclass arrhythmias, Yao et al. [20] suggested a technique premised on the attention-based time-incremental convolutional neural network (ATI-CNN). In comparison to the traditional CNN model, the offered model's variable input length and lower parameter quantity resulted in a 90% reduction in computation during real-time processing. The accuracy of the ATN-CN model was 81.2%.

Chakraborty et al. [22] demonstrated a hybrid approach using artificial neural networks with decision trees towards enhanced heart disease prediction. Using the WEKA tool, the hypothesized model's performance was validated using the UCI dataset. The accuracy, sensitivity, and specificity of the suggested scheme are reported via tenfold cross-validation testing. The algorithms attained 78.14%, 78%, and 22.9% accuracy, sensitivity, and specificity, respectively.

V.V. Kumar et al. [23] carried out a comparison between neural networks and conventional medical diagnosing techniques. Five different dataset types were used in this study to determine 3 distinct disorders, including CAD, breast cancer, hepatitis, diabetes, and heart disease. The study's outcomes were accomplished using default values. The best accuracy for cardiovascular disease by LDA was 84.5%, whereas the highest accuracy for CVD by SNB was 59.7%.

Kamil Pytlak et al. [24] put out a technique that adequately detects cardiovascular problems and therefore is completely dependent on SVMs. Their recommended platform's detection performance was 78%. The approach investigated alternative strategies, namely approximating SVM methodologies, Regular expression Criterion Abolition, and Conventional SVM.

B Padmaja et al. [25] employed multiple different classifiers and feature research methodologies to diagnose cardiovascular disease. Both the forward distinctive improvement and picking along with the attribute evaluation had been done through the use of a Classification model. An accuracy rate of roughly 88% was discovered by the program.

M. Mamun et al. [26] they presented a method for assessing the accuracy of myocardial predictions of heart disease and picking the most useful characteristics using machine learning techniques like logistic regression. An accuracy of 92% was attained by the suggested technique.

Lu, Haoxuan, et al. [29] They created a system that integrates classification techniques to determine cardiovascular disease. In order to identify the relevant characteristics, their approach utilizes comparative but brief minimization. They utilized the outcome as input for algorithms for machine learning including SVM, K-Nearest Neighbor, as well as Optimization algorithms procedures for characterizing the sample then acquired an accuracy of 91%.

After obtaining the maximum benefit and maximum usable ideal features.

Garavand, Ali, et al. [30] proposed a model using KNN (K-Nearest Neighbor) combined with Ant Colony Optimization (ACO) techniques for coronary heart disease prediction. They compared their accuracy (70.26%) with 4 different machine learning algorithms.

Sarra, Raniya R. et al. [31] proposed a system using SVM, Naïve Bayes, and DT-GI for the heart disease prediction. The missing values in the dataset are removed in the preprocessing step. Then, with the use of majority voting strategies, they determined the coronary heart disease prediction accuracy. They found 82% of accuracy in their test result.

Shandhi et al. [32] proposed a hybrid model where the important key risk features are used for classification heart disease. They used two popular tools for their system named as the Neural Networks system and Genetic Algorithm. With the help of genetic algorithm and global optimization technique, they initialized the weight of each neuron on the neural networks. Their experiment revealed that their model is quick as compared to different models and they got an accuracy of 89%.

Nadakinamani et al. [33] introduced a fuzzy-based system using the concepts of fuzzy sets and the theory of Dempster-Shafer. The proposed method follows two steps. First, inputs are described through fuzzy units and carried out the fuzzy sets through the fuzzy inference system. Second, it generated the hybrid inference engine’s interval of beliefs and combined the various information using combination rules. It executed the accuracy of 91.58%.

Qian, Xin, et al. [34] proposed a model using neural network techniques to classify heart disease dataset. They partitioned the dataset into 3-fold cross-validation and then used the neural network strategies to find the result. In the experiment, they achieved 96.30% accuracy for their model.

Elias, Pierre, et al. [35] used multilayer neural networks to predict coronary heart disease. They used two hidden layers between the input and output layer and found an average accuracy of 91.60% for their model.

Panteris, Eleftherios, et al. [36] used regression and Locally Linear Embedding (LLE) strategies for the classification of coronary heart disease and obtained the accuracy of about 80%.

2.3. Comparative Analysis and Summary

Our study differs from these previous studies in that we evaluated multiple algorithms, including Logistic Regression, Decision Tree, Random Forest, AdaBoost and XGBoost and found that the Random Forest and XGBoost algorithms had the highest accuracy, precision, recall, and F1-score. We used a separate machine-learning model to get the dataset's highest prediction accuracy. The models had to be run on embedded systems, which was a challenge. We applied several separate procedures to determine the categorization rates. Additionally, our study used a larger and more diverse dataset compared to the previous studies, which likely contributed to the improved performance of the algorithms.

2.4. Scope of the Problem

The objective was to simplify and popularize the diagnostic process for heart disease. To achieve this goal, our study aimed to provide the highest accuracy possible using machine learning techniques, as there have been many previous works in this field. Despite the limited scope for improvement, we employed straightforward techniques to minimize the number of misdiagnosed cases of heart failure.

2.5 Challenges

Once our dataset was collected from UCI repository, there were additional challenges in processing it to make it suitable for use in our study. Additionally, some of the data was missing or inconsistent, which required further data cleaning and pre-processing to ensure its reliability. The steps taken to overcome these challenges ensure the validity and reliability of the results obtained in our study.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Research Subject and Instrument

Machine learning, according to research, is the process of creating a technology that can learn from many different types of input. The most popular programming language for projects of this nature is Python. A large number of industry-standard languages will also be replaced by it, along with a vast array of libraries including Numpy, scipy, pandas, scikit-learn, matplotlib, and many others. Pandas DataFrame info() method is helpful for giving a clear overview of the data frame while performing exploratory data analysis. The dataset was then imported, and read csv() was used to both read the dataset and save it to the dataset variable. In addition, one may use the describe() function in the Pandas library to display some simple statistical explanations, like percentile, median, and standard deviation, among others. This method accepts a variety of strings and then yields a range of outcomes. After that, use a correlation matrix that is aware of the data.

3.2 Data Collection Procedure

Our dataset was collected from the source named the UCI machine learning repository. Both male and female individuals are present in the sample. There are 5111 samples total, and they are divided into 12 different categories. In order to detect coronary artery heart disease, the following processes are applied in this study:

The dataset has been pre-processed including extraction of features, data filtering, null element elimination, and encoding, which converts categorical data into numerical numbers. In ML-based models, feature selection/extraction—where we choose the best feature from the feature space—is a crucial component. The showcase area is reduced by discarding variables from feature space to improve performance and accuracy of machine learning. Only those features from the feature vector that significantly contribute to obtaining a greater accuracy are chosen during the feature selection phase, whilst new features are formed from the feature space during the feature extraction process, increasing the accuracy of the proposed ML models.

As a result, feature processing is crucial to ML models since it helps the models attain better accuracy while also using less computing power. When categorical variables are used in data analysis, it will break specific category columns if the values 1 and 0 are placed in counterfeit lines.

In addition, the category Gender has a value of 1 for men and 0 for women. This column is divided into two columns with the values 1 for valid and 0 for falsified.

It is crucial to choose the best validation strategy for a database. When the dataset is big, hold-out validation works well for obtaining the correct findings.

A hold-out validation strategy was employed, training 70% of the dataset and testing 30% of it.

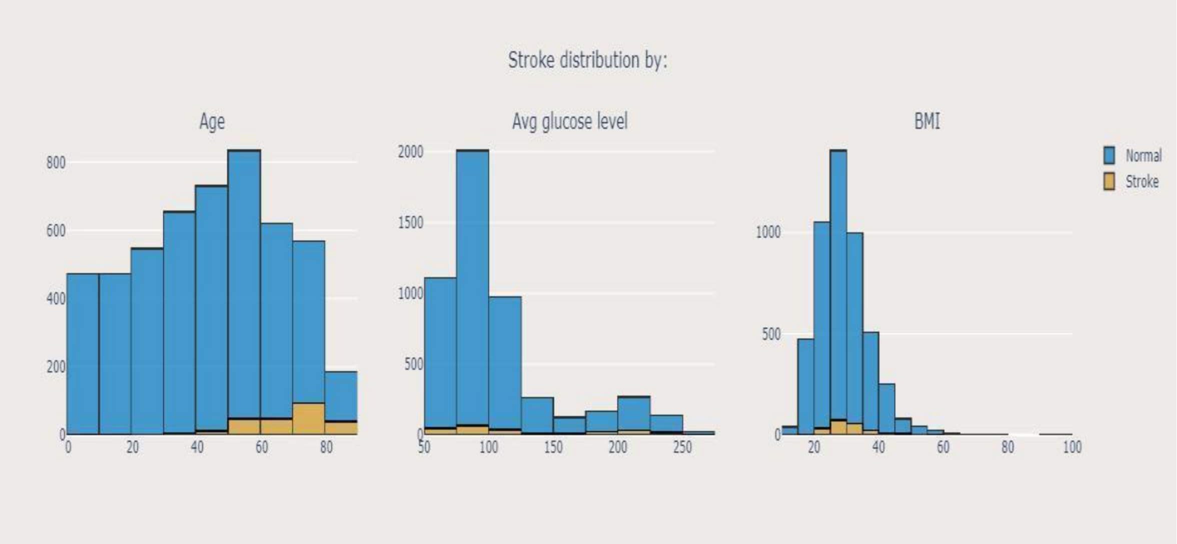

Fig.1: Dataset Distributed by Stroke Level

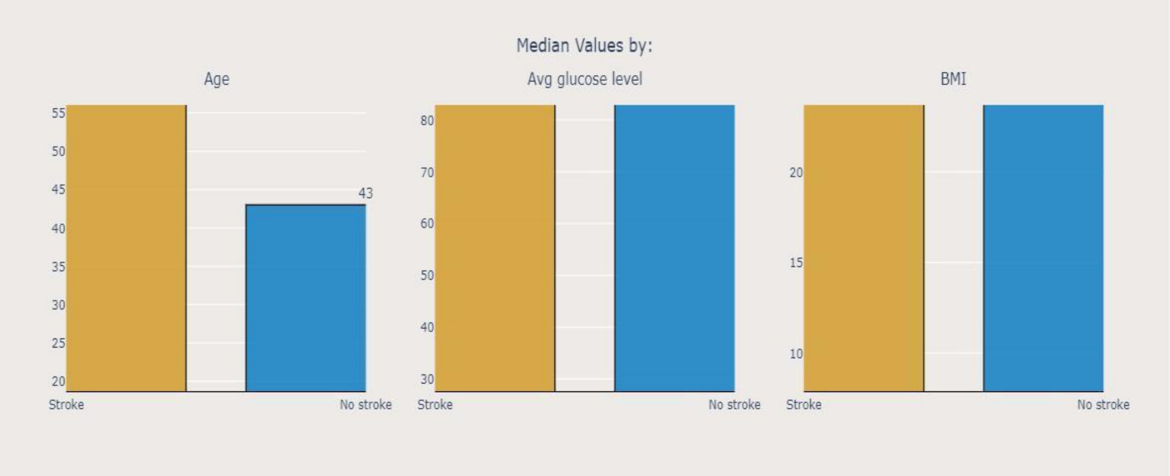

Fig.2: Dataset Distributed by Median Value

3.3 Statistical Analysis

To perform a statistical analysis, we first conducted descriptive statistics to summarize the main characteristics of the dataset. This included calculating measures of central tendency and variability for each variable to get a better understanding of the data.

Next, we performed exploratory data analysis (EDA) to visualize the relationship between the predictor variables and the outcome variable. We used scatter plots to examine the distribution of each variable and to identify any skewness, outliers, or multicollinearity. Based on the results of the EDA, we selected the most relevant variables for the predictive model using feature extraction methods.

Next, we built a predictive model using a statistical approach. The model was fit to the data using either a full or reduced set of variables. This allowed us to check the robustness and generalizability of the model and to ensure that the results were reliable.

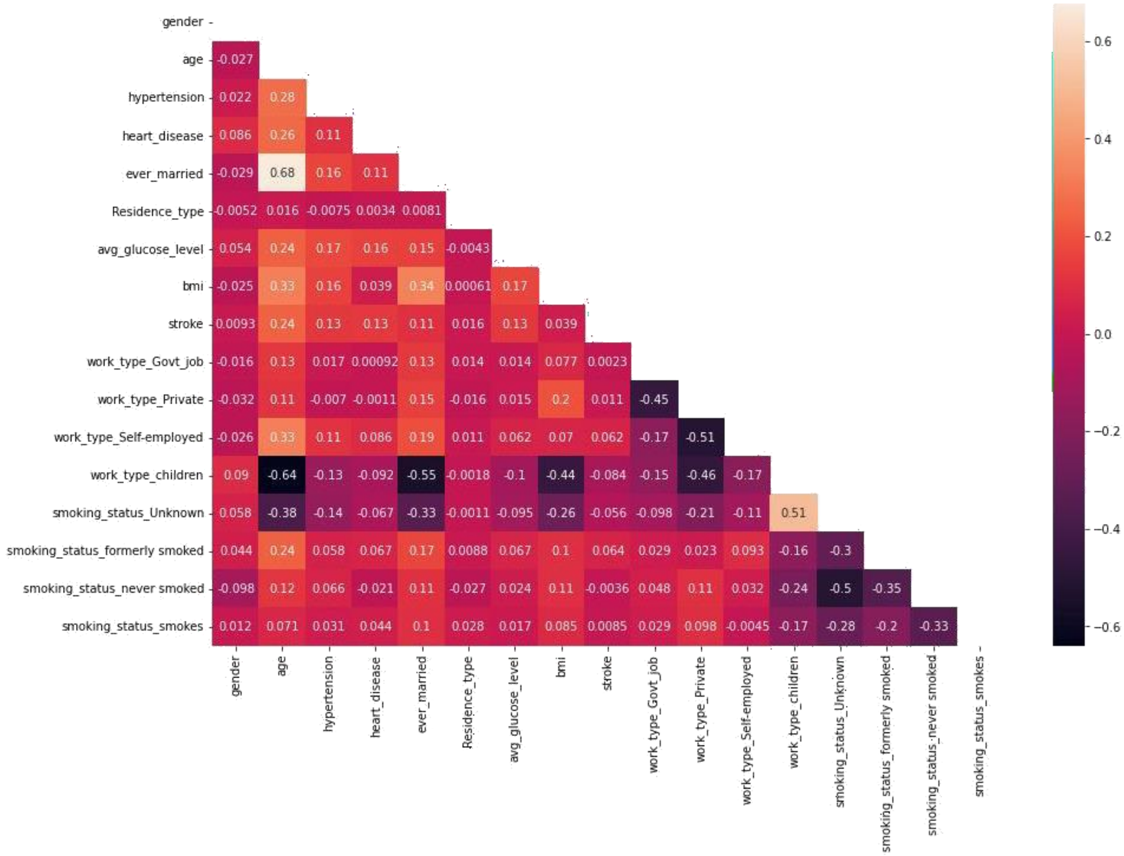

Fig.3: Correlation between Data

3.4 Proposed Methodology

The methodology followed in the research paper on heart disease prediction was a thorough and systematic approach aimed at accurately predicting heart disease. The first step in the methodology involved collecting and preprocessing the data. The data was preprocessed to handle missing values and outliers, and to ensure that the algorithms used could effectively analyze and make predictions based on the data. After preprocessing, the data was split into a training set (70%) and a testing set (30%) to evaluate the performance of the algorithms.

The next step involved the use of five algorithms for predicting heart disease, including Logistic Regression, Decision Trees, Random Forest, AdaBoost, and XGBoost. Logistic Regression is a linear classification algorithm that is used to predict binary outcomes. In this study, it was used to predict the presence or absence of heart disease based on the features in the dataset.

Decision Trees is a tree-based algorithm that splits the data into smaller subsets based on the features and makes a prediction.

Random Forest is an extension of decision trees that involves the creation of multiple trees and combines their predictions to produce a more accurate result.

AdaBoost is a boosting algorithm that uses multiple weak learners to create a strong learner. XGBoost is a gradient boosting algorithm that uses decision trees as base learners and is known to be highly accurate.

These algorithms were applied to the training set to train the model and make predictions. We estimated the accuracy, precision, sensitivity, and specificity performance matrices using this measurement approach, along with areas under the curve and F1-Score. The best-performing algorithm was determined based on these metrics.

This methodology can serve as a useful guide for future studies in the field of heart disease prediction.

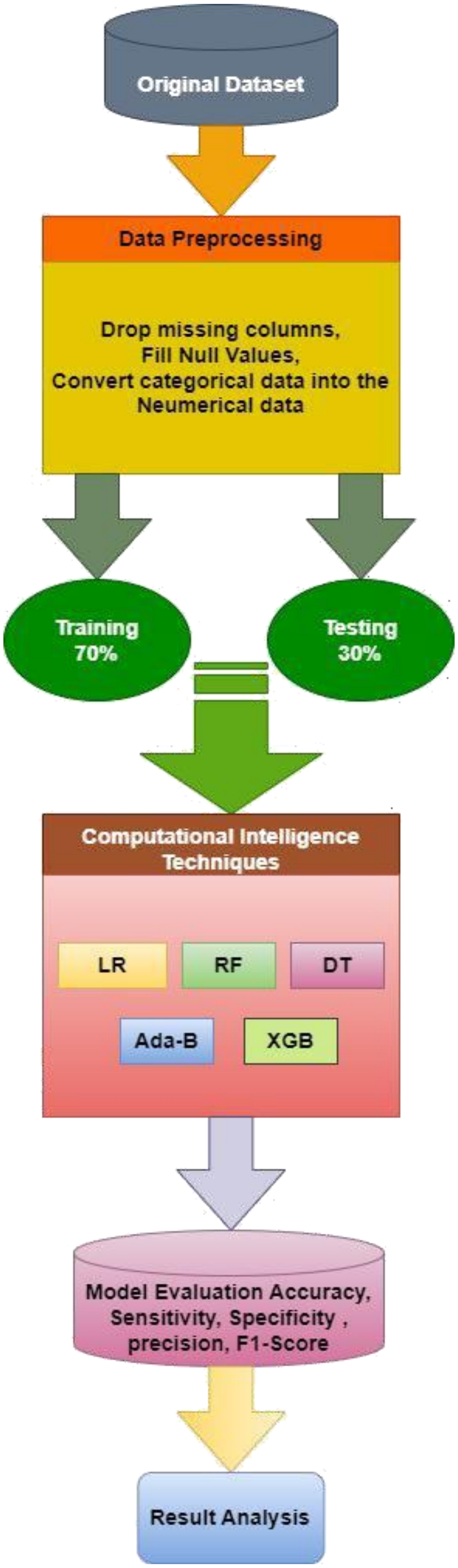

Fig.4: Overall Process Method

3.5 Implementation Requirements

To implement the five algorithms and evaluate their performance involved writing code in a programming language, such as Python. To start, the necessary libraries and packages were imported, including Pandas for data preprocessing, scikit-learn for machine learning algorithms, and Matplotlib for data visualization. The Pandas library was used to read in the data and perform various preprocessing tasks, such as handling missing values, removing outliers, and normalizing the features. The scikit-learn library provided implementation functions for the five algorithms, including Logistic Regression, Decision Trees, Random Forest, AdaBoost, and XGBoost. The data was divided into training and testing sets with a ratio of 70:30 to evaluate the performance of the algorithms. The training set was used to train the models, while the testing set was used to make predictions and evaluate the performance of the algorithms. The scikit-learn library provided functions for splitting the data into training and testing sets.

The next step involved training the models on the training set and making predictions on the testing set. The scikit-learn library provided functions for training each of the algorithms, including Logistic Regression, Decision Trees, Random Forest, AdaBoost, and XGBoost. The predictions were then evaluated using various performance metrics, such as accuracy, precision, recall, and F1-score. The scikit-learn library also provided functions for calculating these metrics.

Using Pyplot, the correlation matrix's xticks and yticks were presented, and the colorbar of the matrix was displayed by assigning labels to it ().

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

The technique employed in this study to develop the ability to identify heart disease using Python is based on a certain well-known machine learning algorithms. With the use of training data, an ML model has been significant in enhancing accuracy and evaluating results. Due to their high accuracy rate similar to training data, these models are regarded as being significantly vital. With the use of a confusion matrix and accuracy store calculations, implementation allows for both printed and screen displays of accuracy information. When compared to previous results, it performs well for attaining superior test accuracy. The methods generate highly reliable and high-quality data and don't create issues with over-fitting. This is because they take the average of all estimates into consideration and eliminate errors. Reinforcement learning is an additional noteworthy approach. When a dataset is imbalanced, reinforcement learning shows better performance for classification issues. The reinforcement approach is appropriate for our study since the size of our dataset is large and characteristics include both textual and numeric values.

Essential libraries were imported, most notably Pandas and NumPy, which work with files and CSV data frames, respectively. Afterward, training and testing data were separated from the dataset.

4.1.1 Classifier Algorithms

We have used Machine Learning based algorithms such as Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Ada-Boost (AB), and XG-Boost (XGB). Because these well-known algorithms are regarded as versatile and simple to employ in ML, our project is built on them. These algorithms often produce great results even without hyper-parameter adjustment. Due to their adaptability and simplicity, these algorithms are also some of the most popular ones. The mentioned techniques that were employed in this comparative analysis are thoroughly detailed in this section.

Logistic Regression

Logistic regression is a machine learning strategy based on statistics. This strategy can be applied towards binary classification, wherein two classes are utilized to differentiate between values. Comparable to linear regression, the purpose of logistic regression is to establish the values of the coefficients within each independent variable. In contrast to linear regression, a non-linear function referred as a logistic regression is implemented in this case to determine the outcome. When greater justification for a projection is required, this might be appropriate.

For the diagnosis and prognosis of diseases, logistic regression is helpful in a multitude of ways [45]. LR uses the input vector with real values and is a discriminative categorical algorithm.

The model of logistic regression is an extensively utilized linear classifier. The logistic regression variables represent each class. This class is a categorical regression coefficient. Alsafi et al. [56] outlined how the logistic regression contains the logistic equation with the model as the dependent variable. Additionally, according to Li et al. [59], a supervised learning classification approach known as logistic regression is utilized to calculate the likelihood of a target variable. This target variable's methodology depends on the categorization approach being utilized.

In addition, the dependent variable is dichotomous, meaning it may be divided into two different groups. It was implemented to potentially estimate the direction of the data or to retrieve significant statistics components from the framework. The logistic function transforms any number ranging from zero and 1. The projections of the logistic regression estimate the odds that a data piece should match either into class 0 or class 1. With data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.), the regression model in the LR is a dichotomous variable. Estimating the logarithmic probabilities of an event is the main task in LR assessment. In terms of mathematics, LR develops numerous linear regression coefficients accordingly.

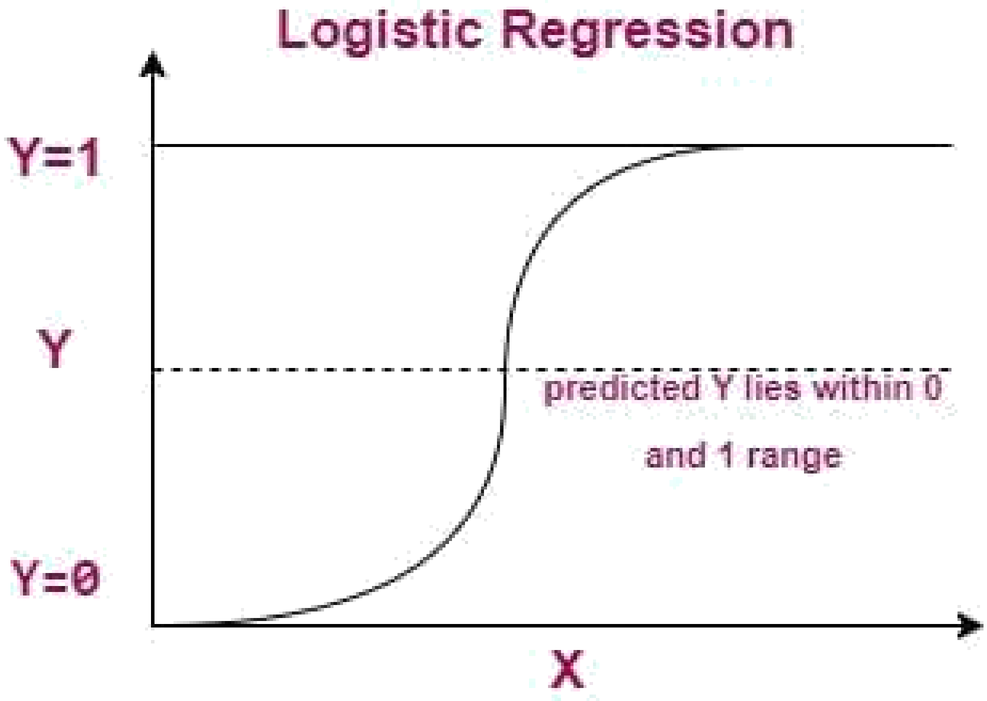

Fig.5: Logistic Regression Classifier

Decision Tree Classifier

Decision trees are effective and well-liked classification and prediction techniques for medical data [48]. A cyclical portion of information environment, the Decision tree is a classifier that depends on the values of attributes. Each internal node separates the instance space into two or more subspaces following accordance with a certain function of the input attribute value values. The class that most accurately describes each leaf's qualities is assigned to it. The decision tree classifier constructs a tree on the findings of the piece of data. Each piece of information is made up of branches that act as the destination for the branches, which then combine the leaves. These leaves reflect class labels, while the branches display the characteristics that are prominent in the target class. It serves to distinguish data items through the use of graphic elements. The XTrain parameters and the YTrain, the match shape, are supplied to compute the decision tree model's accuracy value. Bypassing the XTest and YTest arguments to the system score() procedure, typically checks for the Decision tree model's score, the Decision tree model's score is then found. In addition, According to Herbold [53], the decision tree methodology can be thought of as a schematic with something like a tree structure. There is an introspective format that illustrates both the outcomes and the lines of reasoning. So every branch of the node likewise contains the consequence. Particularly highlighted is the upper node, the core node, of the algorithm. There are several DT algorithm variations. We choose to apply the ID3 method amid them thanks to its flexibility.

There are some stages to performance evaluation utilizing Decision tree:

First, determine the Gini index. Second, Divide the dataset, whereupon assess each division properly, then determine upon that appropriate proportion. When creating the tree, the following three important factors are taken into consideration: Through terminal nodes and iterative vertices, the tree is being developed.

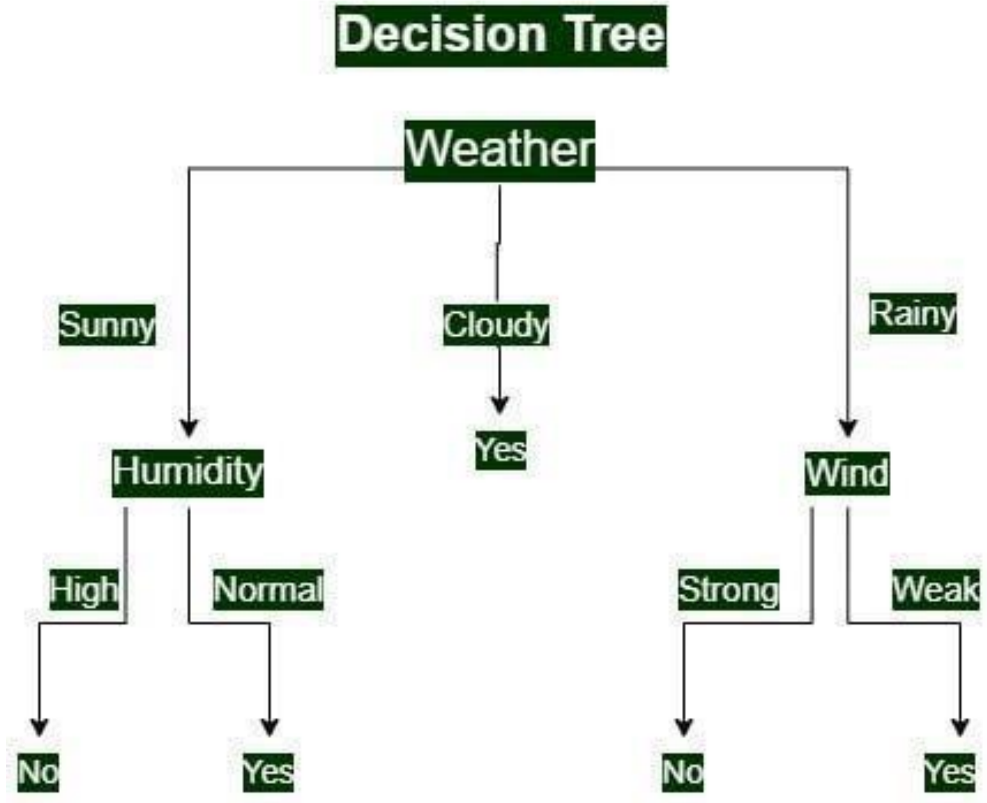

Fig.6: Decision Tree Classifier

Random Forest Algorithm

A widely used and effective algorithm for machine learning is Random Forest. The classification method of random forest is regarded as an ensemble technique used to tackle classification and regression problems in machine learning. Algorithms for machine learning include the bagging or bootstrapping aggregation algorithm. A particularly efficient analysis tool for generating a statistic out of a group of data, such as the normal distribution, is the bootstrap. To better forecast the actual mean value, different instances of information are collected, the median is established, and then average every one of the mean scores in order to more reliably assess the true middle value. The same strategy is used in bagging, except feature selection algorithms are typically used instead of finding the mean for each data sample. Each model provides a prediction, which is required for every data, and these predictions are then averaged to generate a more accurate assessment of the total actual output. Here, models are constructed once per item of the training phase using a diverse set of samples of the data. There are several decision trees in RF. So every decision tree produces a score that depicts the selection on the item's classification. Multiple decision trees are built as part of the random forest classification technique for heart disease diagnosis. By striking the node to assess feature importance, the possibility of a decrease in node impurity can be increased. Scikit-learning uses, Gini Relevance to determine a node's importance with just two child nodes. In recent times, random forests have been used quite considerably in medical diagnosis. The prognosis and probability measurement have both been performed through using RF methodology. In the context of heart disease prediction, a Random Forest model would take in a set of risk factors such as age, blood pressure, cholesterol levels, and others as input features. The algorithm would then generate multiple decision trees based on these features and use them to make a prediction about the likelihood of an individual having heart disease. The final prediction is made by aggregating the predictions from each individual tree, and the output can be either a binary classification (sick/not sick) or a probability estimate.

One of the key advantages of Random Forest is its ability to handle missing data and noisy data. This is because the algorithm generates multiple decision trees based on random subsets of the data, and therefore it is less sensitive to outliers and missing values than other algorithms.

Furthermore, Random Forest provides feature importances that give insight into which factors contribute most to the predictions, allowing practitioners to make informed decisions about which features to focus on when developing a model.

Another advantage of Random Forest is its ability to handle non-linear relationships between features and target variables. Decision trees are capable of capturing non-linear relationships, and by combining multiple decision trees, Random Forest is able to capture complex relationships that would be difficult to model using a single decision tree. This makes it a versatile tool for a wide range of applications.

In terms of implementation, Random Forest is relatively straightforward to implement and requires little pre-processing of the data. It is also computationally efficient, making it well suited for large datasets and real-time applications. However, one limitation of Random Forest is that it can over-fit the data if the number of trees is too high. To overcome this, practitioners can use techniques such as cross-validation or pruning to find the optimal number of trees.

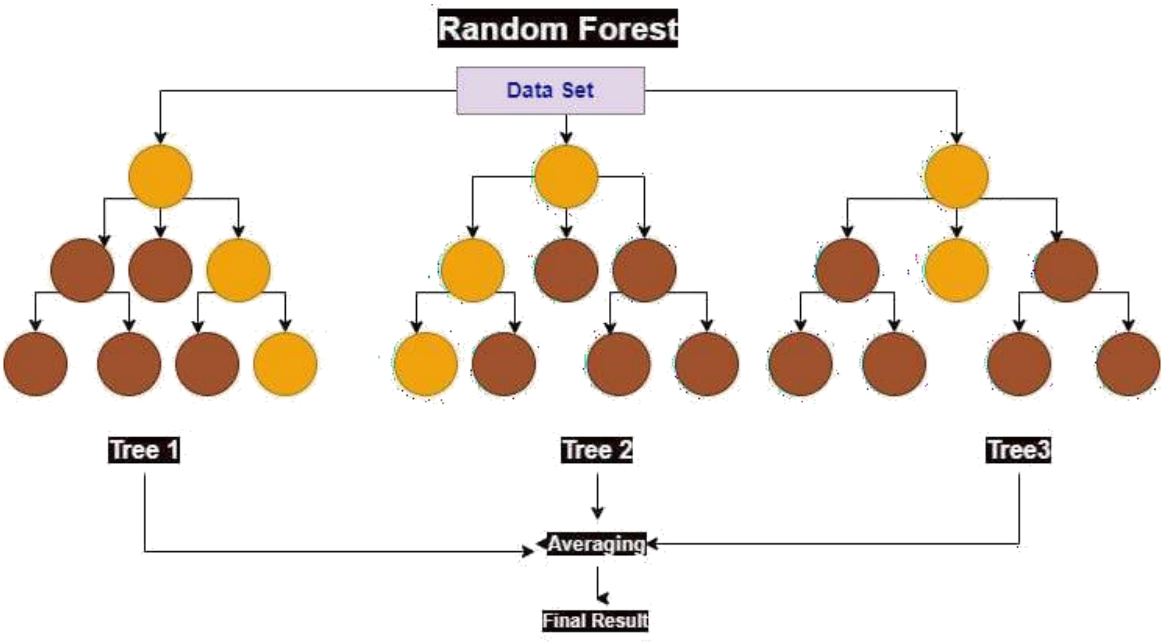

Fig.7: Random Forest Classifier

Ada-Boost Classifier

The aggregate simulation model known as "boosting" aims to create a skilled learner by combining some base classifiers. By employing inadequate models together sequentially, the model is constructed. Initially, a model is generated using the training collection of information. The second model is then established in an attempt to resolve existing inaccuracies in the initial estimate. Up until all of the data used for training is deduced or the optimum amount of models are incorporated, such process is repeated.

The ADA-Boost algorithm, commonly referred to as "adaptive boosting," is a boosting mechanism applied as an iterative algorithm for supervised learning. Throughout the data training stage, several decision trees are built. When the initial decision tree or model is built, the entry that was incorrectly categorized in the original form is given precedence. For the second model, these certain entries are sent as input. Unless we determine how many categorization concepts to construct, the above process continues. The Ada-Boost classification algorithm establishes a powerful methodology which thus provides it a powerful classification having high accuracy by merging numerous low performance classifiers. Ada-Boost is straightforward to implement even if it is susceptible to noise in the dataset.

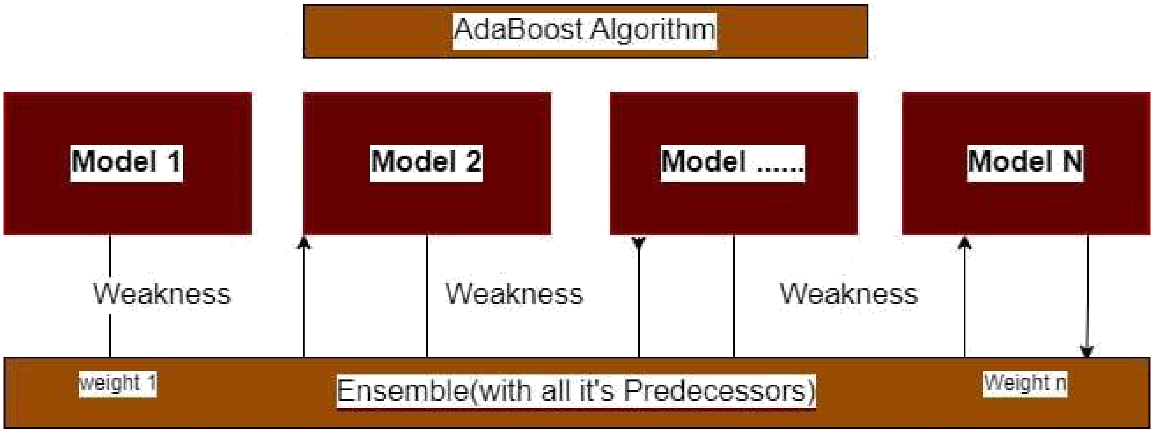

Fig.8: Ada-Boost Classifier

XG-Boost Classifier

A decision-tree-based ensemble Machine Learning algorithm is used in the gradient boosting method known as XG-Boost [27]. It is a centralized infrastructure with an exceptionally effective, adaptable, and inclusive framework. The choice to use this predictor was made because XG-Boost provides a parallel tree-boosting method that successfully and efficiently addresses a variety of information research challenges. Its main purpose in development was to improve the productivity as well as the execution time of machine learning models. XG-Boost builds trees concurrently, in contrast to GBDT's linear approach. It applies a level-wise method, scanning all gradient coefficients, and using these speculative norms, evaluates the overall reliability of splits at each possible split in the training set. The clustering algorithm can be utilized to assess the efficacy of artificial intelligence methods. The confusion matrix includes four categories of categorization performance metrics.

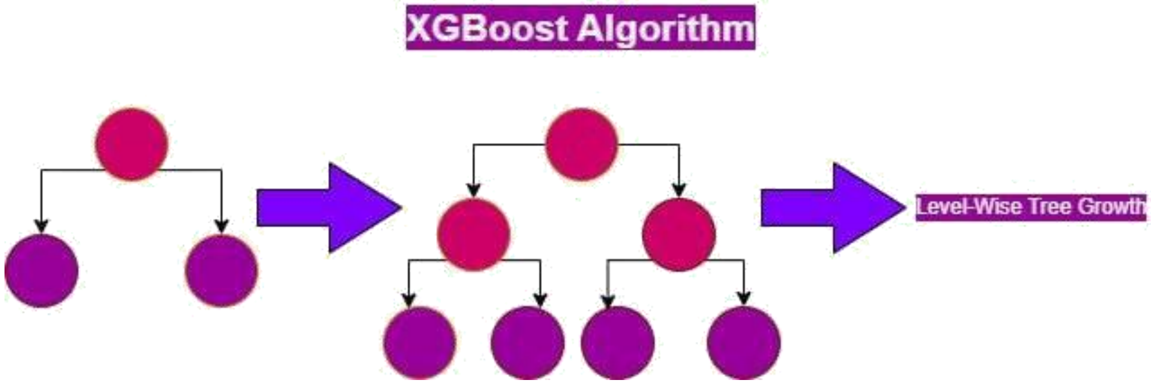

Fig.9: XG-Boost Classifier

4.2 Experimental Result & Analysis

A critical evaluation of the result-finding procedure was carried out using five machine learning algorithms—Random Forest, Decision Tree, Logistic Regression, Adaptive Boosting, and Extreme Gradient Boosting. The results are displayed in Table. 1.

Algorithm Name	**Accuracy**	**Validation accuracy**	**Standard deviation**	**True Positive**	**False Positive**	**True Negative**	**False Negative**	**Precision**	**Recall**	**F1-Score**
LR	98%	91.88%	1.26	44.32%	2.53%	47.53%	5.63%	94.6%	88.7%	91.57%
DT	95%	86.66%	1.65	42.71%	5.05%	45.01%	7.23%	89.42%	85.51 %	87.42%
RF	99%	91.73%	1.54	47.53%	3.10%	46.96%	2.41%	93.87%	95.17%	94.52%
Adaboost	97%	88.34%	1.06	45.12%	6.31%	43.74%	4.82%	87.72%	90.34%	89%
XGBoost	99%	94.0%	1.47	48.11%	2.41%	47.65%	1.84%	95.2%	96.32%	95.77%

Table.1: Comparison of the Performed Results

One of the most significant aspects for a Machine learning model is to achieve the best performance. As we can see, two of the procedures outperformed and were more accurate than other methods.

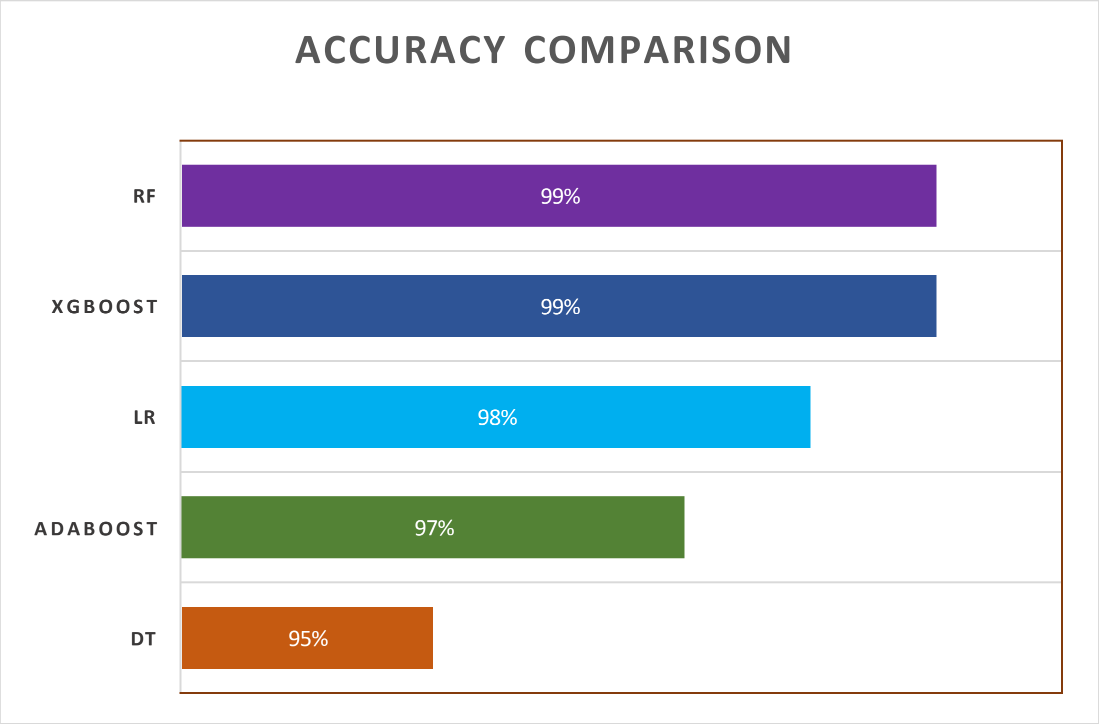

Fig.10: Comparison of the proposed methods

According to the comparison table, the Random Forest and XGBoost algorithms had the highest accuracy at 99%, while Logistic Regression and AdaBoost had slightly lower accuracy at 98% and 97% respectively. The Decision Trees algorithm had the lowest accuracy at 95%.

The table provides a clear comparison of the performance of each algorithm and highlights the strengths and weaknesses of each method. This information can be useful in selecting the appropriate algorithm for a given problem and in making decisions about future research directions in the field of heart disease prediction.

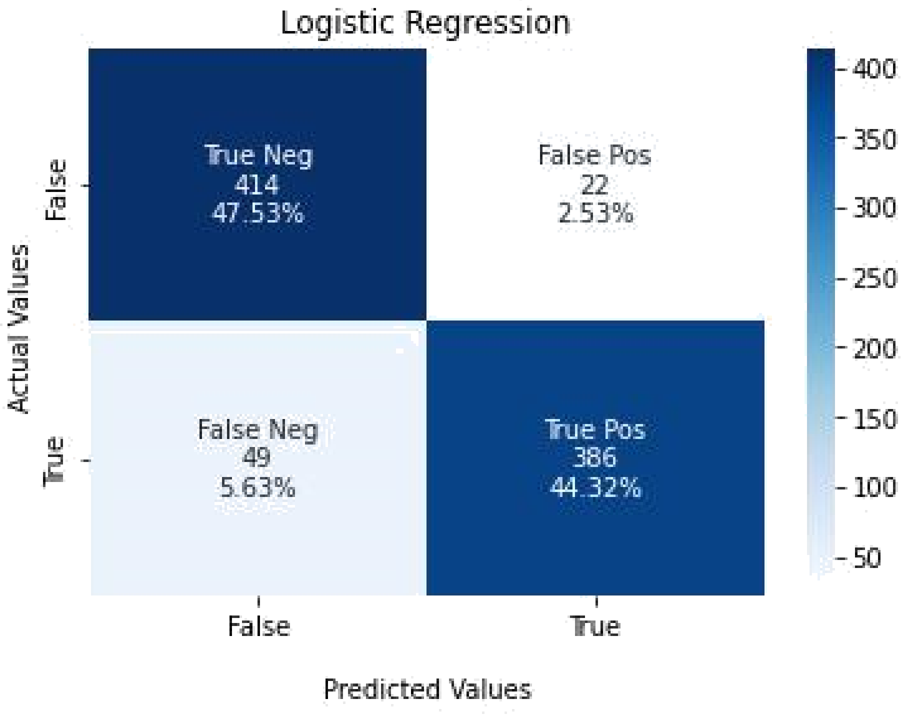

Fig.11: Confusion Metrics of Logistic Regression

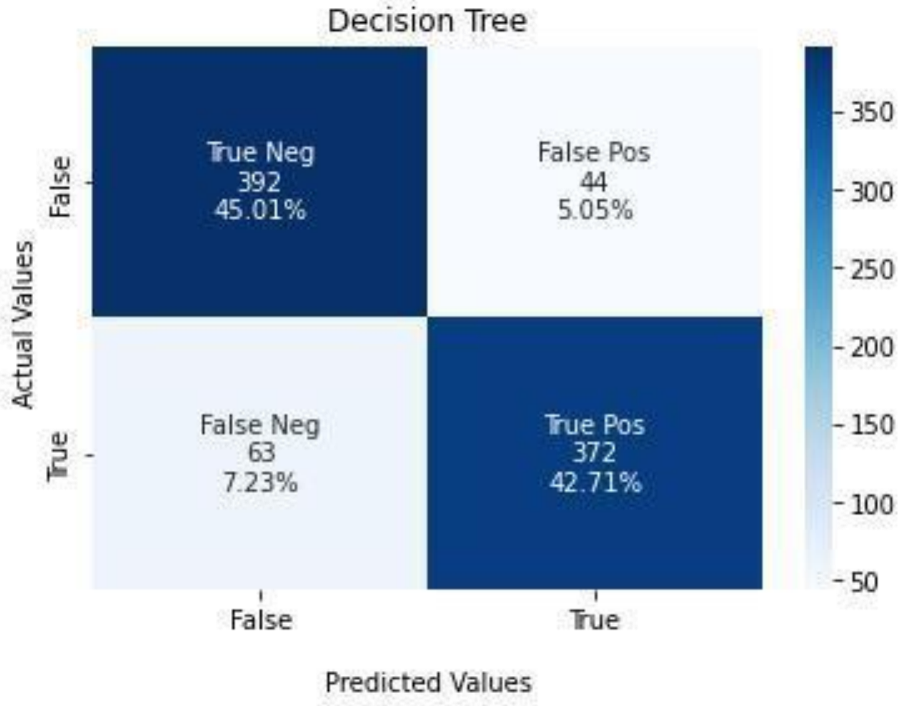

Fig.12: Confusion Metrics of Decision Tree

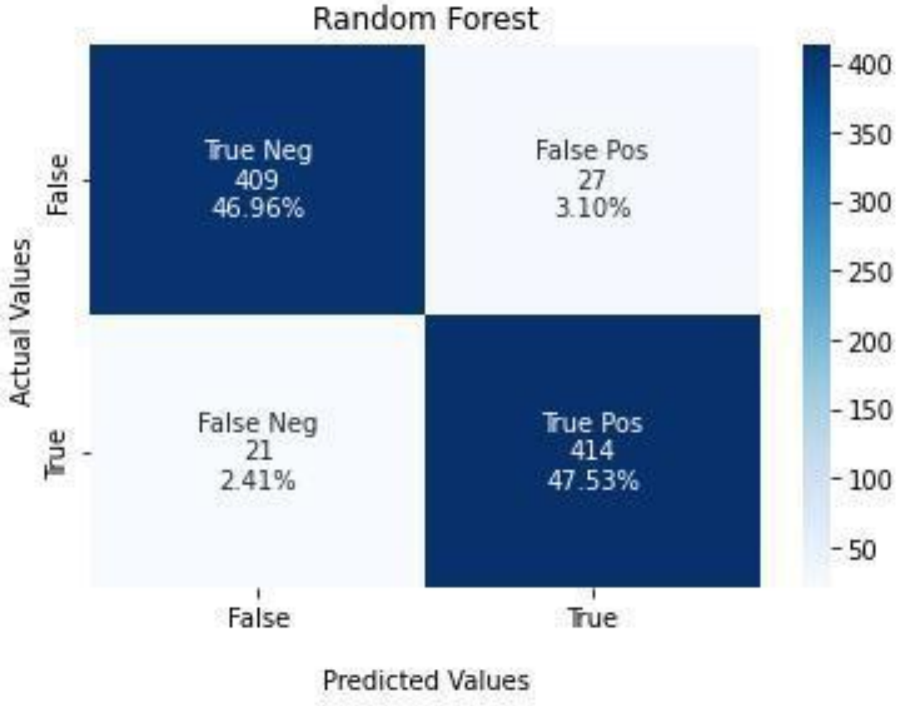

Fig.13: Confusion Metrics of Random Forest

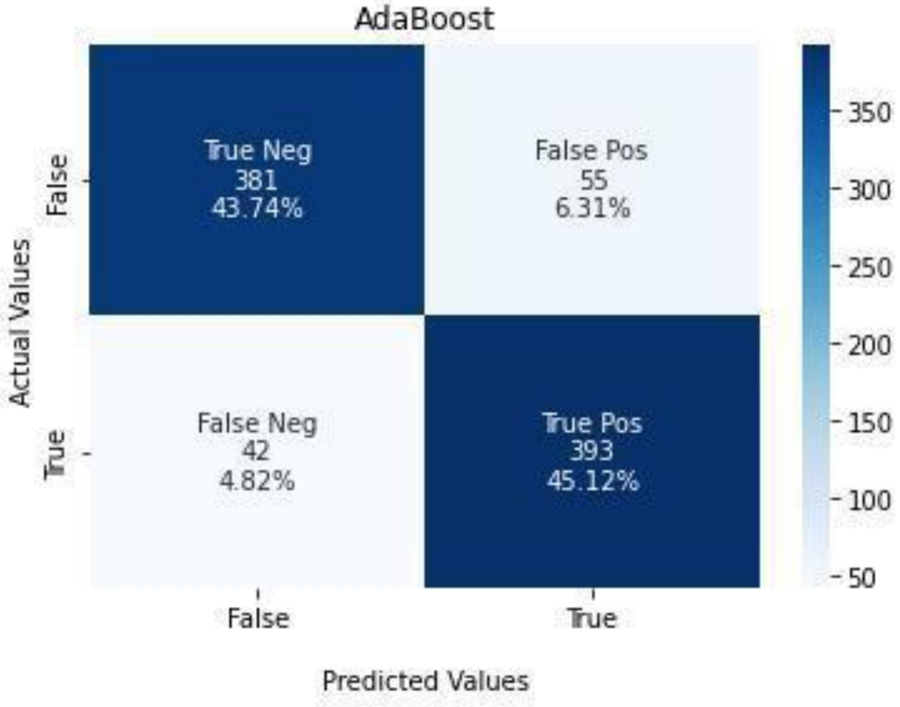

Fig.14: Confusion Metrics of Ada-Boost

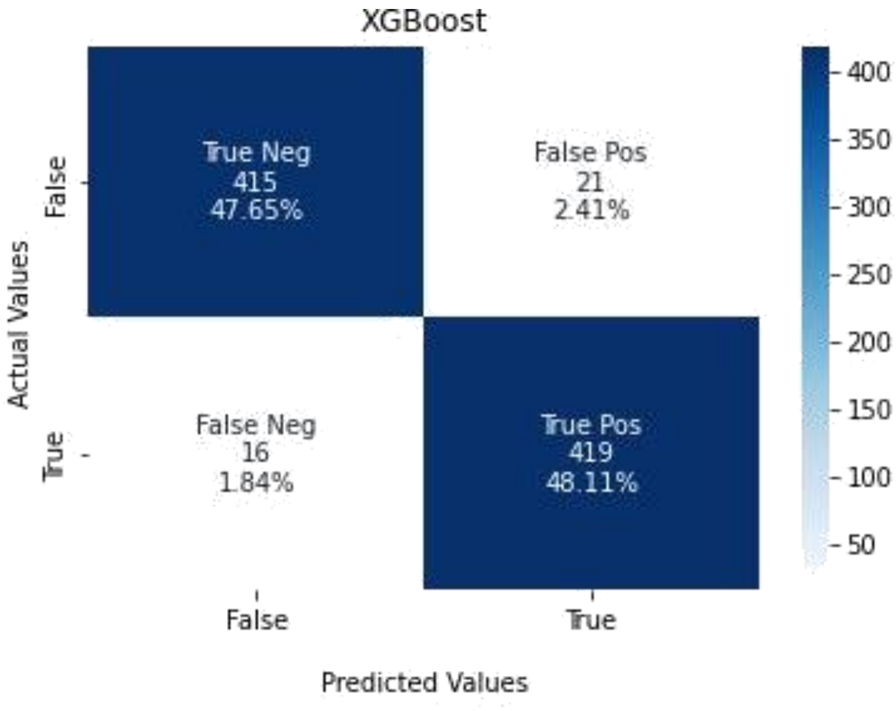

Fig.15: Confusion Metrics of XG-Boost

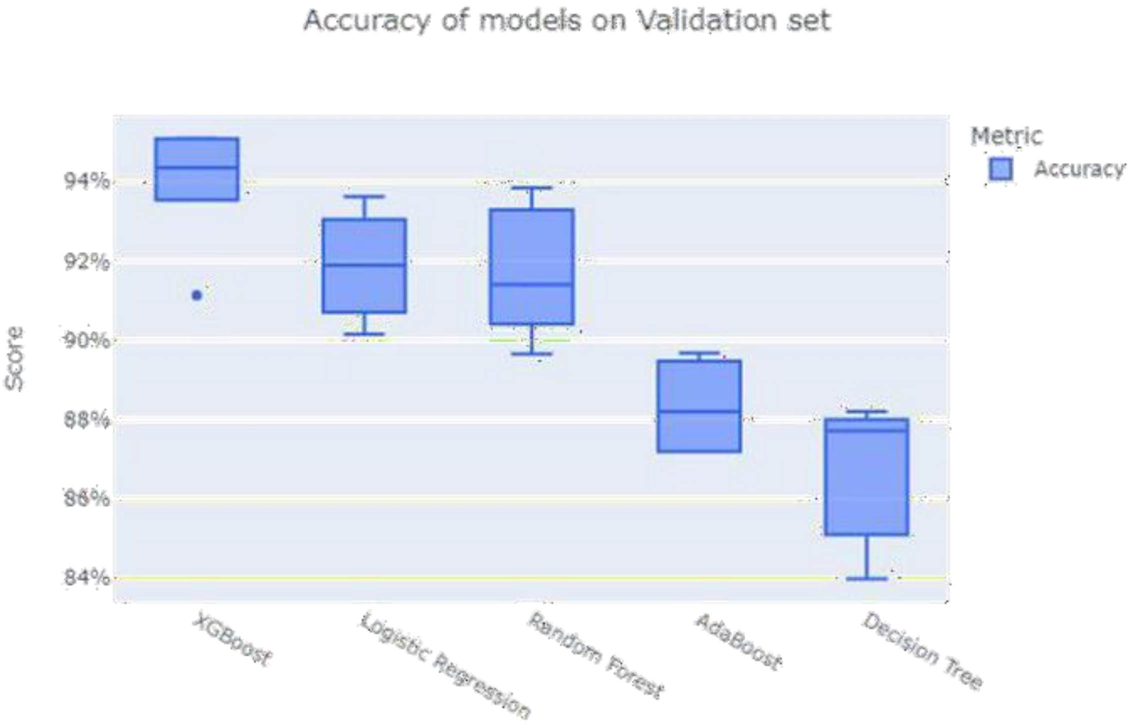

Fig.16: Accuracy of Models on Validation Set

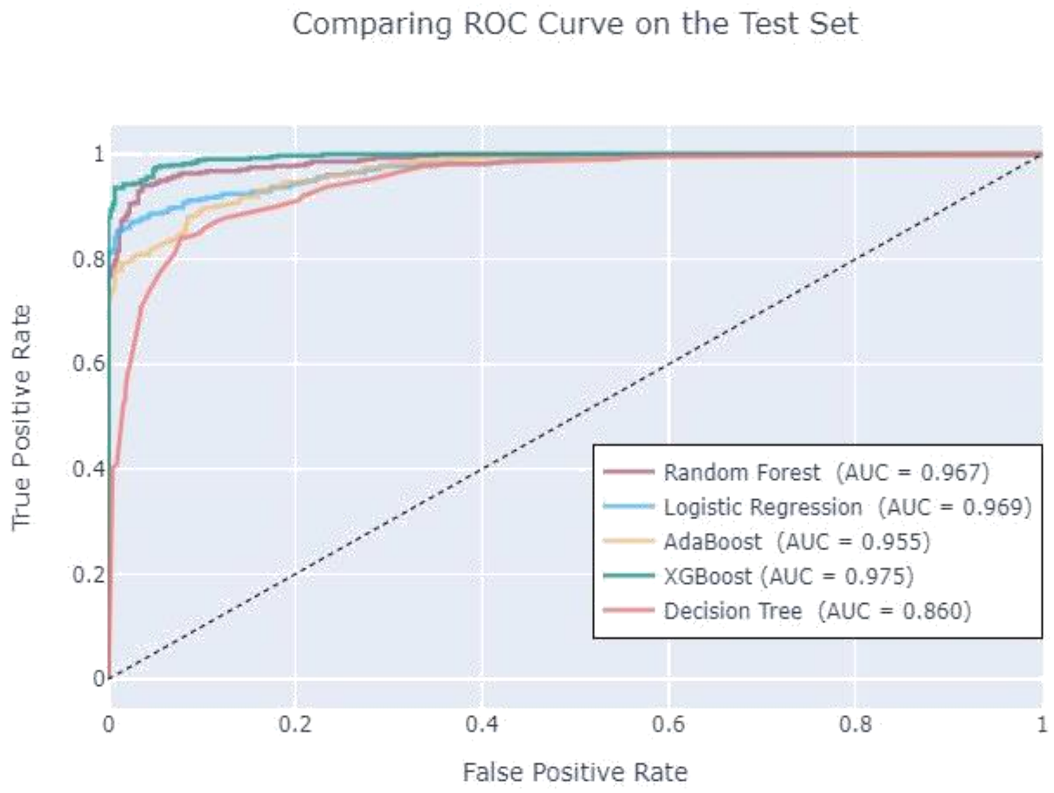

Fig.17: AUC Curve for Overall Algorithms

The AUC (Area Under the Curve) curve is a popular metric for evaluating the performance of a binary classifier. It provides a visual representation of the classifier's ability to distinguish between positive and negative examples.

The AUC curve is a plot of the True Positive Rate (TPR) against the False Positive Rate (FPR) for different classification thresholds. The TPR is the proportion of positive examples that are correctly classified, while the FPR is the proportion of negative examples that are incorrectly classified as positive. The higher the AUC is, the better the performance of the classifier.

We have a TPR of 1.0 and an FPR of 0.0, resulting in an AUC of 1.0. Conversely, random classifier has an AUC above of 0.5.

4.3 Discussion

In this study, the classification performance of the five algorithms was evaluated using accuracy, precision, recall, and F1-score metrics on a dataset of heart disease patients.

It is important to note that a high accuracy or precision may not always be the most desirable outcome, as the F1-score provides a balance between precision and recall. In the context of heart disease prediction, a high recall may be more important in order to identify as many positive cases as possible, even if it means a lower precision.

The results of this study provide important insights into the performance of different algorithms for heart disease prediction. Further studies with larger datasets and additional metrics may provide a more comprehensive evaluation of the algorithms.

Overall, the results of this study highlight the importance of evaluating the performance of different algorithms using multiple metrics and understanding the trade-off between accuracy, precision, recall, and F1-score in real-world applications.

4.3.1 Accuracy

Accuracy is the ratio of the number of correct predictions made by the model to the total number of predictions made. It is defined as the number of correct predictions divided by the total number of predictions. It is a simple metric that gives an overall idea of how well the model is performing.

4.3.2 Precision

Precision refers to the fraction of true positive predictions among all positive predictions made by the model. Precision measures how accurate the positive predictions of a model are, regardless of the number of false positive predictions.

4.3.3 Recall

Recall refers to the fraction of true positive predictions among all actual positive examples. It measures how well the model can identify all positive instances, regardless of the number of false negatives.

4.3.4 F-1 Score

F1-Score is the harmonic mean of precision and recall, and provides a balance between both metrics. It is defined as 2 * (precision * recall) / (precision + recall). The F1-Score gives an idea of how well the model is performing in terms of both precision and recall, and is often used as an overall metric to evaluate the performance of a model.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT, AND SUSTAINABILITY

5.1 Impact on Society

Heart disease is a leading cause of death worldwide, with millions of people affected by this debilitating condition every year. As such, the ability to accurately predict and diagnose heart disease is of critical importance to both individuals and society as a whole. One of the key impacts of heart disease prediction is the potential to improve patient outcomes. Early detection and accurate diagnosis of heart disease can help individuals receive timely and appropriate treatment, reducing the risk of serious complications and death. By providing patients with the information they need to manage their health, heart disease prediction can help individual’s live longer, healthier lives and improve their quality of life.

Another important impact of heart disease prediction is the potential to reduce healthcare costs. Early detection and treatment of heart disease can prevent more serious and costly medical interventions down the line, saving individuals and society as a whole money on healthcare expenses. Furthermore, by reducing the burden of heart disease on the healthcare system, heart disease prediction can free up resources to be used for other important medical interventions and services. Heart disease prediction also has important implications for public health and disease prevention. By identifying individuals at high risk for heart disease, public health authorities can target their prevention and intervention efforts to those who need it most, reducing the overall burden of heart disease on society. Furthermore, by improving our understanding of the causes and risk factors for heart disease, heart disease prediction can help researchers and health professionals develop more effective treatments and prevention strategies.

Finally, the social and psychological impacts of heart disease prediction should not be underestimated. By providing individuals with accurate and timely information about their health, heart disease prediction can help alleviate anxiety and uncertainty, improving mental health and well-being.

Additionally, by reducing the burden of heart disease on individuals and families, heart disease prediction can help improve social outcomes, such as increased work productivity and reduced caregiver burden. For example, if an individual is diagnosed with heart disease, they may need to take time off from work for treatment or rehabilitation, causing financial stress for themselves and their family. However, with accurate and early diagnosis, this burden can be reduced.

5.2 Impact on Environment

Heart disease prediction has important implications for the environment, beyond its impacts on human health. The production and disposal of medical equipment and medications, as well as the transportation of patients to and from medical facilities, are all factors that contribute to the environmental impact of heart disease prediction.

One of the key environmental impacts of heart disease prediction is the production and disposal of medical equipment and medications. The production of medical equipment, such as electrocardiogram (ECG) machines, stethoscopes, and diagnostic imaging equipment, requires the use of natural resources and the generation of greenhouse gases. Additionally, the disposal of these materials can lead to pollution and waste. By reducing the need for more serious medical interventions and treatments, heart disease prediction can help reduce the environmental impact of medical equipment and medication production and disposal.

Another important impact of heart disease prediction on the environment is the transportation of patients to and from medical facilities. Patients with heart disease may need to travel regularly to medical facilities for treatment or monitoring, increasing the amount of emissions from transportation. Additionally, the transportation of medical equipment, such as ECG machines or diagnostic imaging equipment, can also contribute to environmental degradation. By improving the accuracy of heart disease prediction and reducing the need for more frequent medical visits, we can help reduce the environmental impact of patient transportation.

The production and disposal of medical waste is also an important consideration when discussing the environmental impact of heart disease prediction. Medical waste, such as used syringes, gloves, and bandages, can pose serious health risks to both patients and the environment if not properly disposed of. By reducing the need for medical interventions and treatments, heart disease prediction can help reduce the amount of medical waste generated, protecting both human health and the environment.

Finally, the energy consumption of medical facilities is another important environmental impact of heart disease prediction. Medical facilities, including hospitals and clinics, use large amounts of energy to maintain operations, including lighting, heating and cooling, and medical equipment. By improving the accuracy of heart disease prediction and reducing the need for more serious medical interventions, we can help reduce the energy consumption of medical facilities, reducing their environmental impact.

5.3 Ethical Aspects

The development and deployment of heart disease prediction algorithms raise important ethical considerations. These include issues related to privacy, bias, and accuracy of predictions.

Privacy is a major concern when it comes to heart disease prediction algorithms. Health information is highly personal and sensitive, and individuals have a right to keep it confidential. When collecting and storing data for heart disease prediction, it is essential to ensure that individuals' personal information is protected and not shared without their consent. Furthermore, there should be clear policies in place to ensure that the data collected is used only for the purpose of heart disease prediction and is not sold or used for other purposes.

Bias is another important ethical consideration in heart disease prediction. The algorithms used to make predictions can sometimes perpetuate existing biases in the data used to train them. It is important to ensure that the data used to train heart disease prediction algorithms is diverse and representative of the population as a whole, to minimize the potential for bias in predictions.

Accuracy of predictions is another important ethical consideration in heart disease prediction. False positive or false negative predictions can have serious consequences for individuals and their health. False positive predictions can lead to unnecessary medical interventions and treatments, while false negative predictions can result in delayed or missed diagnoses. It is important to ensure that heart disease prediction algorithms are tested and validated to ensure their accuracy, and that individuals are made aware of the limitations and potential consequences of the predictions made by these algorithms.

It is important to consider the potential impact of heart disease prediction algorithms on access to health care. These algorithms may be used to identify individuals at high risk of heart disease, but if access to health care is limited, they may not receive the necessary treatments and interventions to manage their condition. It is important to ensure that heart disease prediction algorithms are used in conjunction with access to appropriate health care services, to maximize their potential to improve health outcomes.

5.4 Sustainability Plan

The development of an accurate heart disease prediction algorithm has been the main focus in our work. To ensure the long-term success and impact of this work, a sustainability plan has been put in place. This plan includes regularly updating the algorithm with new data to maintain its accuracy, collaborating with healthcare providers and patient organizations to integrate it into clinical practice, securing ongoing funding for maintenance and updates, and educating the public and healthcare providers about the benefits of heart disease prediction. By implementing this sustainability plan, we aim to make the algorithm widely available and usable for many years to come, ultimately improving health outcomes for those at risk of heart disease.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

In this study, we aimed to address the critical issue of heart disease and its impact on public health by developing a highly accurate algorithm for heart disease prediction. To achieve this goal, we evaluated the performance of five different machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, AdaBoost, and XGBoost. Our results showed that Random Forest (RF) and XGBoost obtained the highest accuracy of 99%, demonstrating their potential in accurately predicting heart disease.

This research sheds light on the significance of using machine learning algorithms in healthcare, specifically for heart disease prediction. The high accuracy achieved by RF and XGBoost algorithms highlights the potential for these algorithms to be used in clinical settings to improve health outcomes for those at risk of heart disease. Moreover, the results of this study can be useful for healthcare providers, patients, and policymakers in developing effective strategies for the early detection and prevention of heart disease.

6.2 Conclusion

Heart disease is difficult, and it creates several losses annually. The person may swiftly encounter serious repercussions unless the initial symptoms of cardiovascular disease are overlooked. Five machine-learning techniques were applied in our system to forecast coronary heart disease utilizing the data we have gathered. The deep neural network fared better in the comparative analysis had acquired a higher accuracy using our dataset. Random Forest and XGBoost outperformed the others with an accuracy of 99% when applied to the dataset. These computational intelligence approaches are crucial for making diagnosis in the medical field. These methods ought to make the doctor's job easier.

The same methods can be used in the future to predict different diseases, and the diagnosis of myocardial infarction would be performed utilizing highly sophisticated technologies. Improved performance will result from the diversity of resources in extraction of information and comprehension of the issues involved in measuring and gathering data. The production of a legitimate website for the assessment of cardiovascular disease would be accomplished in conjunction with refreshing existing practical applications using innovative features. Feature selection techniques may be used to select the optimum model parameters, which will enhance the overall predictability of the model. The initial segment of the article covers how to use Python to anticipate heart illness according to given circumstances. Python is an advanced language of programming that has fast processing intervals, is object-oriented, and is capable of dynamically constructing options. It is simpler to predict the course of cardiac disease using this language. The cardiac care industry collects data from several organizations and patients in order to adopt the optimum data strategy. The whole healthcare delivery system would improve as a result of doctors being able to easily show this improved treatment paradigm. This heart disease prediction model is primarily used to treat cardiac conditions, different organizations and initiatives could provide improved patient outcomes through scalable and dynamic environmental applications and reach a decision regarding the model's problem.

In addition to this, chapter two covered using Python to find the presence of cardiac disorders. The data collection, which comprises patient information including age, gender, cholesterol, hypertension, and various other aspects, is fundamental to this approach. The dataset visualization tool, however, makes use of a wide range of separately imported libraries. One of the potent computer programs is Python, which supports analytical abilities to generate significant insights regarding health information concerning cardiovascular events. Nevertheless, it also adheres to HIPAA regulations, ensuring medical data confidentiality. The approach for using machine learning to identify cardiac illness is also explored in chapter 3, which includes a variety of algorithms. The ability to predict when a threat may manifest requires machine learning.

The random forest technique seems to be best for this project to produce heart disease diagnosis. The Python approach for detecting heart disease uses this algorithm. This application's accuracy percentage over training data qualifies as substantial.

A model for machine learning has also been crucial in developing accuracy and predicting outcomes using training data. The scenario illustrates how the decision to use random forest classification is made using both the precise dataset and the decision tree. The analysis process is the principal subject of discussion in this segment. Furthermore, it functions with categorical variables; it changes a set of categorical values into fictitious values through the use of binary numbers 1 and 0. The information generated by this application is utilized to predict and establish the specifications of the testing database's functionality; a section notably addresses several health-related markers. The decision tree is the only exception, and other than that, the random forest serves the best values. This is the quickest and most accurate way to anticipate heart disease.

6.3 Implication for Further Study

In our study, we created a medical technique to aid patients and those exhibiting symptoms identify cardiac disorders. The random forest procedure, which produces improved accuracy, is the foundation of our strategy. Its cost of development is minimal. A useful source of data for further analysis may also be found in research outcomes. Therefore, using data and explanations, we can work on developing treatments for cardiovascular disorders in order to carry out logical and rational thinking. This experiment's output would be beneficial in evaluating service providers in a standard manner. Our research aims to provide healthcare with theoretical and practical advances. This study also considers angiography and conventional invasive-based approaches to be appropriate and well-known diagnostic methods for diagnosing heart problems. It is a flaw in the foretelling of cardiac disease.

Additionally, for intelligent learning, precise and effective computational approaches for related disease forecasting are required. This prediction technique is described here to anticipate as well as treat the intention of anticipating as well as treating cardiovascular dysfunction.

This classification method relies on pruning and data-cleansing methods. This method develops a sample collection appropriate for data extraction, chooses an effective strategy, and significantly increases the application's accuracy.

The results of this research on heart disease prediction using machine learning algorithms provide important insights into the potential of these algorithms for improving health outcomes. However, there is still room for further study to optimize their accuracy and implementation. This can include expanding the dataset to be more diverse and inclusive, combining different algorithms, incorporating other relevant factors, evaluating the cost-effectiveness of implementation, and integrating the algorithms into clinical practice. These steps will help to fully realize the potential of these algorithms and improve the lives of those at risk of heart disease. Further research in this field is essential to ensure the continued development and success of these algorithms.

References:

[1] Safial Islam Ayon, Md. Milon Islam & Md. Rahat Hossain (2022) Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques, IETE Journal of Research, 68:4, 2488-2507

[2] Chang, Victor, et al. "An artificial intelligence model for heart disease detection using machine learning algorithms." Healthcare Analytics 2 (2022): 100016.

[3] Ahsan, Md Manjurul, and Zahed Siddique. "Machine learning-based heart disease diagnosis: A systematic literature review." Artificial Intelligence in Medicine (2022): 102289.

[4] Ketu, S., Mishra, P.K. Empirical Analysis of Machine Learning Algorithms on Imbalance Electrocardiogram Based Arrhythmia Dataset for Heart Disease Detection. Arab J Sci Eng 47, 1447–1469 (2022)

[5] MAlnajjar, Mohammad Khaleel, and Samy S. Abu-Naser. "Heart Sounds Analysis and Classification for Cardiovascular Diseases Diagnosis using Deep Learning." (2022).

[6] M. Mamun, M. M. Uddin, V. Kumar Tiwari, A. M. Islam and A. U. Ferdous, "MLHeartDis:Can Machine Learning Techniques Enable to Predict Heart Diseases?," 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2022, pp. 0561-0565

[7] Konstantonis, George, et al. "Cardiovascular disease detection using machine learning and carotid/femoral arterial imaging frameworks in rheumatoid arthritis patients." Rheumatology International 42.2 (2022): 215-239.

[8] Brites, Ivo SG, et al. "Machine learning and iot applied to cardiovascular diseases identification through heart sounds: A literature review." International Conference on Information Technology & Systems. Springer, Cham, 2022.

[9] Gulati, Seema, Kalpna Guleria, and Nitin Goyal. "Classification and Detection of Coronary Heart Disease using Machine Learning." 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). IEEE, 2022.

[10] agavelli, Umarani, Debabrata Samanta, and Partha Chakraborty. "Machine Learning Technology-Based Heart Disease Detection Models." Journal of Healthcare Engineering 2022 (2022).

[11] Mamun, Muntasir, et al. "Heart failure survival prediction using machine learning algorithm: am I safe from heart failure?." 2022 IEEE World AI IoT Congress (AIIoT). IEEE, 2022.

[12] Saboor, Abdul, et al. "A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms." Mobile Information Systems 2022 (2022).

[13] Cenitta, D., R. Vijaya Arjunan, and K. V. Prema. "Ischemic heart disease multiple imputation technique using machine learning algorithm." Engineered Science (2022).

[14] Mhamdi, Lotfi, et al. "Artificial Intelligence for Cardiac Diseases Diagnosis and Prediction Using ECG Images on Embedded Systems." Biomedicines 10.8 (2022): 2013.

[15] Nadakinamani, Rajkumar Gangappa, et al. "Clinical Data Analysis for Prediction of Cardiovascular Disease Using Machine Learning Techniques." Computational Intelligence and Neuroscience 2022 (2022).

[16] Kedia, Shyam, and Megha Bhushan. "Prediction of mortality from heart failure using machine learning." 2022 2nd International Conference on Emerging Frontiers in Electrical and Electronic Technologies (ICEFEET). IEEE, 2022.

[17] Absar, Nurul, et al. "The efficacy of machine-learning-supported smart system for heart disease prediction." Healthcare. Vol. 10. No. 6. MDPI, 2022.

[18] El-Hasnony, Ibrahim M., et al. "Multi-label active learning-based machine learning model for heart disease prediction." Sensors 22.3 (2022): 1184.

[19] Wang, Peng, et al. "A wearable ECG monitor for deep learning based real-time cardiovascular disease detection." arXiv preprint arXiv:2201.10083 (2022).

[20] Q. Yao, R. Wang, X. Fan, J. Liu, and Y. Li, “Multi-class arrhythmia detection from 12-lead varied-length ecg using attention-based time-incremental convolutional neural network,” Information Fusion, vol. 53, pp. 174–182, 2020.

[21] Javeed, Ashir, et al. "Machine learning-based automated diagnostic systems developed for heart failure prediction using different types of data modalities: a systematic review and future directions." Computational and Mathematical Methods in Medicine 2022 (2022).

[22] Chakraborty, Aritra, et al. "A comparative study of myocardial infarction detection from ECG data using machine learning." Advanced Computing and Intelligent Technologies. Springer, Singapore, 2022. 257-267.

[23] V.V. Kumar, Healthcare Analytics Made Simple: Techniques in Healthcare Computing using Machine Learning and Python, Packt Publishing Ltd., 2018, [Accessed on 5th March, 2021].

[24] Kamil Pytlak.(2022, February).Personal Key Indicators of Heart Disease. Version 1. Retrieved July 3, 2022 from https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-ofheart-disease.

[25] B Padmaja, E. (2022). Early and Accurate Prediction of Heart Disease Using Machine Learning Model. Retrieved 13 July 2022, from https://turcomat.org/index.php/turkbilmat/article/view/8438

[26] M. Mamun, S. B. Shawkat, M. S. Ahammed, M. M. Uddin, M. I. Mahmud, A. M. Islam, "Deep Learning Based Model for Alzheimer's Disease Detection Using Brain MRI Images", 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2022

[27] M. Mamun, A. Farjana, M. A. Mamun, M. S. Ahammed and M. M. Rahman, "Heart failure survival prediction using machine learning algorithm: am I safe from heart failure?," 2022 IEEE World AI IoT Congress (AIIoT), 2022, pp. 194-200, doi: 10.1109/AIIoT54504.2022.9817303.

[28] G. N. Ahmad, H. Fatima, S. Ullah, A. Salah Saidi and Imdadullah, "Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV," in IEEE Access, vol. 10, pp. 80151-80173, 2022, doi: 10.1109/ACCESS.2022.3165792.

[29] Lu, Haoxuan, et al. "Research Progress of Machine Learning and Deep Learning in Intelligent Diagnosis of the Coronary Atherosclerotic Heart Disease." Computational and Mathematical Methods in Medicine 2022 (2022).

[30] Garavand, Ali, et al. "Efficient Model for Coronary Artery Disease Diagnosis: A Comparative Study of Several Machine Learning Algorithms." Journal of Healthcare Engineering 2022 (2022).

[31] Sarra, Raniya R., et al. "Enhanced heart disease prediction based on machine learning and χ2 statistical optimal feature selection model." Designs 6.5 (2022): 87.

[32] Shandhi, Md Mobashir Hasan, et al. "Estimation of changes in intracardiac hemodynamics using wearable seismocardiography and machine learning in patients with heart failure: a feasibility study." IEEE Transactions on Biomedical Engineering 69.8 (2022): 2443-2455.

[33] Nadakinamani, Rajkumar Gangappa, et al. "Clinical Data Analysis for Prediction of Cardiovascular Disease Using Machine Learning Techniques." Computational Intelligence and Neuroscience 2022 (2022).

[34] Qian, Xin, et al. "A cardiovascular disease prediction model based on routine physical examination indicators using machine learning methods: A cohort study." Frontiers in cardiovascular medicine 9 (2022).

[35] Elias, Pierre, et al. "Deep learning electrocardiographic analysis for detection of left-sided valvular heart disease." Journal of the American College of Cardiology 80.6 (2022): 613-626.

[36] Panteris, Eleftherios, et al. "Machine Learning Algorithm to Predict Obstructive Coronary Artery Disease: Insights from the CorLipid Trial." Metabolites 12.9 (2022): 816.

[37] Whig, Pawan, Ketan Gupta, and Nasmin Jiwani. "Real-Time Detection of Cardiac Arrest Using Deep Learning." AI-Enabled Multiple-Criteria Decision-Making Approaches for Healthcare Management. IGI Global, 2022. 1-25.

[38] Hussain, Lal, et al. "Detecting congestive heart failure by extracting multimodal features with synthetic minority oversampling technique (SMOTE) for imbalanced data using robust machine learning techniques." Waves in Random and Complex Media 32.3 (2022): 1079-1102.

[39] Gupta, Ankur, et al. "C-CADZ: computational intelligence system for coronary artery disease detection using Z-Alizadeh Sani dataset." Applied Intelligence 52.3 (2022): 2436-2464.

[40] Ghosh, Pronab, et al. "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques." IEEE Access 9 (2021): 19304-19326.

[41] Jindal, Harshit, et al. "Heart disease prediction using machine learning algorithms." IOP conference series: materials science and engineering. Vol. 1022. No. 1. IOP Publishing, 2021.

[42] Ali, Md Mamun, et al. "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison." Computers in Biology and Medicine 136 (2021): 104672.

[43] Bharti, Rohit, et al. "Prediction of heart disease using a combination of machine learning and deep learning." Computational intelligence and neuroscience 2021 (2021).

[44] Kumar, Yogesh, et al. "Heart failure detection using quantum-enhanced machine learning and traditional machine learning techniques for internet of artificially intelligent medical things." Wireless Communications and Mobile Computing 2021 (2021).

[45] Kavitha, M., et al. "Heart disease prediction using hybrid machine learning model." 2021 6th International Conference on Inventive Computation Technologies (ICICT). IEEE, 2021.

[46] Kishor, Amit, and Wilson Jeberson. "Diagnosis of heart disease using internet of things and machine learning algorithms." Proceedings of second international conference on computing, communications, and cyber-security. Springer, Singapore, 2021.

[47] Senan, Ebrahim Mohammed, et al. "Score and Correlation Coefficient-Based Feature Selection for Predicting Heart Failure Diagnosis by Using Machine Learning Algorithms." Computational and Mathematical Methods in Medicine 2021 (2021).

[48] Diwakar, Manoj, et al. "Latest trends on heart disease prediction using machine learning and image fusion." Materials Today: Proceedings 37 (2021): 3213-3218.

[49] Singh, Harinder, et al. "Accuracy detection of coronary artery disease using machine learning algorithms." Applied Nanoscience (2021): 1-7.

[50] Asif, Md, et al. "Performance Evaluation and Comparative Analysis of Different Machine Learning Algorithms in Predicting Cardiovascular Disease." Engineering Letters 29.2 (2021).

[51] Flores, Alyssa M., et al. "Leveraging machine learning and artificial intelligence to improve peripheral artery disease detection, treatment, and outcomes." Circulation Research 128.12 (2021): 1833-1850.

[52] Su, Yu-Sheng, Ting-Jou Ding, and Mu-Yen Chen. "Deep learning methods in internet of medical things for valvular heart disease screening system." IEEE Internet of Things Journal 8.23 (2021): 16921-16932.

[53] Salhi, Dhai Eddine, Abdelkamel Tari, and M. Kechadi. "Using machine learning for heart disease prediction." International Conference on Computing Systems and Applications. Springer, Cham, 2021.

[54] Gürfidan, Remzi, and Mevlüt Ersoy. "Classification of death related to heart failure by machine learning algorithms." Advances in Artificial Intelligence Research 1.1 (2021): 13-18.

[55] Plati, Dafni K., et al. "A machine learning approach for chronic heart failure diagnosis." Diagnostics 11.10 (2021): 1863.

[56] Alsafi, Haedar Emad Sharef, and Osman Nuri Ocan. "A novel intelligent machine learning system for coronary heart disease diagnosis." Applied Nanoscience (2021): 1-8.

[57] Ibrahim, Ibrahim, and Adnan Abdulazeez. "The role of machine learning algorithms for diagnosing diseases." Journal of Applied Science and Technology Trends 2.01 (2021): 10-19.

[58] Arumugam, K., et al. "Multiple disease prediction using Machine learning algorithms." Materials Today: Proceedings (2021).

[59] Li, Jian Ping, et al. "Heart disease identification method using machine learning classification in e-healthcare." IEEE Access 8 (2020): 107562-107582.

[60] Singh, Archana, and Rakesh Kumar. "Heart disease prediction using machine learning algorithms." 2020 international conference on electrical and electronics engineering (ICE3). IEEE, 2020.

Heart disease paper report

ORIGINALITY REPORT

26%	20%	16%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.tandfonline.com Internet Source	6%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	5%
3	Submitted to Intercollege Student Paper	3%
4	Submitted to Coventry University Student Paper	2%
5	www.researchgate.net Internet Source	2%
6	Muntasir Mamun, Md. Milon Uddin, Vivek Kumar Tiwari, Asm Mohaimenul Islam, Ahmed Ullah Ferdous. "MLHeartDis:Can Machine Learning Techniques Enable to Predict Heart Diseases?", 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2022 Publication	1%
7	Submitted to University of Sunderland Student Paper	1%