

# **PREDICTION OF HEART DISEASE USING MACHINE LEARNING ALGORITHMS**

**BY**

**MD. Hasnat Hasnine Shovon**

**ID: 182-15-11744**

**AND**

**Tahmina Akter**

**ID: 182-15-11775**

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

Supervised By

**Dr. Md Zahid Hasan**

Associate Professor & Coordinator MIS

Department of CSE

Daffodil International University

Co-Supervised By

**Israt Jahan**

Lecturer

Department of CSE

Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

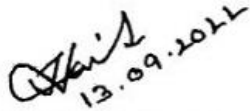
**DHAKA, BANGLADESH**

**SEPTEMBER 2022**

## APPROVAL

This Project/internship titled “**Prediction Of Heart Disease Using Machine Learning Algorithms**”, submitted by Md. Hasnat Hasnine Shovon, ID 182-15-11744 and Tahmina Akter, ID 182-15-11775 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 12th September 2022.

### BOARD OF EXAMINERS

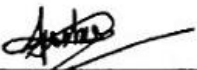
  
13.09.2022

**Chairman**

---

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


**Internal Examiner**

---

**Raja Tariqul Hasan Tusher**  
**Assistant Professor**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology

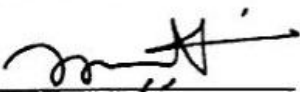
**Internal Examiner**

---

**Md. Sabab Zulfiker**  
**Senior Lectuer**  
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 **Dr. Md Zahid Hasan, Associate Professor & Coordinator MIS, 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:**



---

**Dr. Md Zahid Hasan**  
Associate Professor & Coordinator MIS  
Department of CSE  
Daffodil International University  
**Co-Supervised by:**



---

**Israt Jahan**  
Lecturer  
Department of CSE  
Daffodil International University  
**Submitted by:**



---

**Md. Hasnat Hasnine Shovon**  
ID: 182-15-11744  
Department of CSE  
Daffodil International University

*Tahmina Akter*

---

**Tahmina Akter**  
ID: 182-15-11775  
Department of CSE  
Daffodil International University

## ACKNOWLEDGEMENT

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

We are really grateful and wish our profound indebtedness to **Supervisor Md. Zahid Hasan, Associate Professor, and Co-Supervisor Israt Jahan, lecturer**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “Heart Disease Prediction” to carry out this project. His and her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them have made it possible to complete this project.

We would like to express our heartiest gratitude to **Prof. Dr. Touhid Bhuiyan** and the Head, of the Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of the CSE department of Daffodil International University.

We would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the course work.

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

## **ABSTRACT**

Heart disease similarly known as cardiovascular disease which indicates any uncomfortable disorder affecting the heart. This disease has been the far most common reason of death, premature aged heart failure and influences all types of humans. Our heart's oxygen level drops as a result of blood vessel breakage, and heart attacks follow the cessation of blood pumping. Most common reason of heart attacks or any other heart diseases commonly indicates high blood pressure, high cholesterol, smoking, mental stress and so on. According to reports, nowadays babies also have a variety of heart conditions. A prominent characteristic of acquired heart disease is a coronary artery disease (CAD). Some coronary arteries carry nutrients, oxygen and most importantly- blood to our heart. If those arteries get damaged somehow, a blockage is created in the vessels and blood supply through veins towards the heart gets affected dangerously. The two most frequent reasons are defective cardiac valves and weak or damaged heart tissue are responsible of these premature deaths or early heart diseases. This entire situation can make the heart to beat either too fast or too slow which is not normal pumping at all. Contrastingly, if the beats are too slow, there could not be enough cardiac reductions to provide the body with the blood it actually needs. Chest ache, palpitations, asphyxia, swelling, cyanosis and even sudden death are most known result. Regular heartbeat is always mandatory for the heart to appropriately pump blood and to lead a healthy life.

## TABLE OF CONTENTS

| <b>CONTENTS</b>                          | <b>PAGE NO.</b> |
|--|-----------------|
| Acknowledgments                          | III             |
| Abstract                                 | IV              |
| Table of Contents                        | V-VI            |
| List of Figures                          | VIII            |
| List of Tables                           | IX              |
| <br>                                     |                 |
| <b>CHAPTER</b>                           |                 |
| <b>CHAPTER 1: INTRODUCTION</b>           | <b>01-04</b>    |
| 1.1 Introduction                         | 1               |
| 1.2 Motivation                           | 1               |
| 1.3 Rationale of the Study               | 2               |
| 1.4 Objectives                           | 3               |
| 1.5 Research Questions                   | 3               |
| 1.6 Expected Outcome                     | 3               |
| <br>                                     |                 |
| <b>CHAPTER 2: BACKGROUND</b>             | <b>04-08</b>    |
| 2.1 Introduction                         | 4               |
| 2.2 Related Works                        | 4               |
| 2.3 Research Summery                     | 6               |
| 2.4 Scope of Problems                    | 7               |
| 2.5 Challenges                           | 7               |
| <br>                                     |                 |
| <b>CHAPTER 3: RESEARCH METHODOLOGY</b>   | <b>09-16</b>    |
| 3.1 Introduction                         | 9               |
| 3.2 Research Subject and Instrumentation | 10              |
| 3.3 Data Collection Procedure            | 10              |
| 3.4 Dataset                              | 11              |

|   |              |
|---|--------------|
| 3.5 Class Labels  | 12           |
| 3.6 Data Preprocessing  | 12           |
| 3.7 Data Storing  | 13           |
| 3.8 Machine Learning Algorithms   | 13           |
| 3.9 Implementation Requirements   | 16           |
| <b>CHAPTER 4: EXPERIMENTAL RESULTS &amp; DISCUSSION</b>                           | <b>18-24</b> |
| 4.1 Experimental Setup  | 17           |
| 4.2 Model Summery   | 18           |
| 4.3 Experimental Result and Analysis  | 18           |
| 4.4 Discussion  | 21           |
| <b>CHAPTER 5: SUMMARY, CONCLUSION,<br/>RECOMMENDATION, IMPLICATION FOR FUTURE</b> | <b>24-25</b> |
| 5.1 Summery of the Study  | 24           |
| 5.2 Conclusion  | 24           |
| 5.3 Recommendation  | 25           |
| 5.4 Implication for Further Study   | 25           |
| <b>REFERENCES</b>   | <b>26-27</b> |
| <b>PLAGIARISM REPORT</b>  | <b>28</b>    |

## LIST OF FIGURES

| <b>FIGURES</b>   | <b>PAGE NO</b> |
|--|----------------|
| Figure 3.1.1: Methodology at a Glance                          | 10             |
| Figure 3.3.2: Data info at a glance                            | 11             |
| Figure 3.6.1: HD distribution                                  | 12             |
| Figure 3.3.3: Null Values Handling                             | 13             |
| Figure 3.8.4: XGB model decision strategy                      | 15             |
| Figure 3.8.5: SVM model decision strategy                      | 16             |
| Figure 4. 2. 1: SVM ROC Curve and Confusion matrix             | 19             |
| Figure 4.2.2: Voting Classifier ROC Curve and Confusion matrix | 19             |
| Figure 4.2.3:LR ROC Curve and Confusion matrix                 | 20             |
| Figure 4.2.4: NB ROC Curve and Confusion matrix                | 20             |
| Figure 4.2.5: RF ROC Curve and Confusion matrix                | 21             |
| Figure 4.2.5: XGB ROC Curve and Confusion matrix               | 21             |



## **LIST OF TABLES**

| <b>TABLES</b>                       | <b>PAGE NO</b> |
|-------------------------------------|----------------|
| Table 3.4.1: Attribute Information  | 11             |
| Table 4.3.1: Performance comparison | 22             |
| Table 4.3.2: Performance evaluation | 23             |

# CHAPTER 1

## INTRODUCTION

### 1.1 Introduction

In this modern world, cardiovascular disease is the leading cause of mortality. It's difficult to identify heart disease due to various contributory risk elements, such as smoking, tobacco, alcohol, high blood cholesterol, diabetes mellitus, overweight, high blood pressure, abnormal pulse rate and many other factors [1]. Several methods have been employed to determine the extent of cardiac disease in individuals. Cardiovascular disease has a very complex character, and therefore, the sickness must be treated with care. Failure to do so could harm the heart and perhaps result in a premature death. The primary obstacle to modern healthcare is the provision of top-quality services as well as effective exact diagnosis. Even while cardiac problems have been identified as the primary cause of mortality in recent years, they can also be efficiently managed and controlled. Various machine learning and neural NN algorithms have been utilized to assess heart disease severity illness in people. Several techniques, including the K-Nearest Neighbor Algorithm, Decision Trees, and Genetic Algorithm, are used to categorize the severity of the condition. Furthermore, Naive Bayes [2, 3]. Heart disease's characteristics is intricate, so the sickness must be treated with caution.

Because of numerous contributing risk factors, including diabetes, high blood pressure, high cholesterol, irregular pulse rate, and many other factors, it is problematic to diagnose cardiovascular diseases [4]. Numerous studies show that, in comparison to other classification models in the research community, evolutionary computing techniques have attracted enormous quantities of thought and can offer extreme accuracy in the classification difficulties. It has generally been recommended to minimize the fatality rate as well as to strengthen the decision-making for subsequent prevention and all types of treatment through the timely identification of heart disease in high-risk patients and the improved diagnosis utilizing a prediction model. The appropriate timeframe of disease

detection influences how successfully a disease will be managed overall. The proposed research seeks to discover these health complications as promptly as possible. Using machine learning approaches, we offer a novel approach in this study that intends to identify important characteristics, hence increasing the forecasting accuracy of cardiovascular disorders.

## **1.2 Motivation**

Human heart is the first and foremost important component of the human being. In essence, it controls blood circulation throughout our rest of the body. Any oddity or disfunction to heart can lead to disfunction in other organs. Heart diseases can be characterized as any impairment to the heart's regular function, no matter how slight.

World Health Organization estimations serve as the basis for this task's motivation. The World Health Organization (WHO) predicts that through 2030, roughly 23.6 million individuals will pass away from heart disease. In order to minimize the danger, expectation of coronary illness needs to finish. Every single year, different heart problems cause very early deaths around the entire planet. In the world today, cardiovascular disease is one of the most frequent reasons for mortality in the majority of instances. Early identification and maintaining a healthy, disciplined lifestyle are two important tactics for preventing heart-related illnesses.

## **1.3 Rationale of the Study**

People are dying more regularly these days as a consequence of numerous cardiac illnesses. The primary factor in the large number of deaths among people is heart attacks. As a result of numerous causes and circumstances that have a significant chance of premature death, many cardiac illnesses arise in young patients. Due to male smoking and drinking habits, the death rate is significantly higher in men than in women. The heart's capacity to beat depends greatly on the health of the human body. Because the heart circulates blood to the body's many organs. These conditions include excessive blood pressure, heart attacks,

angina, heart failure, and heart valve disease, among others. It's crucial to diagnose heart related problems at an early stage and commence the suitable treatment immediately.

Classification is one of the most crucial, important, and popular decision-making tools in medical science. Numerous contemporary technologies have previously been established in order to precisely and consistently predict coronary heart disease. Supervised learning eliminates noisy input and learns improved features for classification problems. The performance of classification is much enhanced. However, this particular method performs better with huge datasets and image datasets. The Cleveland and Statlog datasets are not particularly large, and all of the attributes are entirely numerical, therefore the reinforcement technique is not well suited to this task. Therefore, in this case, we'll employ the computational intelligence technique, which actually consists of a number of computational models. It makes it simple to solve many types of issues that are extremely challenging to answer with traditional algorithms.

#### **1.4 Objectives**

- Demonstrating the accuracy of different algorithms and detection of most effective algorithm with the highest accuracy
- Increasing the accuracy of applicable algorithms than previous researches
- using Cleveland dataset. Then previous researches.
- Picking up the most standard algorithm among Logistic Regression, Random Forest, Naive Bayes, XGB classifier and Voting classifier algorithms that can be used in detecting heart diseases at a very preliminary stage to take proper treatment.

#### **1.5 Research Questions**

1. Does it predict an actual output by given sample data with the system?
2. What is the purpose of the thesis?
3. How all the data were collected?

4. Does every algorithm work perfectly (yes/no)?
5. How was the accuracy?

### **1.6 Expected Outcome**

The expectation towards the study is to determine the most efficient Machine Learning (ML) strategy for cardiovascular disease diagnosis. This specific paper compares the accuracy score of different computational intelligence technique such as Logistic Regression, Random Forest, Naive Bayes, XGB classifier and Voting classifier algorithms for predicting heart disease using proper dataset. There are already so many proposed systems to predict the heart diseases, yet are doing this because some of the proposed systems' accuracy is not quite well and also some of them do not handle missing values properly. Again, almost all of the research papers have worked on Statlog dataset. Our target is to work on Cleveland dataset and apply these algorithms to get the maximum accuracy which can be updated.

## **CHAPTER 2**

### **BACKGROUND**

#### **2.1 Introduction**

In recent times, many associations involved in human research applied these algorithms in a big scale. In information mining, we can change these measures of data into useful data for strong and accurate choice making purpose. An alternative cause is that the insurance exchanges made by this part are too much big and complicated, which makes it impenetrable to be broken and also build up by conventional systems. So, it can be growing majorly by using many applications in finding sequences and examples in such big volumes of regular data.

In late patterns, inquiry on the extensive data got fundamental due to monetary weights on medicinal services. This specific separated data can be used for choices taking into account the relapse test of restorative as well as currency related data.

Records of large set of medical data entry which are done by experts in medical field are free for evaluating as well as gathering important knowledge from them. Machine learning techniques are the means of harvesting invisible and cryptic information from humongous scale of available data. Most of the health-related data contains various information. It makes decision making hard and time-consuming task using discrete data. Machine Learning (ML) is basically a part of data mining which efficiently works great on huge scale well-configured dataset. In health service field, machine learning has been used for identification, discovery and prophecy purpose of different illness.

#### **2.2 Related Works**

Fahad Saleh Alotaibi made use of the Heart Disease Dataset, a dataset with 14 variables that was obtained through the Kaggle platform. He tested SVM against Decision Tree,

Logistic Regression, Random Forest, Naive Bayes, and SVM, and found that SVM had the highest accuracy of 84.85% for predicting heart disease [5].

Evanthia E. et al. have provided a comprehensive review and comparison, in terms of advantages and disadvantages and of the methods reported in the literature that address all the types of the HF management involving ML techniques through their research [6].

In the year 2021, Rohit Bharti and others used the Public Health Dataset having four databases, containing 76 attributes [7].

Khaled Mohamad Almustafa conducted a study of various classification algorithms for the coronary disease dataset in order to correctly categorize or forecast HD cases with few attributes. which includes 1025 people's class information taken from Cleveland, Hungary, Switzerland, and Long Beach among its 76 features. Although he only divided the qualities into 14 categories in his study. He improved his accuracy with a decision tree classifier by 93.85% [8].

The author Shu Jiang used downloaded data set “heart disease” from UCI Machine Learning Repository webpage and got Extreme gradient boosting’s accuracy 84% and accuracy of Neural network as 85% [9].

Apurb Rajdhan et al. used the HD Dataset which is a union of 4 different dataset. They recommended Random Forest algorithm for reaching the accuracy score of 90.16% in April,2020 [10].

Purushottama, Prof. (Dr.) Kanak Saxena and Richa Sharma have presented an Efficient Cardiovascular Disease Prediction System using data mining in their paper. They have used 10-fold method and also found the accuracy of 86.3 % in validation data and 87.3 % in training data [11].

Senthilkumar Mohan and others proposed hybrid HRFLM approach which is used in combination of the methods of Random Forest (RF) and Linear Method (LM). HRFLM had 88.4% accuracy [12].

In 2020, Norma Latif Fitriyani et al. proposed an effective heart disease prediction model (HDPM) for HD identification by using DBSCAN, SMOTE-ENN and XGBoost-based Machine Learning Algorithm [13].

Mohammad Ayoub Khan proposed MDCNN model showing 92.2% accuracy [14].

Safial Islam Ayon and others used seven computational intelligence techniques to detect heart disease using the Statlog and Cleveland heart disease dataset [15].

Resul Das et al. Used neural network ensemble on above mentioned dataset and got 89.01% classification accuracy [16].

In their research, Evanthia E. and colleagues have offered a thorough analysis and comparison of all the features of HF treatment using machine learning approaches, as well as the benefits and drawbacks of the various methods that have been previously published [17].

Anjan Nikhil Repaka and his colleagues used Gaussian naïve bayes 91.56% accuracy[18].

### **2.3 Research Summery**

The major reason that we have worked on the topic “Heart Failure Prediction” is to detect heart diseases based on 14 attributes at very early stage to avoid disastrous consequences as well as premature death. We collected recent datasets as much as we could and studied them first. Several studies published accuracy of different algorithms but we only picked some selected latest algorithms like Logistic regression (94.56% accuracy), Naïve Bayes



(94.02% accuracy), Random Forest (94.0% accuracy), XGBoost Classifier (95.65% accuracy), SVM (96.19) and Voting Classifier (95.35% accuracy) .

Our goal was to apply these 6 algorithms on our combined specified datasets. Many researches are done on this topic lately but most of them are based on Statlog dataset. So, our target was to work on Cleveland datasets. We tried to get as much as high accuracy because the higher the accuracy, the better for our study. The specification is also a vital point of the task. Here, specification means detecting which algorithm or model works in more accurate and easier way to predict heart diseases within the shortest time with the highest accuracy.

## **2.4 Scope of the Problems**

Our main goal is to give a system for doctors and medical concerns to detect heart disease at primary stage to prevent heart attack, heart failure and other long-term diseases. This prediction system will deliver necessary and reliable therapy to patients and avoid risky consequences like paralysis or death. ML does a very important role to find the invisible different patterns and thereby analysis the given information. After examining the given information, ML techniques help in heart disease detection as well as early prediction.

A large number of researches mentioned that in the investigation community, computational intelligence approaches have encouraged a large level of thoughtfulness. It also can deliver maximum accuracy in the classification related problems, compared to many data classifier models. This research presents performance evaluation of various ML techniques known as Naive Bayes, Logistic Regression, Random Forest, SVM, XGBoost Classifier and Voting Classifier for predicting heart disease at primary stage. In this paper, we have used six computational intelligence algorithms to detect coronary disease using the Combined heart disease dataset. We have enormous scope to work on other latest algorithms in the future.

## 2.5 Challenges

Collecting data in our country's subcontinent is too challenging. Because accuracy as well as prediction depends on the accurate dataset. For the dataset, there is no individual source, for this reason, we had to struggle a lot. And as our goal was to find out the most eligible algorithms with maximum accuracy, so we needed to collect both online and offline data.

As most of the researches on this topic are based on either Statlog dataset or Cleveland dataset, so there are only a few works on both, so we didn't too much combined dataset online. Here we have used 303 instances from Cleveland dataset source, 294 instances from Hungarian source, 123 instances from Switzerland, 200 from Long Beach and 270 instances from Statlog dataset. For highest accuracy, latest and resourceful dataset is obviously compulsory.

We had faced many problems in our code. In the primary stage, we had few datasets. Many different datasets were being added gradually. Our accuracy in code was incorrect for several weeks due to some wrong values in dataset. By time, we were able to fix our code with the help of more effective and realistic data.

# CHAPTER 3

## RESEARCH METHODOLOGY

### 3.1 Introduction

In the specific part, we will discuss the research methodology and procedures. On the contrary, tools that have been used for the project, data collection, research topic, processing, and pre-processing. analysis statistically, and its implementation will be discussed in this chapter. The full methodology is shown in figure 3.1.1

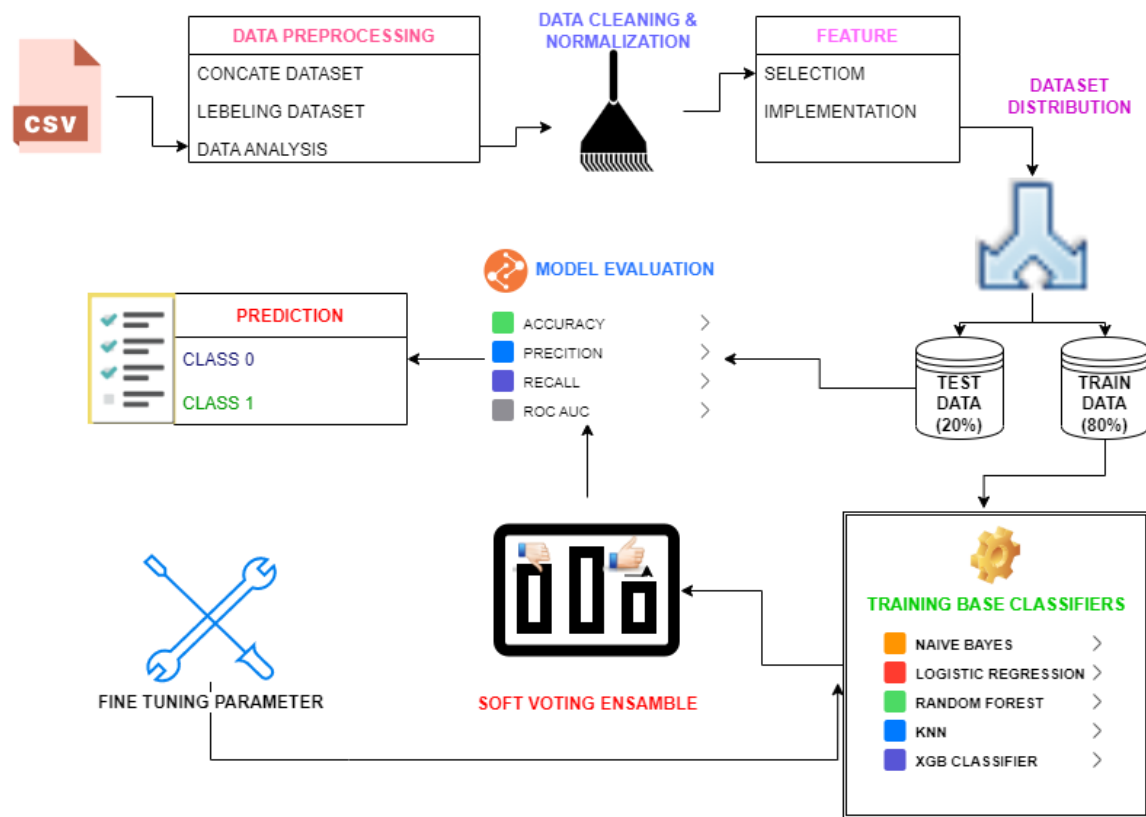


Figure 3.1.1: Methodology at a Glance

## 3.2 Research Subject and Instrumentation

We noticed that choosing the most appropriate algorithm for precise heart disease prediction from so many eligible algorithms is undoubtedly a tough task. We had to be careful about some facts which are-

- Which type of dataset has to be collected
- How many of dataset will be enough to get appropriate result
- Ensuring that the collected offline data are recorded carefully
- Ensuring that the collected information was accurate
- Labeling the dataset in proper way
- Work on necessary graphs to make the concept easily understandable
- Organizing the dataset properly first and then working on code

## 3.3 Data Collection Procedure

We gathered enough data from both online sources. This work is mostly done using a dataset. This dataset was collected from Kaggle [20] which is an online repository for data science and ML datasets. In this data file, 5 data-sets are merged containing 11 attributes.

These five data-sets are:

- |      |             |          |
|------|-------------|----------|
| i.   | Cleveland   | 303 data |
| ii.  | Hungarian   | 294 data |
| iii. | Switzerland | 123 data |
| iv.  | Long Beach  | 200 data |
| v.   | Statlog     | 270 data |

In total there are 1190 patient's observation, of which 270 are duplicates. After data cleaning and null value handling, we get 918 patient data to work with. The dataset comes in (.csv) format.

```

df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Age              918 non-null   int64
1   Sex              918 non-null   object
2   ChestPainType    918 non-null   object
3   RestingBP        918 non-null   int64
4   Cholesterol      918 non-null   int64
5   FastingBS        918 non-null   int64
6   RestingECG       918 non-null   object
7   MaxHR            918 non-null   int64
8   ExerciseAngina   918 non-null   object
9   Oldpeak          918 non-null   float64
10  ST_Slope         918 non-null   object
11  HeartDisease     918 non-null   int64
dtypes: float64(1), int64(6), object(5)
memory usage: 86.2+ KB

```

Figure 3.3: Data info at a Glance

### 3.4 Dataset

This HD dataset used in various was evaluated in the detection of heart disease by various classifier models. A large number of them predicted higher accuracy in last ten years. In dataset there are 918 entries of 12 attributes.

| Attribute number | Description    | Type    |
|------------------|----------------|---------|
| 1                | Age            | Int64   |
| 2                | Sex            | Object  |
| 3                | ChestPainType  | Object  |
| 4                | RestingBP      | Int64   |
| 5                | Cholesterol    | Int64   |
| 6                | FastingBS      | Int64   |
| 7                | RestingECG     | Object  |
| 8                | MaxHR          | Int64   |
| 9                | ExerciseAngina | Object  |
| 10               | Oldpeak        | Float64 |
| 11               | ST_Slope       | Object  |
| 12               | HeartDiseas    | Int64   |

Table 3.4.1: Attributes information

### 3.5 Class Labels

Here we have two(2) classes. The classes are :

- HeartDisease = 1
- HeartDisease = 0

The amount of HD patient are 410 and 508 patient are healthy.

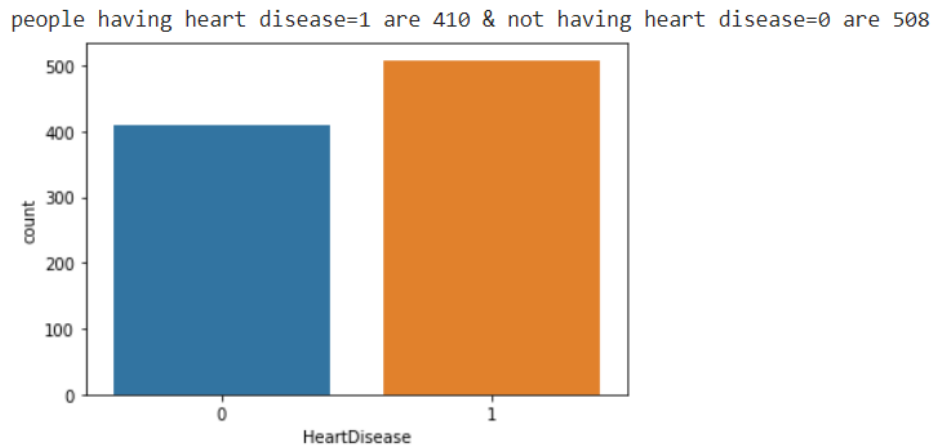


Figure 3.3.2: Heart disease distribution

### 3.6 Data Preprocessing

After importing the data to preprocess we tried to handle the null values. But, we got lucky there as there were no null values to be found.

#### Checking for null values

```
heart.isnull().sum()/len(heart)*100
```

|                |         |
|----------------|---------|
| Age            | 0.0     |
| Sex            | 0.0     |
| ChestPainType  | 0.0     |
| RestingBP      | 0.0     |
| Cholesterol    | 0.0     |
| FastingBS      | 0.0     |
| RestingECG     | 0.0     |
| MaxHR          | 0.0     |
| ExerciseAngina | 0.0     |
| Oldpeak        | 0.0     |
| ST_Slope       | 0.0     |
| HeartDisease   | 0.0     |
| dtype:         | float64 |

No null values found

Figure 3.3.3: Null values handling

After that Label Encoder library was used to convert object datatype to machine readable (value between 0 and n\_classes-1) format.

### 3.7 Data Storing

After data organization data storing process has been started. In this section, we stored the dataset both in the local PC's directory and google drive. As we used the Jupyter notebook so local PC directory was compulsory. For security, we also used google colab. For this reason, we also uploaded those datasets into google drive. And if we need to use those datasets in the future for further improvement of the project uploading google drive was the safest way. Because there might be any technical fault in the local pc, all the datasets would be lost. For this reason, we uploaded the dataset to google drive also.

### 3.8 Machine Learning Algorithms

#### 3.8.1 Logistic Regression

The classification procedure known as logistic regression is most frequently used for binary classification issues. The LR algorithm uses the logistic function to compress the validation data of a linear equation between 0 and 1 instead of projecting a straight line or hyper plane. Its 13 independent parameters show that logistic regression is effective at making predictions. The algorithm aids in the diagnosis of diseases in numerous ways. It uses the real-valued input vector and a discriminative category technique. Assuming the tendency

of information or removing significant statistical values from the model were the true purposes of this. The reliant variable in the LR is a binary consisting of data indicated as 1 (yes) or 0 (no). The primary goal is to mimic an anomaly's logarithmic

### 3.8.2 Naive Bayes (NB)

Naive Bayes is a probabilistic classifier which relies on the Bayes Theorem. It makes powerful independency between features assumption. Using this theorem, one can decode the conditional property. The following is a theorem (1):

$$P(C_k | X) = \frac{P(C_k)P(X|C_k)}{P(X)} \dots \dots \dots (1)$$

Where  $X = (x_1 \dots, x_n)$  denotes an n-dimensional vector.  $C_k$  symbolizes each class and characteristics. If we assume those features. It is not correct to assume that ages are unrelated to one another.in various difficulties, and it can also have an impact on precision the classifier's The key advantage of Naive Bayes is that it is simple to implement. It is an online algorithm, and it can be trained in a matter of minutes. Time is linear.

### 3.8.3 Random Forest

Random forest is an integrated learning classification approach. It can employ a resampling strategy in the training process, in this every instance returned from the main train data is randomly picked from the existing number of instances, resulting in a whole new train data-set, and several decision trees are constructed separately. In each decision tree, the best test from the fresh train data-set is chosen as the decision point to carry out the split test, and after that the outcome of the single unique decision tree is created; the concluding decision is made by calculating the mode of classification results of all decision trees. It follows (2):

$$Gain(D) = 1 - \sum_{i=1}^k p_i^2 \dots \dots \dots (2)$$

Here,

k is the number of unique labels



$p_i$  is flow of  $p$ .

### 3.8.4 XGB Classifier

XGBoost works evaluating decision trees, that creates a graph that evaluates the input in various “if” conditions . Where “if” function is met requirements affects the following “if” function and eventual classification. XGBoost the Algorithm continuously attaches more and more “if” functions to the decision tree to build a strong model.



Figure 3.8.4: XGB model decision strategy

This model learns a quickly than many machine learning classifications and works well on multi class data and small datasets.

### 3.8.5 Support Vector Machine

Svm operates by fitting info to a hyper-dimensional plane so that some points can be classified even where the data are not otherwise linearly separable a separator between the classes is found then the data are altered in a system that a line could be drawn as a hyperplane.

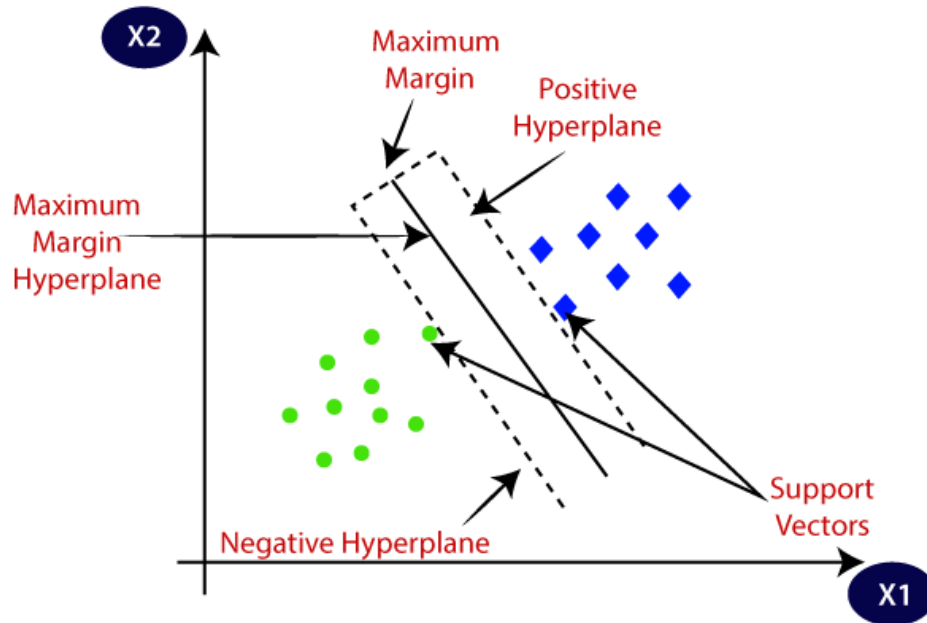


Figure 3.8.5: SVM model decision strategy

### 3.9 Implementation requirement

Hardware/Software requirement:

- Operating system (minimum windows 10)
- Hard disk (256 GB minimum)
- Ram (10 GB minimum)

Required Developing Apparatus:

- Python Installed System
- Anaconda
- Google Collaboratory

## CHAPTER 4

### EXPERIMENTAL RESULTS AND DISCUSSION

#### 4.1 Experimental Setup

For our desired system and code implementation, first task was to gather proper and enormous datasets. Here, we have used Python environment, Anaconda and Google Collab. Python is an open to all, high-level programming language that is not very difficult to use and it is supposed to be the strongest language among all programming languages. We chose it because python commands and its in-built functions are so easy that it reduces the time of implementation. Colab is a hosted Jupiter, installed and configured so that we do not have to do anything on the computer or laptop but only work from the browser to the resources of the cloud. In this Python step, they are based on notebooks which can be text, image, or code. Because only Python kernel can be used as opposed to Jupiter Collab at the moment.

In this approach, various referred classification algorithms were utilized to compare these algorithms performance in term of the prediction of the above referred HD data-set. First of all, data(.csv) file was imported from local machine. A dataframe named “heart” was created to for the machine to read easily. Some of the data in dataset were object type variable. It is essential to use an encoding technique. The data-set attributes required to be transformed earlier before using in the model in our approach. Those data are encoded using “LabelEncoder” but it leads to a lower accuracy. So dummy variables are used instead. After further finding training phase ended, we found out that the model's accuracy had superior values.

Adding dummy-variable the dataset records generally results in much faster training. This behavior is evidenced in our model.

The dataset was split in two parts, with one indicating to information for use in training, which was 80 percent of the total information; the other 20 percent was utilized during the model testing phase. After all, training dataset were trained using previously

mentioned Naïve Bayes, Logistic regression, Random forest and XGBoost classifiers and evaluated by test data-set. Then, a Voting classifier is build using previously mentioned algorithms. Soft voting method is used and gained a great accuracy score.

#### **4.2 Model Summary:**

The Heart Disease Prediction System finds out the most efficient Machine Learning (ML) algorithm for detection of heart diseases. This specific system detects heart diseases at a very early stage working on given attributes using different computational intelligence technique such as Logistic Regression, Random Forest, Naive Bayes, XGB classifier and Voting classifier algorithms. It gives the output either in positive or in negative. Our recommended system has a very major role to find the invisible discrete patterns and thereby analysis the provided data. After brief comparison of the data, this system helps in heart disease prediction as well as early diagnosis.

#### **4.3 Experimental Result and Analysis**

1. This system predicts an output by given test data. Our recommended system evaluates various guesses and gives precise output by utilizing Voting classifier.
2. All the selected algorithms which we chose for our research, all 5 works perfectly in this purpose.
3. Basic dataset was collected from Kaggle [20] which is an online community for data science and Machine Learning datasets. In this data file, 5 data-sets are merged containing 11 attributes. In total there are 1190 patient data, of which 270 are duplicates. After data cleaning and null value handling, we get 918 patient data to work with.
4. The purpose of this thesis is to make a heart-disease detection environment using the mentioned dataset. As in real life data, it serves the purpose of this report and allows the prediction system to generalize to any new data. By using this system, we can easily detect any cardiovascular disease before it gets deadly for the patient. Thus, this system can save lives.

5. Among the algorithms SVM has the best accuracy rate of 96.19%.

Here is the confusion metrics and ROC Curve:

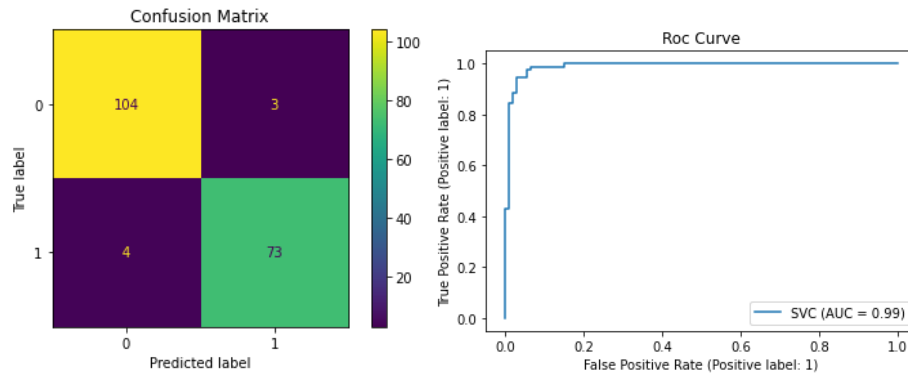


Figure 4. 2. 1: SVM ROC Curve and Confusion matrix

Soft Voting classifier gained 95.65% accuracy.

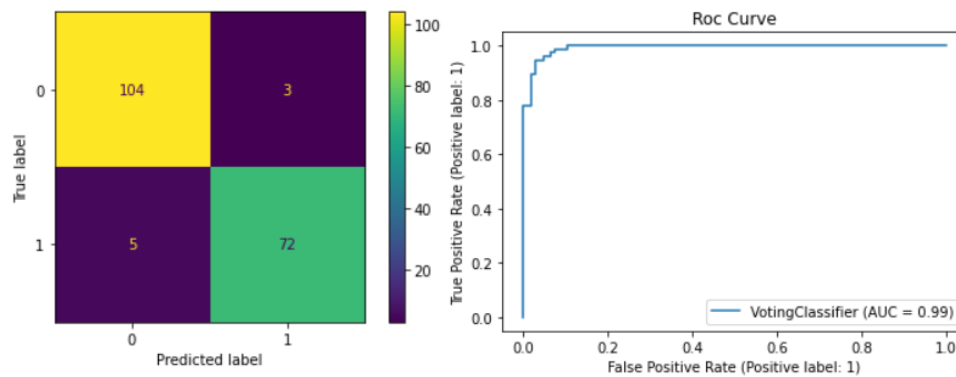


Figure 4.2.2: VotingClassifier ROC Curve and Confusion matrix

Logistic regression has the accuracy rate of 94.56%. Here is the confusion metrics and ROC Curve:

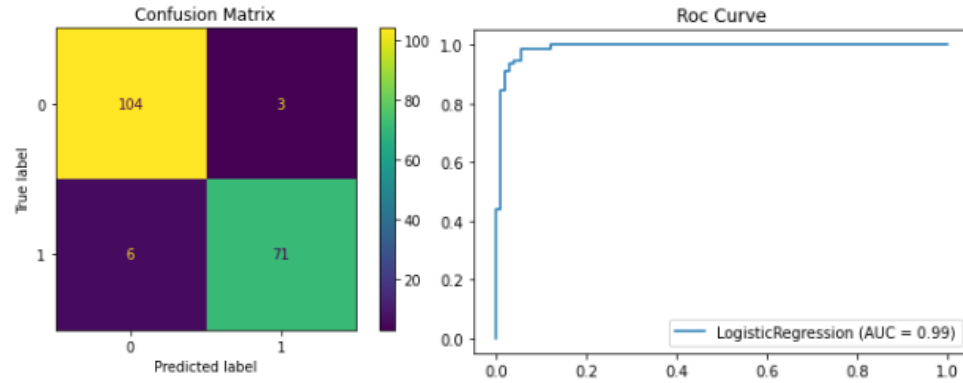


Figure 4.2.3:LR ROC Curve and Confusion matrix

Naïve Bayes has the accuracy rate of 94.02%. Here is the confusion metrics and ROC Curve:

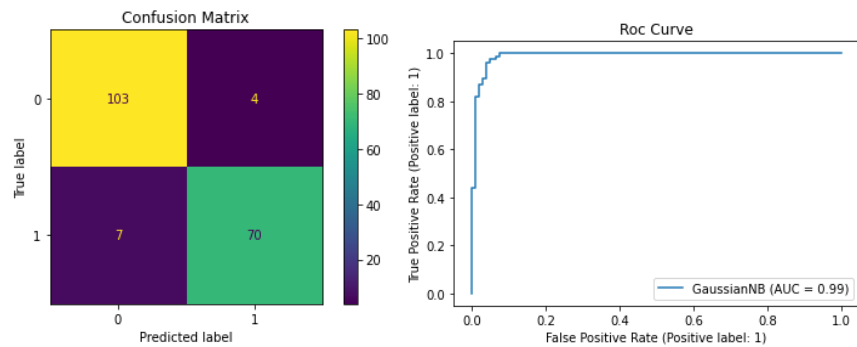


Figure 4.2.4: NB ROC Curve and Confusion matrix

Random Forest has the accuracy rate of 94.02%. Here is the confusion metrics and ROC Curve:

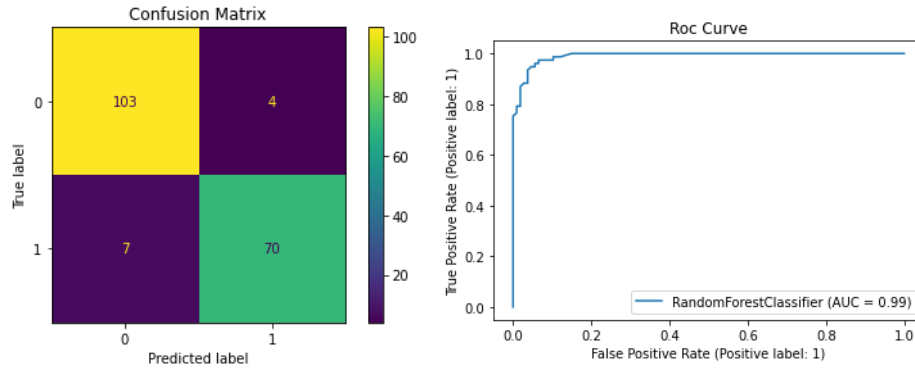


Figure 4.2.5: RF ROC Curve and Confusion matrix

XGB has the accuracy of 95.65%.

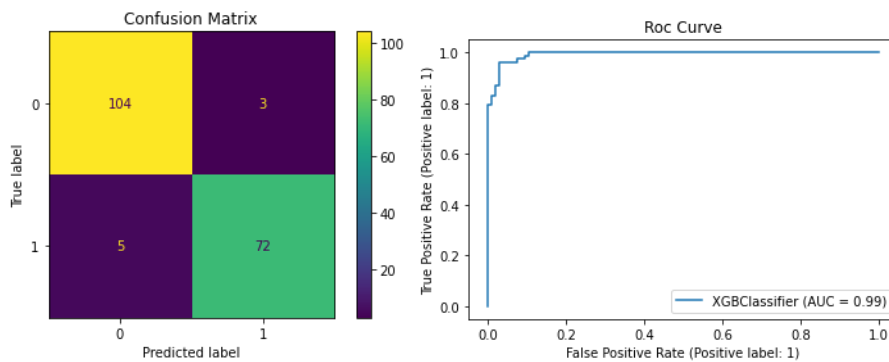


Figure 4.2.5: XGB ROC Curve and Confusion matrix

#### 4.4 Discussion

Identifying the processing of raw healthcare data of heart information will save human lives in the long term and will help in early prediction of abnormal activity in heart. In this effort, raw data was processed quickly using machine learning techniques to provide a new discernment. The ability to forecast heart disease is extremely difficult and important in the medical community. However, if the condition is discovered in the primary or secondary stage and preventative measures are implemented as soon as feasible, the mortality rate can be significantly reduced. In order to focus the research on actual datasets rather than just theoretical frameworks or computer simulations, further extension of this study is very desirable. Random Forest (RF), Support Vector Machine

(SVM), Logistic Regression, naive Bayes, XGBoost Classifier, and Voting classifier properties are used to create the recommended unique system.

Using each algorithm, we get different precision, recall and f1-score. The metrics used for evaluating parameter choices and scenarios, and hence the model metrics, are specified as follows:

$$accuracy = \frac{t_p + t_n}{number\ of\ samples}$$

$$precision = \frac{t_p}{t_p + f_p}$$

$$recall = \frac{t_p}{t_p + f_n}$$

$$f1\ score = 2 \times \frac{precision \times recall}{precision + recall}$$

where  $t_p$  = true positives,  $t_n$  = true negatives,  $f_p$  = false positives and  $f_n$  = false negatives.

| Algorithm                     | Precision   | Recall      | F1-score    | Accuracy (%) |
|-------------------------------|-------------|-------------|-------------|--------------|
| Naive Bayes                   | 0.94        | 0.93        | 0.94        | 94.02        |
| Logistic Regression           | 0.94        | 0.94        | 0.95        | 94.65        |
| Voting Classifier             | 0.96        | 0.955       | 0.96        | 95.35        |
| Random Forest                 | 0.94        | 0.94        | 0.94        | 94.00        |
| XGBoost                       | 0.96        | 0.96        | 0.95        | 95.65        |
| <b>Support Vector Machine</b> | <b>0.96</b> | <b>0.96</b> | <b>0.96</b> | <b>96.19</b> |

Table 4.3.1: Performance comparison

This system proved to be quite accurate in the prediction of heart disease. The future course of this research can surely be performed with diverse mixing of machine learning techniques for betterment of prediction techniques. New feature-selection methods can be developed and improved to get a broader perception of the most important features to maximize the performance of heart disease prediction.



From The above Table SVM has the best accuracy among them. By comparing other works we found this result:

| Approach/References          | Applied Algorithm | Accuracy      |
|------------------------------|-------------------|---------------|
| Khaled Mohamad Almustafa [8] | Decision tree     | 93.85%        |
| Apurb Rajdhan [10]           | Random Forest     | 90.16%        |
| NORMA LATIF FITRIYANI [13]   | MLP               | 85.68+4.21%   |
| Mohammad Ayoub Khan [15]     | MDCNN             | 92.2%         |
| <b>Our Approach</b>          | <b>SVM</b>        | <b>96.15%</b> |

Table 4.3.2: Performance evaluation

## **CHAPTER 5**

### **SUMMARY, CONCLUSION, RECOMMENDATION, IMPLICATION FOR FUTURE RESEARCH**

#### **5.1 Summary of the Study**

This specific system detects heart diseases at a very early stage working on given attributes using different computational intelligence techniques such as Logistic Regression, Random Forest, Naive Bayes, XGB classifier and Voting classifier algorithms. It gives the output either in positive or in negative. Our recommended system plays a very major role to detect the hidden discrete patterns and thereby analyze the given data. After analysis of the data, this system helps in heart disease prediction as well as early diagnosis.

#### **5.2 Conclusion**

Building a system to reliably and effectively anticipate cardiovascular (heart) disorders has become crucial due to the rising death rate from cardiovascular diseases. This study was motivated by the need to identify an accurate and efficient machine learning system for diagnosing cardiovascular disorders. With the use of the gathered dataset, our study analyzes the prognostic power of the Logistic Regression, SVM, Random Forest, Naive Bayes, XGBoost Classifier, and Voting classifier algorithms for heart disease. The results of our experiments indicate that the Support Vector Machine method is the most effective one for detecting heart-related disease, with an accuracy score of 96.19%. In future this work can be elaborated by making a web-application based tester on the SVM algorithm as well as using a larger dataset. It will help to deliver effective and more precise results and help medical specialists by predicting the cardiovascular disease effectively and accurately in primary stage.

#### **5.3 Recommendation**

The use of ML is an endorsement of modern computational science, where human brilliance and several algorithms are applied. Simply, the prediction of heart disease and accuracy can be figured out through this system. This system is recommended as it provides an instrument for medical specialist to detect heart disease at when it only starts to develop. In addition to avoiding serious effects and premature deaths, this aids in giving patients effective care.

#### **5.4 Implication for Further Study**

In the machine learning-based classification technique we suggest, there are high detection rates on heart disease prediction. We tested our techniques on UCI Heart Disease dataset. Other datasets can be used to test this strategy, and results can be improved by applying different machine learning methods to this dataset. Additional classifiers could be added to enhance heart attack detection. So, anyone can use our approaches for getting better results in future work.

## REFERENCES

- [1] Heart Disease Statistics 2022. Available online: <https://www.singlecare.com/blog/news/heart-disease-statistics/>
- [2] Durairaj, M. Revathi, V., 2015. Prediction Of Heart Disease Using Back Propagation MLP Algorithm. , 4(08), pp.235–239.
- [3] Gavhane, A., 2018. Prediction of Heart Disease Using Machine Learning. 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), (Iceca), pp.1275–1278.
- [4] Heart Disease Risk Factors. Available online: <https://www.texasheart.org/heart-health/heart-information-center/topics/heart-disease-risk-factors/>
- [5] Alotaibi, Fahd Saleh. "Implementation of machine learning model to predict heart failure disease." *International Journal of Advanced Computer Science and Applications* 10.6 (2019).
- [6] Evanthia E. Tripoliti, Theofilos G. Papadopoulos, Georgia S. Karanasiou, Katerina K. Naka, Dimitrios I. Fotiadis, Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques, *Computational and Structural Biotechnology Journal*, Volume 15, 2017, Pages 26-47, <https://doi.org/10.1016/j.csbj.2016.11.001>.
- [7] Rohit Bharti, Aditya Khamparia, Mohammad Shabaz, Gaurav Dhiman, Sagar Pande, Parneet Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning", *Computational Intelligence and Neuroscience*, vol. 2021, ArticleID 8387680, 11 pages, 2021. <https://doi.org/10.1155/2021/8387680>
- [8] Almustafa, K.M. Prediction of heart disease and classifiers' sensitivity analysis. *BMC Bioinformatics* 21, 278 (2020). <https://doi.org/10.1186/s12859-020-03626-y>
- [9] jiang, S. (2020). Heart Disease Prediction Using Machine Learning Algorithms. UCLA. ProQuest ID: jiang\_ucla\_0031N\_18588. Merritt ID: ark:/13030/m52v7pfc. Retrieved from <https://escholarship.org/uc/item/7977j5cf>
- [10] Rajdhan, Apurb, et al. "Heart disease prediction using machine learning." *International Journal of Research and Technology* 9.04 (2020): 659-662.
- [11] Saxena, Kanak, and Richa Sharma. "Efficient heart disease prediction system." *Procedia Computer Science* 85 (2016): 962-969.
- [12] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." *IEEE access* 7 (2019): 81542-81554.
- [13] Fitriyani, Norma Latif, et al. "HDPM: an effective heart disease prediction model for a clinical decision support system." *IEEE Access* 8 (2020): 133034-133050.
- [14] Khan, Mohammad Ayoub. "An IoT framework for heart disease prediction based on MDCNN classifier." *IEEE Access* 8 (2020): 34717-34727.

- [15] 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, DOI: 10.1080/03772063.2020.1713916
- [16] Resul Das, Ibrahim Turkoglu, Abdulkadir Sengur. "Effective diagnosis of heart disease through neural networks ensembles." Expert Systems with Applications, Volume 36, Issue 4, 2009, Pages 7675-7680, ISSN 0957-4174, <https://doi.org/10.1016/j.eswa.2008.09.013>
- [17] N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," in IEEE Access, vol. 8, pp. 133034-133050, 2020, doi: 10.1109/ACCESS.2020.3010511.
- [18] Soni, Jyoti, Ujma Ansari, Dipesh Sharma, and Sunita Soni. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." International Journal of Computer Applications 17, no. 8 (2011): 43-48.
- [19] A. N. Repaka, S. D. Ravikanti and R. G. Franklin, "Design And Implementing Heart Disease Prediction Using Naives Bayesian," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 292-297, doi: 10.1109/ICOEI.2019.8862604.
- [20] Heart Failure Prediction Dataset. Available online: <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction>
- [21] Heart Disease individual Dataset. Available at: <https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/>

# PLAGIARISM REPORT

## Heart Disease Prediction using machine learning

### ORIGINALITY REPORT

|                                |                                |                            |                              |
|--------------------------------|--------------------------------|----------------------------|------------------------------|
| <b>29%</b><br>SIMILARITY INDEX | <b>24%</b><br>INTERNET SOURCES | <b>20%</b><br>PUBLICATIONS | <b>18%</b><br>STUDENT PAPERS |
|--------------------------------|--------------------------------|----------------------------|------------------------------|

### PRIMARY SOURCES

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>www.ijert.org</b><br>Internet Source  | <b>3%</b> |
| <b>2</b> | <b>Submitted to Daffodil International University</b><br>Student Paper   | <b>2%</b> |
| <b>3</b> | <b>dspace.daffodilvarsity.edu.bd:8080</b><br>Internet Source   | <b>2%</b> |
| <b>4</b> | <b>www.tandfonline.com</b><br>Internet Source  | <b>1%</b> |
| <b>5</b> | <b>Purushottam, , Kanak Saxena, and Richa Sharma. "Efficient Heart Disease Prediction System", Procedia Computer Science, 2016.</b><br>Publication | <b>1%</b> |
| <b>6</b> | <b>jcreview.com</b><br>Internet Source   | <b>1%</b> |
| <b>7</b> | <b>www.sersc.org</b><br>Internet Source  | <b>1%</b> |
| <b>8</b> | <b>bmcbioinformatics.biomedcentral.com</b><br>Internet Source  | <b>1%</b> |