# Heart Disease Prediction System

by

Tayab Al Azad Amit
18201197
Raida Nawar Fullkoli
17301204
Niloy Palit
17301094
Farhana Khan Nafisa
19101654
MD Muntakim Ahmed Binoy
17201048

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
January, 2022

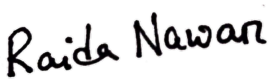
# Declaration

It is hereby declared that

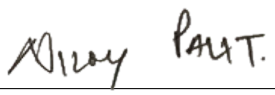
1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

**Student's Full Name & Signature:**

Tayab Al Azad Amit
18201197



Raida Nawar Fullkoli
17301204



Niloy Palit
17301094

Farhana Khan Nafisa
19101654

MD Muntakim Ahmed Binoy
17201048

# Approval

The thesis/project titled "Heart Disease Prediction System" submitted by

1. Tayab Al Azad Amit (18201197)
2. Raida Nawar Fullkoli (17301204)
3. Niloy Palit (17301094)
4. Farhana Khan Nafisa (19101654)
5. MD Muntakim Ahmed Binoy (17201048)

Of Fall, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 16,2022.

**Examining Committee:**

Supervisor:
(Member)

Moin Mostakim
Lecturer
Department of Computer Science and Engineering
Brac University

Co- Supervisor:
(Member)

Faisal Bin Ashraf
Lecturer
Department of Computer Science and Engineering

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

# Ethics Statement

We, hereby declare that this thesis paper is completely based on the findings of our enormous research. In this research paper, we ensure the highest level of plagiarism-free environment. We are confirming the total security of the informants or any relevant sources

# Abstract

[1]According to the World Health Organization (WHO), 17.9 million people die each year due to cardiovascular diseases (CVDs), almost 31% of all deaths worldwide. This single piece of evidence is strong enough to describe the lethal nature of cardiovascular diseases or, as we know, heart diseases. There is no denying that different medical sectors using the help of high-end technologies, now have figured out ways to tackle serious CVDs. However, then again, we indeed cannot rule out the amount of distress these CVDs bring. We need to know how to prepare ourselves to face different heart diseases. One of the many ways can be implementing different Machine Learning and Neural Network algorithms. Say, for example, in this paper; we will discuss algorithms like Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), ConvMLP, and ANN; on how each of these techniques can be applied to find out a better way to predict the availability of heart disease in a particular individual depending on few given factors. Our main goal is to make the course easy to detect diseases that belong to the heart and enriches the medical sector. In our country, the medical sector is improving day by day. We aim to boost this improved significantly. By using Machine Learning and Neural Network algorithm, we are optimistic about implementing this idea.

**Keywords:** Random Forest Classifier; Decision Tree; Logistic Regression; Support Vector Machine; MLP Classifier; Conv-MLP; ANN; Machine Learning; Neural Network; Heart Disease Detection;

# Dedication

Firstly, we would like to dedicate our paper to our beloved parents. Without their boundless sacrifices and motivation, we would not have been able to come this far. Secondly, Our supervisor and co-supervisor sir for their smart guidance which has a significant role to improve the quality of this research.

# Acknowledgement

To begin with, all praise to the Great Allah for whom our thesis have been completed without any vital interruption. Along with, we would like to thank our honorable and conscientious supervisor Moin Mostakim sir and co-supervisor Faisal Bin Ashraf sir for their kind support, guidance, and feedback in our work. He motivated us to conduct this research and always gave his valuable time to make us understand the research works. Furthermore, We would like to thank our respected BRAC University for providing essential resources and facilities during our research phase. To conclude, a big thank to our parents without their perpetual support may not be possible. With their mentionable support and prayer, we are now on the brink of our graduation.

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

A vast number of people are getting affected every day by the causes of heart disease. Every day the number of affected people is increasing speedily. In today's world, the chief cause of death is heart disease. It is the primary basis of death.[1] WHO claims that because of the cementation of cardiovascular diseases, over 17.9 million people die. Coronary artery diseases and cerebral strokes are the common reasons for death. Some factors lead a person to be affected by heart diseases. This can be personal or professional habitude as example smoking, indiscipline lifestyle. Some known factors are responsible for heart diseases. Addiction to alcohol, caffeine, and mental stress are the main risk factors for heart disease. [2]Obesity, HBP, excessive cholesterol are predisposing factors of heart disease. Nowadays, the medical sector is growing faster with lots of efficient equipment. However, when there is a matter of heart disease, people have to bear a significant amount of costs because of many diagnostic tests. Medical diagnosis should be serviceable, trustworthy and help to reduce the high costs for diagnostic tests by using computer techniques. Early and exact diagnosis plays a significant preface in preventing death by various heart diseases. Heart disease prediction by using machine learning is a significant idea to augment the speed of medical science. In Computer Science, Machine Learning tends to be one of the most efficient branches. Since we are trying to predict a particular result in this paper, we believe algorithms like Logistic Regression, Decision Tree, Support Vector Machine, Random Forest Classifier,Conv-MLP, and ANN are the most fruitful way to predict the desired result. As we can see, Machine Learning and Neural Network have higher potential in all sorts of medical research. [3] Implementation of medical diagnosis with ANN, a method suggested by Frank Rosenblatt in 1958, can give the best possible outcomes. The Artificial Neural Network(ANN) is commonly used to predict heart disease. Recent research has shown that ANN-based classification generates promising results in which the neural networks represent the problem layer by layer by using an N-dimensional tensor. ANN can learn by processing labeled training datasets because ANN can identify data trends and form a model after going through layers of functions. ANN can use this to perform classification with input testing data. Eventually, it will increase the accuracy and the quality of the model. In this research, we use a computerized method to get a better result called ConvMLP. This contains multiple layers, and each layer is connected with each layer. There are layers of nodes: an input layer, a hidden layer, and an output layer. Our primary purpose of using ConvMLP classifiers is to get better results of accuracy, and its performance is up to the mark doubtlessly. Machine learning is an

outgoing branch that belongs to Artificial intelligence. All the ML techniques follow different ideologies for implementation. This is a bright side as our goal is to find out the best possible predicting rate. We believe it is essential in medical industries to have precise ways of predicting particular diseases, let alone serious ones like CVDs or, as we know, heart diseases. In this paper, we will figure out through different steps which Machine Learning model will give us the highest accuracy to predict heart disease with few attributes in a particular dataset.[4]The dataset in this particular paper has been used from Kaggle .Throughout the paper, we will discuss the different Machine Learning and Neural Network techniques we mentioned before. Once we are done with clarifying the methods, we will discuss the results of each technique and later will compare them to previous research.Finally conclude with the particular technique that will provide us with the highest accuracy for prediction.

## 1.1 Research Problem

In this day and age, Machine Learning has spread its divine touch through every aspects of our lives. It has become a key strategy and procedure for solving problems like computational biology, which are being intricately dealt with through the use of algorithms. Over the years, heart diseases like cardiovascular disease, coronary heart disease and heart failure have been identified as one of the primary causes of early death. Many data analytics tools using machine learning have been applied to look for early indications to lessen the impact of having such a disease. Combining the amalgamation of these key identifications based on the individual's age, sex, diabetes status, and past hospitalization record, we can get a fair estimation of the severity of the disease. On top of that, several datasets based on the estimation of future heart disease incidence greatly increases the chances of early diagnosis depending upon the circumstances.[5] Various machine learning algorithms (Support-vector Machine Algorithm, Naive Bayes Algorithm, KNN Algorithm, Decision Tree Model, Logistic Regression Model, and Random Forest Classifier Model and so on) are applied before to predict heart disease cases with minimal attributes. Countless numbers of models have been put forward, in which the accurate data of affected patients have been used to modify the models. As per the research, Our primary concern is to take safety measures to decrease the impact of this disease. Even though using the best possible procedures to obtain a maximal result has some significant drawbacks due to uncertain characteristics of this disease. So, in this work, we tried to present a system that can improve the work efficiently based on available datasets. We did compare previously used algorithms and find a better accuracy rate, leading to a satisfactory result.

## 1.2 Research Objectives

This study has the following aims:

• To find out a reliable early detection method to take immediate actions at early stages of heart diseases.

• To report a comparative analysis of heart disease datasets using various machine learning and neural network algorithms.

• To predict Heart Disease cases with best performance-based algorithm.

• To compare efficiency of machine learning and neural networks in medical science.

## 1.3 Thesis Structure

**Chapter 1:** Introduction where motivation, problem statement, research objectives and contribution are discussed.

**Chapter 2:** Related works where background, literature review and algorithms are discussed. In the literature review part, we reviewed the previous work.

**Chapter 3:** System Model represents the view of our methodology.

**Chapter 4:** Experiment Analysis and Data Visualization, where we visualized our data and analyzed the result occurred from our algorithms.

**Chapter 5:** Conclusion and Future works where the conclusion consists of our work till now and future works include the scope of improvement.

# Chapter 2

# Related Work

## 2.1 Literature Review

[6] This paper presented three supervised machine learning algorithms, KNN, Naive Bayes, and SVM, to predict heart diseases and compare their performances. The dataset consists of 302 records in the Heart disease database. The dataset was divided into a training dataset, and the testing dataset also, it was split into 70% and 30%. In machine learning, the confusion matrix is a particular table design where it permits the perception of the execution of a calculation and confusion matrix obtained by three different algorithms. Among three algorithms, Naive Bayes predicts the diseases with the highest accuracy of 86.6% compared to KNN and SVM.

[7] In this paper, it is presented that machine learning is used to increase the accuracy rate. It proposed a logistic regression-based approach of machine learning for heart disease prediction. Also, other algorithms were Naive Bayes and Comparing and confusion matrices. According to them, the experimental result of LR was 86.89%, and Logistic Regression performed better at 86.89% accuracy while other algorithms performed 77.85% for KNN, 86% for NB, 78.69% for DT, and 82% for SVM. They concluded their paper by explaining that we can use a logistic regression algorithm to predict if a patient has heart disease or not.

[8] This paper explained the heart diseases like cardiovascular diseases (CVD), Coronary artery diseases (CAD), Coronary heart disease (CHD). They presented data mining algorithms and popular data mining algorithms: decision trees, Naive Bayes, K-means, SVM, and artificial neural networks. They also showed classification accuracy and time complexity of the algorithms where the Naive Bayes algorithm plays a key role in shaping improved dataset classification. They conclude their paper by saying that electing the suitable technique for data cleaning and proper classification algorithms can lead to the development of prediction systems.

[9]This paper explained the Neural Network-based prediction of coronary heart disease, and it proposed a Neural Network-based CHD risk prediction method based on feature correlation analysis (NN-FCA) which includes two processes, and these are feature selection and feature correlation analysis. In the first system process, the KNHANES-VI dataset was examined, then statistical analysis was performed to identify features related to CHD risk.In the fourth step, using feature correlation

analysis NN-based coronary heart disease risk predictors were trained,Performance measurements were taken in the fifth phase to confirm NN-based CHD risk prediction using feature correlation analysis. They also showed the input variables for the model training were age, sex, cholesterol, blood pressure, and other features. The output variables were high blood pressure, dyslipidemia, stroke, myocardial infarction, and angina.The experimental results also showed that the NN-FCS model and FRS model were equally good. After comparing the validation of the FRS for the Korean population, the NN-FCA model resulted in more accurate coronary heart disease risk prediction. They concluded by suggesting that similar clinical decision support could be developed and NN-FCA can be used for diseases other than coronary heart disease.

[10] In this paper, the researchers explained that for small datasets Naïve-Bayes algorithm is suitable, and for larger datasets, the decision tree is suitable. In their paper, they used a small dataset. So, their preferred algorithm was Naïve-Bayes. They explained the datasets that they were working with. Then they gave some ideas about Naïve-Bayes and K-means. After that, they discussed the methodologies for the implementation. They conclude their paper by explaining their proposed methods, how problems can be handled and how more accurate results can be obtained.the authors have applied four machine learning supervised models (SVM, NB, KNN, tree) on a heart disease dataset. Then tries to predict the possible coronary heart disease patient in the next ten years. This dataset contains 16 attributes. They start with input heart disease data then build the model and finally calculate ta accuracy for the constructed model. In conclusion, we can say it is a system, which only can find out the accuracy in the percentage of those models.

[11] The researchers proposed learning vector Quantization neural system calculation to diagnose heart disease in this paper. The researchers applied KNN on a benchmark dataset to investigate its efficiency in the diagnosis of heart disease. The research outcomes show that applying KNN resulted in 97.4%, which is higher than discoveries on that benchmark dataset. The neural system in this framework recognizes 13 clinical incorporates as data and predicts predicts the proximity or absence of coronary sickness in the patient, close by different execution measures.

[12] The authors of this paper compared algorithms with different performance measures using machine learning. All the data were pre-processed, and after testing, different algorithms worked well on different datasets. K-Nearest Neighbor K-NN, Random Forest RF, and Artificial Neural Network MLP are the algorithms that work well for most datasets. The writers have proposed an improved model using FCBF. Multilayer perception approaches, Particle Swarm Optimization, and Ant Colony Optimization. The existing methods and their researched results found that the proposed optimized method beats different coronary illness prediction and classification models.

[13] In this paper, decision trees, Naïve Bayes, and neural networks are analyzed for heart disease prediction. Based on their accuracy, the execution of these strategies is looked at. This work shows the precision of Neural Networks, Decision Trees, and Naive Bayes is 100%, 99.62%, 90.74%, respectively. By analyzing this research

work out of these classification models, Neural Networks outflanked the other two techniques in heart disease prediction accuracy.

[14] The authors of this paper investigated ensemble classification to improve the performance accuracy of weak algorithms by combining multiple classifiers. They implemented this algorithm with a medical dataset to show its functionality to forecast disease at an early stage. This study finds that the ensemble technique is able to improve the weak classifiers at a rate of 7%. They also improvised the process with feature selection and got a satisfactory improvement.

[15] In this paper, for heart disease classification, the authors proposed ANN and Neuro-Fuzzy systems. They used the Cleveland dataset. They used 80% of data for training and 20% of data for testing. They got 87.04% accuracy for ANN.

[16] In this paper, they used SVM and ANN for heart disease prediction. They have used 200 data for training and 103 data for testing. They got 80.41% accuracy in ANN.

[17] In this paper, MLP with Back Propagation (BP) is used. 182 data is used for training and 121 data is used for testing. They got 80.99% accuracy for MLP.

## 2.2 Methodology

Here, we used six different models. The project will figure out the best possible model among the six chosen. It is necessary because, as users, data pre-processing techniques determine sufficient input data, we can assume that based on the best model. In the following section in figure 2.1, the work plan of the six models used are,

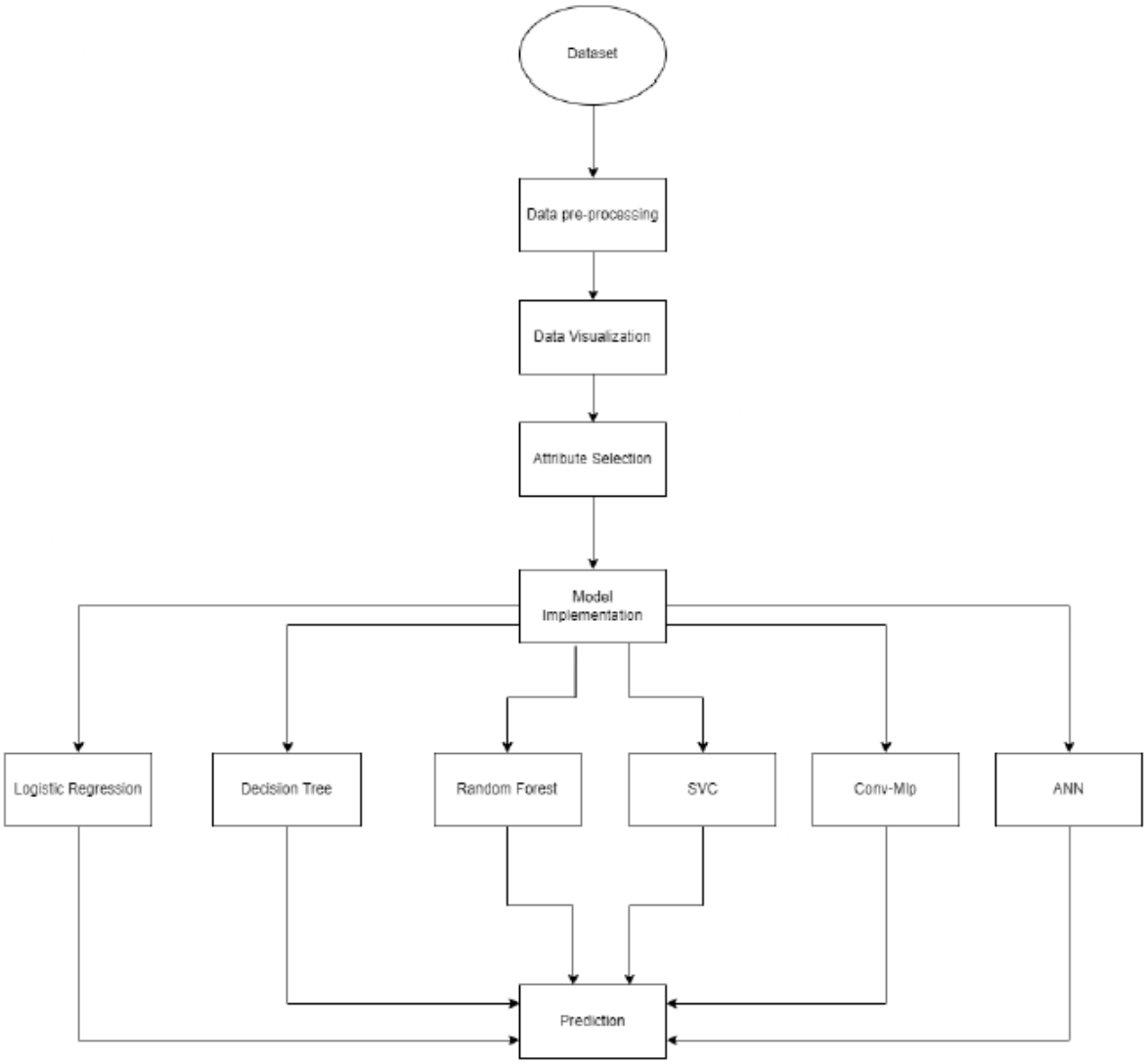



Figure 2.1: The flow chart of the proposed Heart Disease Prediction System

### 2.2.1 Dataset

[4]Our paper consists of 14 different Attributes for predicting heart disease. The attribute is described briefly below in table 2.1:

| **Attributes** | **Values and Meaning** |
|---|---|
| Age | Age (Years) |
| Sex | Value 1 and 0 for male and female |
| Cp | Chest pain of a patient ranges(0-3) |
| Trestbps | resting blood pressure on hospital admission (mm Hg) |
| Chol | cholesterol measurement( mg/dl). |
| Fbs | Fasting blood sugar measured within 120 mg/dl. Values are in values of 0 and 1. 0 for less than 120 mg/dl, and 1 for more than 120 mg/dl. |
| Restecg | Resting Electrocardiographic Results within the range of 0-2 |
| Thalach | Highest heart rate (71-202) |
| Exang | Exercise induced angina(exang), ST depression induced by exercise relative to rest (oldpeak),the slope of the peak |
| Oldpeak | ST depression induced by exercise relative to rest.This value is in the range of 0-6.2 |
| Slope | This is the slope of the peak exercise ST segment |
| Ca | Number of major vessels (0-3) colored by fluoroscopy |
| Thal | 3=normal;6=fixed defect;7=reversible defect |
| Target | predicted result in the form of binary values (0 and 1), 0 means the person doesn't have heart disease and 1 means the person has heart disease . |

Table 2.1: Dataset

### 2.2.2 Data Preprocessing

There are several methods for Data Preprocessing depending on the data condition. For our system, we are using Data Cleansing and Create Dummy Variables. Both two will be described below:

#### Data Cleansing

In the dataset, we find many irrelevant data and null values. We see many outlier values that may negatively impact our model accuracy by visualizing our dataset. That's why we use the cleansing process to detect the data from the dataset and then correct those complications by replacing, modifying, deleting the complex data.

#### Create Dummy Variables

Generally, what data we get from the real world has a significant difference between them and directly impacts the performance of the model. By following the dataset, we get numerous categorical attributes named CP (Chest Pain Type), thal ($\beta$- Thalassemia Cardiomyopathy), slope (The peak exercise ST segment). That absolute attribute can impact the model's stability and the coefficients. So, we can encode the categorical variables as dummy variables that allow easy calculation of the odds ratios, increasing the model's ability and better significance in coefficients.

### 2.2.3 Data Visualization

#### Visualization of heart disease frequency for Age

According to the visualization from figure 2.2, we find out that 41-54 shows a higher chance of having heart diseases. For the age range of 55-70, we see a tendency of the population not to have heart diseases.

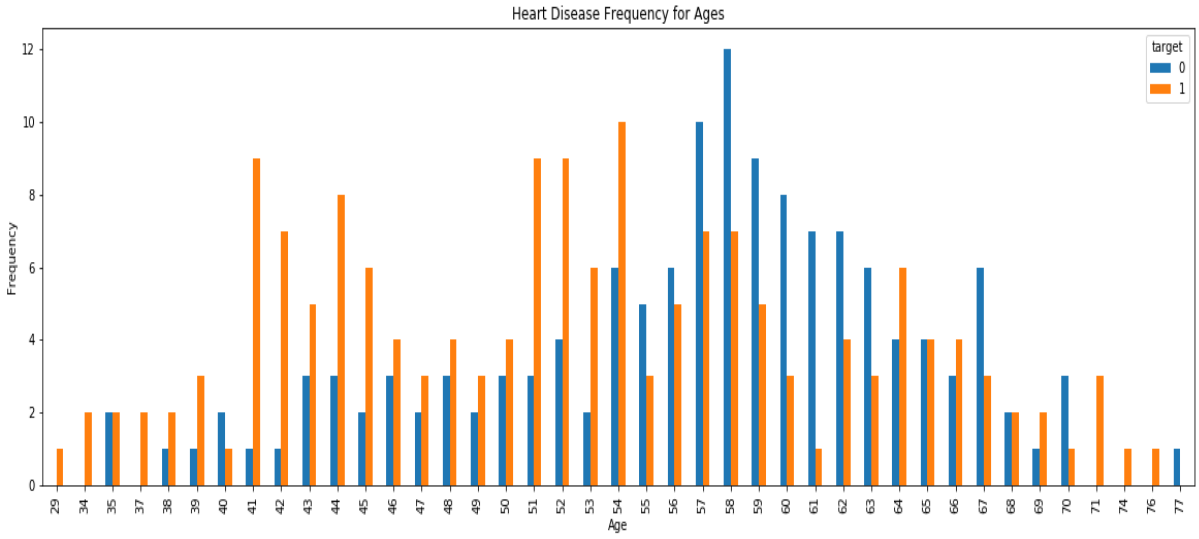



Figure 2.2: Heart disease frequency for ages

### Visualization of heart disease frequency for Sex

In figure 2.3,the graphical representation shows that females have fewer chances of being diagnosed with heart diseases.

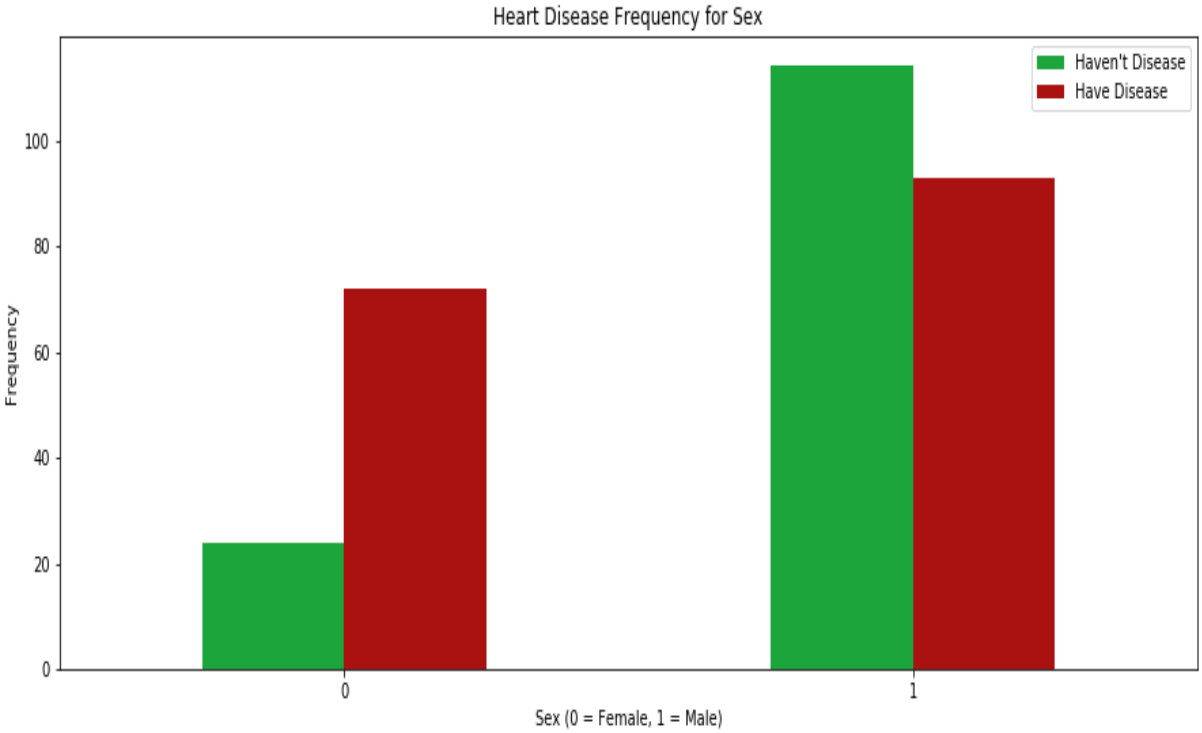



Figure 2.3: Heart disease frequency for sex

### Heart Disease Frequency according to FBS

The dataset has been divided into two sections of Fasting blood sugar higher than 120 and one lower than 120.From figure 2.4,it is evident that the graph shows a higher tendency of heart diseases to be present if the fasting blood sugar remains under 120 mg/d.

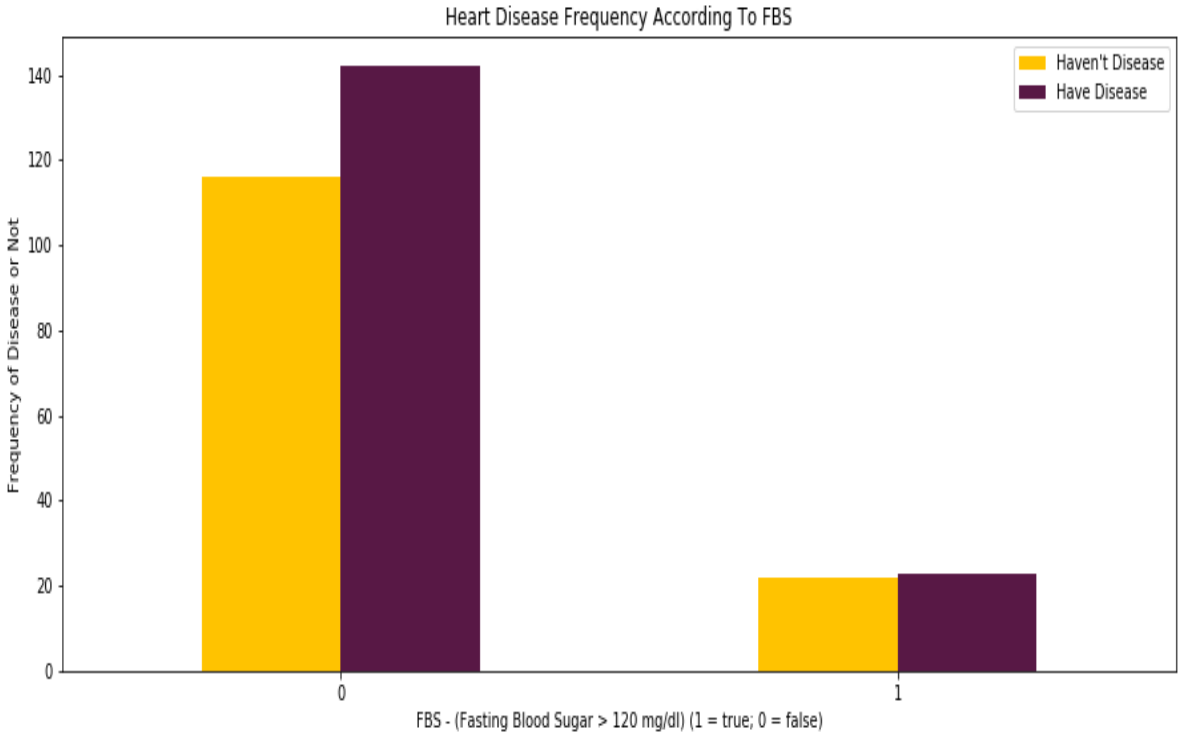



Figure 2.4: Heart disease frequency according to fbs

### 2.2.4 Attribute Selection

After data analysis by visualization, we find some irrelevant data null values. For this reason, we get some outlier, occurs the curse of dimensionality, increases overfitting, and complex the model calculation. So, we use some techniques of data pre-processing and then select the relevant features which are strongly correlated.

## 2.3 Algorithm Descripction

We have used different kinds of Algorithms for our research purpose. The following are explained in this section:

### 2.3.1 Logistic Regression

Logistic regression is a mathematical modeling approach that can be used to describe the relationship of several Xs to a dichotomous dependent variable where X is considered to be independent variables. The most common uses of logistic regression are predicting and assessing the chances of success. It underlies the sector of classification. Simply put, logistic regression is used to find out the chances of a designated class or events existing such as pass/fail, win/lose etc. [18] Binary logistic regression, multinomial logistic regression, and ordinal logistic regression are the three types of logistic regression are discussed here. The cost function is limited between 0 and 1 through logistic regression.[19] Attributes with p-values less than alpha are found (5%). P is the probability of an event that is the risk of CHD. Also, P always stays between 0 and 1
*Sex

*Age

*cigsPerDay

*totChol

*sysBP

*glucose

$\text{logit}(p) = \log\left(\frac{p}{(1-p)}\right)$
$= \beta_0 + \beta_1 \times \text{ sex } + \beta_2 \times \text{ age } + \beta_3 \times \text{ cigsPerDay } + \beta_4 \times \text{ totChol } + \beta_5 \times \text{ sysBP } + \beta_6 \times \text{ glucose}$

### 2.3.2 Support Vector Machine

[20]Support Vector Machine (SVM) is a well-known supervised machine learning algorithm which can be used for either classification or regression challenges. However, it is mostly used in classification problems. The SVM algorithm's purpose is to generate optimal lines or decision boundaries that divide n-dimensional space into classes, allowing additional data points to be conveniently placed in the correct category in the future which is called the hyperplane. The SVM selects vectors to help create the hyperplane. There are two lines parallel to the both sides of the hyperplane which is called a parallel hyperplane. Also the marginal line is created along the nearest positive point and most nearest negative point. The algorithm is called a support vector machine because these extreme cases are called support vectors. It helps to find the expected classification mistake for the unseen patterns.

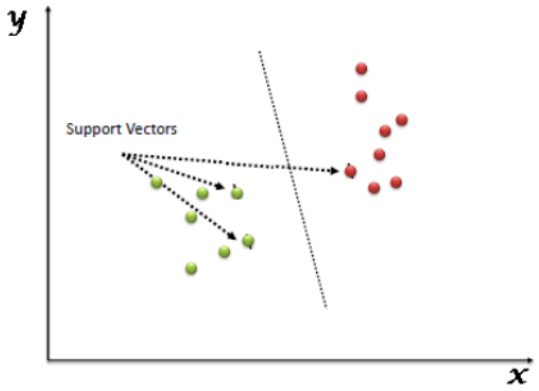



Figure 2.5: Support Vector Machine

### 2.3.3 Random Forest

[21]Random forest is a supervised machine learning technique that is widely used in classification and regression. It is a classification model that creates a forest with many trees. This classifier provides quality results without hyper-parameter tuning. It is renowned for making more precise predictions by creating multiple decision trees and generally assembling them. It is also called the supervised learning algorithm. In machine learning, it is an absolute privilege to use the random forest in both the case of classification and regression in this research. The training samples are picked up randomly with a proper replacement at the tree construction phase. With the help of Bootstrap aggregating, this is edified. A forest consists of trees and more trees form the forest burlier. The forest has concurred with an arrangement of decision trees. This bootstrap aggregating method plays a vital role in the overall result. It has indistinguishable parameters like bootstrap aggregating classifiers and decision trees. On data samples, the random forest algorithm generates and, after getting the result, chooses the most acceptable solution with voting. Random forest merges and constructs several decision trees to pursue detailed and stable forecasts.

By using random forest classifiers, merging with bootstrap aggregating classifiers and decision trees is unnecessary because we can use the random forest classifiers class precisely. It has a quality that the relative significance of each function on the forecast is very normal to calculate. We can obtain the regression task in the random forest with the help of the regression algorithm.

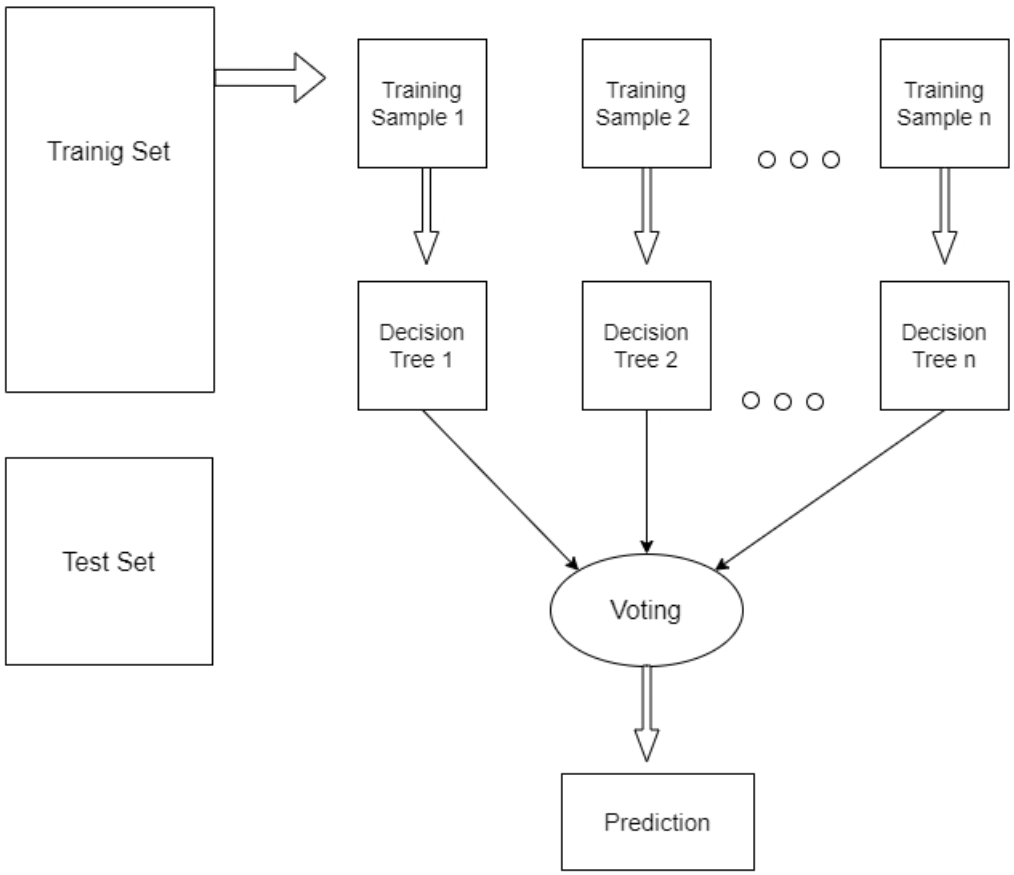



Figure 2.6: Random Forest Classifier

The concept of RF is ensemble learning, which may be a handle of combining different classifiers to fathom a complex issue and make strides in the execution of the demonstration. It is known that high accuracy comes from more trees, and that handles the missing values.

### 2.3.4 Decision Tree

A decision tree is a useful algorithm for prediction and classification in the form of a tree structure.There are 3 components of a decision tree-paths,branches and nodes. The path is a set of branches where the attribute's values are called. Leaves represent the class values. After that, their node called root represents the total dataset. [22]A decision tree is a non-parametric algorithm without distributional assumption.It can also handle large and complex data sets. For large datasets, the study data will be divided into training and validation sets. To arrive at the best fi-

nal model, training data sets create decision tree models, and validation data groups determine the right tree size.
There are some algorithms in the decision tree, and these are ID3, C4.5, CART, CHAID, MARS. We have used the CART algorithm in this paper. This algorithm applies to both classification and regression. The CART algorithm employs the Gini Index criterion to break a node into sub-nodes. The CART algorithm combines testing with a test data set with cross-validation to quantify fit quality.

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \tag{2.1}$$

Here in equation 2.1, E denotes entropy, pi denotes the probability of randomly picking an element of i, c denotes the number of classes.

### 2.3.5 MLP Classifier

[23] MLP stands for multi-layer perceptron, the most basic neural network.MLPs are best suitable for the classification prediction issues.It has hidden layers which helps to do harder prediction also it helps to understand more complex data and it has a higher predictive power. In Mlp we can add multiple layers as much as we need. We can add more hidden layers and add more neurons to each layer so we can have a higher predictive power. [24]They're also helpful for predicting situations like regression in which a real-valued quantity is predicted from a collection of inputs. [25]ConvMLP is a hierarchical convolutional MLP for visual recognition, consisting of a convolutional layer and a step-by-step code design of the MLP. Experiments with object recognition and semantic segmentation have shown that visual representations learned through ConvMLP can be seamlessly transferred to achieve competitive results with fewer parameters.

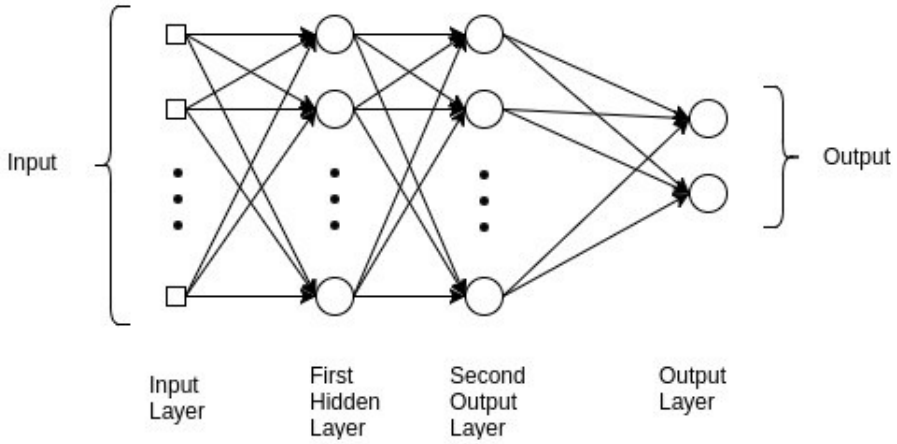



Figure 2.7: MLP Classifier

### 2.3.6 Artificial Neural Network(ANN)

Artificial neural networks (ANNs) are a one of a kind technology that is used in research on the brain and nervous system.[26] ANN is structured like neurons in the brain, just like the brain the neurons are structured in layers. [26]ANN portrays three layers of neural network in which: The input layer is first. The input neurons that are on duty to send data to the hidden layer are situated in this layer. The data that will be input is computed in the hidden layer, The output which consists of activation, weight, and cost functions is transferred to the output layer from the hidden layer. ANN(artificial neural network) uses pattern recognition or data classification with the help of a learning process in biological systems that sometimes requires adjustments to the synaptic connection between neurons to feed its information-processing system. Here, by neuron it means neural net which consists of a huge number of processing units. Each neuron has a connection with other neurons that helps in transmitting data from one to another. Each and every connection is associated with a weight.[26] ANNs stand on the assumptions that: Information processing happens at neurons, through connection links the information is passed between neurons.In a neural net every single connection link connects with each other with weight that multiplies the signal transmitted through it and an activation function is applied by each neuron to the network input.

# Chapter 3

# System Model

Figure 3.1 represents the graphical view of our methodology. Firstly, we collected a dataset. Then we used data pre-processing to process the data. We used data cleansing and created dummy variables to detect the irrelevant data from the dataset and then correct those complications by replacing, modifying, deleting the complex data. After that, we visualized the data from the dataset. We found categorical and continuous values from the dataset and scaled the dataset using standard scaler in data-processing. After using the batch normalization function and dropout function to solve overfitting issues, training data was in a batch form in batch normalization to normalize the value. After collecting the data, we applied some algorithms. After the data gets processed by the algorithms we mentioned above, we got an output. Hence, we compared our data to standard algorithm data. After reaching, we got our predicted success rate.

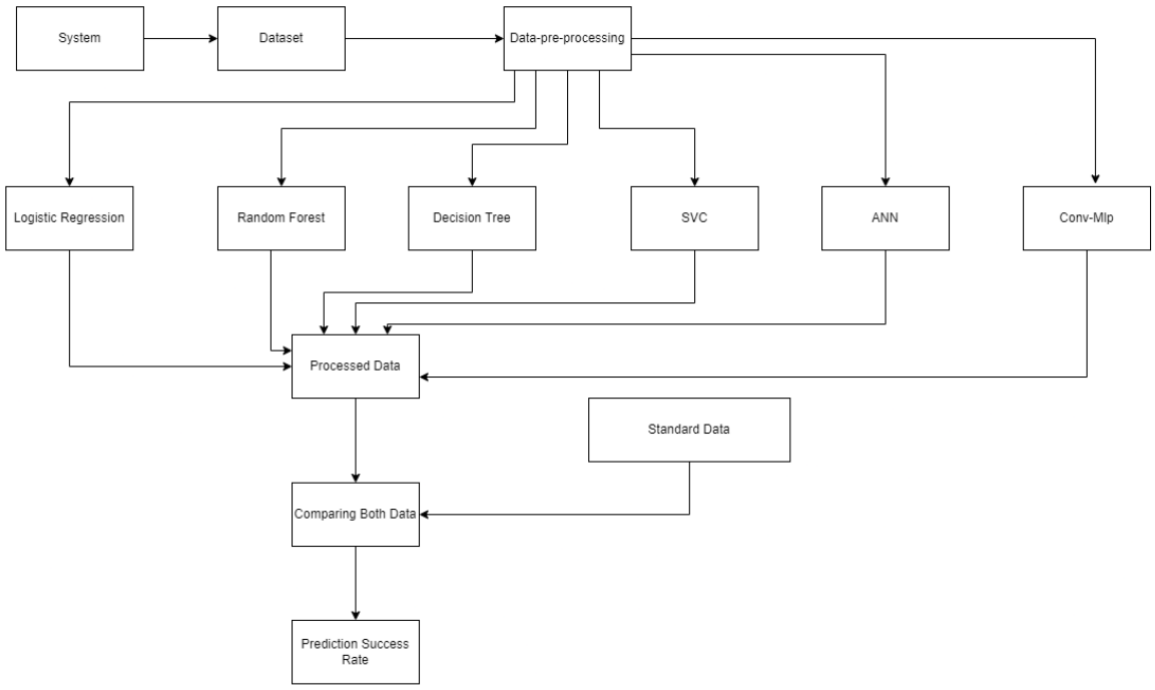



Figure 3.1: System Design

# Chapter 4

# Experiment Analysis

Our thesis aims to predict heart disease cases with minimal attributes by using available technological advancements that develop prediction models for heart diseases survivability. The accuracy of the model defines the performance of the model. According to the training and testing dataset, many of the model's predictions are correct.

This paper used machine learning and neural networks to predict heart diseases with minimum attributes. We used logistic regression, support vector machine, decision tree, and random forest in machine learning and from the Neural Networks, we used ANN and Conv-MLP to predict heart disease. In logistic Regression, we get 88.52% accuracy, means that this model will correctly predict 88.52 out of 100 samples.

| Test | Actual | Predicted |
|---|---|---|
| 0 | 0 | 0 |
| 1 | 1 | 1 |
| 2 | 0 | 1 |
| 3 | 1 | 1 |
| 4 | 1 | 1 |
| 5 | 0 | 1 |
| 6 | 0 | 0 |
| 7 | 1 | 1 |
| 8 | 1 | 1 |
| 9 | 1 | 1 |

Table 4.1: Test comparison

In Support Vector Machine, we get 90.16% accuracy. We get close accuracy concerning logistic regression in Random Forest Classifier. The accuracy of this model is 86.88%. It says that this model will correctly predict around 86.88 out of 100 samples. We get 78.68% accuracy in the Decision Tree Classifier. This model will

correctly predict the accuracy of 78.68 out of 100 samples represented by this model. This model is not suitable for high performance according to our attributes. In the Neural Network, we used Conv-mlp and Artificial Neural Network. In the Conv-mlp, we get 87.20% accuracy. This means this algorithm will correctly predict around 87.20 out of 100 samples. Again in Artificial Neural Network, the accuracy is 91.24%. This algorithm so far gives the highest accuracy. It will predict around 91.24 out of 100 samples correctly. We can consider the ANN as the fittest function to obtain optimal weights for our attributes.

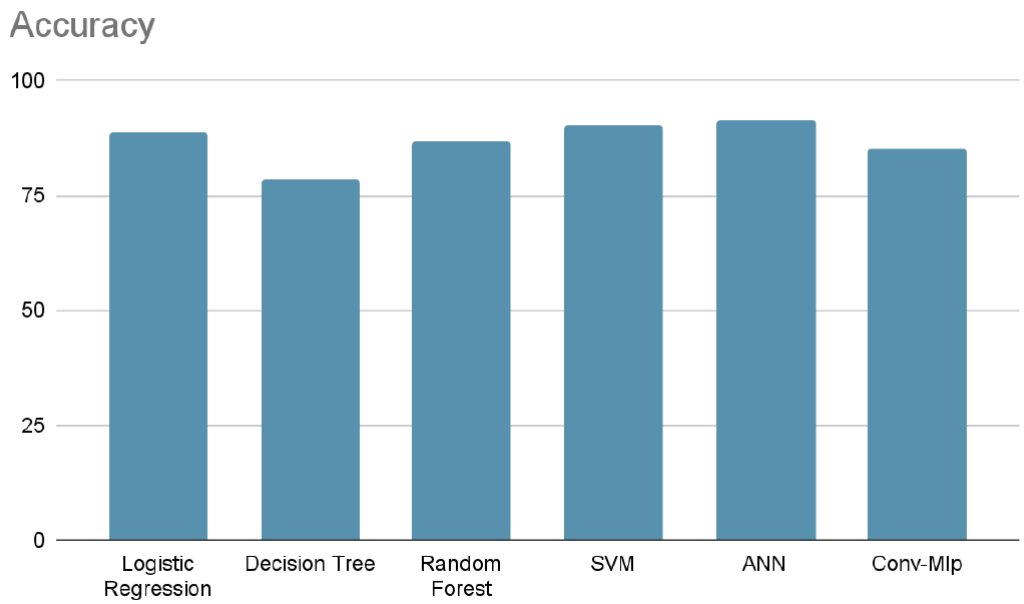



Figure 4.1: Accuracy Chart

Precision and recall are the best way to evaluate a model. Precision is a metric that identifies the number of correct optimistic predictions out of all positive predictions. On the other hand, recall is also a metric determining the number of accurate optimistic predictions and identifying missed positive predictions. That is why recall is also known as sensitivity.

In the Random Forest Classifier model, the precision value is 90%, which means the confirmation of heart disease is 90% correct. On the other hand, the value of recall is 84.37% that means it correctly identifies 84.37% of all who have heart disease. In the Decision Tree, we get lower values of precision and recall, which are 82.75% and 75%. That provides lower accuracy of rectification and sensitivity. In the Support Vector Machine, we get a precision value of 93.33% and an 87.5% recall value. In logistic regression, the precision value is 90.32% which means the confirmation of heart disease is 90.32% correct. And the value of recall is 87.5% that means it correctly identifies 87.5% of all who have heart disease. In ANN, the precision value is 89.77%, which means Heart disease is 89.77 percent accurate. Similarly, its recall value is 91.78 %, which means it correctly identifies 91.78% of all who have heart disease. In Conv-Mlp, we get 86.81% precision and 90.69% recall value. In SVM and ANN, we get the highest precision and recall value.

F1 score is calculated from the test of precision and recall. The highest value of the f1 score is 1, indicating the perfect precision and recall result. On the other hand, the lowest value of the f1 score is 0, which shows one of the results of precision or recall is zero. But in Random Forest Classifier, we get an f1 score of 87.09%, which means precision and recall are commendable. We obtained the f1 score of 78.68% in the Decision Tree, which means that accuracy and recall don't get effective results. That's why it affects the outcome of the f1 score. In the Support Vector machine, we get the better value of the f1 score, which is 90.32%. It indicates that recall or precision can retrieve the relevant data correctly. In Logistic Regression, the f1 score is 88.88%, among those three Machine Learning models. On the other hand, in neural network conv-mlp scores 88.31%, and in ANN, the f1-score is 90.72% which is the highest f1-score among all algorithms. This result indicates better identification of relevant data from the dataset.

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
|---|---|---|---|---|
| Random Forest Classifier | 86.88% | 90% | 84.37% | 87.09% |
| Decision Tree | 78.68% | 82.75% | 75% | 78.68% |
| Support Vector Machine Classifier | 90.16% | 93.33% | 87.5% | 90.32% |
| Logistic Regression | 88.52% | 90.32% | 87.5% | 88.88% |
| Conv MLP | 87.20% | 86.81% | 90.69% | 88.31% |
| ANN | 91.24% | 89.77% | 91.78% | 90.72% |

Table 4.2: Result Analysis

Root Mean Square Error (RMSE) is the standard way to detect the model's error. RMSE is useful when significant errors occur, affecting the model's performance. The lower value of RMSE indicates the difference between predicted values and testing values is little. The higher value of RMSE suggests the difference between predicted values and testing values is umpteen.

In the below table the Root Mean Square Error has been shown.According to table 4.3, We get an 18.57% RMSE value in the Random Forest Classifier model means high error. In the Decision Tree model, we get the highest value of RMSE, which is 21.31% which indicates the highest error for the predicted value. In the Support Vector Machine, we obtain a value of RMSE, which is 9.83%. That means this model is suitable for this dataset and gives a low error of predicted value. In Logistic Regression, we achieve an RMSE value of 11.47%, which means this model will predict standard error according to parameter values. In Conv-Mlp we get a 12.78% error. In ANN, we get 8.74% RMSE which is the lowest value and gives the lowest error.

| **Model Name** | **Root Mean Square Error(RMSE)** |
|---|---|
| Random Forest Classifier | 18.57% |
| Decision Tree | 21.31% |
| Support Vector Machine Classifier | 9.83% |
| Logistic Regression | 11.47% |
| Conv MLP | 12.78% |
| ANN | 8.74% |

Table 4.3: Root Mean Square Error(RMSE)

Below figure 4.2 to 4.7 shows the ROC curve of six different models. According to our research, we got the best accuracy in ANN, and also the RMSE value is lowest in this model, which means this model will give the lowest error. From this information, we would suggest that the ANN model is best suited for our dataset to predict heart disease.

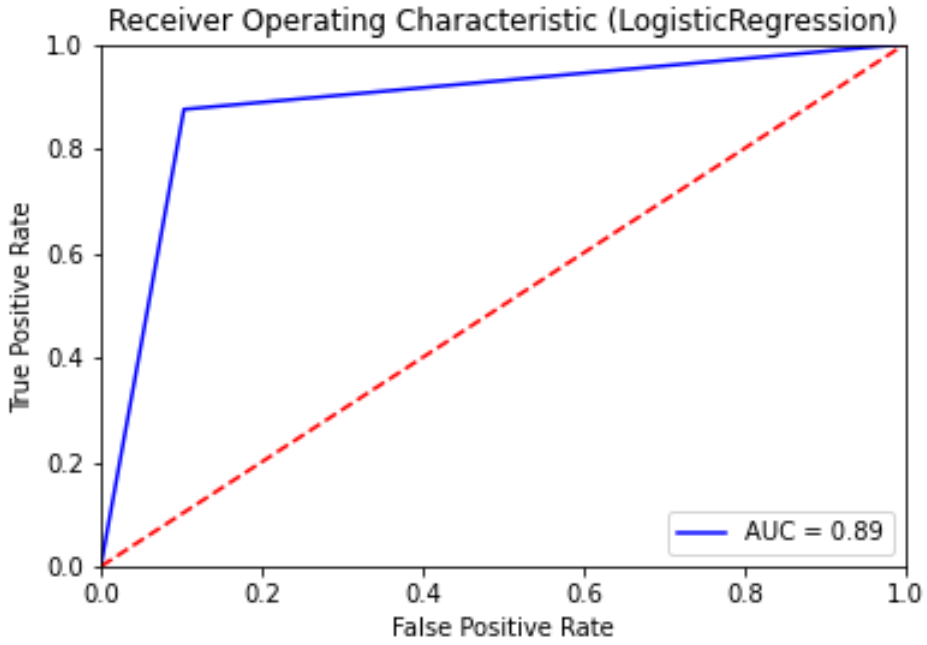



Figure 4.2: ROC curve for Logistic Regression

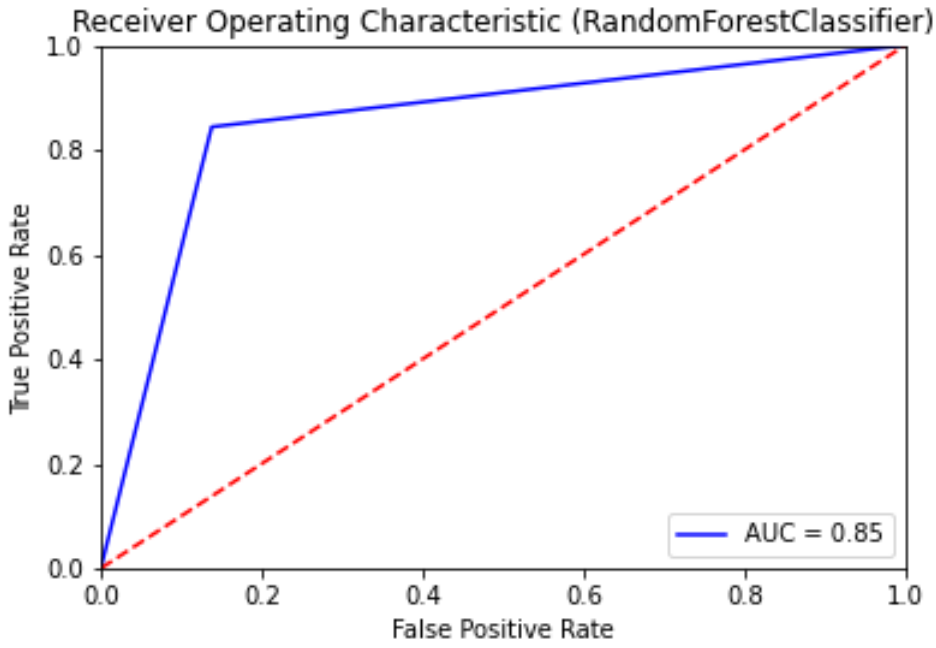



Figure 4.3: ROC curve for Random Forest Classifier

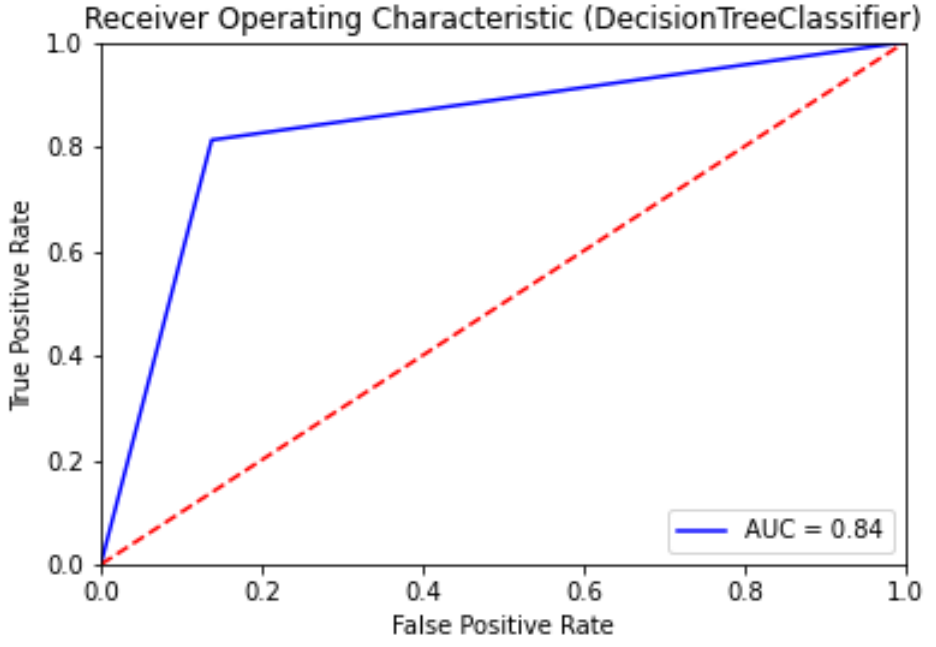



Figure 4.4: ROC curve for Decision Tree Classifier

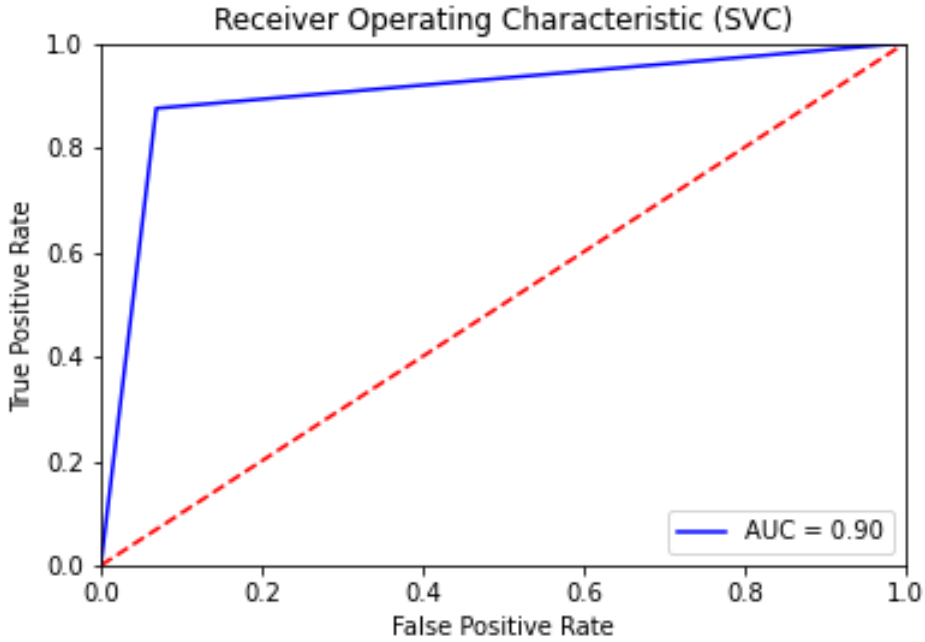



Figure 4.5: ROC curve for SVM

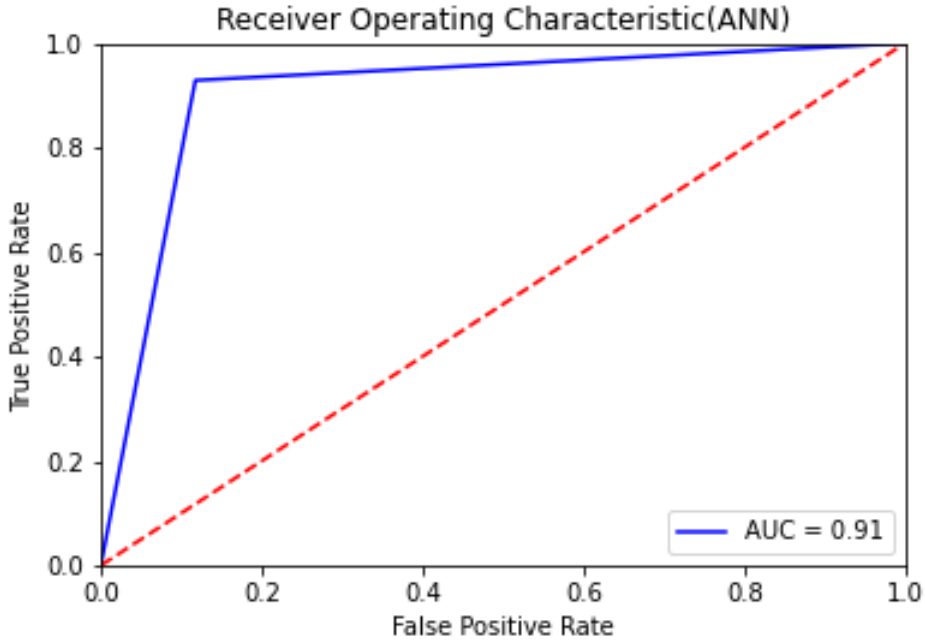



Figure 4.6: ROC curve for ANN

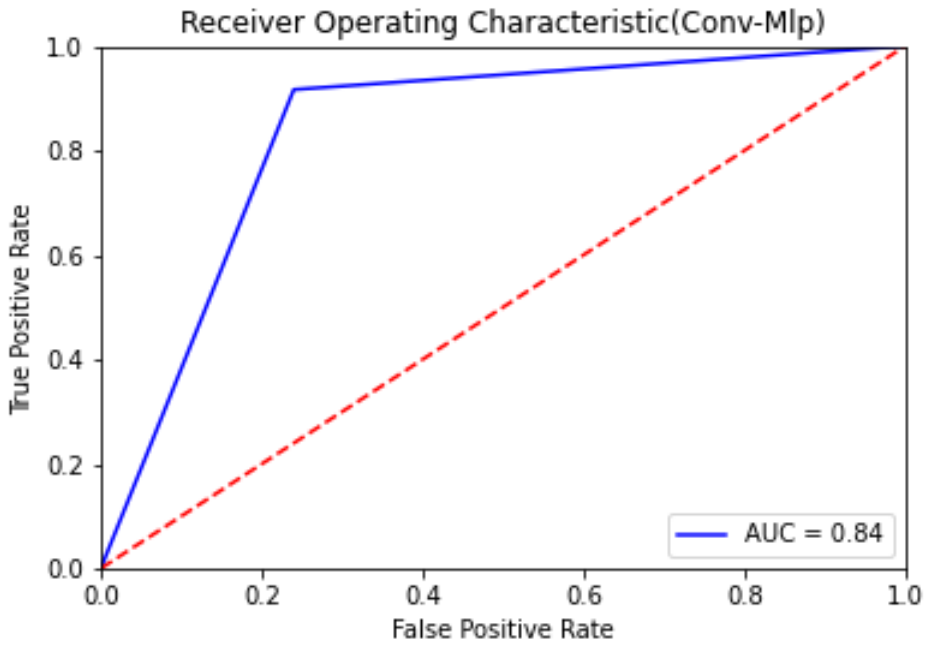



Figure 4.7: ROC curve for Conv-MLP

Here, below figure 4.8 shows the training accuracy and validation accuracy of ANN, and Figure 4.9 shows the training loss and validation loss of ANN.Figure 4.10 shows training accuracy and loss of Conv-MLP.

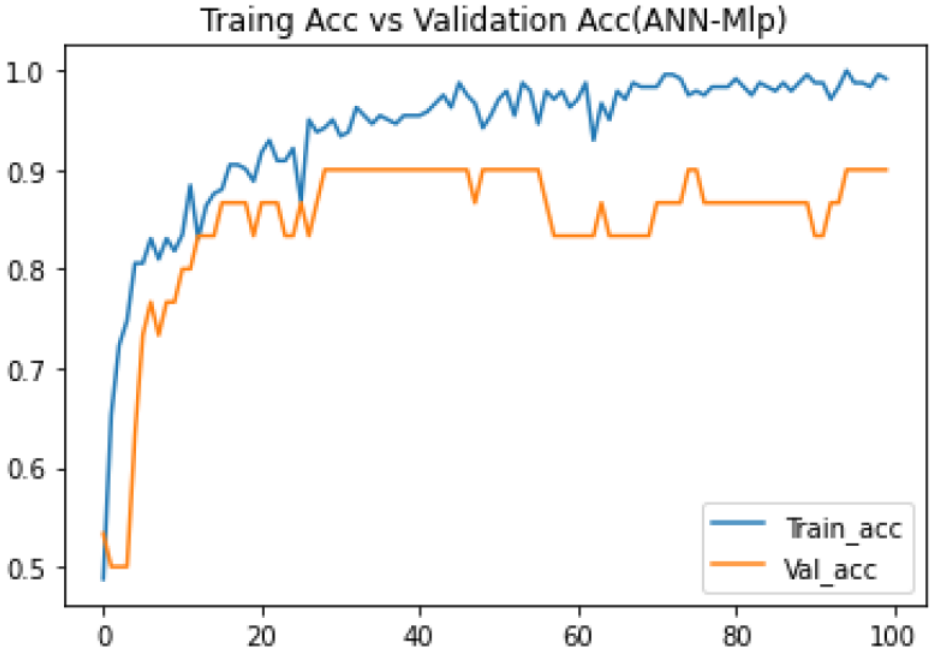



Figure 4.8: Training accuracy vs Validation accuracy

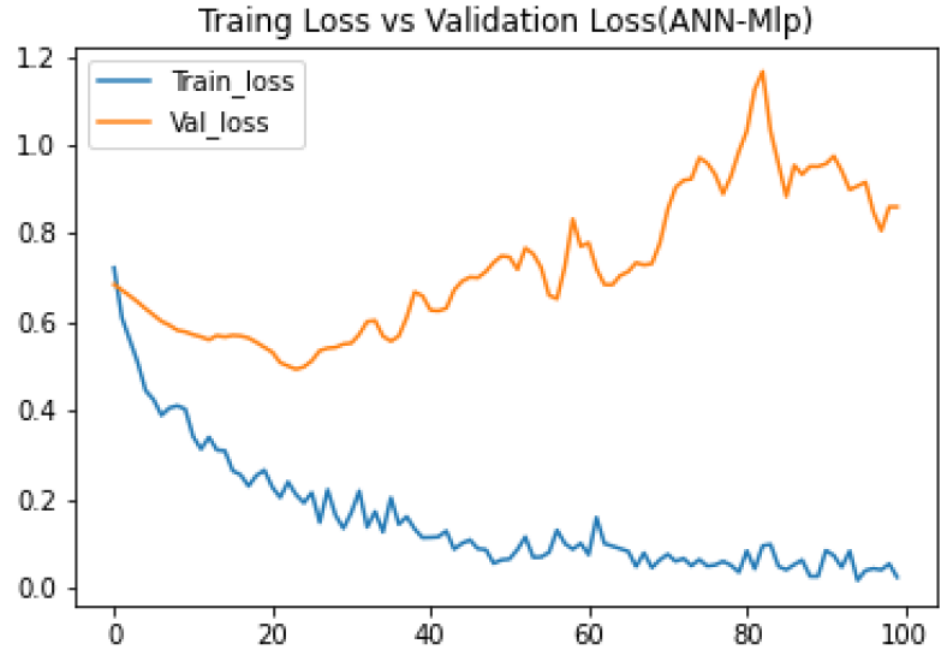



Figure 4.9: Training loss vs Validation loss

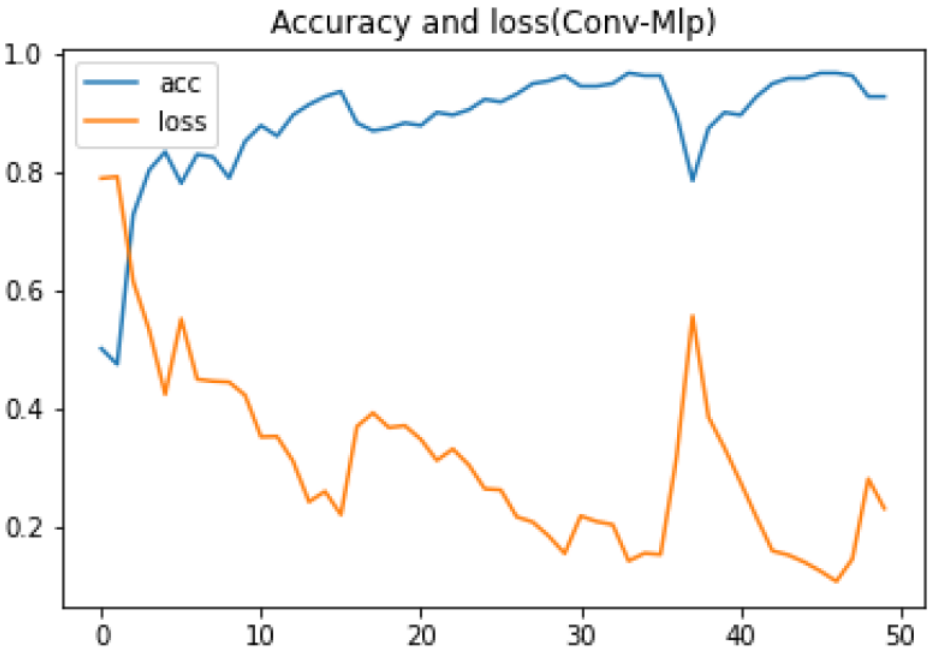



Figure 4.10: Accuracy and Loss of Conv-Mlp

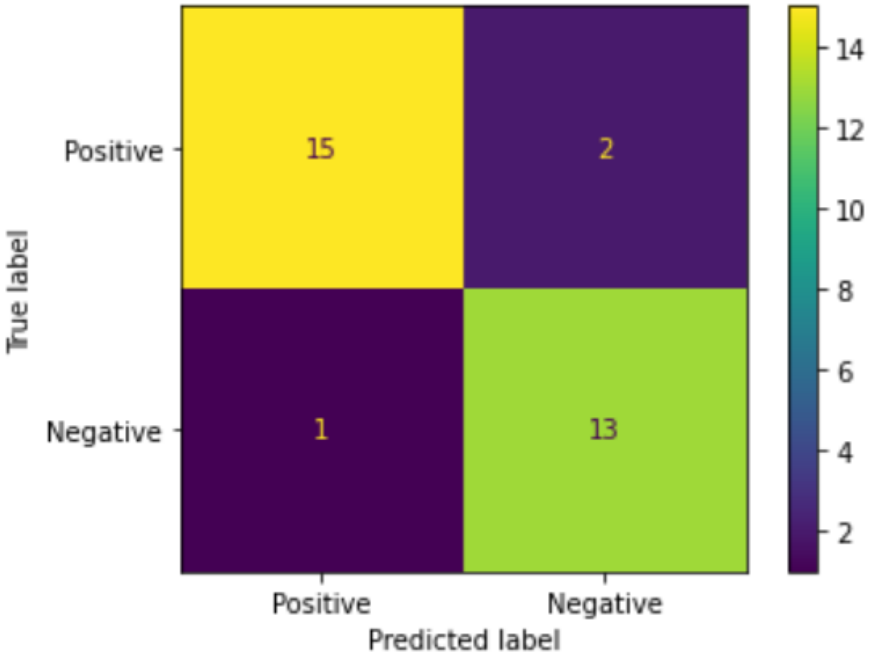



Figure 4.11: shows the confusion matrix of ANN

Here, figure 4.11 shows the confusion matrices of ANN. In table 4.4, we can see the comparison of different accuracy values from different papers.

| **Writer** | **Year** | **Approach** | **Accuracy** | **Accuracy of our work (ANN)** | **Accuracy of our work (Conv-MLP)** |
|---|---|---|---|---|---|
| S. Anitha and N. Sridevi | 2019 | (KNN+NB+SVM) | 86.6% | 91.24% | 87.20% |
| R. Prasad, P. Anjali, S. Adil, and N. Deepa | 2019 | LR | 86.89% | 91.24% | 87.20% |
| M..A. M. et al | 2014 | ANN | 87.04% | 91.24% | 87.20% |
| M. Gudadhe et al. | 2010 | SVM | 80.41% | 91.24% | 87.20% |
| A. K. et al. | 2011 | MLP | 80.99% | 91.24% | 87.20% |

Table 4.4: Comparison results with previous works

We compared the accuracy of our algorithms for detecting heart disease to earlier research. S. Anitha and N. Sridevi, 2019 achieved an accuracy of 86.6% using Naive Bayes. R. Prasad, P. Anjali, S. Adil, and N. Deepa, 2019 achieved an accuracy of 86.6% using Logistic Regression. M. A. M. Abusharian et al., 2014 achieved accuracy 87.04% using ANN. M. Gudadhe et al., 2010 achieved an accuracy of 80.4% using SVM. A. khemphilia et al., 2011 achieved an accuracy of 80.99% using MLP. In all cases, we can see our proposed algorithm gives the highest accuracy of 91.24% which is ANN and we also get very good performance in Conv-Mlp which is 87.20%.It is evident that our proposed approach ANN has the best accuracy of the previous works.

# Chapter 5

# Conclusion and Future Works

## 5.1 Conclusion

Our overall aim is to define various helpful machine learning algorithms to predict various cardiovascular diseases. This is a common health issue in Bangladesh and in the whole world. Because of the changed food habits of people in today's world, people are more likely to get heart disease. Also, smoking, indisciplined lifestyle, addiction to alcohol, caffeine, and mental stress are chronic risk factors for heart disease. There is still time for improvements, the medical sector is growing faster, and machine learning is most efficient. Heart diseases are increasing rapidly day by day. This is our matter of concern to predict any such diseases. In this research, Using the patient's medical history, we made a system for predicting whether the patient is likely to be diagnosed with heart disease or not. We are using different algorithms like Random Forest, Decision Tree, Support Vector Machine, Logistic Regression, ConvMLP, Artificial Neural Network(ANN) to predict and classify patients with heart diseases. This prediction system can increase medical care and also reduce costs. To conclude, there are a few statements we can give. The system that we discussed in this paper can be helpful in many different ways. For example, if any research institute wishes, they can use our system to find prediction-based results for future references. Our system will provide them with a reasonable accuracy rate. Also, medical institutions can use this system as one other way to predict their patients' availability of any heart diseases. Indeed, this will increase their efficiency. Our main motive stands to help these sorts of organizations, let it be medical or research.

## 5.2 Future Work

From our above analysis, we achieved great accuracy from our models, and our future goal is to extend the research and use it for the betterment of medical science. We would like to implement Basic-Mlp, Axial-Mlp, and Deep Residual Network for better accuracy in the future. Moreover, this type of research can be used for other disease detection. We also want to work with more datasets in the future, as for better performance, we need more datasets.

# Bibliography


[1] *Cardiovascular diseases (cvds)*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] A. Must, J. Spadano, E. H. Coakley, A. E. Field, G. Colditz, and W. H. Dietz, "The disease burden associated with overweight and obesity," *Jama*, vol. 282, no. 16, pp. 1523–1529, 1999, Accessed: 2022-1-05.

[3] P. A. Hebbar, M. M. Kumar, and A. Mathur, "Theory, concepts, and applications of artificial neural networks," *Applied Soft Computing: Techniques and Applications*, pp. 153–176, 2022, Accessed: 2022-1-03.

[4] Ronit, *Heart disease uci*, Accessed: 2021-9-28, Jun. 2018. [Online]. Available: https://www.kaggle.com/ronitf/heart-disease-uci.

[5] V. Ramalingam, A. Dandapath, and M. K. Raja, "Heart disease prediction using machine learning techniques: A survey," *International Journal of Engineering & Technology*, vol. 7, no. 2.8, pp. 684–687, 2018, Accessed: 2022-1-03.

[6] S. Anitha and N. Sridevi, "Heart disease prediction using data mining techniques," *Journal of analysis and Computation*, 2019, Accessed: 2022-1-08.

[7] R. Prasad, P. Anjali, S. Adil, and N. Deepa, "Heart disease prediction using logistic regression algorithm using machine learning," *IJEAT) ISSN*, pp. 2249–8958, 2019, Accessed: 2021-10-18.

[8] A. Hazra, S. K. Mandal, A. Gupta, A. Mukherjee, and A. Mukherjee, "Heart disease diagnosis and prediction using machine learning and data mining techniques: A review," *Advances in Computational Sciences and Technology*, vol. 10, no. 7, pp. 2137–2159, 2017, Accessed: 2022-1-03.

[9] J. K. Kim and S. Kang, "Neural network-based coronary heart disease risk prediction using feature correlation analysis," *Journal of healthcare engineering*, vol. 2017, 2017, Accessed: 2022-1-13.

[10] N. Rajesh, T. Maneesha, S. Hafeez, and H. Krishna, "Prediction of heart disease using machine learning algorithms," *International Journal of Engineering and Technology (UAE)*, vol. 7, no. 2.32, pp. 363–366, 2018, Accessed: 2021-9-19.

[11] M. Shouman, T. Turner, and R. Stocker, "Applying k-nearest neighbour in diagnosing heart disease patients," *International Journal of Information and Education Technology*, vol. 2, no. 3, pp. 220–223, 2012, Accessed: 2021-11-24.

[12] Y. Khourdifi and M. Bahaj, "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization," *International Journal of Intelligent Engineering and Systems*, vol. 12, no. 1, pp. 242–252, 2019, Accessed: 2021-12-22.

[13] H.-Y. Kim, "Analysis of variance (anova) comparing means of more than two groups," *Restorative dentistry & endodontics*, vol. 39, no. 1, pp. 74–77, 2014, Accessed: 2021-12-25.

[14] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked*, vol. 16, p. 100 203, 2019, Accessed: 2022-1-02.

[15] M. A. Abushariah, A. A. Alqudah, O. Y. Adwan, R. M. Yousef, *et al.*, "Automatic heart disease diagnosis system based on artificial neural network (ann) and adaptive neuro-fuzzy inference systems (anfis) approaches," *Journal of software engineering and applications*, vol. 7, no. 12, p. 1055, 2014, Accessed: 2021-09-17.

[16] M. Gudadhe, K. Wankhade, and S. Dongre, "Decision support system for heart disease based on support vector machine and artificial neural network," in *2010 International Conference on Computer and Communication Technology (ICCCT)*, Accessed: 2021-10-12, IEEE, 2010, pp. 741–745.

[17] A. Khemphila and V. Boonjing, "Heart disease classification using neural network and feature selection," in *2011 21st International Conference on Systems Engineering*, Accessed: 2021-11-02, IEEE, 2011, pp. 406–409.

[18] A. T. Nishadi, "Predicting heart diseases in logistic regression of machine learning algorithms by python jupyterlab," *International Journal of Advanced Research and Publications*, vol. 3, pp. 69–74, 2019, Accessed: 2022-1-09.

[19] N. K. Nissa, *How to predict coronary heart disease risk using logistic regression?* Apr. 2021. [Online]. Available: https://medium.com/analytics-vidhya/how-to-predict-coronary-heart-disea%20se-risk-using-logistic-regression-c069ab95cbec.

[20] Y. Liu, Z. Xu, and C. Li, "Online semi-supervised support vector machine," *Information Sciences*, vol. 439-440, pp. 125–141, 2018, Accessed: 2022-1-04. DOI: 10.1016/j.ins.2018.01.048.

[21] *Understanding random forest classifiers in python.* [Online]. Available: https://www.datacamp.com/community/tutorials/random-forests-classifier-python.

[22] S. Radhika, S. R. Shree, V. R. Divyadharsini, and A. Ranjitha, "Symptoms based disease prediction using decision tree and electronic health record analysis," *European Journal of Molecular & Clinical Medicine*, vol. 7, no. 4, pp. 2060–2066, 2020, Accessed: 2021-12-26.

[23] Uniqtech, *Multilayer perceptron (mlp) vs convolutional neural network in deep learning*, Jun. 2019. [Online]. Available: https://medium.com/data-science-bootcamp/multilayer-perceptron-mlp-vs-convolutional-neural-network-in-deep-learning-c890f487a8f1.

[24] J. Brownlee, "When to use mlp, cnn, and rnn neural networks (2018)," *URL: https://machinelearningmastery. com/when-to-use-mlp-cnn-and-rnn-neural-networks*, 2018, Accessed: 2021-9-16.

[25] J. Li, A. Hassani, S. Walton, and H. Shi, "Convmlp: Hierarchical convolutional mlps for vision," *arXiv preprint arXiv:2109.04454*, 2021, Accessed: 2021-10-20.

[26] M. H. Hassoun *et al.*, *Fundamentals of artificial neural networks*. MIT press, 1995.