AN ANALYTICAL HEART DISEASE PREDICTION SYSTEM USING MACHINE LEARNING

BY

Zahidul Islam Rakib
ID: 181-15-1972

Mohammad Nadim Mahmud Neon
ID: 181-15-1746

AND

Sadiqur Rahman
ID: 181-15-1782

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

Supervised By
Mohammad Jahangir Alam
Lecturer
Department of Computer Science & Engineering
Daffodil International University

Co-Supervised By
S.M Aminul Haque
Associate Professor
Department of Computer Science & Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

OCTOBER 2021

APPROVAL

This Project titled "Heart Disease: An Analytical Prediction System Using Machine Learning", submitted by Zahidul Islam Rakib, Mohammad Nadim Mahmud Neon and Sadiqur Rahman 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 18.09.2021.

BOARD OF EXAMINERS

Tania Khatun

Tania Khatun — **Internal Examiner**
Senior Lecturer
Department of Computer Science and Engineering
Daffodil International University

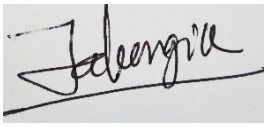
Mohammad Jahangir Alam — **Internal Examiner**
Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Dewan Md. Farid — **External Examiner**
Associate Professor
Department of Computer Science & Engineering
United International University, Bangladesh

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Mohammad Jahangir Alam, Lecturer, Department of Computer Science & Engineering** 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:

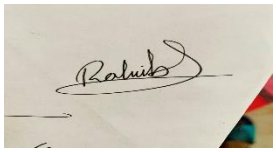


Mohammad Jahangir Alam
Lecturer
Department of Computer Science & Engineering
Daffodil International University

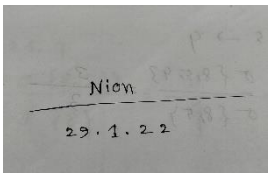
Co-Supervised by:

S.M Aminul Haque
Associate Professor
Department of Computer Science & Engineering
Daffodil International University

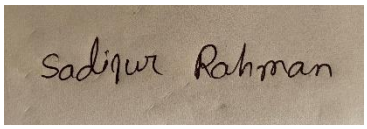
Submitted by:



Zahidul Islam Rakib
ID: 181-15-1972
Department of Computer Science & Engineering
Daffodil International University



Mohammad Nadim Mahmud Neon
ID: 181-15-1746
Department of Computer Science & Engineering
Daffodil International University



Sadiqur Rahman
ID: 181-15-1782
Department of Computer Science & Engineering
Daffodil International University

ACKNOWLEDGEMENT

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

We really grateful and wish our profound our indebtedness to **Mohammad Jahangir Alam**, **Lecturer**, Department of Computer Science & Engineering, Daffodil International University, Dhaka. Deep Knowledge & 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, reading many inferior draft and correcting them at all stage havemade it possible to complete this project.

We would like to express our heartiest gratitude to Mohammad Jahangir Alam, S.M Aminul Haque and Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

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

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

ABSTRACT

Heart disease is the top cause of mortality in every country on the planet, according to the World Health Organization. Every year, hundreds of individuals all over the world lose their lives as a result of this disease. Due to the fact that the procedure is both time-consuming and costly, physicians may still predict heart illness with reasonable accuracy. As a result, the authors propose a technique to help physicians in diagnosing and making better decisions by anticipating the beginning of cardiac disease in order to aid them in their decisions. This research utilizes risk factor data collected from medical records to train five distinct models using five different machine learning techniques. It was decided to utilize Logistic Regression, K-Nearest Neighbours, Gaussian Naive Bayes, Decision Tree, and AdaBoost as the best method for predicting heart disease based on the dataset. A number of variables are considered in order to choose the most appropriate algorithm. With the UCI dataset, the greatest accuracy was achieved by Naïve Bayes with 90.32% accuracy, followed by KNN and Logistic Regression with 87.1% accuracy. In this article, the accuracy is predicted, and additional variables such as the Jaccard Score and the Cross Validated Score are shown. The authors conclude by recommending that various validation techniques be used on prospectively acquired data in order to support the suggested methodology.

Keywords: Heart disease, Machine learning, Classification, UCI heart disease data set.

Table of Contents

APPROVAL II
DECLARATION III
ACKNOWLEDGEMENT IV
ABSTRACT V
LIST OF FIGURES IX
LIST OF TABLES… V

CHAPTER 1: INTRODUCTION 1
1.1 Introduction 1
1.2 Motivation 3
1.3 Rational of the Study 3
1.4 Research Questions 4
1.5 Expected Outcome 4
1.6 Project Management and Finance 5
1.7 Project Layout 5
CHAPTER 2: BACKGROUND & LITERATURE REVIEW 5
2.1 Preliminaries 5
2.2 Machine Learning Techniques 5
2.2.1 Supervised Learning 5
2.3 Classification Techniques 5
2.3.1 Learning 6
2.3.2 Classification 6
2.4 Related Works 6
2.5 Comparative Analysis 10
2.6 Challenges 12
CHAPTER 3: RESEARCH METHODOLOGY & SYSTEM ARCHITECTURE 13
3.1 Algorithmic Details 13
3.1.1 Logistic Regression 13
3.1.2 K-Nearest Neighbors 13
3.1.3 Gaussian Naïve Bayes 14
3.1.4 Decision Tree 14

3.1.5 Adaptive Boosting (AdaBoost) 14
3.2 Proposed System 15
3.2.1 Data Collection....15
3.2.2 Dataset....15
3.2.3 Data Pre-processing....16
3.2.4 Data Normalization 16
3.2.5 Splitting Dataset 16
3.2.6 Implement Different Algorithms....16
3.2.7 Result Analysis....17
3.2.8 Extract Appropriate Algorithm....17
3.2.9 Predict....17
CHAPTER 4: EXPERIMENTAL RESULTS & DISCUSSION....18
4.1 Experimental Results....18
4.1.1 Data Acquisition....18
4.1.2 Data Utilization 19
4.2 Result & Discussion 20
4.2.1 Confusion Matrix....20
4.2.2 Classification Report 23
4.3 Result Analysis....24
4.3.1 Accuracy....24
4.3.2 Jaccard Score 24
4.3.3 Cross Validated Score 25
4.3.4 Misclassification & Error 26
CHAPTER 5: FUTURE SCOPE & CONCLUSION 28
5.1 Future Scope....28
5.2 Conclusion....28
References 29

LIST OF FIGURES

Figure 3. 1: Proposed Method to Predict Heart Disease 15

Figure 4. 1: Accuracy Chart.....24
Figure 4. 2: Jaccard Score Chart.....25
Figure 4. 3: Cross Validated Score 26

LIST OF TABLES

Table 4. 1: Data Acquisition & Null Value Percentage.....18
Table 4. 2: Dataset Description 19
Table 4. 3: Confusion Matrix Resutl 21
Table 4. 4: Classification Report 23
Table 4. 5: Accuracy, Jaccard, Cross Validated and AUC Score.....26
Table 4. 6: Miss Classification & Error.....27

CHAPTER 1
INTRODUCTION

1.1 Introduction

Almost everyone in this time period is so concerned with their everyday lives and occupations that they do not have enough time to take care of their own personal well-being. Because of a lack of time, they are unable to maintain a balanced diet and, in order to save time, resort to fast food restaurants. They are also unable to participate in regular physical exercise because of their condition. They have a busy lifestyle and are prone to become irritated at certain points. When they are upset, it is possible that their blood pressure may be raised as well. The consequence is that individuals get ill and suffer from various cardiac issues as the result of their actions. Heart disease is a condition that affects the heart or the blood vessels in the body. Coronary artery disease is the most prevalent form of heart disease, and it is characterized by symptoms such as chest discomfort, heart attacks, and stroke.

The heart is a very important organ in our body. People die when the system stops working properly. The heart is in charge of pumping blood throughout our bodies and into our veins. Our circulatory system revolves around the heart, which acts as its focal point. The heart is in charge of the circulation of blood to and from all of the organs and organ systems in your body. Our blood is responsible for transporting the oxygen and nutrients that our organs need in order to operate correctly. The valves in our hearts are important for ensuring that blood flows in the correct direction throughout our bodies. Electricity regulates the pace and rhythm of our heartbeats, among other things, since our hearts have electrical circuitry built in to them. It is possible that failure of the heart to function correctly may result in the failure of other organs in the body, including the liver. As a consequence, it becomes more difficult to maintain the health of the heart and other organs.

A number of risk factors for cardiovascular disease include cigarette smoking, a poor diet, high cholesterol, high blood pressure, a lack of physical activity, and obesity, to name a few. Behavioral risk factors may cause high blood pressure, overweight, elevated lipids in the bloodstream, elevated glucose in the bloodstream, and obesity in certain people. These "intermediate-risk variables" may be detected in primary care settings and suggest a greater likelihood of having a heart attack, having a stroke, having heart failure, or having

other health consequences. In many instances, the underlying disease of the blood arteries shows itself without the presence of any signs or symptoms. A heart attack may be the first sign of a more serious issue that is about to manifest itself. The middle of your chest may be uncomfortable or excruciatingly painful, and discomfort or agonizing pain may also occur in your upper arms, lower arms, jaw, left shoulder and elbows, and lower back when you have a heart attack. In addition, the patient may have trouble breathing, nausea, or vomiting, light-headedness or faintness, paleness, and a cold sweat, among other symptoms. In addition to these symptoms, women may suffer shortness of breath, nausea, back or jaw pain, and vomiting during pregnancy. The most frequent symptom of a stroke is sudden weakness of the face, arm, or leg, which happens on just one side of the body the majority of the time, according to the American Heart Association.

The possibility of preventing a heart illness from developing entirely exists if the condition can be anticipated and treated at its earliest stage. It is hoped that researchers from all around the globe would collaborate to attempt to anticipate cardiac disease in its very early stages. Computer-aided design (CAD) and machine learning are becoming more significant tools in medical research. Training and testing datasets are used in Machine Learning, and they are split into two categories: training and testing. There are many different Machine Learning Algorithms that can be used to predict anything. A number of algorithms, including the Logistic Regression, K-Nearest Neighbours, Gaussian Naive Bayes, Decision Tree, and AdaBoost, were used in this investigation. In order to train these algorithms, they must be fed a relevant dataset acquired from the University of California, Irvine Machine Learning Repository. The authors make an effort to predict whether or not a person has coronary artery disease by using the algorithms described above. To gather data for the model's training, biological variables such as gender, age, blood pressure, and cholesterol were used to collect information. The cost of detecting and treating heart illness is too high in poor nations. If a person has been suffering from cardiac disease for an extended period of time, they may need surgical intervention. Consequently, the author of this article selected the algorithm that he felt would provide the most accurate results for this investigation

1.2 Motivation

There are many variables that influence a person's heart on a daily basis. Many issues are developing at an alarming rate, and new cardiac illnesses are being discovered at an alarming rate. In today's stressful environment, the heart, as an important organ in the human body that pumps blood throughout the body for blood circulation, is critical, and its health must be preserved in order to maintain a healthy lifestyle. The health of a person's heart is determined by the experiences that person has had throughout his or her life, and it is entirely reliant on the professional and personal behaviors of that individual. There may also be a number of hereditary variables involved in the transmission of a particular kind of heart disease from generation to generation. Worldwide, more than 12 million fatalities occur each year as a result of different kinds of heart disease, which is often referred to as cardiovascular disease, according to data from the World Health Organization. Heart disease is a broad phrase that encompasses a wide range of illnesses that affect the heart and arteries of a human being in a number of different ways. Heart disease affects individuals of all ages, including children and teenagers who are between the ages of 20 and 30 years old. It is possible that the increase in the likelihood of heart disease among young people is due to poor eating habits, lack of sleep, a restless nature, depression, and a variety of other factors such as obesity, poor diet, family history, high blood pressure, high blood cholesterol, sedentary behavior, family history, tobacco use, and hypertension.

1.3 Rational of the Study

Assuming Heart Disease can be analyzed and anticipated early, it tends to be relieved by appropriate treatment rapidly. In reality, individuals need to test numerous things in the lab to turn out to be certain that they have Heart Disease or not, and it is extremely tedious. That is the reason, a model has been proposed and prepared by a significant dataset that can foresee Heart Disease at any stage. Primary reasoning spotlight to deal with this review to make individuals' life simpler and to save time. A web interface has been created where patients can give information and can foresee Heart Disease at home. It will be useful for specialists and Heart Disease patients

1.4 Research Questions

There are various inquiry can be posed with regards to this review. From various people, a bunch of inquiry has been separated to make this concentrate more minimized.

- Why Heart Disease Prediction was the objective of this review?

  Heart Disease is perhaps the most concerning issue on the planet. It deteriorates after some time. It has many stages and at the last stage the heart quits working by any means. Then, at that point, it drives an individual absurdly. Along these lines, on the off chance that Heart Disease can be anticipated at beginning phase, it tends to be relieved with legitimate treatment and keeping up with many guidelines and guidelines. That is the reason Heart Disease was the objective of the review.

- Why machine learning approach? Does it reliable?

  ML is a most famous procedure which is fundamentally utilized for any sort of expectation. Utilizing an immense number of information, a model can prepare itself and can anticipate any result. Utilizing ML approach in clinical dataset it tends to be anticipated Heart Disease without any problem. In the present circumstances, the world is going through a 4 modernization period. Assuming a situation can be considered around 10 years before when Artificial Intelligence or Machine Learning was not fostered that much, around then these superb focuses were only a name with some of numerical rationale. Yet, presently, a big part of the entire planet's innovation relies upon Artificial Intelligence. Subsequently, appropriate practice and more accuracy on this field can make it more solid however it is dependable enough at this point.

- What are the motivations to utilize 5 separate calculations?

  Utilizing 5 unique calculations, one appropriate calculation was designated that suits the Heart Disease dataset. Separating 5 calculations and investigating them the most appropriate calculation has been acquired that will have less blunder rate and most elevated precision rate. If by some stroke of good luck one calculation was chosen, it is difficult to track down the best calculation as nobody don't know which calculation will suit the dataset in the most proficient and powerful manner.

1.5 Expected Outcome

There have been a few times throughout this study when the primary goal or anticipated result has been altered. It contributes to the clarification of the precise result of this research. This study offers the potential to anticipate the onset of Heat Disease in its early stages. Heart disease may be predicted with precision using mathematical calculations and algorithms. While heart disease has become a significant threat to today's population, this method of forecasting heart disease will help in the identification of risk factors and the determination of whether a patient is at danger of having a heart attack or not. The most important result of this article is the discovery of a perfect algorithm that is appropriate for predicting cardiac disease with the aid of machine learning techniques.

1.6 Project Management and Finance

This project was managed using Python language with with a minimum PC requirement.

Components	**Price (Current Market) [in Tk]**
CPU: Intel i5 7th Gen	18,550
RAM: DDR4 8 GB	3,450
ROM: 1 TB	3,800
GPU: GTX 1050 TI 4 GB	18,170
Monitor 22 inch	10,500
Mouse	500
Keyboard	500
Grand Total	**55,470**

1.7 Project Layout

In this report, A total of five separate chapters are addressed in order to make this research report more compact and efficient for any readers or researchers.

Chapter 1 gives an important introduction about this research work. This is linked to Heart Disease and provides a short overview of the condition. This chapter describes the research motivation, the reason for the investigation, the pertinent research questions, the anticipated result, and the overall management information, as well as the financial implications of the study.

Chapter 2 gives the detailed report about background of this study. Based on the findings of this research study, machine learning systems, categorization information, and associated work have been developed. This chapter also includes descriptions of comparative analysis and the breadth of the issue statement, as well as perceived difficulties.

Chapter 3 gives the descriptive information about methodology, proposed system for this research study. Algorithmic details for each used algorithms are described from mathematical scratch to current condition is discussed.

Chapter 4 gives the complete result analysis for each steps result. In concludes with best accuracy score with best algorithm, jaccard score, cross validated score, confusion matrix. Misclassification, Mean Absolute Error and Mean Squared Error are described in the last segment of this chapter.

Chapter 5 shows the future scope of this research work where it is briefly described as the extension of this research study. This chapter concludes the entire research report with useful conclusion where core findings of this research is briefly discussed.

CHAPTER 2

BACKGROUND & LITERATURE REVIEW

2.1 Preliminaries

In this chapter, a wide discussion about the background of "Heart Disease: An Analytical Prediction System Using Machine Learning" entitled research study.

2.2 Machine Learning Techniques

Category classification may be accomplished via the use of supervised machine learning and categorization techniques. A machine learning system is described as one that is self-contained and capable of collecting and integrating data on a continuous basis for the goal of making judgments. It is possible to learn from previous events, make analytical observations, and use other techniques in order to build a system that is always improving. A wide range of machine learning techniques are available in different forms and sizes.

2.2.1 Supervised Learning

Supervised machine learning methods are extensively utilized in this study. Supervised machine learning methods create future predictions by analyzing labeled samples of previous occurrences. The learning method constructs an inferred function from an examination of a well-known training dataset in order to produce predictions about the output values. If the learning algorithm compares its output to what was intended, it may identify and correct any mistakes. This thesis is concerned with the first stage categorization of Heart Disease using a variety of supervised learning methods.

2.3 Classification Techniques

Classification is a kind of data analysis that generates models that describe key data classes from large amounts of data. It is the most well-known and extensively utilized machine learning method available today. Classifier models, also known as classification models, are capable of predicting categorical class labels when trained under supervision. The predictions are discrete and not in any particular sequence. There is no way to get an intermediate value from a classifier. For example, a classifier may be created to determine

if an image includes the image of a frog or the image of a fish. Either "frog" or "fish" will be predicted as the outcome. There is no way to get an intermediate value from a classifier. On labeled data, a classification learning method may be applied to improve accuracy. When it comes to categorization learning, there are two kinds of data. One kind of data is referred to as training data, while another is referred to as test data. Training data are used to construct the model, while test data are used to verify the model's correctness and accuracy. A two-step classification procedure may be used to describe the classification process.

2.3.1 Learning

After an appropriate method and training data have been identified, a classifier is constructed and tested against real-world data during the learning phase. In the event that a classification algorithm and training data are combined, the outcome is a classifier. A classifier is basically a collection of rules that can be applied to a number of different scenarios in a variety of different contexts.

2.3.2 Classification

Using the classifier or model created during the prediction phase, it is feasible to anticipate which class of unknown data will be predicted as a consequence of the learning phase's results. This part makes use of the test data in order to determine whether or not the predictions made by a model are correct or incorrect.

2.4 Related Works

Using different methods, in this paper of Motarwar et al. [1] the researchers present a machine learning framework to predict the likelihood of cardiac disease. It uses five algorithms: Random Forest, Naive Bayes, SVM, Hoeffding Decision Tree, and Logistic Model Tree (LMT). The Cleveland dataset is used to train the model. The dataset is preprocessed, then selected for prominence. On uses the resulting dataset for framework training. The findings indicate that Random Forest provides the best accuracy. This paper's

main goal is to better forecast heart disease risk. It is 95.08 percent for Random Forest. The accuracy of each method was compared using the Cleveland dataset.

Using machine learning to identify risk variables is promising. In this paper of Ghosh et al. [2] researchers present a model that combines several techniques to effectively predict cardiac disease. Using machine learning to identify risk variables is promising. This study utilized a merged dataset. In addition, the researcher used machine learning techniques to determine the model's accuracy, sensitivity, error rate, precision, and F1 score, along with NPR, FPR, and FNR. The findings are separated for comparison. Using RFBM and Relief feature selection techniques, the suggested model achieved the best accuracy (99.05%).

Healthcare experts that specialize in the field of cardiac disease have their own limitations, and they are unable to predict with high accuracy the likelihood of developing heart illness. It is the goal of this paper of Saw et al. [3] to improve the accuracy of Heart Disease predict accuracy using the Logistic Regression model of machine learning while taking into consideration a health care dataset that classifies patients as having heart diseases or not based on the information contained in their medical records.

With the use of data analysis, doctors may be able to save more lives by better diagnosing their patients. Researchers now have access to data mining tools thanks to advances in software engineering. In this research of Tougui et al. [4] the researchers have chosen to compare six common data mining tools: Orange, Weka, RapidMiner, Knime, Matlab and Scikit-Learn, while employing six machine learning approaches: Logistic Regression, Support Vector Machine, K Nearest Neighbors, Artificial Neural Network, Naïve Bayes and Random Forest on a dataset of heart disease patients in order to achieve this. While RapidMiner's Support Vector Machine (SVM) had the best specificity, Matlab's Artificial Neural Network model (85.86 percent) had the highest accuracy and sensitivity (94.38%). However, the accuracy, sensitivity, and specificity of Knime's K Nearest Neighbors model were all below the industry average, with a combined score of 63.64%.

In this study of Ahmed et al. [5] proposed a real-time system for forecasting cardiac illness using medical data streams that characterize the current health state of a patient. The system is implemented in Matlab. The primary aim of the proposed system is to identify the most effective machine learning algorithm for predicting heart disease with the highest degree of accuracy. This study made use of four different machine learning algorithms. In

order to improve accuracy, they utilized hyperparameter tweaking and cross-validation in conjunction with machine learning. Results showed that the random forest classifier outperformed the other models by obtaining the greatest accuracy of 94.9 percent, outperforming them all.

In this paper Li et al. [6] presented a machine learning-based method for diagnosing cardiac disease. The system uses standard feature selection algorithms such as Relief, Minimal redundancy maximal relevance, least absolute shrinkage selection operator, and Local learning to remove irrelevant and redundant features. A new rapid conditional mutual information feature selection method is presented. The suggested feature selection technique (FCMIM) improves SVM accuracy by 92.37 percent compared to other proposed approaches.

In this paper Singh et al. [7] calculated the accuracy of machine learning algorithms for predicting heart disease. The algorithms considered are k-nearest neighbor, decision tree, linear regression, and support vector machine (SVM), and the dataset used for training and testing is from the University of California, Irvine (UCI). The k-nearest neighbor method achieved the greatest accuracy of 87%, which is greater than the accuracy achieved by the other algorithms tested in this article.

Various characteristics of heart disease are presented in this study of Shah et al. [8], as well as a model based on supervised learning techniques such as Naïve Bayes, decision trees, K-nearest neighbors, and a random-forest-based method. It makes use of preexisting data from the UCI collection of heart disease patients in Cleveland. According to this study, individuals are at risk of getting heart disease. With an accuracy of 90.789 %, K-nearest neighbor has the greatest accuracy score out of all of them.

This article of Samhitha et al. [9] investigates the use of outfit characterization to improve the accuracy of fragility estimates by combining various classifiers. The equipment was tested on a heart disease dataset. This paper's focus is not just on improving feeble order computations, but also on using a therapeutic dataset to show their usefulness in predicting infection early on. The study's findings indicate that group techniques like stowing and boosting may help weak classifiers predict heart disease risk. The forecast model is shown with several highlights and grouping methods. The researchers create an

improved exhibition level with an accuracy level of 88:7% using the half breed irregular woods with a straight model.

In this research Princy et al. [10], a cardiovascular dataset is categorized using various state-of-the-art Supervised Machine Learning algorithms that are specifically utilized for illness prediction and are applied to a variety of problems. The findings show that the Decision Tree classification model outperformed other methods such as Naive Bayes, Logistic Regression, Random Forest, SVM, and KNN in terms of predicting cardiovascular illnesses. It was the Decision Tree that provided the best outcome, with an accuracy rate of 73%. This method may be beneficial for physicians in terms of predicting the development of cardiac problems and providing suitable therapy in the future.

Using the Naive Bayes classification algorithm and the random forest classification technique, Ansari et. al [11] determined whether or not the patient had the disease under investigation. There was also a thorough evaluation of the two algorithms' overall performance. A dataset's classification techniques and general structure and complexity may be analyzed using the simulation results.

The neural network and logistic regression are two of the most widely used machine learning techniques for identifying cardiovascular disease. These researchers look at neural networks, K-nearest neighbor, naive bayes, and logistic regression, among other techniques, as well as composite methods that combine the aforementioned heart disease diagnostic algorithms, among other techniques. They built their system on the benchmark dataset from the University of California, Irvine Machine Learning Repository. In this research, participants with and without cardiac problems were compared. This study of Ali et. al [12] may prove to be a helpful tool for doctors and other health care professionals in providing appropriate patient advice. It may also aid in the prediction of potentially severe issues.

By using fingertip video, the ‘Modified Artificial Plant Optimization algorithm (MAPO)' was described in this paper of Yadav et. al [13]. On the heart disease dataset, the severity of heart illness was calculated using Logistic Regression, Naive Bayes, XGB (Extreme Gradient Boosting), and ANN (Artificial Neural Network). The MAPO given surpasses previous similar studies, with maximum relative errors (84 out of 100) compared to less than 5%. This study presents MAPO, a modified version of APO, and assesses its

effectiveness in predicting heart disease. Relational MAPO has been shown to be very accurate in choosing optimum features from relational datasets such as the heart disease, MNIST, and Framingham datasets.

Samhitha et. al [14] of this work provide a complete overview of the most commonly utilized predicted heart disease classification systems: Naive bayes (NB), Artificial Neural Network (ANN), K-nearest Neighbor (KNN), Support vector machine (SVM). In the medical field, machine learning and image fusion have been outlined, particularly for the prediction of heart disease. The algorithm's conclusion has been decided; the proposed system can be deployed in a practical setting. To improve algorithm precision and yield dependable results, more specialized feature selection procedures are applied.

This method of Princy et. al [15] predicts the likelihood of the development of heart disease. The findings of this method offer a percent-based assessment of the probability of getting heart disease in a given individual. The datasets are divided into categories based on medical criteria. In order to evaluate such factors, this system makes use of a data mining classification method. Using Python programming, the datasets are analyzed using two major Machine Learning Algorithms: The Decision Tree Method and the Naive Bayes Method, with the Decision Tree Method showing to be the most accurate algorithm of the two in terms of heart disease accuracy. Heart disease patients were correctly identified by the Naive Bayes classifier with an accuracy level of 87 percent, while the Decision tree model correctly forecasted heart disease patients with an accuracy level of 91%.

2.5 Comparative Analysis

In this research Princy et al. [16], a cardiovascular dataset is categorized using various state-of-the-art Supervised Machine Learning algorithms that are specifically utilized for illness prediction and are applied to a variety of problems. The findings show that the Decision Tree classification model outperformed other methods such as Naive Bayes, Logistic Regression, Random Forest, SVM, and KNN in terms of predicting cardiovascular illnesses. It was the Decision Tree that provided the best outcome, with an accuracy rate of 73%. This method may be beneficial for physicians in terms of predicting the development of cardiac problems and providing suitable therapy in the future.

An ECG classifier for left ventricular hypertrophy was developed using deep neural networks and ECG data in this research. In order to quickly evaluate the body's health, the classification technique uses data preprocessing to fill in missing values from ECG data received from IoT sensors. The forecast model's accuracy, sensitivity, and specificity were all greater than the precision of two clinical techniques, according to the findings of the experiments [17]. In our paper we had use the UIC dataset to define the best algorithm for diagnostics. Five algorithms were selected for creating five models and finding out the best algorithm according to its performance and other features.

According to the findings of this study, a modified algorithm based on logistic regression and principal component analysis will be developed for more accurately predicting heart disease based on various characteristics such as age, blood pressure, chest pain, serum cholesterol levels, and heart rate. Patients will be classified according to the severity of coronary artery disease. An accuracy of 86% was achieved by using a logistic regression model with all variables, and an accuracy of 68% was achieved by using a logistic regression model with PCA. A precision of 77% was achieved by using a logistic regression model with PCA [18]. In out paper Logistic Regression took the second place on accuracy and Jaccard Score. And Naïve Bayes was selected as the best algorithm.

The fast growth of the elderly population and the difficulties associated with providing health and social care have become a focal focus for business and researchers in the modern era. The purpose of this study is to provide a framework for heart disease prediction utilizing key risk variables and a variety of classifier configurations, including K-nearest neighbors, Naive Bayes, support vector machine, Lasso, and ridge regression methods. Apart from these data classification techniques, researchers performed linear discriminant analysis and principal component analysis. The support vector machine has a 92% accuracy rate, whereas the F1 machine has an accuracy rate of 85%. Precision, accuracy, and sensitivity are used to assess the performance of the planned study activity [19]. In our study the both accuracy and the F1-Score was high for Naïve Bayes (90%). The performance of this algorithm is overall high for Precision and Recall too for Negative and Positive values and the Macro and Weighted Average.

Researchers propose a two-stage decision support system to reduce over-fitting and maximize generalization. It has two stages: one based on statistical mutual information and

the other on neural networks. This method was tested on the Cleveland heart disease database's HF subgroup. Their findings indicate that the proposed decision support system has improved generalization and decreased mean percent error (MPE) to 8.8%, considerably less than previous research. The model's 93.33 % accuracy rate outperforms twenty-eight previously established HF risk prediction models with 57.85 to 92.31 %. Our decision assistance technology may aid cardiologists if used in a clinical setting [20]. In our paper the error rate was carefully measured the error rate for Naïve Bayes was also less than other algorithms too. The Misclassification, Mean Absolute Error and Mean Squared Error was also less (9.68%).

2.6 Challenges

Cardiac disease is a broad term that refers to a wide range of various cardiac conditions and illnesses of the heart. It is the most common kind of heart disease in the world, with coronary artery disease (CAD) being the most common. It is characterized by decreased blood flow to the heart muscle. As a consequence of reduced blood flow, a heart attack may develop. One of the most difficult aspects of this study has been identifying the proper factors that must be taken into account when evaluating whether or not someone has heart disease. The task of removing all of the null values from the dataset that has been compiled is a difficulty. This may be a time-consuming procedure. It was also difficult to come up with an algorithm that was appropriate for the circumstance. The dataset was trained using a variety of algorithms, and the scientists then selected the algorithms that performed the best when it came to detecting heart disease from the training data they had collected.

CHAPTER 3

RESEARCH METHODOLOGY & SYSTEM ARCHITECTURE

3.1 Algorithmic Details

A total of five of the finest algorithms from the Supervised Machine Learning algorithms are used in order to accomplish this study effort. According to the most basic definition, an algorithm is an ordered collection of instructions that guides computer software on how to convert an input set of data into useable data. The truth is that statistics are facts, and any information that is helpful to people, robots, or algorithms is valuable information. Machine learning algorithms operate in a similar manner, with some elements of mathematics thrown in for good measure. For all Machine Learning Algorithms, mathematical transformations are not created in exactly the same way. However, this research study covers the most significant machine learning algorithms, each of which includes key algorithmic processes across the whole system architecture.

3.1.1 Logistic Regression

In statistics, a classification method known as Logistic Regression may predict a binary result of either 0 or 1. Due to the fact that this system predicts a binary form of Heart Disease detection as positive or negative for heart disease, this algorithm can be used to comprehend and anticipate the result with relative simplicity, and it is simple to apply. By integrating a range of characteristics rather than a single feature, these models may be able to handle issues that are much more difficult to solve. As the Y-axis moves from 0 to 1, the value of the variable increases. In fact, these two numbers serve as the maximum and minimum values in the sigmoid function, which is perfect for our purposes of classifying data into two categories. By calculating the sigmoid function of X, this system obtains a probability (between 0 and 1 obviously) of an observation belonging to one of the two categories.

3.1.2 K-Nearest Neighbors

Furthermore, by using a simple supervised machine learning technique, it is feasible to apply the k-nearest neighbors' method (KNN) to both classification and regression

issues. KNN is an easy concept to grasp and put into practice. The fundamental theorem of KNN is the Euclidean distance. Because the dataset is divided into two categories, KNN is used in this classification.

3.1.3 Gaussian Naïve Bayes

Gaussian Naive Bayes is a version of the Naive Bayes algorithm that allows for a Gaussian normal distribution and continuous data. Naive Bayes is a set of supervised algorithms based on the Bayes theorem for classification machines, and it is one of the most widely used algorithms today. The classification method is simple, yet it is very effective. If you are working with continuous data, it is common to assume that the values associated with each class are distributed according to a normal (also known as Gaussian) distribution.

3.1.4 Decision Tree

Decision trees may be used to solve classification and regression issues in supervised learning situations. But they are most often used in the solution of classification issues. Internal nodes contain dataset characteristics, branches represent decision rules, and each leaf node provides the conclusion across this tree-structured classifier. It is necessary to utilize the Decision Tree to construct a training model that uses basic decision rules in order to predict the class or value of the input variables based on the training dataset in order to do this.

3.1.5 Adaptive Boosting (AdaBoost)

A method to group learning known as adaptive classification boosting (also known as AdaBoost classification) is used to increase the classification accuracy of a classifier. To create a powerful classifier, it starts with a collection of weak classifiers. When using this technique, a bad classifier learns from the errors of its preceding classifiers. For example, consider the dataset with n samples. A one-to-one weighting scheme is used for each sample at the start. And with this dataset, a mediocre classifier is constructed. The total error of this classifier is calculated. This overall error in the categorization of the data samples serves as a measure of the effect of such a classificatory criterion.

3.2 Proposed System

It is shown in the suggested system the steps that must be taken in order to extract an acceptable algorithm for predicting Heart Disease in patients. Following the collection of the dataset, the data was preprocessed before being used to build and observe several Machine Learning algorithms. Following that, additional essential components were carried out in order to determine the best algorithms, such as Accuracy, Misclassification, Jaccard Score, Cross validation, and Confusion matrix. It was decided on the best algorithm once all of the required components had been completed and explained, and then all of the algorithms were rated.

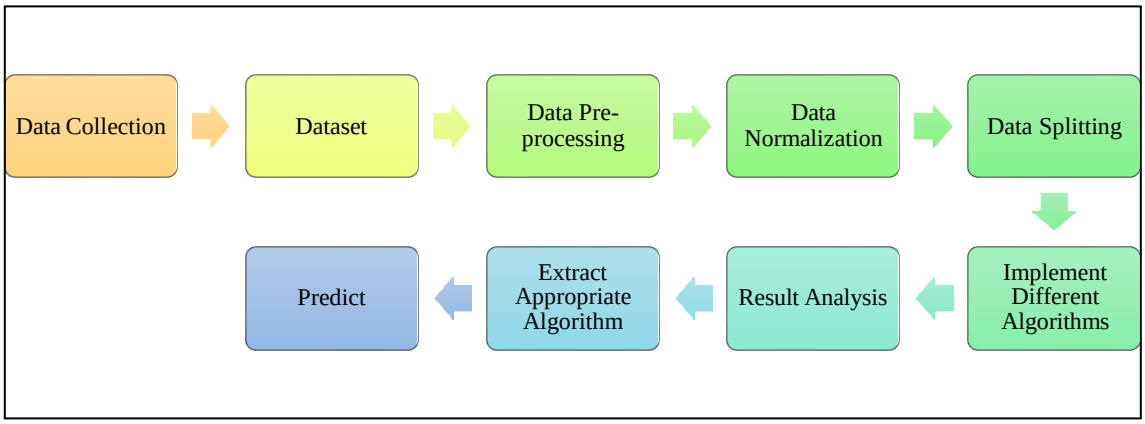


FIGURE 3. 1: PROPOSED METHOD TO PREDICT HEART DISEASE

3.2.1 Data Collection

In order to evaluate heart disease, the system required real-world data that had to be collected. The data was obtained from the UCI, in order to train the model. A total of 303 samples of data were gathered. Each sample includes 14 predictors of future behavior. There are 303 entries with a total of 76 characteristics, however only the 14 most important features were included in the analysis.

3.2.2 Dataset

Combining the data set into a single CSV file reduced the number of rows to a level that was suitable for applying various Machine Learning Algorithms. In order to anticipate anything, machine learning algorithms need a large amount of data. The raw dataset includes 303 rows and 14 columns, with some missing values in some of the rows and columns. There are some missing values in the data set, as well as some random noise.

3.2.3 Data Pre-processing

When developing a prediction model, the first step is to prepare the data for analysis. When data is transformed into an understandable format, the model's efficiency is increased as a result of this assistance. Medical data are often inaccurate, lacking in attribute values, and noisy as a result of the inclusion of outlives or unnecessary information.

When processing this dataset, the initial step was to look for any missing values. The missing data were then replaced with mean values, and so on. This is how the dataset was handled prior to the use of any algorithmic techniques.

3.2.4 Data Normalization

The normalization of data is a critical component of any large dataset. When it comes to preparing data for machine learning applications, normalization is a technique that is often employed as part of the data preparation process. Normalizing numeric columns in a dataset result in a consistent scale for all of the numeric columns in the dataset, which does not distort variances in value ranges. Prior to doing any statistical analysis, normalization should be carried out. In the context of machine learning, there is no need for normalization of any dataset. It is only required when the features have values that fall inside a certain range. The data in this example was standardized before it was pleated.

3.2.5 Splitting Dataset

To use any machine learning technique, the dataset must be split into two parts: one for training the model and another for testing it. This is called data partitioning. This must be completed prior to the use of any machine learning techniques. 90% of the data in this dataset was utilized to train the models, while the remaining 10% of the data was used to test the models, according to the authors.

3.2.6 Implement Different Algorithms

In order to write this article, five machine learning methods were chosen. The algorithms' names are as follows: Logistic Regression, K-Nearest Neighbours, Gaussian

Naive Bayes, Decision Tree, and AdaBoost. Logistic Regression is the most often used method. Using this information, six different models were developed and tested. The algorithms were meticulously executed, and the output results were meticulously examined and evaluated.

3.2.7 Result Analysis

To extract a best algorithm the results the outcomes need to be extracted. It was necessary to take thorough measurements of the Confusion matrix, accuracy, Precision, Recall, and F1-Score. This are a part of result that can give us an over view of the performance of the algorithms. By monitoring the result one can select a best algorithm for prediction.

3.2.8 Extract Appropriate Algorithm

To extract a best algorithm all of the algorithms need to be executed with different performance values. The algorithms need to be analyzed with the dataset. Finding a suitable algorithm is really tough matter as different features need to be performed. The Confusion Matrix plays a very important role in this analysis. As it can give a clear view what is going on.

3.2.9 Predict

By extracting we can come to a solution that it is the best algorithm among the selected one. The algorithm then can be used as a model for the database and for farther use. By inserting values in the model, the model has the ability go an appropriate solution.

CHAPTER 4

EXPERIMENTAL RESULTS & DISCUSSION

4.1 Experimental Results

Following the successful deployment of Machine Learning Model Creation, each algorithm demonstrated its own accuracy and scores, which were used to determine which algorithm was the most accurate in predicting heart disease. Consequently, the outcomes of experimentation are an analytical segment in which each and every potential score for each and every algorithmic application and process may be investigated.

4.1.1 Data Acquisition

The data was collected from the University of California, Irvine, in order to train the model. A total of 303 samples of data were collected for this study. Each sample contains 14 predictors of future behavior that may be used to predict future behavior. However, only the 14 most significant qualities were included in the study since there are 303 entries with a total of 76 characteristics in total. The following are the features of each of the 14 qualities are given in the TABLE 4.1.

TABLE 4. 1: DATA ACQUISITION & NULL VALUE PERCENTAGE

Attributes	Description
Age	Age in years [29-77]
Sex	Male =1, Female=0
Fbs	Fasting blood sugar >120 mg/dl, Value 1 = yes, Value 0 = no
Cp	Chest pain types Value 1 = typical angina, Value 2 = atypical angina Value 3 = non-angina, Value 4 = asymptomatic
Trestbps	Resting blood pressure in mm Hg [94-200]
Chol	Serum cholesterol in mg/dl [126--564]
Restecg	Resting electrocardiographic, Value 0=normal, Value1= having ST-T wave abnormality Value2 = left ventricular hypertrophy by Estes' criteria
Thalach	Maximum heart rate achieved, [71-202]
Exang	Exercise induced angina value 0 = no, 1 = yes
Oldpeak	Measure of ST depression induced by exercise relative to rest [0--6.2]

Slope	Measure of slope for peak exercise ST segment, Value 1= up sloping, Value 2= flat, Value 3= down sloping
Ca	Number of major vessels colored by fluoroscopy [03]
Thal	Thallium stress test, Value 3= normal, Value 6= fixed defect, Value 7=reversible defect
Num	Value 1 = presence of HD Value 0= absence of HD

4.1.2 Data Utilization

The first step in processing this information was to search for any missing values. Following that, missing data were replaced with mean values, and so on. This is the manner in which the dataset was treated prior to the use of any computational methods. After that, the data description for each of the 14 characteristics was retrieved in order to get a better knowledge of the dataset. The Max, 75%, 50%, 25%, Min Std, Mean Count was extracted from the dataset. The whole utilization of the dataset is given in the Table II.

TABLE 4. 2: DATASET DESCRIPTION

	count	**mean**	**std**	**min**	**0.25**	**0.50**	**0.75**	**max**
age	303	54.37	9.08	29	47.5	55	61	77
sex	303	0.68	0.47	0	0	1	1	1
cp	303	0.97	1.03	0	0	1	2	3
trestbps	303	131.62	17.54	94	120	130	140	200
chol	303	246.26	51.83	126	211	240	274.5	564
fbs	303	0.15	0.36	0	0	0	0	1
restecg	303	0.53	0.53	0	0	1	1	2
thalach	303	149.65	22.91	71	133.5	153	166	202
exang	303	0.33	0.47	0	0	0	1	1
oldpeak	303	1.04	1.16	0	0	0.8	1.6	6.2
slope	303	1.40	0.62	0	1	1	2	2
ca	303	0.73	1.02	0	0	0	1	4
thal	303	2.31	0.61	0	2	2	3	3
target	303	0.54	0.50	0	0	1	1	1

4.2 Result & Discussion

The value of Heart Disease was made positive whereas the value of heart disease was detected positive and for negative result negative is returned. When demonstrating particular findings and evaluating the efficiency of machine learning algorithms, the confusion matrix has been used. The template for the confusion matrix for different algorithm types is shown in Table IV (below).

4.2.1 Confusion Matrix

A confusion matrix must be generated in order to verify the results obtained from the implementation aspect. A Confusion matrix is a N x N matrix that is used to evaluate the performance of a classification model when there are N target classes. The accuracy of the machine learning model is evaluated by comparing the actual target values to the predicted values in the matrix. It becomes clear how well the algorithmic model is doing and what errors it is making as a result of this process. With the help of certain mathematical equations, it will be easy to determine the Precision, Recall, and Accuracy of a system based on binary classification. Furthermore, for multi-class categorization, it is necessary to go through these average values using either a micro average or a macro average, depending on the situation. Before going into detail about these, it is necessary to understand four fundamental building elements that are utilized in the computation of different assessment measures. True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) are the four types of false positives and false negatives (FN).

TABLE I. CONFUSION MATRIX

		Positive	**Negative**
Actual	Positive	TP	FP
	Negative	FN	TN
		Predicted	

TABLE 4. 3: CONFUSION MATRIX RESUTL

Algorithm	Confusion Matrix		
		Positive	**Negative**
Logistic Regression	**Positive**	14	1
	Negative	3	13
K - Nearest Neighbours	**Positive**	14	1
	Negative	3	13
Gaussian Naive Bayes	**Positive**	14	1
	Negative	3	13
Decision Tree	**Positive**	14	1
	Negative	3	13
AdaBoost	**Positive**	14	1
	Negative	4	12

A. True Positive (TP)

Positive tuples are those that the classifier correctly categorized. It is denoted by the acronym's letter TP.

B. True Negative (TN)

Negative tuples are positive tuples that were erroneously classified by the classifier. The letter TN may be used to indicate these occurrences.

C. False Positive (FP)

The erroneous classification of these negative labeled tuples as positive has piqued our attention today. This kind of connection may be denoted by the usage of FP.

D. False Negative (FN)

These positive tuples were misclassified as negative by the classifier. It is designated by the abbreviation FN.

E. Precision

Precision may be seen as a measure for assessing the degree to which something is correct (i.e., what percentage of tuples labeled as positive are actually such). In other words, it refers to the proportion of recovered occurrences that are really significant. The mathematical formula for calculating precision is shown in Equation (i).

$$Precision = \frac{TP}{TP + FP} \dots\dots\dots\dots\dots\dots (i)$$

F. Recall

A measure of completeness (how many positive tuples are recognized as such) is used in machine learning to assess the quality of a dataset. A relevant instance is defined as the proportion of relevant examples that have been retrieved as a percentage of the total number of relevant instances that were found. Calculated in Equation (ii), the mathematical formula to measure Recall.

$$Recall = Sensitivity = \frac{TP}{TP + FN} \dots\dots\dots\dots\dots\dots (ii)$$

G. F1-Measure

The weighted harmonic mean is a method of evaluating a test's accuracy and recall, and this measurement is referred to as the F measure (for precision and recall). The mathematical formula for measuring is given in Equation (iii). F1-Measure.

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \dots\dots\dots\dots\dots\dots (iii)$$

H. Accuracy

An accurate classifier on a given test set is defined as the percentage of test set tuples that are correctly classified by the classifier on that test set in a given test set. It is easier to understand how to evaluate accuracy from equation (vi).

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \dots\dots\dots\dots\dots\dots (vi)$$

4.2.2 Classification Report

A classification report is a statistic that is used in machine learning to assess the overall performance of the system. In order to show the accuracy, recall, F1 Score, and support of a trained classification model, it is utilized in conjunction with a training dataset. Performance statistics for a classification-based machine learning model are represented by this metric. Accuracy, recall, F1 score, and model support are all shown in this table. It provides a more accurate view of the overall performance of the trained model. In order to comprehend classification reports generated by machine learning models, it is necessary to be familiar with all of the metrics presented in the research study.

TABLE 4. 4: CLASSIFICATION REPORT

Algorithm	Class	Precision	Recall	F1-Score	Accuracy (%)
Logistic Regression	Negative	0.93	0.81	0.87	87.10
	Positive	0.82	0.93	0.87	
	Macro Avg.	0.88	0.87	0.87	
	Weighted Avg.	0.88	0.87	0.87	
K - Nearest Neighbours	Negative	0.93	0.81	0.87	87.10
	Positive	0.82	0.93	0.87	
	Macro Avg.	0.88	0.88	0.87	
	Weighted Avg.	0.88	0.87	0.87	
Gaussian Naive Bayes	Negative	1.00	0.81	0.90	90.32
	Positive	0.83	1.00	0.91	
	Macro Avg.	0.92	0.91	0.90	
	Weighted Avg.	0.92	0.90	0.90	
Decision Tree	Negative	0.86	0.75	0.80	80.65
	Positive	0.76	0.87	0.81	
	Macro Avg.	0.81	0.81	0.81	
	Weighted Avg.	0.81	0.81	0.81	
AdaBoost	Negative	0.92	0.75	0.83	83.87
	Positive	0.78	0.93	0.85	
	Macro Avg.	0.85	0.84	0.84	
	Weighted Avg.	0.85	0.84	0.84	

4.3 Result Analysis

Here the five algorithms were used to analyze the data set. Based on this five algorithms accuracy and other theoretical approach one suitable algorithm was selected and used to predict the Hearth Disease.

4.3.1 Accuracy

Accuracy is a measure of an algorithm's optimum performance, or how well it works when given a set of instructions. A probabilistic method may be used to measure performance, and accuracy can be utilized to do so. Among the algorithms tested, Naïve Bayes obtained an accuracy of 90.32%, followed by Logistic Regression with an accuracy of 87.10 percent. The Decision Tree method has the lowest accuracy of all of the algorithms tested, with an accuracy of just 80.65 %, which is still very high.

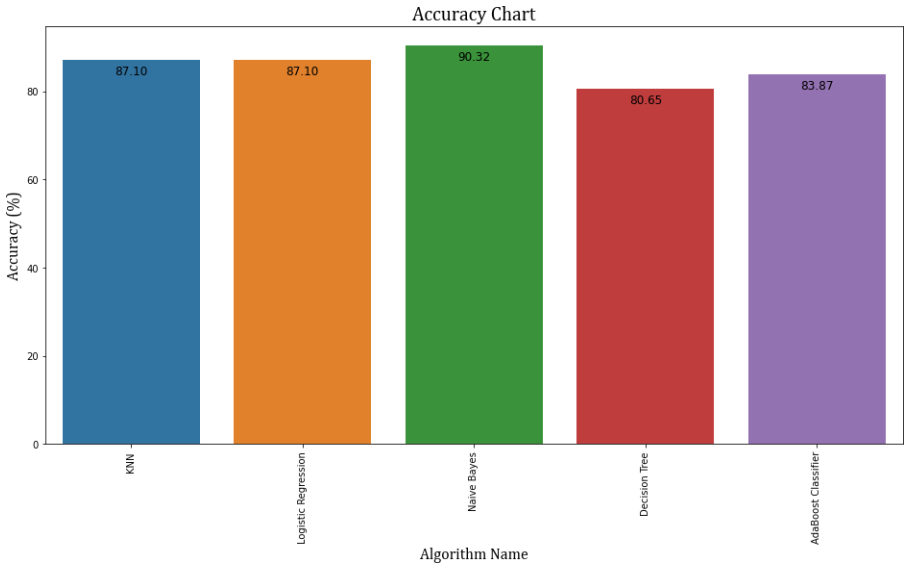


FIGURE 4. 1: ACCURACY CHART

4.3.2 Jaccard Score

The Jaccard score is a number that is used to compare and contrast the similarity and variety of samples. They are equivalent in terms of the intersection to the union ratio. It is possible to compare two finite sample sets of similar size using the Jaccard coefficient, which is a statistical statistic. It is determined by dividing the cross-sectional area of the sample sets by the total area of the sample sets when they are joined together.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

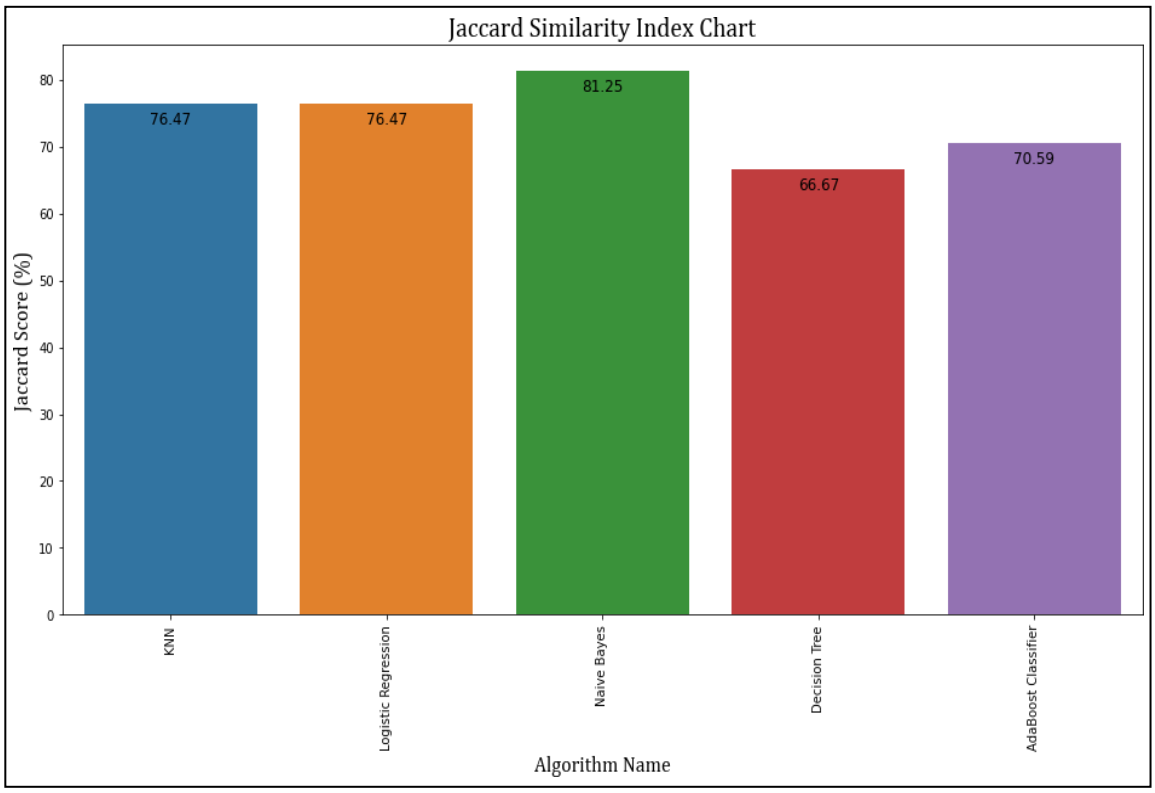


FIGURE 4. 2: JACCARD SCORE CHART

Here again Naïve Bayes got the height value with 80.25% followed by KNN and Logistic Regression with 76.47% accuracy. Here the lowest accuracy was obtained by Decision tree with only 66.67% accuracy.

4.3.3 Cross Validated Score

Cross-validation is a statistical method that is used to evaluate the competency of machine learning models. Cross Validation begins with the data being shuffled and split into k folds, after which it is repeated a number of times more. As a consequence, k models are fitted to (k-1)/k of the data, and as a result of the fitting, 1/k of the data is evaluated to determine its significance. So that the final score can be used in the actual implementation, the findings from each assessment are averaged, and the resultant model is fitted to the whole dataset before it can be used.

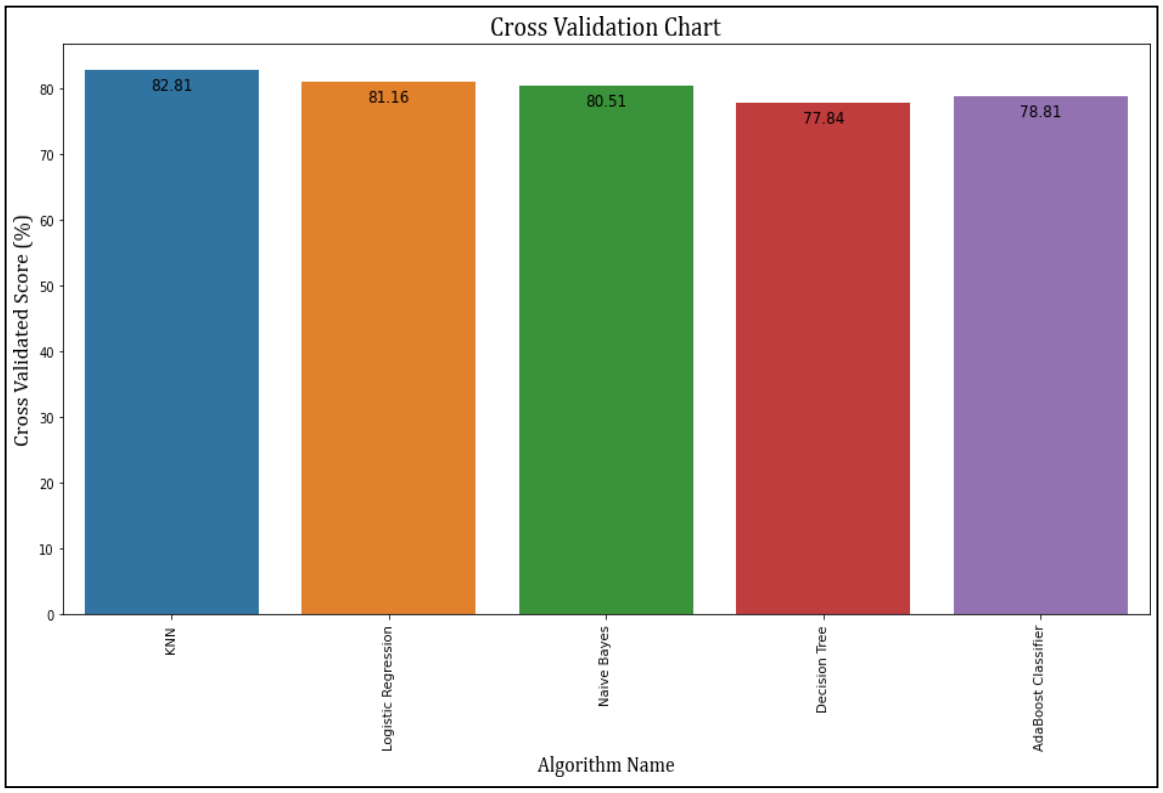


FIGURE 4. 3: CROSS VALIDATED SCORE

Among the models tested for Cross Validated Score, KNN had the best score of 82.81%, followed closely by Logistic Regression (81.16%). And Naive Bayes won third place with an 80.51% score, beating out the competition. Once again, decision tree comes in bottom place, this time with an accuracy of just 77.84%.

TABLE 4. 5: ACCURACY, JACCARD, CROSS VALIDATED AND AUC SCORE

Algorithm Name	Accuracy Score (%)	Jaccard Score (%)	Cross Validated Score (%)
Naive Bayes	90.32	81.25	80.51
KNN	87.1	76.47	82.81
Logistic Regression	87.1	76.47	81.16
AdaBoost Classifier	83.87	70.59	78.81
Decision Tree	80.65	66.67	77.84

4.3.4 Misclassification & Error

It may be difficult to determine whether or not a certain algorithm is accurate. Misclassification often leads in absolute error and mean square error, both of which are components of an accuracy score for a machine learning model. When an incorrect attribute is chosen, it is possible that misclassification may result. When the error rate for all classes,

groups, or categories of a variable is the same as the error rate for the variable itself, misclassification occurs.

The absolute error is the degree of inaccuracy in a measurement. The Mean Absolute Error (MAE) is a statistical term that refers to the average of all absolute mistakes in a measurement.

$$\text{MEA} = \frac{1}{n}\sum_{i=1}^{n}|x_i - x|$$

The mean squared error indicates how near a regression line is to a collection of points (MSE).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - y|^2$$

The Error rate for the Random Forest was also less for Misclassification with 6.45%, Mean Absolute Error with 6.45% and Mean Squared Error 6.45%. Which is less than all other algorithms.

TABLE 4. 6: MISS CLASSIFICATION & ERROR

Algorithm Name	Misclassification (%)	Mean Absolute Error (%)	Mean Squared Error (%)
Naive Bayes	9.68	9.68	9.68
KNN	12.9	12.9	12.9
Logistic Regression	12.9	12.9	12.9
AdaBoost Classifier	16.13	16.13	16.13
Decision Tree	19.35	19.35	19.35

CHAPTER 5

FUTURE SCOPE & CONCLUSION

5.1 Future Scope

Using additional data mining methods such as time series analysis, clustering and association rules, support vector machines, and evolutionary algorithms, this research can be developed in the future. In light of the study's limitations, it is necessary to develop more sophisticated and combination models in order to achieve better accuracy in the early detection of heart disease. Future developments in IoT will allow for the development of a web-based platform for sharing best model data, which will allow the model to learn from fresh data and train itself for greater accuracy, since entering new data into the database will allow the machine to learn more quickly.

5.2 Conclusion

In this study, the goal is to forecast whether or not a patient will develop cardiac disease in the future. The main goal is to identify different data mining methods that may be used in the successful prediction of cardiac disease. In this study, Naive Bayes, KNN Logistic Regression, AdaBoost Classifier, and Decision Tree were used to classify data in the UCI repository using supervised machine learning methods, which were performed on the UCI repository.

The findings of this study may aid physicians in the development of a more accurate and efficient prediction system for heart disease. Validation of the model is accomplished via the use of cross-validation and train-test split data. Through the use of Naïve Bayes, the best accuracy is achieved in this case. With an accuracy percentage of 90.32 percent. For the accuracy to be justified, other metrics such as the Confusion matrix, Precision, Recall (F1-score) were utilized to evaluate the data.

Because of the significance of the results of this research, it is anticipated that improved diagnoses and interpretations will lead to better treatment options based on our machine learning-based prediction model.

References

[1] Motarwar, P., Duraphe, A., Suganya, G., & Premalatha, M. (2020, February). Cognitive Approach for Heart Disease Prediction using Machine Learning. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-5). IEEE.

[2] Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. J. M., Ignatious, E., ... & De Boer, F. (2021). Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques. IEEE Access, 9, 19304-19326.

[3] Saw, M., Saxena, T., Kaithwas, S., Yadav, R., & Lal, N. (2020, January). Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning. In 2020 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-6). IEEE.

[4] Tougui, I., Jilbab, A., & El Mhamdi, J. (2020). Heart disease classification using data mining tools and machine learning techniques. Health and Technology, 10, 1137-1144.

[5] Ahmed, H., Younis, E. M., Hendawi, A., & Ali, A. A. (2020). Heart disease identification from patients' social posts, machine learning solution on Spark. Future Generation Computer Systems, 111, 714-722.

[6] Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart disease identification method using machine learning classification in e-healthcare. IEEE Access, 8, 107562-107582.

[7] Singh, A., & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. In 2020 international conference on electrical and electronics engineering (ICE3) (pp. 452-457). IEEE.

[8] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. SN Computer Science, 1(6), 1-6.

[9] Samhitha, B. K., Priya, M. S., Sanjana, C., Mana, S. C., & Jose, J. (2020, July). Improving the Accuracy in Prediction of Heart Disease using Machine Learning Algorithms. In 2020 International Conference on Communication and Signal Processing (ICCSP) (pp. 1326-1330). IEEE.

[10] Princy, R. J. P., Parthasarathy, S., Jose, P. S. H., Lakshminarayanan, A. R., & Jeganathan, S. (2020, May). Prediction of Cardiac Disease using Supervised Machine Learning Algorithms. In 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 570-575). IEEE.

[11] Ansari, M. F., AlankarKaur, B., & Kaur, H. (2020, May). A prediction of heart disease using machine learning algorithms. In International conference on image processing and capsule networks (pp. 497-504). Springer, Cham.

[12] Ali, L., & Bukhari, S. A. C. (2020). An approach based on mutually informed neural networks to optimize the generalization capabilities of decision support systems developed for heart failure prediction. Irbm.

[13] Yadav, S. S., Jadhav, S. M., Nagrale, S., & Patil, N. (2020, March). Application of machine learning for the detection of heart disease. In 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (pp. 165-172). IEEE.

[14] Samhitha, B. K., Priya, M. S., Sanjana, C., Mana, S. C., & Jose, J. (2020, July). Improving the Accuracy in Prediction of Heart Disease using Machine Learning Algorithms. In 2020 International Conference on Communication and Signal Processing (ICCSP) (pp. 1326-1330). IEEE.

[15] Princy, R. J. P., Parthasarathy, S., Jose, P. S. H., Lakshminarayanan, A. R., & Jeganathan, S. (2020, May). Prediction of Cardiac Disease using Supervised Machine Learning Algorithms. In 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 570-575). IEEE.

[16] El Hamdaoui, H., Boujraf, S., Chaoui, N. E. H., & Maaroufi, M. (2020, September). A Clinical support system for Prediction of Heart Disease using Machine Learning Techniques. In 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP) (pp. 1-5). IEEE.

[17] Wu, J. M. T., Tsai, M. H., Xiao, S. H., & Liaw, Y. P. (2020). A deep neural network electrocardiogram analysis framework for left ventricular hypertrophy prediction. Journal of Ambient Intelligence and Humanized Computing, 1-17.

[18] Ansari, M. F., AlankarKaur, B., & Kaur, H. (2020, May). A prediction of heart disease using machine learning algorithms. In International conference on image processing and capsule networks (pp. 497-504). Springer, Cham.

[19] Patro, S. P., Padhy, N., & Chiranjevi, D. (2021). Ambient assisted living predictive model for cardiovascular disease prediction using supervised learning. Evolutionary Intelligence, 14(2), 941-969.

[20] Ali, L., & Bukhari, S. A. C. (2020). An approach based on mutually informed neural networks to optimize the generalization capabilities of decision support systems developed for heart failure prediction. Irbm.