

# “Heart Disease Prediction using Stacked Generalization Ensemble Technique”

A THESIS REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE AWARD OF THE DEGREE  
OF

MASTER OF TECHNOLOGY  
IN  
SOFTWARE ENGINEERING

Submitted by

**ACHINT SINGH (2K19/SWE/20)**

Under the supervision of

**DR. DIVYASHIKHA SETHIA**



**Department of Software Engineering**  
**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Bawana Road, Delhi 110042

**JULY, 2021**

**DEPARTMENT OF SOFTWARE ENGINEERING**  
**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Bawana Road, Delhi-110042

**CANDIDATE'S DECLARATION**

I hereby declare that the work presented in this report entitled “Heart Disease Prediction using Stacked Generalization Ensemble Technique”, in fulfillment of the requirement for the award of the MASTER OF TECHNOLOGY degree in Software Engineering submitted in Software Engineering Department at DELHI TECHNOLOGICAL UNIVERSITY, New Delhi, is an authentic record of my own work carried out during my degree under the guidance of Dr. Divyashikha Sethia.

The work reported in this has not been submitted by me for the award of any other degree or diploma.



Date: 22/10/2021

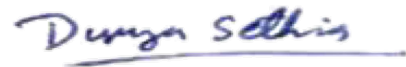
Place: Delhi

**Achint Singh(2K19/SWE/20)**

**DEPARTMENT OF SOFTWARE ENGINEERING**  
**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Bawana Road, Delhi-110042

**CERTIFICATE**

This is to certify that Achint Singh (2K19/SWE/20) has completed the research titled “Heart Disease Prediction using Stacked Generalization Ensemble Technique” under my supervision in fulfillment of the MASTER OF TECHNOLOGY degree in Software Engineering at DELHI TECHNOLOGICAL UNIVERSITY.



Place: Delhi

Dr. Divyashikha Sethia

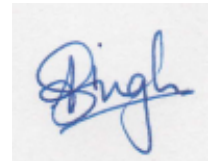
Date: 22.10.2021

**SUPERVISOR**

**DEPARTMENT OF MECHANICAL ENGINEERING**  
**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Bawana Road, Delhi-110042

**ACKNOWLEDGEMENT**

I am very thankful to Dr. Divyashikha Sethia (Assistant Professor, Department of Software Engineering) and all the faculty members of the Department of Software Engineering at DTU. They all provided me with immense support and guidance for the project. I would also like to express my gratitude to the University for providing me with the laboratories, infrastructure, testing facilities, and environment which allowed us to work without any obstructions. I would also like to appreciate the support provided to us by our lab assistants and our peer group who aided us with all the knowledge they had regarding various topics.



ACHINT SINGH

Place: Delhi

Date: 22.10.2021

2K19/SWE/20

## Abstract

The Myocardium, also known as the Heart, is responsible for pumping oxygenated blood to and deoxygenated blood from other human body parts. All the other organs in the human body are dependent on the coherent working of this organ. Cardiovascular diseases are some of the deadliest diseases, which have caused millions of deaths worldwide. Early detection of heart diseases is an ongoing and crucial problem in medical science, and various researchers are attempting to improve physician's ability by developing an intelligent medical decision support system. Through this research work, we submit a systematic ensemble-based approach for heart disease prediction. It uses a hybrid combination of Support Vector Machine (SVM) algorithm, Artificial Neural Network (ANN), and Extreme Gradient Boost (XGB) algorithms, which are then combined using the Stacking ensemble method. This system has provided an accuracy of 92% for the prediction of heart disease.

# Contents

Candidate’s Declaration	i
Certificate	ii
Acknowledgement	iii
Abstract	iv
Content	vi
List of Tables	vii
List of Figures	viii
List of Symbols, Abbreviations	ix
<b>1 INTRODUCTION</b>	<b>x</b>
1.1 General	x
1.2 Problem Statement	xi
1.3 Research Gaps	xi
1.4 Objectives of the Project	xi
1.5 Proposed Solution	1
<b>2 BACKGROUND</b>	<b>2</b>
2.1 Algorithms Used	2
2.1.1 Artificial Neural Network	2
2.1.2 Support Vector Machine	3
2.1.3 XGBoost	4
2.1.4 Ensemble Learning	4
2.2 Performance Metrics	5
2.2.1 Accuracy	6
2.2.2 Sensitivity/Recall	6
2.2.3 Precision	7
2.2.4 Specificity	7
2.2.5 F1 Score	7
<b>3 RELATED WORKS</b>	<b>8</b>
<b>4 PROPOSED METHODOLOGY</b>	<b>10</b>
4.1 Data acquisition	10
4.2 Data Pre-Processing	11

4.3	Base models . . . . .	12
4.3.1	Artificial Neural Networks . . . . .	12
4.3.2	Support Vector Machine (SVM) . . . . .	12
4.3.3	Extreme Gradient Boost . . . . .	13
4.4	Stacking Ensemble Model . . . . .	13
<b>5</b>	<b>RESULTS and DISCUSSION</b>	<b>15</b>
<b>6</b>	<b>CONCLUSION</b>	<b>22</b>

## List of Tables

3.1	Comparison of previous research works . . . . .	9
4.1	Attributes Used In This Paper . . . . .	11
5.1	Results of proposed model . . . . .	20
5.2	Comparison of our results with previous research works . . . . .	20



## List of Figures

2.1	Basic structure of Multi-Layer Perceptron . . . . .	3
2.2	Example of SVM algorithm . . . . .	4
2.3	Architecture of Stacked Generalization Technique . . . . .	5
2.4	Structure of Confusion Matrix . . . . .	6
4.1	Proposed Model . . . . .	10
4.2	Code Snippet for Artificial Neural Network . . . . .	12
4.3	Code Snippet for Support Vector Machine . . . . .	13
4.4	Code Snippet for Extreme Gradient Boost Algorithm . . . . .	13
4.5	Code Snippet for Stacking Ensemble Model . . . . .	14
5.1	Extreme Gradient Boost Confusion Matrix Heatmap . . . . .	16
5.2	Multi Layer Perceptron Confusion Matrix Heatmap . . . . .	17
5.3	Support Vector Machine Confusion Matrix Heatmap . . . . .	18
5.4	Stacking Ensemble classifier Confusion Matrix Heatmap . . . . .	19
5.5	Comparison of results . . . . .	21

## List of Symbols

ML - Machine Learning  
XGB - Extreme Gradient Boost  
SVM - Support Vector Machine  
BP - Back Propagation  
ANN - Artificial Neural Network  
MLP - Multi-Layer Perceptron

# Chapter 1

## INTRODUCTION

### 1.1 General

The Heart is an essential organ of the human body. It pumps to circulate the blood to and from the body by pushing it from the Heart's left chambers through arteries of decreasing length and finally reaching the capillaries throughout the body. Proper working of the Heart is essential for almost all the work a person has to do. Any disease of the Heart can cause a tremendous amount of discomfort and even prove deadly. According to the past surveys conducted by the World Health Organization (WHO), around 17 million deaths transpire worldwide yearly due to cardiovascular diseases. Heart disease consists of a range of cardiovascular problems. Many diseases and ailments are possible symptoms of heart disease. Types of cardiac diseases include:

- *Atherosclerosis*: Condition in which arteries get hardened. item *Cardiomyopathy*: Heart's muscles harden or grow weak in this condition.
- *Arrhythmia*: It is a heart rhythm abnormality.
- *Congenital heart defects*: These heart irregularities are present at the time of birth.
- *Coronary artery disease (CAD)*: The effect of plaque buildup in the Heart's arteries. It is sometimes called ischemic heart disease.
- *Heart infections*: Causes maybe some bacteria, viruses, or parasites.

The term cardiovascular disease refers to heart-based ailments that are specific to the blood vessels. Cardiovascular Disease (CVD) or Heart Attack is a common complication observed among all humans. Various factors, like smoking, blood pressure is too high, consumption of alcohol, high cholesterol level, obesity, diabetes, and others, are responsible factors increasing the risk of heart problems. However, recent studies suggest that with the introduction of artificial intelligence and its applications in medical sciences, we can help prevent any such disease. Several factors are involved in analyzing and diagnosing these heart diseases, for which physicians generally make decisions based on evaluations of the newest test results. The preliminary conclusions made other patients with similar conditions get examined by the doctors. These intricate procedures are complicated when considering the number of factors that the physician has to evaluate. Diagnosing heart diseases involves a skilled team of physicians. In order to help in this diagnosis, various researchers are applying artificial intelligence and machine learning approaches to predict and detect heart diseases.

## 1.2 Problem Statement

Heart disease is the a significant problem for medical practitioners since it has a high mortality rate. Early detection is essential for adequate treatment. The growth of technology and computers becoming a norm in all hospitals and clinics, which produces a massive amount of medical data every year. Analysis of such data can give us insights into diseases and their causes. This data can also help in prediction of disease by applying artificial intelligence techniques. Data Mining and Machine learning are emerging fields of immense significance for providing prognosis and a deeper comprehension of the medical data. Prediction by utilizing data mining systems provides precise results related to heart diseases. The prediction can solve complex queries related to detecting heart diseases and assisting medical practitioners in making accurate and feasible clinical decisions.

## 1.3 Research Gaps

Various researchers are applying artificial intelligence techniques to predict and detect heart diseases. There have been numerous papers in the past years that study this. For example, Radhimeenakshi et al. [18] have applied SVM and ANN as data mining techniques. It provided an accuracy of 78% and 80%, respectively. This model also implements these algorithms but with better results. The results of same algorithms are improved by using the min-max technique for normalization of the dataset and changing the parameters of the algorithms. Das et al. [2] was the first to create an ensemble of neural networks using SAS-based software, which provided an improved accuracy over single algorithm models, having an accuracy of 89%. Their approach used the SAS software, an enterprise-level software that is not easily accessible. This research work has analyzed many such approaches and found some gaps which can be improved which are listed below.

- Most of the previous works have applied one or two machine learning algorithms to this dataset. Those have an accuracy of 70% or 80% but no higher.
- Lack of techniques that combine accuracies of single ML models such as ensemble techniques.
- Previous works lack open-source techniques for implementing AI models for Heart diseases prediction.
- Data cleaning and pre-processing can be improved to get better results even when using basic algorithms.

## 1.4 Objectives of the Project

This project aims to complete the following objectives:

- Obtaining and analyzing dataset for training and validating the learning models.
- Researching and validating the performance of previous work done on this topic.
- Comparing the performance of different machine learning models available to predict heart disease.

- Develop a new model that combines the capabilities of top performers by using ensemble techniques.
- Comparing the performance of the developed model with other models.

## 1.5 Proposed Solution

This work aims to create an efficient system for heart disease prediction, overcoming the shortcomings of previous works done on this topic. To achieve this we have introduced the Stacking Ensemble method. The stacking ensemble-based model in this research combines the three base models ANN, SVM, and XGboost, and has an absolute accuracy of 92% for heart disease prediction, which is higher than the previous works. The results are generated from the designed systems and compared with several other kinds of research done on the topic over the years, and it offers an improvement. It can take input details of a new patient and tell their likelihood of having heart disease efficiently.

## Chapter 2

### BACKGROUND

#### 2.1 Algorithms Used

This section discusses the algorithms in the proposed approach.

##### 2.1.1 Artificial Neural Network

In Artificial intelligence, an Artificial Neural Network is a method that attempts to mimic the network of neurons that create the human Brain so that computers can learn, make decisions, and behave like a human. Programmers design the artificial neural network to act like brain cells and mimic their properties and behaviors such as learning, observation, and prediction. The Brain comprises billions of nerve cells called 'neurons' used to send messages across the whole body. ANNs are composed of nodes, which mimic the working of a biological neuron in the Brain. These nodes connect in different layers by links. These links aid in communication between these nodes. The nodes take input data and perform operations on it. They transmit these results to other neurons, which can then perform other functions on them. The output is also known as 'node value' or 'activation value' at each node. Every link between 2 nodes A and B has a weight value associated with it. This weight value represents the strength of the connection between them. If the weight from Node A to node B has a higher value, it implies that node A has a more significant effect on node B. The neural network learns by altering the value of these weights and updating them after each iteration.

The two types of artificial neural networks are

- Feedforward Neural Network : feedforward neural network is an artificial neural network in which connections between the nodes do not form a cycle, i.e. the connection is one way and no information is relayed back through the network.
- Feedback neural networks : feedback loops are permitted in such neural networks, i.e. the signals are fed back to the node from which they originated.

A Multi layer perceptron (MLP) is a feedforward ANN architecture for predicting and classifying new observations. The basic structure of an MLP has an input layer, hidden layers (1 or more), and an output layer. Each node, except the input layer, uses a non-linear activation function and acts as a neuron. The activation function matches the weighted inputs to the output of each node. Figure 3.1 shows the basic structure of an MLP with one input layer, one hidden layer, and one output layer. The parameters and values used in this project for each algorithm are explained in Chapter 4.

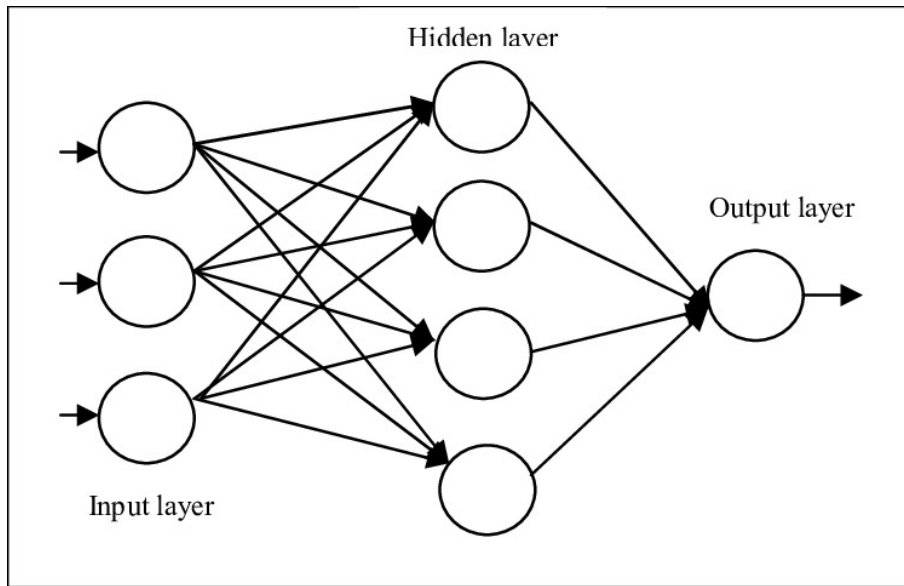


Figure 2.1: Basic structure of Multi-Layer Perceptron

### 2.1.2 Support Vector Machine

SVM is a supervised machine learning algorithm that utilizes classification algorithms for double-group classification problems. As per [7], SVM is an algorithm that employs a non-linear mapping to modify the initial training data into a higher dimension. The newly created dimension looks for the optimal linear hyperplane based on separation, which is a "decision boundary" dividing the tuples among classes. Using proper non-linear mapping to a sufficiently high dimension, we can separate data from two classes using a hyperplane. Using support vectors and margins, SVM searches for the hyperplane. Vapnik [9] first introduced the model for SVM which is a learning method that bargains between accuracy and generalization error. SVMs construct a hyperplane that splits classes on either side of the hyperplane. The Figure 3.2 below shows the working of SVM algorithm on a classification problem. Consider a set of points represented as 'X' and 'O'. The hyperplane classifies these into two classes. It correctly classified 11 out of 12 X's and 8 out of 10 O's. It has a separation percent of 86.63%.

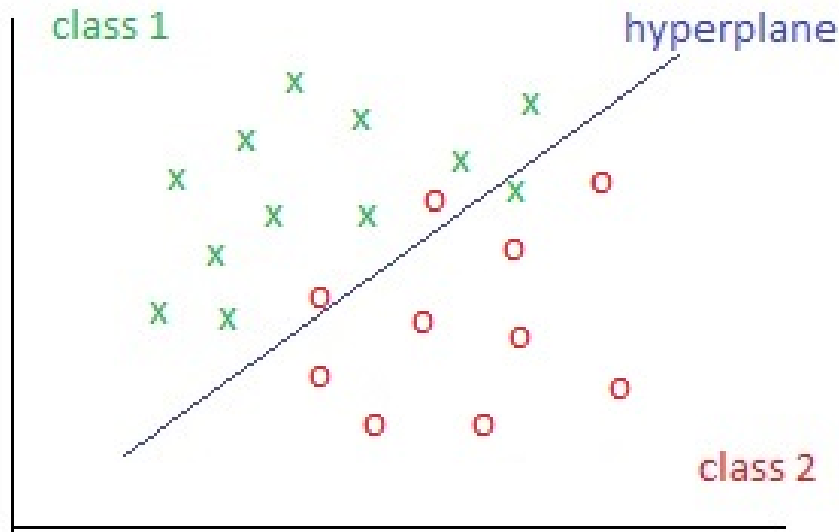


Figure 2.2: Example of SVM algorithm

### 2.1.3 XGBoost

Gradient boosting generates a more reliable model for prediction by creating an ensemble of weak models, typically decision trees. It is a machine learning technique for solving regression or classification problems. Extreme Gradient Boost (XGBoost) is an implementation of the Gradient Boosting concept, but it is made unique by "a more regularized model formalization to control over-fitting, which gives it better performance," as stated by the inventor of the algorithm, Tianqi Chen [22]. Over fitting is less in this algorithm. In this algorithm, decision trees get generated sequentially. Weights play a crucial part in this algorithm. Each independent variable is assigned a weight, and then the weight is fed into the decision tree, predicting results. Then we increase the weight of wrongly predicted variables, which the second decision tree uses as an input. These singular predictors are then combined together to produce a highly accurate model. We can use it to solve regression, classification, and other prediction problems as per the user's need. In this research, we have used this model twice, once in the first part as a base algorithm, and then as the meta-algorithm of the stacking model.

### 2.1.4 Ensemble Learning

Ensemble learning is a machine learning method in which the output of two or more models (named "weak learners" or "base models") are combined to get better results for solving a problem. We do this since correctly combining base models can give much better results. There are three significant kinds of ensemble techniques:

- Bagging: initially, it considers homogeneous base models and learns them separately from each other in parallel. Then it merges them based on some deterministic process.



- **Boosting:** It takes base models, learns them sequentially so that a base model is dependent on the former, and combines them using a deterministic procedure.
- **Stacking:** It considers heterogeneous base models, learns them in parallel, and merges them by training a meta-model to produce predictions based on the different base models. Stacking is helpful since it can utilize the capabilities of various classification or regression models and merge them to have better performance than any single one

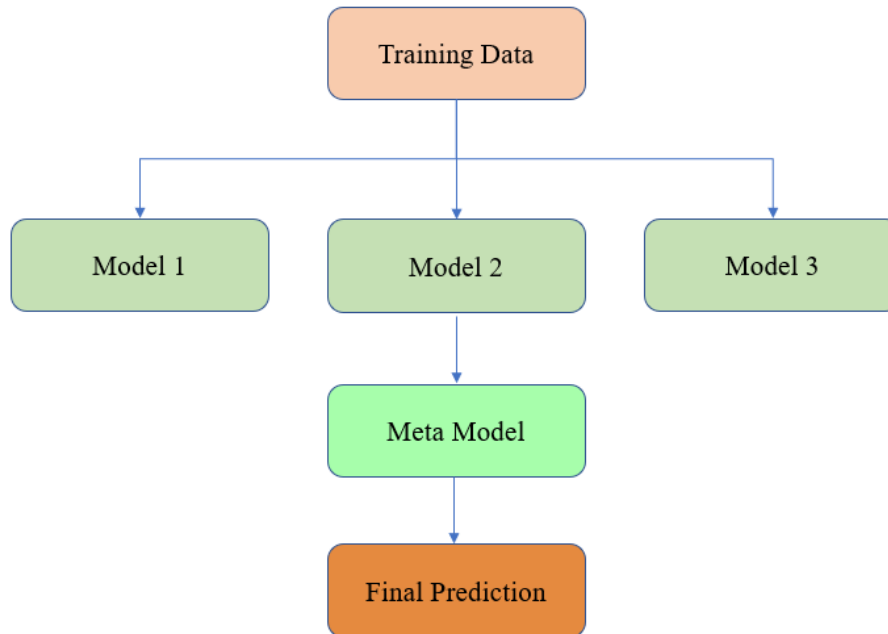


Figure 2.3: Architecture of Stacked Generalization Technique

Figure 3.3 shows the architecture of the stacking ensemble model, which has two levels:

- **Level 0 or Base Models:** These models are directly applied to the training data and give the first part of the results. The results from these models, in the next phase of this technique, are combined.
- **Level-1 Model or Meta-Model:** it learns how to combine the predictions of the base models in the most efficient way. After combining the algorithms, it gives us the final result.

## 2.2 Performance Metrics

To measure the performance of our algorithms, we use a Confusion Matrix, which is a 2x2 matrix with the following values.

- *True Negative (TN):* Correct negative prediction, i.e., predicted no disease and disease does not exist in the person
- *True Positive (TP):* Correct positive prediction, i.e., predicted heart disease present in the patient, and it exists in the patient

- *False-negative (FN)*: Incorrect negative prediction, i.e., predicted no disease when actually heart disease present. It is also known as Type 2 error.
- *False-positive (FP)*: predicted heart disease present but does not exist in the patient. It is also known as Type 1 error

Figure 3.4 shows the basic structure of a Confusion Matrix, with the performance metrics that it can show for an algorithm.

		Predicted Value		
		Positive	Negative	
Actual Value	Positive	True Positive (TP)	False Negative (FN) Type II Error	<b>Sensitivity</b> $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	<b>Specificity</b> $\frac{TN}{(TN + FP)}$
		<b>Precision</b> $\frac{TP}{(TP + FP)}$	<b>Negative Predictive Value</b> $\frac{TN}{(TN + FN)}$	<b>Accuracy</b> $\frac{TP + TN}{(TP + TN + FP + FN)}$

Figure 2.4: Structure of Confusion Matrix

### 2.2.1 Accuracy

It is the ratio of the sum of correct predictions divided by the total number of predictions. It measures the correctness of the algorithm.

Accuracy is calculated using the formula:

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

### 2.2.2 Sensitivity/Recall

Recall or sensitivity is the ratio of correctly classified positive values to the total number of actual positive instances.

The sensitivity is calculated using the following formula:

$$Recall/Sensitivity = \frac{TP}{TP + FN}$$

Recall signifies the positive predictions that are classified incorrectly.

### 2.2.3 Precision

Precision is the ratio of correct positive outcomes to the total positive outcomes for a class.

$$Precision = \frac{TP}{TP + FP}$$

Precision signifies how many positive outcomes are actually correct for a class.

### 2.2.4 Specificity

Specificity is the ratio of the correctly predicted negative values (TN) and the total pessimistic predictions.

It is calculated by the formula

$$Specificity = \frac{TN}{TN + FP}$$

### 2.2.5 F1 Score

F1 score is the harmonic mean of the precision and recall of a model. It is a measure of the accuracy of a model. A higher F1 score implies that the model has fewer incorrect predictions.

The formula for the F1 score is as follows.

$$F1\ Score = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

## Chapter 3

### RELATED WORKS

Das [2] employed SAS software for diagnosing heart disease. By placing a neural network-based ensemble system at the system's center. The ensemble models combine the base model results, thus creating better results. This model reported an accuracy of 89.01%. The SAS-based software has two programs called SAS enterprise guide and miner. These then create an ensemble model for heart disease detection. The limitation being it requires enterprise-level software to give results. The proposed approach suffers no such disadvantage and can be applied easily using any programming environment or a GUI-based user-friendly software.

Olaniyi et al. [3] proposed a feedforward multilayer perceptron and support vector machine system. They reported accuracy of 85% for multilayer perceptron and 87.5% for support vector machine. We have also used these same algorithms in this research and have obtained better results.

Gudadheet al. [4] created a cardiovascular disease classification based on Artificial Neural Network and support vector machine. Results obtained show that an MLP neural network had an accuracy of 80.41%. In this research, we have used both these algorithms and have obtained better results due to better data pre-processing and tweaking the parameters of the algorithms.

Detrano et al. [9] obtained an accurate classification accuracy of almost 77% for predicting heart disease using logistic regression. This was one of the first research conducted using the UCI heart disease dataset.

Mohan et al. [5] proposed a system with a high-performance level, having an accuracy of 88.7% by using the prediction model hybrid random forest with a linear model (HRFLM). In this approach, HRFML had combined the features of Random Forest and linear method.

Vikas et al. [12] proposed three data mining algorithms, CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3), and decision table (DT) extracted from a decision tree to develop their model on a large dataset. They have shown the result of 10-fold cross-validation methods to measure the unbiased estimate.

Kannan et al. [15] compared the accuracy of 4 different machine learning algorithms with receiver operating characteristic (ROC) curve for predicting heart disease by the 14 attributes from UCI Datasets. Their approach is limited since it only compares the basic ML algorithms and does no augmentation.

Akansh et al. [17] compared k nearest neighbors, SVM, and Naive Bayes algorithms for predicting heart disease. In their work, the highest accuracy was by using Naive Bayes at 88%. This work compared three models on the same dataset.

Table 3.1 show the algorithms and accuracy of these works in tabular form.

Research Work	Algorithm Applied	Accuracy
Adnan et al. [1]	Feed forward MLP	85%
Gudadhe et al. [4]	MLP with BP	80%
Mohan et al.[5]	Decision Tree	87%
Detrano et al.[9]	Logistic Regression	77%
Das et al. [3]	Neural network ensemble	89.01%
Akansh et al.[17]	Naive Bayes	88%
Radhimeenkshi et al.[18]	SVM and ANN	78% and 80%

Table 3.1: Comparison of previous research works

In this research, these works are analyzed and used to improve the working and results. Initially, they are used to understand the dataset. Some of the research gaps found in these approaches enabled us to improve upon them. Some of these used the same algorithms applied in this research work but with less accurate results. Chapter 4 discusses the techniques used in this research work in detail. Table 3 in Chapter 5 shows the comparison of results with related research.

## Chapter 4

### PROPOSED METHODOLOGY

This project proposes an ensemble-based approach for the prediction of heart disease. Figure 4.1 shows the flow chart of the model used in this research. This section explains these steps in detail.

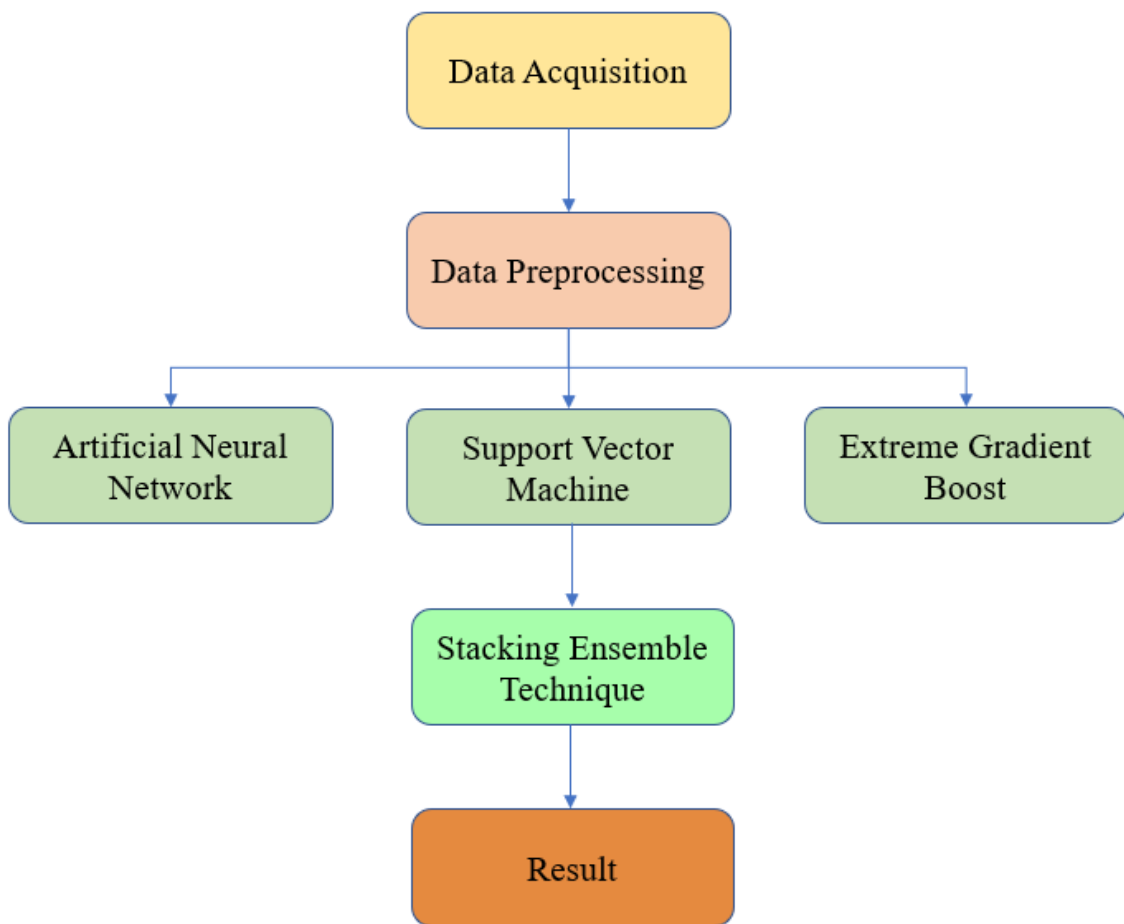


Figure 4.1: Proposed Model

#### 4.1 Data acquisition

This research uses the heart disease database from the UCI machine learning repository [21]. This dataset was created by

- Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
- University Hospital Zurich Switzerland: William Stein-Brunn, M.D.
- University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
- V.A. Medical Center, Long Beach, and Cleveland Clinic Foundation: Robert De-trano, M.D., Ph.D.

The database contains 303 samples of patients, some of whom have heart disease, and some do not. The database has 76 raw attributes. The first step is to take this raw data, clean it and select relevant attributes.

## 4.2 Data Pre-Processing

In this section, first, the tuples which have missing values are removed. There are six such tuples out of the 303 total patient details. The next step is to divide these selected 297 tuples into training and testing sets, using the train test split function with 80% data for training and 20% for testing. The next step is the selection of attributes. From the 76 raw attributes in the heart disease dataset, previous research used 13 attributes as input and the last target attribute as output. Table 1 contains the list of these attributes and their value range.

Attribute	Range
Age	30-70 years
Sex	Male/Female
Chest pain type	typical angina, atypical angina, non-angina pain, asymptomatic
Resting Blood Pressure	Variable
Serum Cholesterol	in mg/d
Fasting Blood Sugar	less than 12mg/dl
Resting Electro-cardiographic Results	0, 1 and 2
Maximum Heart Rate Achieved	Variable
Exercise Induced Angina	Yes/No
Old Peak	ST depression induced by exercise relative to rest
The slope of the peak exercise ST segment	Variable
Number of major vessels	0-3
Thal	3 = normal; 6 = fixed defect and 7 = reversible defect

Table 4.1: Attributes Used In This Paper

The variable to predict is Target, with values ranging from 0 to 4. The assumption is that every value with 0 indicates the Heart is working correctly for the patient. Values 1,2,3,4 indicate that the patient has heart disease. The variables measured on different scales do not contribute equally to the model fitting, and might create a bias. Normalization of data is required to avoid such bias. In this proposed model we have applied Min-Max normalization method. The min max normalization method converts the variables into a range of [0,1].

## 4.3 Base models

This paper uses three models ANN, SVM, XGBoost, which will act as the base models. The stacking ensemble algorithm takes the predictions of these models as input. Combining these models using XGBoost as a meta-model gives the final results.

### 4.3.1 Artificial Neural Networks

In the proposed approach, the artificial neural network created is a multilayer perceptron network with an input layer, two hidden layers with 5000 and 30 nodes, respectively, and one output layer. Figure 5.2 in chapter 5 shows the accuracy and its confusion matrix in the form of a heatmap. The Figure 4.2 below shows the snippet of code for the ANN algorithm.

```
m2 = 'Neural Network'
mlpc = MLPClassifier(solver='lbfgs', alpha=1e-5, hidden_layer_sizes=(7000,30),
                    random_state=1)
mlpc.fit(X_train, y_train)
mlpc_predicted = mlpc.predict(X_test)
mlpc_conf_matrix = confusion_matrix(y_test, mlpc_predicted)
mlpc_acc_score = accuracy_score(y_test, mlpc_predicted)
print("confussion matrix")
print(mlpc_conf_matrix)

print("Accuracy of : MLP Classifier",mlpc_acc_score*100,'\n')
print(classification_report(y_test,mlpc_predicted))
```

Figure 4.2: Code Snippet for Artificial Neural Network

### 4.3.2 Support Vector Machine (SVM)

In SVM, this approach uses C-Support Vector Classification. A kernel has the function of transforming an input data space into the needed structure. SVM uses a technique called the kernel trick. The kernel inputs a low-dimensional input space and modifies it into a higher-dimensional space. The kernel we have used is called the Radial basis function. RBF can map an input space into an infinite-dimensional space. The Figure 4.3 below shows the snippet of code for the SVM algorithm.



```

m3 = 'Support Vector Classifier'
svc = SVC(kernel='rbf', C=2)
svc.fit(X_train, y_train)
svc_predicted = svc.predict(X_test)
svc_conf_matrix = confusion_matrix(y_test, svc_predicted)
svc_acc_score = accuracy_score(y_test, svc_predicted)
print("confussion matrix")
print(svc_conf_matrix)
print("\n")
print("Accuracy of Support Vector Classifier:",svc_acc_score*100,'\n')
print(classification_report(y_test,svc_predicted))

```

Figure 4.3: Code Snippet for Support Vector Machine

### 4.3.3 Extreme Gradient Boost

Extreme Gradient Boost involves creating and adding trees to the model sequentially. The parameters to consider in its application are the number of trees, each tree's depth, and the learning rate. Modifying the values of these parameters changes the accuracy of the model. In this research, the learning rate used is 0.01, the total number of trees is 25, and the depth of the tree is 15. These values have given us the desired results, as shown in the figure 5.1 in the form of a confusion matrix. The figure 4.4 below shows the snippet of code for the XGBoost algorithm.

```

m1 = 'Extreme Gradient Boost'
xgb = XGBClassifier(learning_rate=0.01, n_estimators=25, max_depth=15,gamma=0.6,
                    subsample=0.52,colsample_bytree=0.6,seed=27,
                    reg_lambda=2, booster='dart', colsample_bylevel=0.6,
                    colsample_bynode=0.5)
xgb.fit(X_train, y_train)
xgb_predicted = xgb.predict(X_test)
xgb_conf_matrix = confusion_matrix(y_test, xgb_predicted)
xgb_acc_score = accuracy_score(y_test, xgb_predicted)
print("confussion matrix")
print(xgb_conf_matrix)
print("\n")
print("Accuracy of Extreme Gradient Boost:",xgb_acc_score*100,'\n')
print(classification_report(y_test,xgb_predicted))

```

Figure 4.4: Code Snippet for Extreme Gradient Boost Algorithm

The results from these algorithms will then act as an input for our final step of applying the stacking ensemble method.

## 4.4 Stacking Ensemble Model

Once we have the desired results using the base models, they are combined using the stacking method to improve the accuracy and functioning of the model. A meta-model is

selected, which can then combine the base models. In the proposed approach the meta model used is Extreme Gradient Boost. The figure 4.5 below shows the snippet of code for the Stacking Classifier algorithm.

```
m4=StackingCVClassifier(classifiers=[xgb,svc,mlpc],
                        meta_classifier= xgb,random_state=42)
scv.fit(X_train,y_train)
scv_predicted = scv.predict(X_test)
scv_conf_matrix = confusion_matrix(y_test, scv_predicted)
scv_acc_score = accuracy_score(y_test, scv_predicted)
print("confussion matrix")
print(scv_conf_matrix)
print("\n")
print("Accuracy of StackingCVClassifier:",scv_acc_score*100,'\n')
print(classification_report(y_test,scv_predicted))
```

Figure 4.5: Code Snippet for Stacking Ensemble Model

The next chapter shows the results of this work for each algorithm in the form of confusion matrices and the comparison with previous work referenced in this report.

## Chapter 5

### RESULTS and DISCUSSION

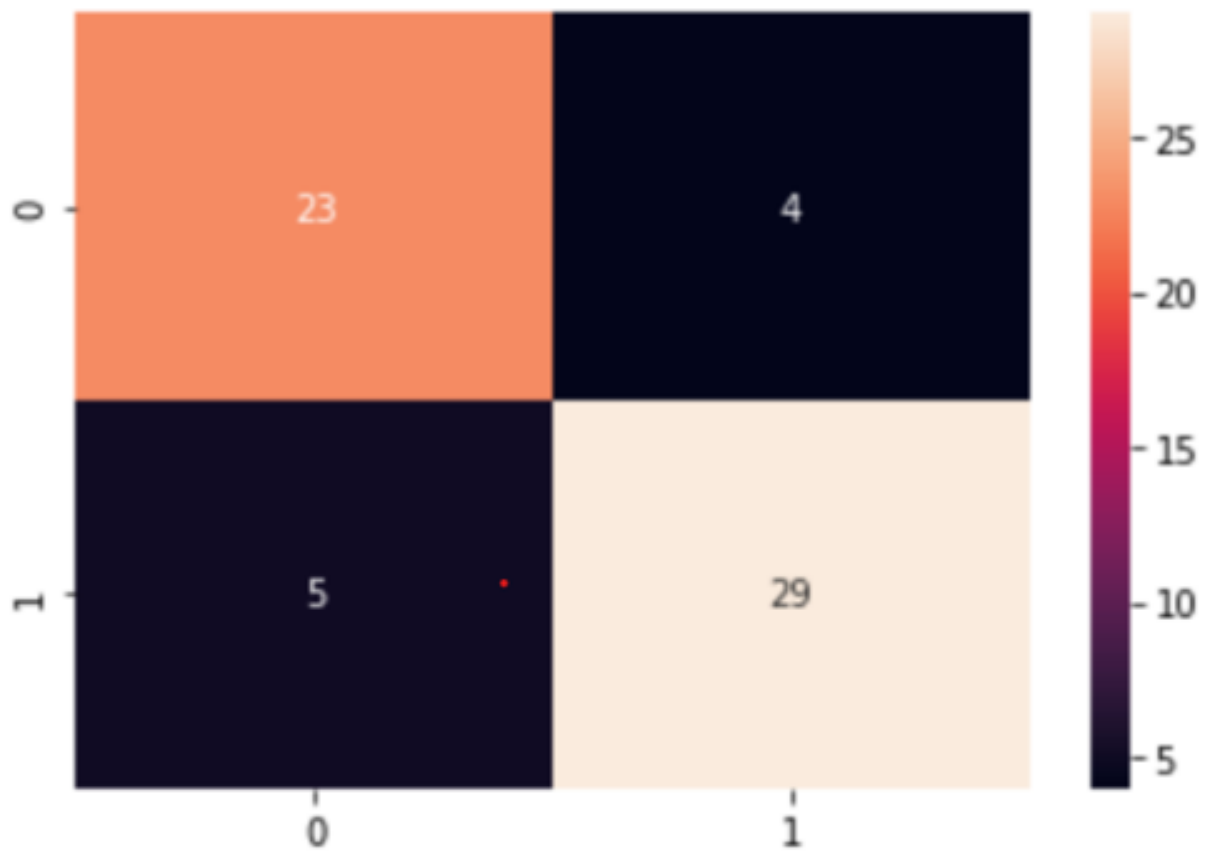
The proposed work uses the heart disease dataset; there were two classes of patients:

- Healthy Patients, i.e., No presence of heart disease.
- Unhealthy Patients, i.e., Heart disease possible in the patient.

This model initially applies Artificial neural networks, Support Vector Machine, and Extreme Gradient Boosts. After getting their output, these models get combined using a stacking ensemble method, with extreme gradient boost as the meta-model. The final result is a prediction value of the target variable in the dataset for each patient. The prediction is heart disease present in the patient or no heart disease present. Figures 1 to 4 show the heatmap of the confusion matrix for these algorithms. A heatmap is a technique for visualization of data.

Then comparison with the previous works mentioned in this paper with our results is shown.

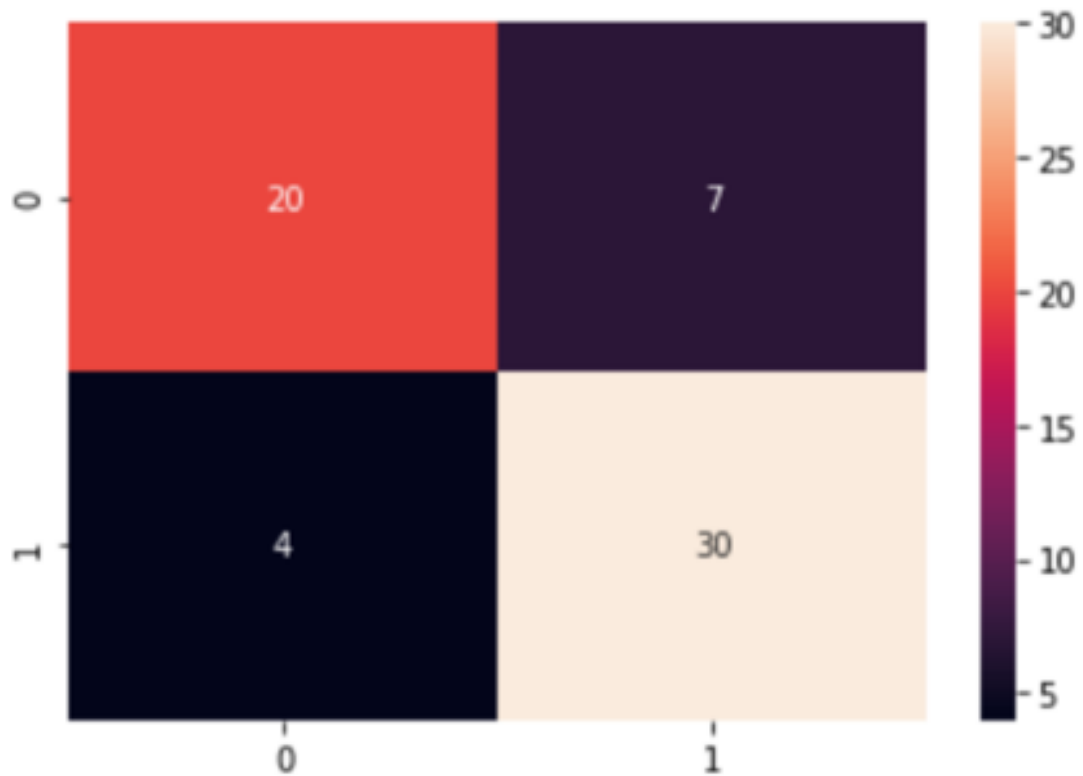
Figure 5.1 shows the heatmap and results for the XGboost algorithm. In this implementation of XGBoost, the accuracy is 85.24%. The figure also shows other metrics for this algorithm such as precision, recall, f1-score and support.



Accuracy of Extreme Gradient Boost: 85.24590163934425

	precision	recall	f1-score	support
0	0.82	0.85	0.84	27
1	0.88	0.85	0.87	34
accuracy			0.85	61
macro avg	0.85	0.85	0.85	61
weighted avg	0.85	0.85	0.85	61

Figure 5.1: Extreme Gradient Boost Confusion Matrix Heatmap

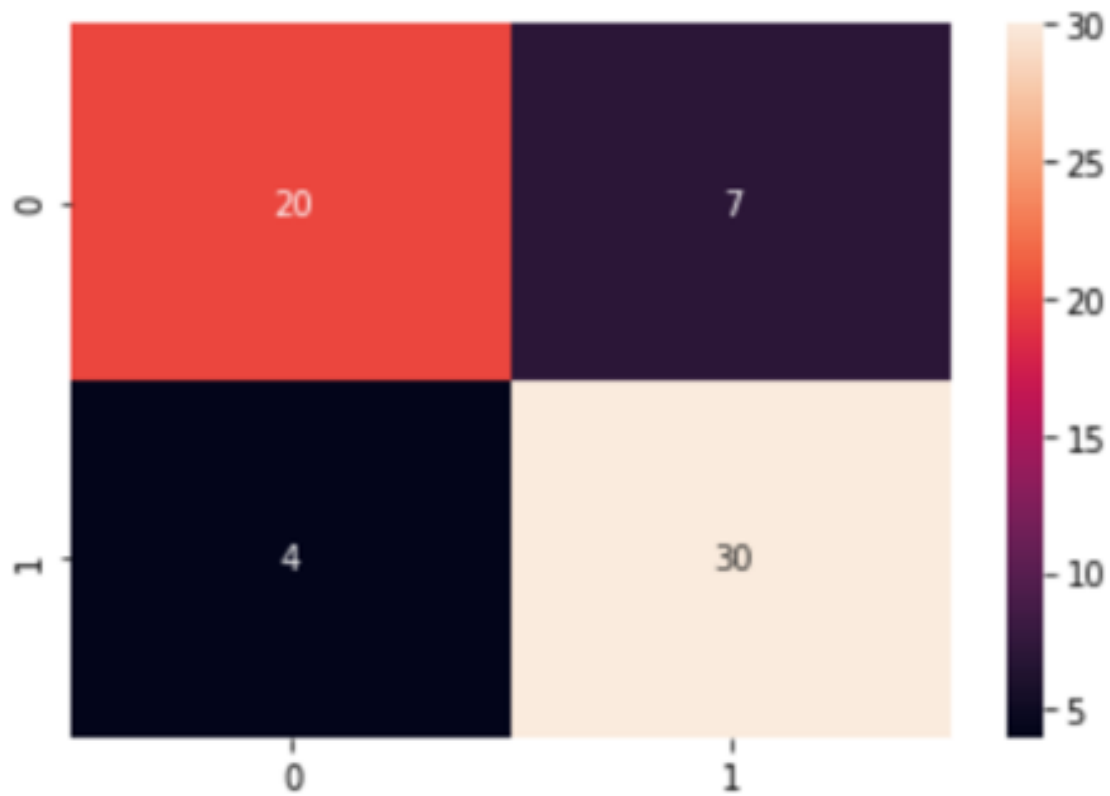


Accuracy of : MLP Classifier 81.9672131147541

	precision	recall	f1-score	support
0	0.83	0.74	0.78	27
1	0.81	0.88	0.85	34
accuracy			0.82	61
macro avg	0.82	0.81	0.81	61
weighted avg	0.82	0.82	0.82	61

Figure 5.2: Multi Layer Perceptron Confusion Matrix Heatmap

Figure 5.2 shows the heatmap and results for the MLP algorithm. In this implementation of MLP, the accuracy is 81.96%. The figure also shows other metrics for this algorithm such as precision, recall, f1-score and support.

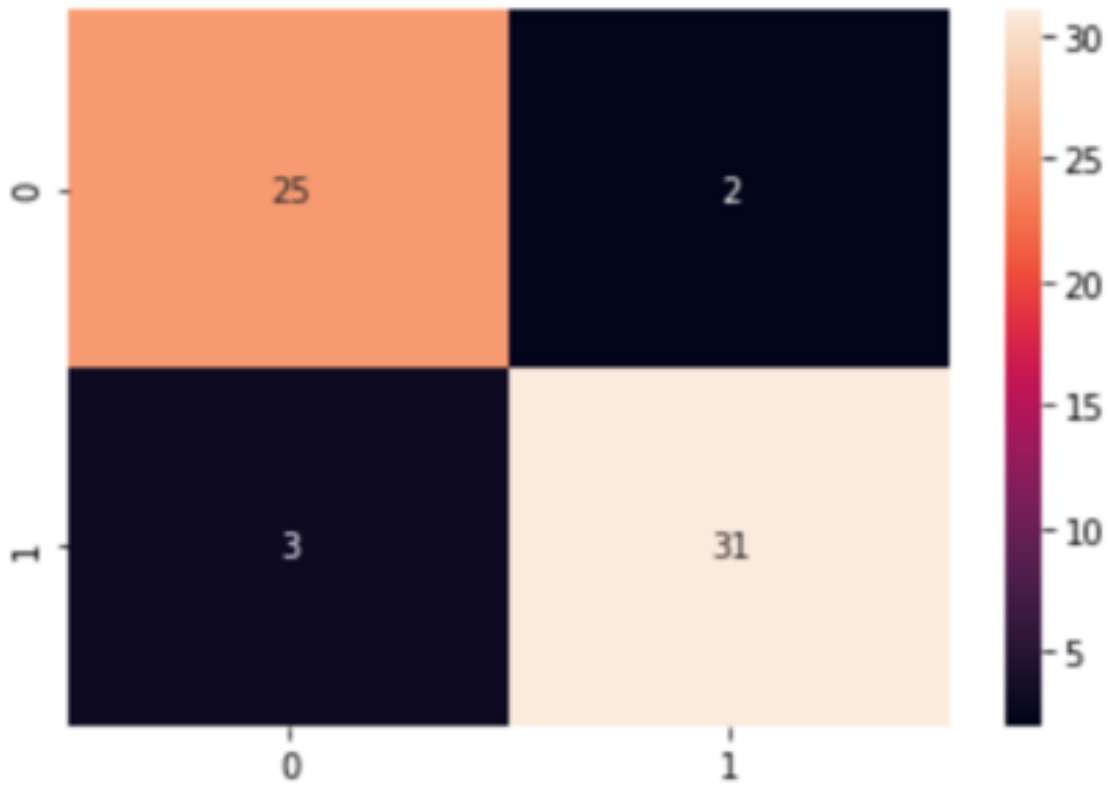


Accuracy of Support Vector Classifier: 88.52459016393442

	precision	recall	f1-score	support
0	0.86	0.89	0.87	27
1	0.91	0.88	0.90	34
accuracy			0.89	61
macro avg	0.88	0.89	0.88	61
weighted avg	0.89	0.89	0.89	61

Figure 5.3: Support Vector Machine Confusion Matrix Heatmap

Figure 5.3 shows the heatmap and results for the SVM algorithm. In this implementation of SVM, the accuracy is 88.52%. The figure also shows other metrics for this algorithm such as precision, recall, f1-score and support.



Accuracy of Stacking Ensemble Classifier: 92.0

	precision	recall	f1-score	support
0	0.89	0.93	0.91	27
1	0.94	0.91	0.93	34
accuracy			0.92	61
macro avg	0.92	0.92	0.92	61
weighted avg	0.92	0.92	0.92	61

Figure 5.4: Stacking Ensemble classifier Confusion Matrix Heatmap

Figure 5.4 shows the heatmap and results after combining the base algorithm using the stacking technique. The accuracy is 92.0%. The figure also shows other metrics for this algorithm, such as precision, recall, f1-score, and support.

Table 5.1 shows the accuracy of these models compiled together.

<b>Algorithm</b>	<b>Accuracy</b>
Multi layer Perceptron	81.94%
Support Vector Machine	88.52%
Extreme Gradient Boost	85.26%
Stacking ensemble method	92%

Table 5.1: Results of proposed model

Individually amongst the three initial algorithms, SVM has the highest accuracy of 85.26%. This research achieves the highest accuracy of the model by combining these three algorithms, using Stacking as portrayed in Table 5.1. Considering Table 5.2, it shows the comparison of results with other researches done on the heart disease dataset.

<b>Research Work</b>	<b>Algorithm Applied</b>	<b>Accuracy</b>
Adnan et al[1]	Feed forward MLP	85%
M. Gudadhe et al[4]	MLP with BP	80%
Mohanet al[5]	Decision Tree	87%
Detrano et al[9]	Logistic Regeression	77%
Resul Das et al[3]	Neural netowrk ensemble	89.01%
Radhimeenkshi et al.[18]	SVM and ANN	78% and 80%
Our approach	Stacking ensemble method	92%

Table 5.2: Comparison of our results with previous research works



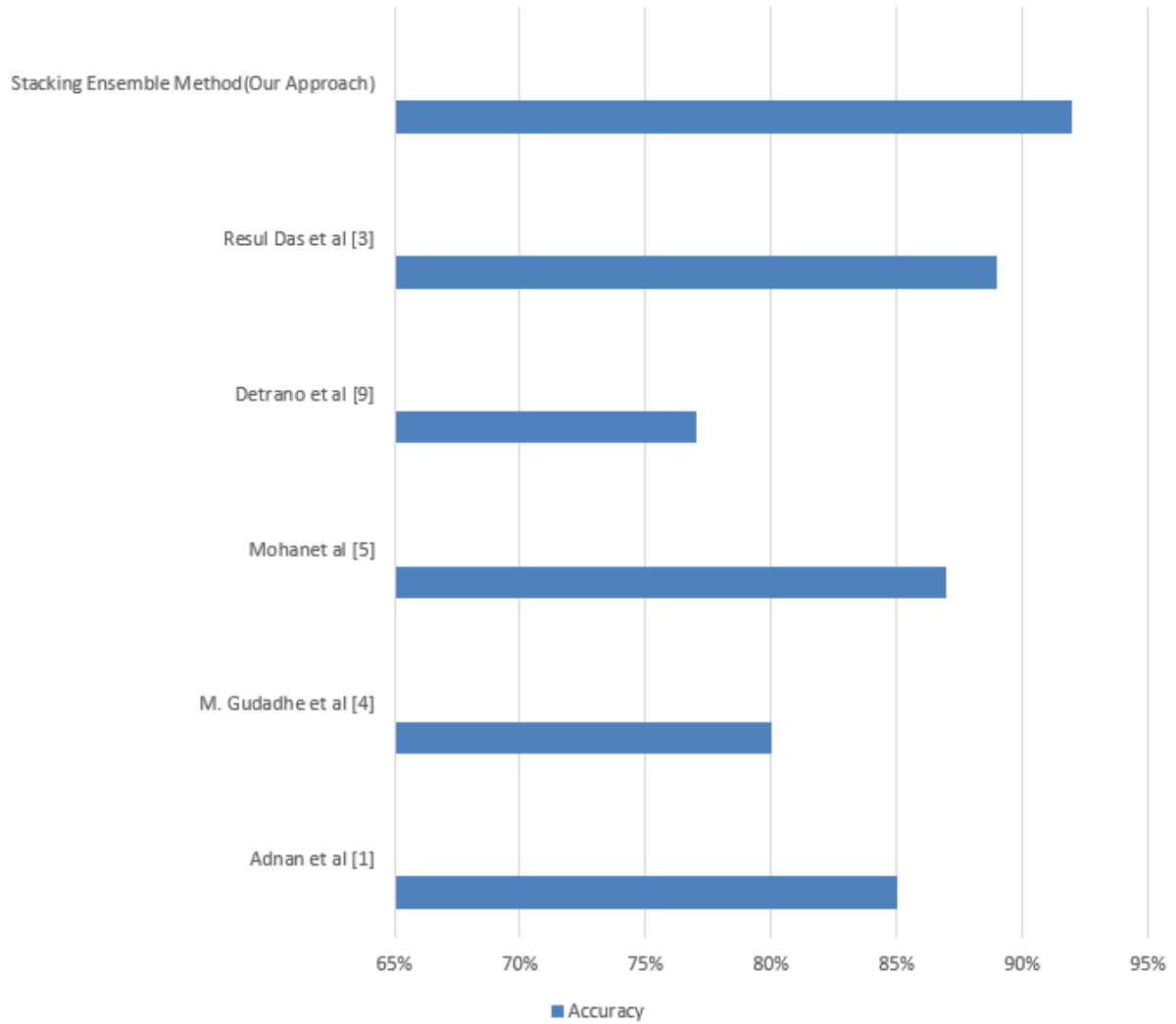


Figure 5.5: Comparison of results

As seen by Table 5.2 and Figure 5.5, the proposed approach of stacking ensemble method has been an improvement over the previous works done on this topic. This model overcomes the shortcomings of these previous works and creates an efficient decision support system for predicting heart diseases.

## Chapter 6

### CONCLUSION

Heart disease is one of the most deadly diseases affecting millions of people worldwide. To treat it effectively, it must get diagnosed at an early stage. Using the large amounts of medical data available, the prediction of such diseases using AI algorithms is increasing. In this research work, the goal is to study such previous works and overcome their shortcomings by introducing a new ensemble-based approach. The new model introduced consists of artificial neural networks, support vector machine, and extreme gradient boost algorithms as base models. Previous works have used these algorithms before, but this model combines them using the Stacked Generalization Ensemble technique. This model is created the UCI Heart Disease Dataset [21]. This model can classify if a person has heart disease or not with an accuracy of 92%, which is higher than the previous works on this topic.

The future work for this project will include applying this model to new patient's data and training it using more extensive and dense data sets to improve it further. This research work aimed to generate a model to predict heart disease, but this model can be applied to other diseases if the relevant data is available. Another future goal is to implement this model in a GUI-based software that any doctor in any clinic or hospital can use without any difficulty.

## Bibliography

- [1] Ebenezer Obaloluwa Olaniyi, Oyebade Kayode Oyedotun, Khashman Adnan, "Heart Diseases Diagnosis Using Neural Networks Arbitration", International Journal of Intelligent Systems and Applications(IJISA), vol.7, no.12, pp.75-82, 2015. DOI: 10.5815/ijisa.2015.12.08
- [2] 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>.  
(<https://www.sciencedirect.com/science/article/pii/S095741740800657X>)
- [3] Ebenezer Obaloluwa Olaniyi, Oyebade Kayode Oyedotun, Khashman Adnan, "Heart Diseases Diagnosis Using Neural Networks Arbitration", International Journal of Intelligent Systems and Applications(IJISA), vol.7, no.12, pp.75-82, 2015. DOI: 10.5815/ijisa.2015.12.08
- [4] M. Gudadhe, K. Wankhade and S. Dongre, "Decision support system for heart disease based on support vector machine and Artificial Neural Network," 2010 International Conference on Computer and Communication Technology (ICCCT), 2010, pp. 741-745, doi: 10.1109/ICCCT.2010.5640377.
- [5] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
- [6] Muhammad, Yar & Tahir, Muhammad & Hayat, Maqsood & Chong, Kil. (2020). Early and accurate detection and diagnosis of heart disease using intelligent computational model. Scientific Reports. 10. 10.1038/s41598-020-76635-9.
- [7] S. Bashir, Z. S. Khan, F. Hassan Khan, A. Anjum and K. Bashir, "Improving Heart Disease Prediction Using Feature Selection Approaches," 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), 2019, pp. 619-623, doi: 10.1109/IBCAST.2019.8667106.
- [8] Malav, Amita & Kadam, Kalyani & Kamat, Pooja. (2017). PREDICTION OF HEART DISEASE USING K-MEANS and ARTIFICIAL NEURAL NETWORK as HYBRID APPROACH to IMPROVE ACCURACY. International Journal of Engineering and Technology. 9. 3081-3085. 10.21817/ijet/2017/v9i4/170904101.
- [9] Anthony P. Morise, Robert Detrano, Marco Bobbio, George A. Diamond, Development and validation of a logistic regression-derived algorithm for estimating the

incremental probability of coronary artery disease before and after exercise testing, *Journal of the American College of Cardiology*, Volume 20, Issue 5, 1992, Pages 1187-1196, ISSN 0735-1097, [https://doi.org/10.1016/0735-1097\(92\)90377-Y](https://doi.org/10.1016/0735-1097(92)90377-Y). (<https://www.sciencedirect.com/science/article/pii/073510979290377Y>)

- [10] Xiao Liu, Xiaoli Wang, Qiang Su, Mo Zhang, Yanhong Zhu, Qiugen Wang, Qian Wang, "A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method", *Computational and Mathematical Methods in Medicine*, vol. 2017, Article ID 8272091, 11 pages, 2017. <https://doi.org/10.1155/2017/8272091>
- [11] Amin Ul Haq, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, Ruinan Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms", *Mobile Information Systems*, vol. 2018, Article ID 3860146, 21 pages, 2018. <https://doi.org/10.1155/2018/3860146>
- [12] Chaurasia, Vikas and Pal, Saurabh, Early Prediction of Heart Diseases Using Data Mining Techniques (2013). *Caribbean Journal of Science and Technology*, Vol. 1, 208-217, 2013 , Available at SSRN: <https://ssrn.com/abstract=2991237>
- [13] Dangare, Chaitrali and Apte, Sulabha, A Data Mining Approach for Prediction of Heart Disease Using Neural Networks (November 14, 2012). *International Journal of Computer Engineering and Technology (IJCET)*, Volume 3, Issue 3, October-December 2012, Available at SSRN: <https://ssrn.com/abstract=2175569>
- [14] Patel, Jaymin & Tejalupadhyay, Samir & Patel, Samir. (2016). Heart Disease prediction using Machine learning and Data Mining Technique. 10.090592/IJCSC.2016.018.
- [15] Kannan R., Vasanthi V. (2019) Machine Learning Algorithms with ROC Curve for Predicting and Diagnosing the Heart Disease. In: *Soft Computing and Medical Bioinformatics*. SpringerBriefs in Applied Sciences and Technology. Springer, Singapore. [https://doi.org/10.1007/978-981-13-0059-2\\_8](https://doi.org/10.1007/978-981-13-0059-2_8)
- [16] D. Verma and N. Mishra, "Analysis and prediction of breast cancer and diabetes disease datasets using data mining classification techniques," 2017 International Conference on Intelligent Sustainable Systems (ICISS), 2017, pp. 533-538, doi: 10.1109/ISS1.2017.8389229.
- [17] Gupta A., Kumar L., Jain R., Nagrath P. (2020) Heart Disease Prediction Using Classification (Naive Bayes). In: Singh P., Pawłowski W., Tanwar S., Kumar N., Rodrigues J., Obaidat M. (eds) *Proceedings of First International Conference on Computing, Communications, and Cyber-Security (IC4S 2019)*. Lecture Notes in Networks and Systems, vol 121. Springer, Singapore. [https://doi.org/10.1007/978-981-15-3369-3\\_42](https://doi.org/10.1007/978-981-15-3369-3_42)
- [18] S. Radhimeenakshi, "Classification and prediction of heart disease risk using data mining techniques of Support Vector Machine and Artificial Neural Network," 2016 3rd International Conference on Computing for Sustainable Global Development (IN-DIACom), 2016, pp. 3107-3111.
- [19] Nahar, Jesmin & Imam, Tasadduq & Tickle, Kevin & Chen, Yi-Ping Phoebe. (2013). Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems with Applications*. 40. 1086â€"1093. 10.1016/j.eswa.2012.08.028.

- [20] A. Dewan and M. Sharma, "Prediction of heart disease using a hybrid technique in data mining classification," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), 2015, pp. 704-706.
- [21] UCI Machine Learning Repository: Heart Disease Data Set.  
<http://archive.ics.uci.edu/ml/datasets/Heart+Disease>  
Creators:  
1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D. 2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D. 3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D. 4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.