IMPROVING HEART DISEASE DIAGNOSIS THROUGH SUPERVISED MACHINE LEARNING TECHNIQUES

BY

Khaled Bin Sijon
ID: 213-15-4554

FINAL YEAR DESIGN PROJECT REPORT

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

Supervised By

Dr. S. M. Aminul Haque
Professor and Associate Head
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By

Dr. Md. Zahid Hasan
Associate Professor
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

July 2024

APPROVAL

This Thesis titled "Improving Heart Disease Diagnosis Through Supervised Machine Learning Techniques", submitted by Khaled Bin Sijon to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 16-07-2024.

BOARD OF EXAMINERS

Dr. Sheak Rashed Haider Noori (SRH) — **Board Chairman**
Professor & Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Zahid Hasan (ZH) — **Internal Examiner 1**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Mr. Abdus Sattar (AS) — **Internal Examiner 2**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Manowarul Islam (MI) — **External Examiner**
Associate Professor
Department of Computer Science and Engineering
Jagannath University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Dr. S. M. Aminul Haque, Professor and Associate Head, Department of Computer Science and Engineering, Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Dr. S. M. Aminul Haque
Professor and Associate Head
Department of CSE
Daffodil International University

Co-Supervised by:

Dr. Md. Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

Submitted by:

Sijon

Khaled Bin Sijon
ID: 213-15-4554
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

We are grateful and wish our profound indebtedness to **Dr. S. M. Aminul Haque, Professor and Associate Head**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Data Mining, and Machine Learning*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to the **Dr. Sheak Rashed Haider Noori Sir**, for his kind help in finishing our project and also to other faculty members and the staff of the Department of CSE, Daffodil International University.

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

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

ABSTRACT

Heart disease is a major risk to public health, accounting for 17.9 million deaths globally annually. Predicting cardiac disease with machine learning (ML) approaches like Linear Regression, Random Forests, Naive Bayes, Decision Tree, Neural Network, XGBoost, AdaBoost, Support Vector Machines, and Catboost models, has shown encouraging results and we use ensemble model from the best model depending on their accuracy and others factors. And we find the best accuracy 97.53% from KNN & RF combined. These algorithms use large medical data to provide personalized risk assessment models for each unique user. As a result, heart disease may have a lessening impact on international health systems and improve methods for risk assessment and treatment. The association between cholesterol and fasting blood sugar levels and heart attacks was shown to be the weakest. The risk factors for heart disease are older adults and men. However, it is important to note that these ML approaches are not foolproof and may have limitations in accurately predicting all cases of heart disease. Additionally, further research and development in this area are necessary to enhance the accuracy and reliability of these personalized risk assessment models. It is crucial for healthcare professionals to continue monitoring and staying updated on the latest advancements in ML technology to effectively prevent and treat heart disease in individuals at risk.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Introduction

Before the early 1900s, heart illness occurred rarely. But it's increasing day by day. In 1900, it was the fourth most common cause of death, but three decades later, heart disease deaths had increased to become the third most common cause of death globally [1]. As of 2023, the global population is 8 billion, and approximately 620 million people are suffering from heart and circulatory diseases. Every year, around 60 million people are diagnosed with heart or circulatory diseases. Shockingly, an estimated 20.5 million people lost their lives due to such diseases in 2021, which means an average of 56,000 people per day, or one death every 1.5 seconds. (Source: The British Heart Foundation) [2]. Some major risk factors for heart disease are age, sex, blood pressure, cholesterol, fasting blood sugar, and maximum heart rate. We will apply machine learning algorithms, including Support Vector Machine, K-Nearest Neighbor, Decision Tree, Random Forest, Genetic Algorithm, Naïve Bayes, and Logistic Regression, to predict and analyze the data [3]. This allows us to evaluate which algorithm provides the highest accuracy. When we find the best algorithm for research, we will be able to identify if the patient has heart disease or not.

1.2 Motivation

Heart disease affects about 60 million people worldwide and has been the top cause of mortality for the past 30 years. Approximately one in three fatalities worldwide is attributed to heart and circulation disorders, with an anticipated 20.5 million deaths in 2021—or 56,000 deaths each day, or one death every 1.5 seconds [2]. However, because they are unaware of the causes of the disease and its treatment options, individuals are not taking enough precautions to prevent it. Everyone could benefit if they were aware of the risk factors for heart disease. While several strategies have been used in the past to forecast heart disorders, we can properly identify the risk of heart disease by preprocessing datasets and using machine learning algorithms

1.3 Rationale of the Study

Machine learning (ML) is used to predict cardiac disease because of its capacity to deal with complicated, high-dimensional datasets and nonlinear relationships. By combining various patient data sources such as demographics, medical history, lifestyle behaviors, genetic predispositions, and clinical test findings, machine learning can reveal hidden patterns and improve predictive accuracy. Emerging data sources and technologies, such as EHRs, wearable devices, and medical imaging, allow ML systems to extract valuable insights from vast amounts of heterogeneous data. ML-based models can aid in customized treatment by enabling targeted preventive efforts and optimal resource allocation. This study intends to increase the accuracy, efficiency, and efficacy of heart disease risk assessment, resulting in better healthcare delivery and patient well-being. Machine learning algorithms, for example, can use a mix of genetic markers, lifestyle factors, and diagnostic tests to forecast an individual's risk of developing cardiovascular disease. The ML system can deliver personalized suggestions for preventive measures and treatment alternatives based on data from electronic health records, wearable devices monitoring physical activity, and medical imaging scans of the heart. This personalized strategy enables healthcare providers to react earlier and more effectively, perhaps preventing or postponing the beginning of cardiovascular disease. Patients can also feel more empowered and engaged in their treatment by learning about their risk factors and how to improve their heart health. Overall, the incorporation of machine learning in heart disease risk assessment seems promising.

1.4 Research Questions

The study topics serve as guiding principles for further investigation into the use of machine learning (ML) techniques to forecast cardiac disease. They capture the core questions that this study aims to answer and give a framework for assessing the efficacy, feasibility, and implications of ML-based prediction models. The following research questions have been formulated:

1. What ML algorithms are most effective in predicting heart disease?

The purpose of this topic is to find the machine learning algorithms that outperform others in forecasting heart disease outcomes. It entails comparing and assessing different machine-learning approaches, such as logistic regression, support vector machines, decision trees, random forests, neural networks, and ensemble methods. This research aims to determine the best modeling strategies for heart disease prediction tasks by methodically evaluating the resilience, computing efficiency, and predictive ability of several algorithms.

2. How do different features and datasets impact the performance of heart disease prediction models?

This question studies how feature selection, feature engineering, and dataset features affect the predictive accuracy and generalization capacity of ML models for heart disease prediction. It comprises determining the significance of specific characteristics (e.g., demographic factors, clinical parameters, biomarkers) in predicting heart disease risk, as well as testing the effect of data preprocessing procedures (e.g., normalization, imputation, dimensionality reduction) on model performance. This study seeks to improve the interpretability and reliability of prediction models by carefully investigating the effects of feature subsets, data representations, and preprocessing procedures.

3. What is the comparative performance of ML-based prediction models against traditional risk assessment methods?

This topic compares the effectiveness of ML-based prediction models to traditional risk assessment tools and scoring systems used in clinical practice (e.g., Framingham Risk Score, SCORE, and ASCVD Risk Estimator). It entails undertaking extensive validation tests and performance evaluations to determine the discriminatory power, calibration, sensitivity, specificity, and clinical value of machine learning models in comparison to traditional risk stratification methods. By comparing ML-based prediction models with conventional criteria, this study hopes to demonstrate their efficacy as additional or alternative tools for assessing heart disease risk and informing clinical decision-making.

These research issues cover fundamental areas of using machine learning to predict cardiac disease, such as algorithm selection, feature engineering, model evaluation, and comparison analysis. By systematically investigating these questions, this study hopes to advance our understanding of ML-based predictive modeling in healthcare, contribute to the development of robust and clinically relevant prediction models, and, ultimately, improve heart disease prevention, diagnosis, and management.

1.5 Expected Output

We expect our study to provide some results if we can do it correctly and appropriately.

1 Increase the accuracy of the early diagnosis of heart disease.
2 Accurately determining the severity of heart disease.
3 The goal is to reduce the number of early deaths from heart disease.
4 Normal people are capable of effectively timing their actions.
5 Increase public knowledge of heart disease after the forecast.

1.6 Project Management and Finance

Timeline of my project:

Figure 1.6.1: Project Timeline Gantt Chart

It is important to have financial resources to motivate researchers in all research work. As I am currently in the process of creating a website, finance is also crucial for me to be able to launch the website publicly for users to access. I have calculated the approximate cost of launching my website, which includes other costs as well as website maintenance and marketing strategies to attract users. To ensure the success of my website, I will need to secure funds to cover these costs. Without the necessary financial resources, it will be challenging to achieve the level of visibility and engagement needed to make the website a valuable resource for users. By meticulously planning and budgeting for these expenses, I am confident that I will be able to create a website that not only draws in users but also provides them with a valuable and engaging experience.

Table 1.6.1: Project Cost Breakdown Table

Category	Item	Description	Unit Cost	Quantity	Total Cost

Web Development	1. Website design & development 2. Domain registration & hosting	1. Basic website with research information and contact form 2. Annual domain registration and website hosting	25,000 tk 5,500 tk /year	1 2 Years	25,000 tk 11,000 tk
Maintenance	1. Website updates & bug fixes 2. Content updates	1. Ongoing maintenance and updates for website functionality 2. Adding new research information and updates to the website	550 tk/ hours 350 tk/article	2hours/month 2 articles	1100 tk/month 700 tk
Computer Configuration	1. CPU 2. CPU Cooler 3.RAM	1. Intel 13th Gen Core i7 13700K Raptor Lake Processor 2. Lian Li Galahad II TRINITY 360 ARGB AIO Liquid CPU Cooler 3. TEAM XTREEM 32GB (2x16GB) DDR5 7600MHz Desktop RAM	47,500 tk 17,500 tk 24,500 tk	1 1 1	47,500 tk 17,500 tk 24,500 tk
T. Project Cost					127,300 tk

1.7 Report layout

Such a proposed report would be organized in detail to guide readers through the research process, its findings, and the implications that arise. Each chapter is suited to a different purpose but all bear together on the cohesiveness and depth of the document as a whole.

Chapter 1: Introduction

The introduction chapter provides the context for the study, outlining the importance of heart disease diagnosis and the need for a comparative comparison of detection and classification methods. It describes the research questions, objectives, and overall framework of the study.

Chapter 2: Background

The background chapter provides a thorough review of relevant literature and previous studies. This section explains the theoretical foundations, techniques, and technological breakthroughs connected to heart disease detection, including the machine learning model algorithm utilized in recent work and knowledge of current technologies. It provides background for the current investigation and identifies gaps in existing knowledge.

Chapter 3: Research Methodology

The methodology chapter explains the approach used to conduct the study. It describes the research design, data gathering methods, model architecture, and the reasoning for selecting various methodologies. The chapter also examines ethical concerns and any restrictions that may affect the study.

Chapter 4: Experimental Result

This chapter shows the experimental results in the domain of heart disease detection. Different studies concerning the performance of detection and classification methods are presented, together with visualization tools and statistical measures that validate these results.

Chapter 5: Impact on Society

The social impact chapter discusses the wider implications of the research findings on healthcare and medical diagnostics in society. It emphasizes how developments in methods for heart disease diagnosis can impact patients, their treatment, and care in hospitals, and general public health. It also considers technologies that might be derived from it, together with their social and economic impacts.

Chapter 6: Summary, Conclusion, Recommendation, and Implications for Future Research

This final chapter provides a summation of the full study. It highlights the important findings, reiterates the study's conclusions, and makes recommendations for practical applications and further research. The study's ramifications for future research are examined, offering a guide for those looking to build on its findings.

This report layout provides a rational flow of material, taking the reader through the study process from introduction to wider suggestions and ideas. It enables a thorough grasp of the research process and its possible impact on both the academic and practical elements of detecting cardiac disease using machine learning models.

CHAPTER 2
Background

2.1 Terminologies

In this section, we introduce the fundamental concepts and terminologies required to comprehend the context of heart disease prediction using machine learning (ML). By defining keywords and elucidating fundamental ideas, we want to bring clarity and coherence to the following debates and analyses. Machine Learning (ML) is an area of AI that develops methods and models that allow computers to learn from data without explicit programming. Machine learning (ML) approaches enable systems to recognize patterns, make predictions, and produce insights from complicated information, simplifying decision-making and problem-solving across multiple domains [5]. Feature selection will be one of the critical steps in the development of a model to enhance model performance and reduce computational complexity. In our case, we use various techniques for feature selection: correlation analysis, RFE, and PCA to identify the most critical features contributing to heart disease prediction [4]. Predictive modeling involves creating and assessing mathematical models to predict future outcomes or interpret unknown attributes from input data. When applied to heart disease prediction, the goal is to evaluate an individual's risk of developing heart disease or experiencing negative consequences by using relevant patient information, biomarkers, and clinical indicators [6]. Cross-validation is a resampling technique that evaluates the robustness and stability of prediction models by dividing available data into subsets (folds) for training and validation. Common cross-validation methods include k-fold, leave-one-out, and stratified cross-validation. Cross-validation aids in estimating model generalization performance and identifying potential sources of bias or overfitting by systematically evaluating them on diverse data sets [7].

2.2 Related Works

These are a few of the decision support systems that have been created by different researchers to forecast heart disease, and they are discussed here. The author Wang, S. et al [8] study used a logistic regression algorithm and 14 physical indicators from 302 patients to investigate heart attacks. Strong correlations were found with chest pain, heart rate, exercise-induced angina, ST depression, age, sex, blood vessel number, thalassemia, and fasting blood sugar. Older people and males were more likely to develop heart disease. The algorithm's accuracy was 85.95%.

Rati Goel et al. [9] This study analyzes and predicts heart disease using a data set consisting of 13 components and 383 individual values using machine learning and Python. The objective is to enhance the precision of cardiac disease detection by machine learning algorithms, tackling problems such as hypertension, hyperglycemia, hypertension, and hypercholesterolemia.

M Kumar et al. [10] The paper aims at predicting whether a person is easily susceptible to heart-related diseases. Most of the research in this area has focused on comparing the commonly known algorithms. A comparison of a majority of algorithms has been done to find the best algorithm. Multiple Machine Learning (ML) classifications were run to determine the outcome. Support Vector Classifier (SVC) and Nu-SVC had the highest accuracy in determining heart-related disease. Due to its high accuracy, it was concluded that ML algorithms are best suited for decision-making purposes concerning clinical scenarios.

Nandal N et al. [11] This study presents a machine learning-based heart attack prediction (ML-HAP) method, focusing on predicting heart attacks using various risk factors like high blood pressure, cholesterol, and diabetes. Data from the UCI ML Repository was analyzed using ML approaches such as Support Vector Machines, Logistic Regression, Naïve Bayes, and XGBoost. The results showed

XGBoost provided the best prediction, with an area under the curve of.94 and a.92. The prediction was evaluated for accuracy, precision, recall, and area under the curve. The study suggests further optimization by focusing on associated risk factors.

Arko Pal Pal et al. [12] This paper proposes a healthcare monitoring system using wearable devices to measure vital signs of heart disease, detect irregularities, and predict future problems, as heart diseases often go undetected until they become a concern.

Maanaav Motiramani et al. [13] The healthcare sector is a significant source of data, particularly in the detection of heart disease. Early diagnosis is crucial for successful treatment. Heart attacks are increasing due to factors like work demands, lack of physical activity, irregular eating, obesity, and high-fat diets. India faces a high burden of cardiovascular disease, with annual deaths predicted to rise from 2.26 million in 1990 to 4.77 million in 2020. This study aims to use machine learning models to detect heart attacks, balancing accuracy and recall for effective prediction.

T A Mohanaprakash et al. [14] Machine Learning (ML) is crucial in healthcare for disease prediction and early patient care. This study evaluates and suggests heart disease prediction using SVM, MLR, and RF algorithms. The proposed method outperforms current methods in accuracy, forecast speed, and consistency of outcomes. It is also suitable for classifying lung cancers using trained datasets for accurate identification.

2.3 Comparative Analysis and Summary

In Bangladesh, 70% of people suffer from heart disease [15]. And they care even less about their bodily wellness than that. Many are suffering from a variety of symptoms as a result of their ignorance about their health. Most people have excessive blood sugar, high cholesterol, and other health problems that increase their chance of developing cardiovascular disease as a result of their consistently poor diets.

Table 2.3.1. Comparative analysis with previous work

SL No	Author Name	Used Algorithm	Best Accuracy with Algorithm
1.	Wang, S. et al [4]	Logistic Regression=85.95%	LR=85.95%
2.	Rati Goel et al. [5]	Logistic Regression=77%, KNN=82%, SVM=86%, Naïve Bayes=68%, Decision Tree=83%, and Random Forest=83%	SVM= 86%
3.	M Kumar et al. [6]	Support Vector Classifier (SVC) and Nu-SVC	SVC = 88%
4.	Nandal N et al. [7]	Logistic Regression 85%, Naïve Bayes 82%, SVM 64%, and XGBoost 92%	XGBoost= 92%
5.	Arko Pal Pal et al. [8]	KNN=80%, Logistic Regression=78%, SVC=78%, Gradient Boosting Classifier=78%, Random Forest Classifier=78%, GradientBoostingClassifier=76%, Convolutional Neural Network (CNN)=83%	CNN = 83%

6.	Maanaav Motiramani et al. [9]	Naive Bayes, Decision Tree, K-Nearest Neighbors Algorithm (KNN), Neural Network	
7.	T A Mohanaprakash et al. [10]	Support Vector Machine (SVM), Multiple linear regression (MLR), Random Forest (RF)	

2.4 Scope of the Problem

In this section, we define the exact scope and focus of the heart disease prediction problem investigated in this study. By establishing the study's boundaries and objectives, we want to provide clarity and context for the methodological and analytical discussions that will follow.

The heart disease prediction problem targets persons at risk of cardiovascular complications, such as coronary artery disease, myocardial infarction, and heart failure. Patients with established cardiovascular risk factors such as hypertension, dyslipidemia, diabetes, obesity, smoking, and a family history of heart disease may be eligible. The study may also include asymptomatic adults who are getting routine cardiovascular risk assessments for primary preventive objectives.
Heart disease prediction models examine various demographic, clinical, lifestyle, and diagnostic parameters that can affect cardiovascular health. These factors may include age, gender, blood pressure, lipid profiles (e.g., cholesterol levels), glycemic parameters (e.g., fasting blood glucose), heart rate, resting electrocardiogram, number of vessels, thalassemia, and relevant medical history.
This study will focus on identifying any cardiovascular events or the development of clinically significant heart disease during the specified follow-up period. Cardiovascular events will help determine if a patient has heart disease. Secondary outcome measures may include intermediate goals such as changes in cardiac biomarkers or disease progression markers. Model creation and validation

CHAPTER 3

Research Methodology

3.1 Research Subject and Instrumentation Research Subject:

Individuals in the study were recruited from UCI Machine Learning Repository databases. Individuals of various ages, genders, and cardiovascular risk profiles may meet the inclusion criteria. Furthermore, research subjects may include both asymptomatic individuals and patients with known cardiovascular disease or risk factors, ensuring that the study group is representative and generalizable.

I've been using machine learning to solve the issue at hand in this particular case. This presentation examines the differences between many unique machine learning techniques, such as Random Forest, Support Vector Machine, Linear Regression, Naïve Bayes, Decision Tree, XGBoost, AdaBoost, CatBoost, Neural Network, and K-Nearest Neighbors Algorithm (KNN). I used a dataset and these algorithms to predict the occurrence of heart disease. I used the Python programming language to run all of the heart disease data through Google Collaboratory for more precise results.

.

3.2 Data Collection Procedure

A. Datasets

The first stage in developing a prediction system is collecting data from the internet [16]. Data is then cleaned and the training and testing datasets are identified. We used 80% of the dataset for training and 20% for testing the system. After cleaning, the data collection has 14 columns and 383 rows created by code (data.info). Figure 3.2.1 illustrates how using the code (data.head()) creates a row in a table.

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Figure 3.2.1: Dataset Info

B. Process of Experiments

Data Source:

In this study we used data from a mid-scale, multi-center UCI Machine Learning Repository database with different patient groups. The UCI database collects de-identified patient information from a variety of medical organizations, including medical centers, hospitals, and medical practices to establish a comprehensive repository of clinical data across several domains [16].

Inclusion Criteria:

To be included in the UCI database, patients must be 18 years of age or older and have pertinent clinical data for cardiovascular health and risk assessment. The study group comprises patients with documented experiences or diagnoses of cardiovascular diseases, including coronary artery disease, hypertension, dyslipidemia, heart failure, and myocardial infarction.

Data Collection:

The collection contains a diverse set of structured and unstructured data items taken from the electronic health record (EHR), including:

1. Demographic Information: Age and gender
2. Clinical measurements: Blood pressure, lipid profiles (total cholesterol, blood vessel count), and glycemic parameters (fasting blood sugar)
3. Medical history: Prior cardiac events (myocardial infarction, stroke), comorbidities (hypertension, obesity),
4. Diagnostic tests and procedures include electrocardiograms (ECGs).

TABLE 3.2.1: Sample attribute with notes and values from the dataset

Attribute	Code given	Note	Values
1. age	Age	in years	Numeric
2. sex	Sex	1 = male 0 = female	Binary
3. chest pain type	cp	0,1,2,3 4	Values
4. resting blood pressure	restbps	in mm Hg	Numeric
5. serum cholesterol	cholesterol	in mg/dl	Numeric
6. fasting blood sugar	fbs	> 120 mg/dl	Numeric
7. resting electrocardiographic results	restecg	0,1,23	Values
8. maximum heart rate achieved	thalach	71-202	Numeric
9. exercise-induced angina	exang	0,1	Binary
10. oldpeak = ST	oldpeak	depression	Numeric
11. the slope of the peak exercise ST segment	slope	0,1,23	Values
12. number of major vessels fluoroscopy	ca	0,1,2,3 4	Values
13 defect: normal; fixed; reversible; non-reversible	thal	0,1,2,3 4	Values
14. class	target	0,1	Binary

Data Preprocessing:

Real-life data contains significant amounts of missing and noisy data. These data are pre-processed to avoid such problems and make strong predictions. Table 3.2.1 shows the sequential chart for our proposed model. Cleaning gathered data frequently results in noise and missing values. To achieve reliable results, data should be cleansed of noise and missing values filled in. Transforming data from one format to another improves comprehension. This involves smoothing, standardization, and formation activities [17].

Data Splitting:

Splitting the dataset into training and testing sets helps with model building and evaluation. The training set, which makes up the majority (70-80%) of the dataset, is used to train machine learning models using historical data. The testing set, which represents the remaining percentage (20-30%) of the dataset, is set aside for independent validation to evaluate model performance on previously unseen data [18].

Data Training:

Machine learning models, such as Linear Regression, Decision Trees, Random Forests, Support Vector Machines, KNN, XGBoost, Neural Networks, AdaBoost, and CatBoost are trained on preprocessed training data. To maximize performance, model hyperparameters are tweaked using approaches like grid search and random search [19].

Ensemble Model:

An ensemble model is a valuable and powerful method in machine learning, as it reduces the variance, combining predictions across multiple individual models to obtain a more accurate and robust overall prediction. Take the scenario whereby one gets different experts with their insights, pools them together, and lets them work collaboratively on decision-making; there are various ensemble models [20].

Classification Methodology:

After training, models categorize individuals as having or not having heart disease based on input features. Metrics used to evaluate classification performance include accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). The trained models are then used in the testing dataset to determine how well they generalize to previously unseen data [21].

Results:

The experiment results are divided into three categories depending on the topologies of the unique specific systems, machine learning models, and ensemble techniques.

3.3 Statistical Analysis

In this study, machine learning algorithms are utilized to predict cardiac disease. The algorithms' performance is assessed using robust statistical approaches such as Linear Regression, Random Forest, Decision Tree, XGBoost, AdaBoost, CatBoost, Neural Network, SVM, and KNN. The primary developments and diversity of the prediction findings projected in this study are fully summarized by descriptive statistics such as mean, median, standard deviation, and range. Inferential statistics, such as hypothesis testing, t-tests, ANOVA, and confidence intervals, demonstrate how various algorithms perform differently and similarly.

Furthermore, Advanced statistical approaches are applied, including ROC analysis and AUC. As a result, the measures assist in identifying how well the algorithms perform in properly identifying individuals with and without cardiac disease. Correlation analysis will aid in identifying important aspects that are most likely to cause heart disease, as well as providing insights into the underlying risk variables that contribute to this. It will also examine the significance score for every attribute to see how it contributed to the prediction process.

Additionally, Cross-validation protects non-generalizing models. A strong relationship between statistical analysis and machine learning leads to the detection of trends and patterns, as well as possible data correlations. It is an effective method for validating models and providing actionable insights pertinent to specific clinical practice applications, but it is also useful for guiding future research.

	count	mean	std	min	25%	50%	75%	max
age	303.000	54.366	9.082	29.000	47.500	55.000	61.000	77.000
sex	303.000	0.683	0.466	0.000	0.000	1.000	1.000	1.000
cp	303.000	0.967	1.032	0.000	0.000	1.000	2.000	3.000
trtbps	303.000	131.624	17.538	94.000	120.000	130.000	140.000	200.000
chol	303.000	246.264	51.831	126.000	211.000	240.000	274.500	564.000
fbs	303.000	0.149	0.356	0.000	0.000	0.000	0.000	1.000
restecg	303.000	0.528	0.526	0.000	0.000	1.000	1.000	2.000
thalachh	303.000	149.647	22.905	71.000	133.500	153.000	166.000	202.000
exng	303.000	0.327	0.470	0.000	0.000	0.000	1.000	1.000
oldpeak	303.000	1.040	1.161	0.000	0.000	0.800	1.600	6.200
slp	303.000	1.399	0.616	0.000	1.000	1.000	2.000	2.000
caa	303.000	0.729	1.023	0.000	0.000	0.000	1.000	4.000
thall	303.000	2.314	0.612	0.000	2.000	2.000	3.000	3.000
output	303.000	0.545	0.499	0.000	0.000	1.000	1.000	1.000

Figure 3.2.2: Basic Descriptive Statistical Analysis

3.4 Proposed Methodology

To find the best classifier with the highest accuracy for heart disease prediction, we used 10 alternative models. I gather user data, incorporate it into the dataset, choose which data to clean, apply each machine learning model to forecast the illness, and then assess the accuracy. K-Nearest Neighbors, Support Vector Machines, Decision Trees, Random Forest, Linear Regression, AdaBoost, XGBoost, CatBoost, Neural Network, and Naive Bayes were the ten classifiers employed. I will utilize each model to provide predictions based on the data that the user-supplied after the data has been cleaned. The best classifier for predicting heart disease was then identified by calculating each model's accuracy using a variety of measures, including accuracy, precision, recall, and F1 score. Based on the assessment of these, the most accurate model will be chosen for the ensemble model of this predictive analysis. We ensemble the model depending on which model has the highest accuracy for this experiment. Then we merge two, three, and four models to get the highest accuracy. This will ultimately improve patient outcomes and quality of care, leading to better overall healthcare delivery. Ultimately, the goal is to save lives and improve the health of individuals.

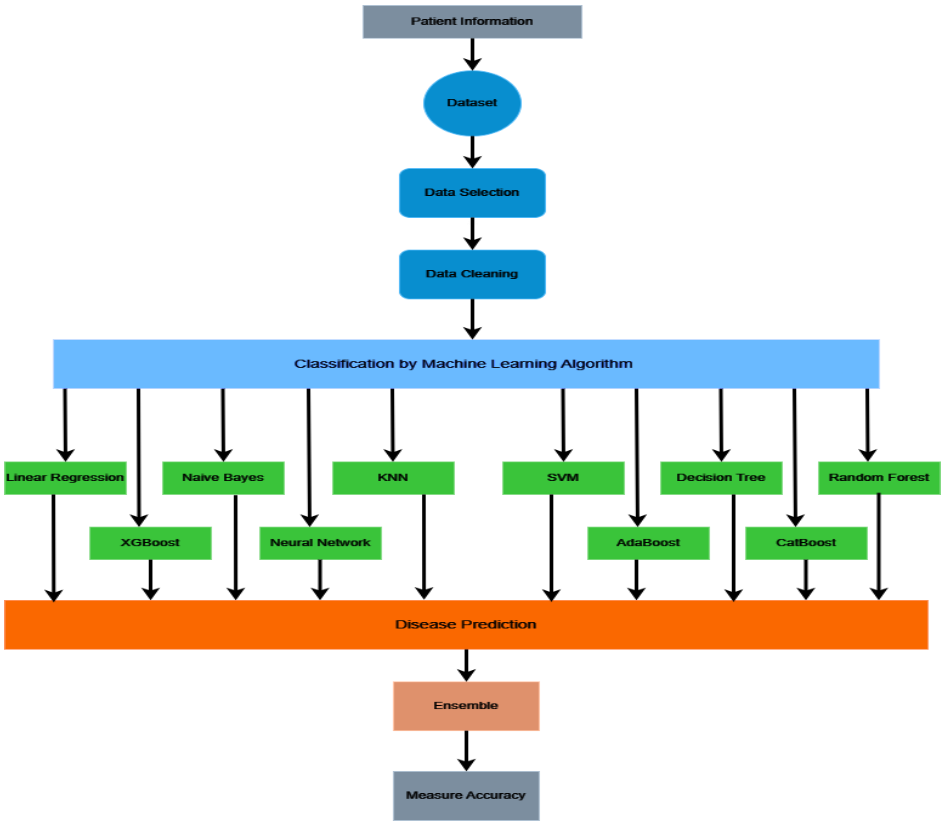

Figure 3.4.1: Proposed Methodologies

3.5 Implementation Requirements

Afterward carefully analyzing all relevant statistical or analytical approaches and methodologies, we have compiled a list of the hardware, software, and development tools required to anticipate cardiac disease. These are, in fact, the products that are most likely to be required.

Hardware Requirements:

- Windows 11 operating system,
- 16 GB of RAM,
- Chrome or Microsoft Web Explorer,

- NVidia GPU.

Software Requirements:

- Jupyter Notebook,
- Anaconda,
- Google Colab and Python 3.8 or higher version.

Ethical Considerations:

- **Patient confidentiality and data safety:** Compliance with ethical principles and legislation to safeguard the confidentiality and safety of patient data utilized in the study.
- **Informed consent:** Compliance with ethical principles and legislation to protect the confidentiality and safety of patient data used in research.

Documentation and Reproducibility:

- **Code documentation:** Thorough documentation of the implementation code is required to ensure transparency, repeatability, and future collaboration.
- **Version control:** Use tools (e.g., Git) to track code changes and ensure an organized development process.

The accomplishment of these implementation requirements will help in carrying out the research methodically and rigorously to ensure the trustworthiness of the results and contributions towards helping in setting advances in the field of heart disease diagnosis using machine learning algorithms.

CHAPTER 4

Experimental Result

4.1 Experimental Setup

The effectiveness of machine learning models in predicting cardiac disease is assessed in this work utilizing a dataset from the UCI Machine Learning Repository.

The dataset includes 303 patient records with 14 parameters such as age, gender, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG, maximum heart rate reached, exercise-induced angina, old peak, slope, number of main vessels, and thalassemia. A computer with an Intel i7 processor, 16GB RAM, and a 500GB SSD was used for the research.

Data preprocessing included dealing with missing data, standardizing features, and identifying important characteristics for predicting heart disease. Several machine learning techniques were tested, including linear regression, decision trees, random forest, support vector machines (SVM), XGBoost, AdaBoost, CatBoost ,and neural networks.

Training and validation included a train-test split, cross-validation, and assessment measures like accuracy, precision, recall, F1-Score, and ROC-AUC. The dataset was collected from the UCI Machine Learning Source, a publicly available dataset appropriate for cardiopathy prediction research.

The studies were conducted in a controlled computing environment using Python and the relevant libraries. We verified that all software and tools were configured and tested to produce accurate and reproducible results.

4.2 Experimental Results & Analysis

Result of Experiment 1: Machine Learning Model

This section compares eight unique machine learning models: Linear Regression, Naive Bayes, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Random Forest, XGBoost, Neural Network, AdaBoost, and CatBoost. Classification performance comes first. After that, the overall metrics of the models are examined. Collection, descriptions, possible causes, and opportunities for improvement in outcomes. We have examined the accuracy of various machine learning models, including AdaBoost, CatBoost, XGBoost, Neural Network, Decision Tree, KNN, Naive Bayes, SVM, and Random Forest, in identifying heart disease in Table 4.2.1. After conducting the investigation, it was found that the Naive Bayes, K-Nearest Neighbors, Random Forest, Neural Network, and AdaBoost models performed the best in recognizing heart disease within medical data, achieving a maximum accuracy of 90.16% on three machine learning model they are: K-Nearest Neighbours, Random Forest, and AdaBoost. On the other hand, Naive Bayes, Neural Network, and CatBoost showed relatively good performance, achieving an accuracy between 85% and 90%. Some models such as Linear Regression, Support Vector Machine, Decision Tree, and XGBoost give a lower accuracy of 85%. This slight difference highlights the significance of choosing the most appropriate machine learning model for heart disease diagnosis, as it significantly effects the accuracy and dependability of the analytic process.

TABLE 4.2.1: Accuracy for Classification of Individual Machine Learning Model in Detecting Heart Disease Prediction

Architecture	**Training Accuracy**	**Model Accuracy**
Linear Regression	**0.86%**	**83.61%**
Naive Bayes	**0.84%**	**85.25%**
SVM	**0.88%**	**83.61%**
KNN	**0.84%**	**90.16%**
Decision Tree	**1.0%**	**81.97%**
Random Forest	**1.0%**	**90.16%**
XGBoost	**1.0%**	**83.61%**
Neural Network	**0.84%**	**85.25%**
AdaBoost	**0.92%**	**90.16%**
CatBoost	**0.87%**	**86.89%**

TABLE 4.2.2: Precision, Recall, F1-Score, and Support for Machine Learning Model

Linear Regression				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.74	0.87	0.80	23
Heart Disease	0.91	0.82	0.86	38
Accuracy			0.84	61
Macro Avg	0.83	0.84	0.83	61
Weighted Avg	0.85	0.84	0.84	61
Naïve Bayes				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.78	0.88	0.82	24
Heart Disease	0.91	0.84	0.87	37
Accuracy			0.85	61
Macro Avg	0.84	0.86	0.85	61
Weighted Avg	0.86	0.85	0.85	61
SVM				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.70	0.90	0.79	21
Heart Disease	0.94	0.80	0.86	40
Accuracy			0.84	61
Macro Avg	0.82	0.85	0.83	61
Weighted Avg	0.86	0.84	0.84	61

KNN				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.81	0.96	0.88	23
Heart Disease	0.97	0.87	0.92	38
Accuracy			0.90	61
Macro Avg	0.89	0.91	0.90	61
Weighted Avg	0.91	0.90	0.90	61
Decision Tree				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.81	0.79	0.80	28
Heart Disease	0.82	0.85	0.84	33
Accuracy			0.82	61
Macro Avg	0.82	0.82	0.82	61
Weighted Avg	0.82	0.82	0.82	61

Random Forest				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.85	0.92	0.88	25
Heart Disease	0.94	0.89	0.91	36
Accuracy			0.90	61
Macro Avg	0.90	0.90	0.90	61
Weighted Avg	0.90	0.90	0.90	61

XGBoost				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.81	0.81	0.81	27
Heart Disease	0.85	0.85	0.85	34
Accuracy			0.84	61
Macro Avg	0.83	0.83	0.83	61
Weighted Avg	0.84	0.84	0.84	61

Neural Network				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.78	0.84	0.81	25
Heart Disease	0.88	0.83	0.86	36
Accuracy			0.84	61
Macro Avg	0.83	0.84	0.83	61
Weighted Avg	0.84	0.84	0.84	61

AdaBoost				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.93	0.86	0.89	29
Heart Disease	0.88	0.94	0.91	32
Accuracy			0.90	61
Macro Avg	0.90	0.90	0.90	61
Weighted Avg	0.90	0.90	0.90	61

CatBoost				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.93	0.86	0.89	29
Heart Disease	0.88	0.94	0.91	32
Accuracy			0.90	61
Macro Avg	0.90	0.90	0.90	61
Weighted Avg	0.90	0.90	0.90	61

In section 4.2.2, a comprehensive breakdown of the training percentages and sample accuracies resulting from eight distinct supervised machine-learning models is presented in Table 4.2.2. This table illustrates the precision, recall, F1-score, and support values for each category. The machine-learning models featured in the table encompass Logistic Regression, Naive Bayes, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, Random Forest, XGBoost, and Neural Network.

Remarkably, the KNN, Random Forest, and Neural Network models exhibit impressive performance when subjected to the test dataset, particularly in terms of precision. These models have demonstrated their ability to effectively identify heart disease, consistently displaying superior precision, recall, and F1-score across multiple classes.

Training and Validation Accuracy and loss of Machine Learning Model:

This plot illustrates, with the base model, training and validation accuracy. The epochs are on the x-axis, and accuracy percentage or loss on the y-axis. The figure is that of one where training and validation data get well split and there isn't any overfitting.

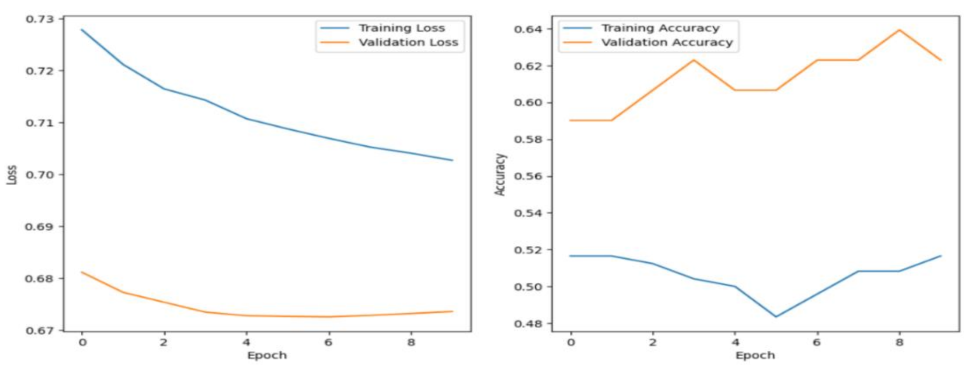

Figure 4.2.1: Training and validation accuracy and loss over the epochs (Naive Bayes)

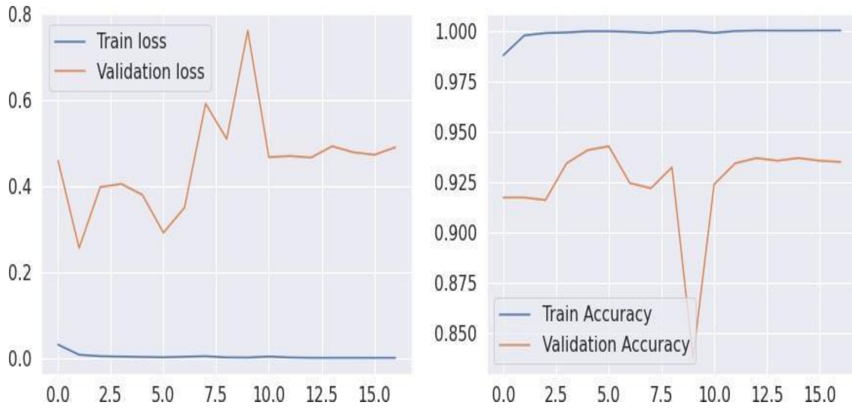

Figure 4.2.2: Training and validation accuracy and loss over the epochs (K-Nearest Neighbor).

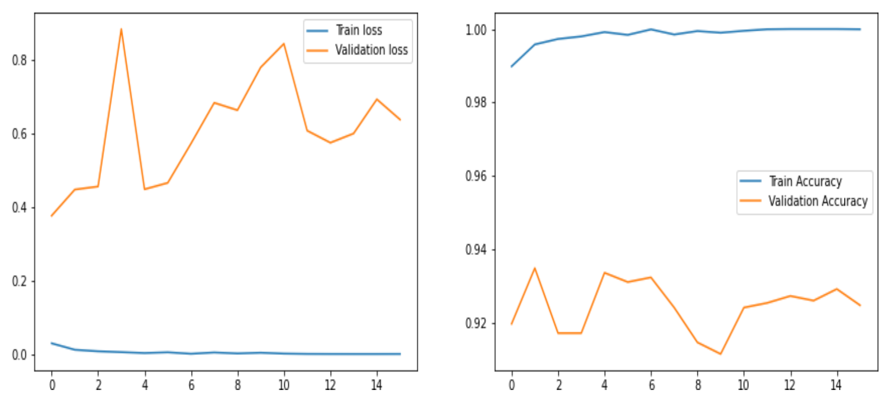

Figure 4.2.3: Training and validation accuracy and loss over the epochs (Random Forest)

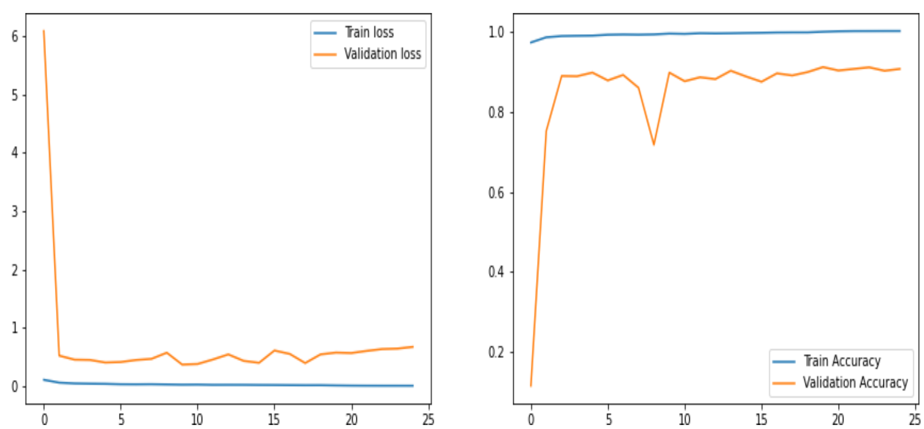

Figure 4.2.4: Training and validation accuracy and loss over the epochs (Neural Network).

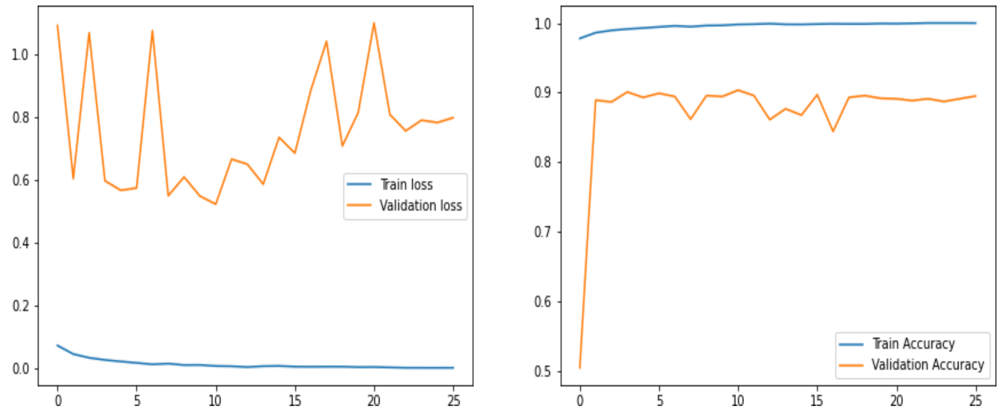

Figure 4.2.5: Training and validation accuracy and loss over the epochs (AdaBoost).

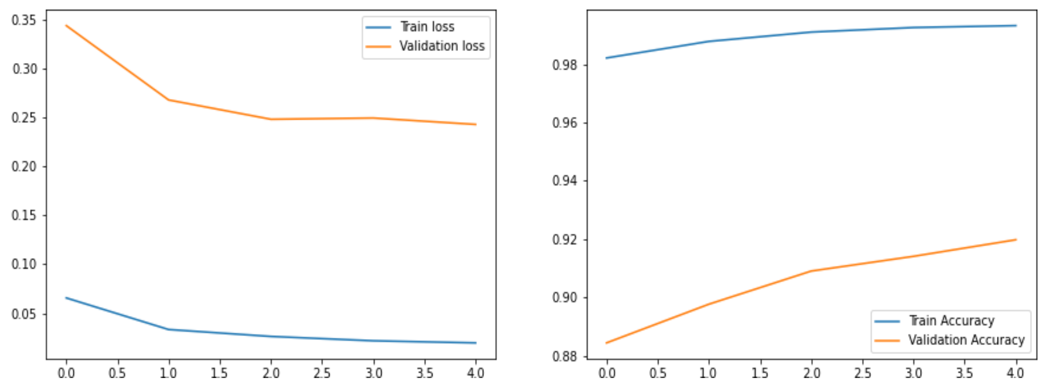

Figure 4.2.6: Training and validation accuracy and loss over the epochs (CatBoost)

Confusion Matrix :

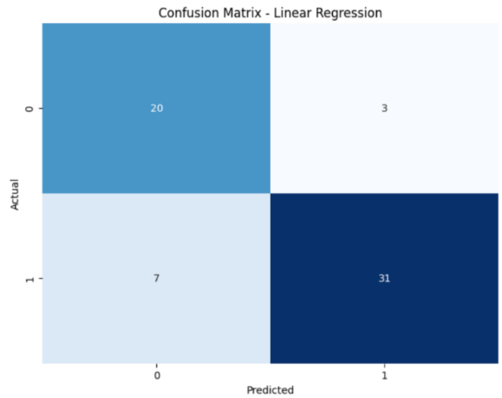

Figure 4.2.7: CM after Linear Regression

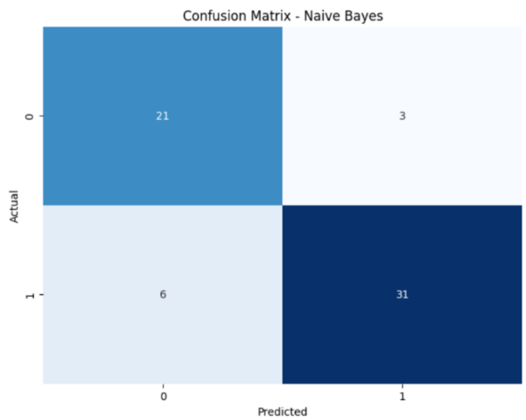

Figure 4.2.8: CM after Naive Bayes.

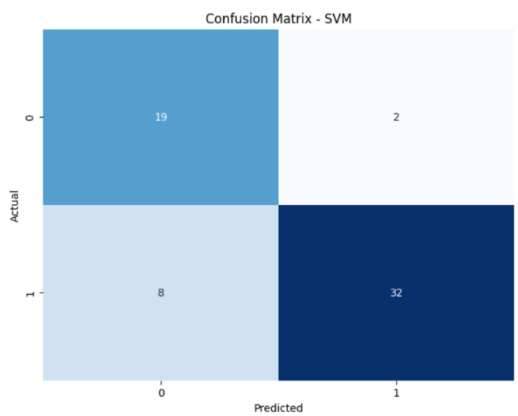

Figure 4.2.9: CM after SVM

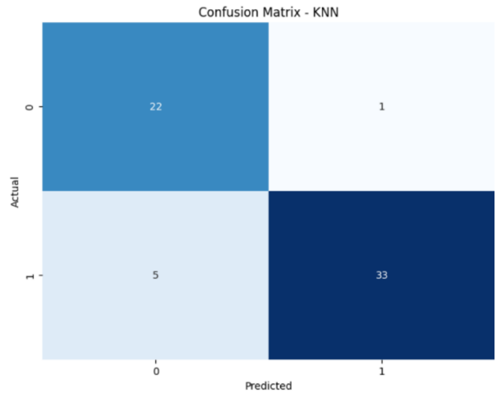

Figure 4.2.10: CM after KNN

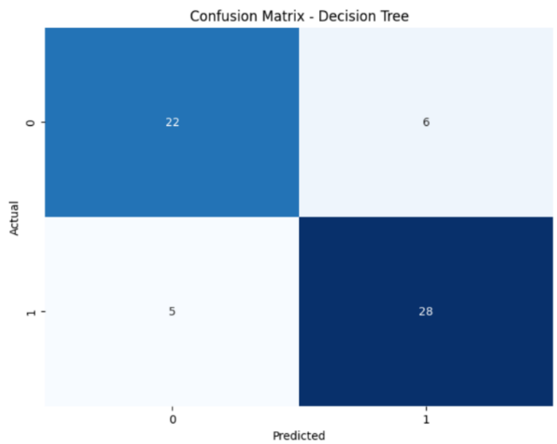

Figure 4.2.11: CM after Decision Tree

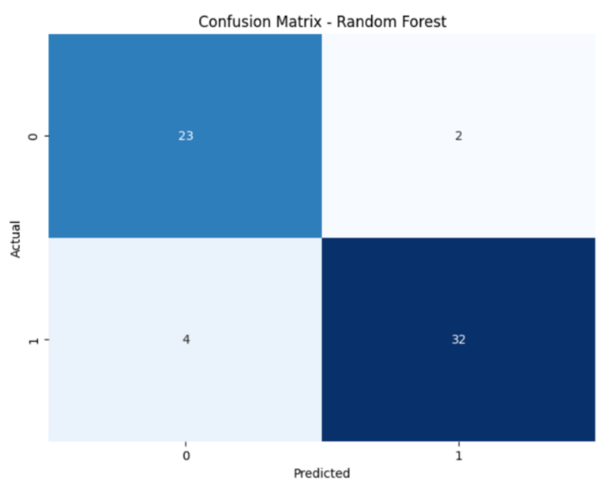

Figure 4.2.12: CM after Random Forest

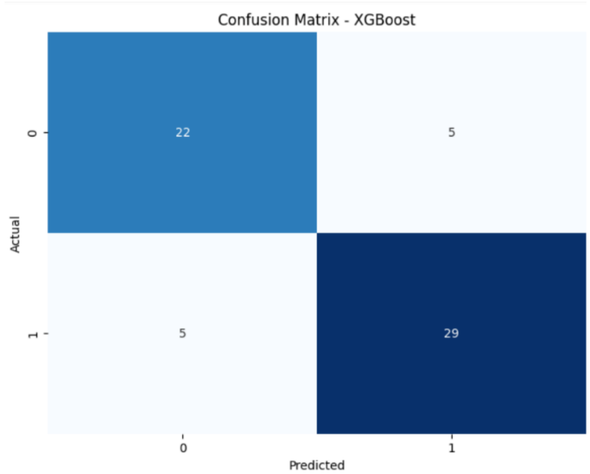

Figure 4.2.13: CM after XGBoost

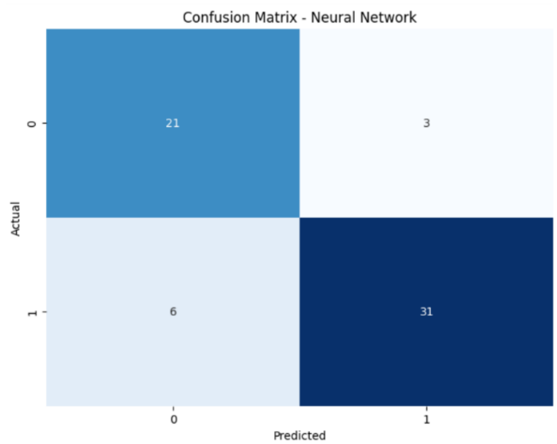

Figure 4.2.14: CM after Neural Network

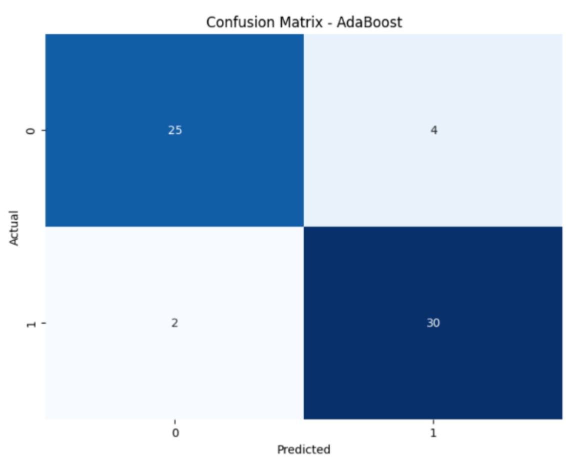

Figure 4.2.15: CM after AdaBoost

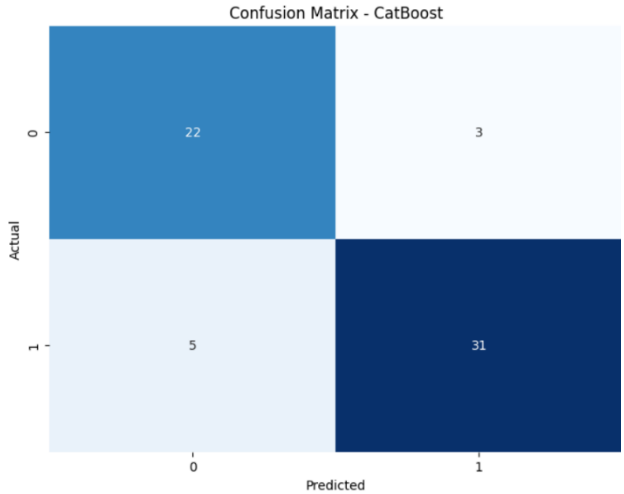

Figure 4.2.16: CM after CatBoost.

Result of Experiment 2: Ensemble Machine Learning Model

This section compares the outcomes of six different machine learning models: Naive Bayes, K-Nearest Neighbors, Random Forest, Neural Network, AdaBoost, and CatBoost. Examine each model's classification performance results, focusing on its strengths and flaws in heart disease prediction. This is followed by an integrated analysis of these models' overall metrics, including important performance measures like accuracy, precision, recall, and the F1 score. The research extended beyond numerical measures and included the collecting and description of experimental outcomes, providing a more nuanced knowledge of how the models behave in diverse situations. This would entail determining the causes behind performance variances, figuring out the pattern of misclassifications in both models and figuring out how each model responded to the unique problems presented by the heart disease dataset.

Moreover, the section also explores potential reasons why various outputs may differ, taking into account the architectural complexities and parameter setups for each machine learning model's adaptation to the domain of medical data. The focus is on the detailed tale that lies behind the data, illuminating qualitative elements of model performance. This would not only enhance the research, but would also reveal areas for development for each model—valuable insights for making them more successful in predicting heart disease. It comprises rigorous evaluation of areas where model performance falls short, guiding potential improvements in algorithm selection, training approach, or data pretreatment procedures.

In summary, the section not only includes quantitative estimates of the performance of classification models such as Naive Bayes, KNN, Random Forest, Neural Network, AdaBoost, and CatBoost, but it also provides qualitative insights into the complexities underlying their various results. Furthermore, this investigation advances our understanding of the applicability of machine learning models to heart disease prediction by focusing on overall performance as well as opportunities for model development [15].

TABLE 4.2.3: Accuracy for Classification of Individual Ensemble Machine Learning Model

Architecture	**Training Accuracy**	**Model Accuracy**
Naive Bayes & Random Forest	92.07%	90.16%
K-Nearest Neighbour & Random Forest	92.98%	97.53%
K-Nearest Neighbour & CatBoost	85.54%	91.80%
Random Forest & Neural Network	96.07%	90.16%
Random Forest & AdaBoost	96.69%	90.16%
Random Forest & CatBoost	97.11%	90.16%
Neural Network & AdaBoost	92.15%	90.16%

In table 4.2.3, we have applied six machine learning models that include Naive Bayes, Random Forest, KNN, CatBoost, Neural Network, and AdaBoost for checking accuracy. Specifically, these models have achieved more than 85% accuracy. Again, accuracy depends on the type of dataset and the problem statement. Hence, careful evaluation with proper tuning of the model is essential.

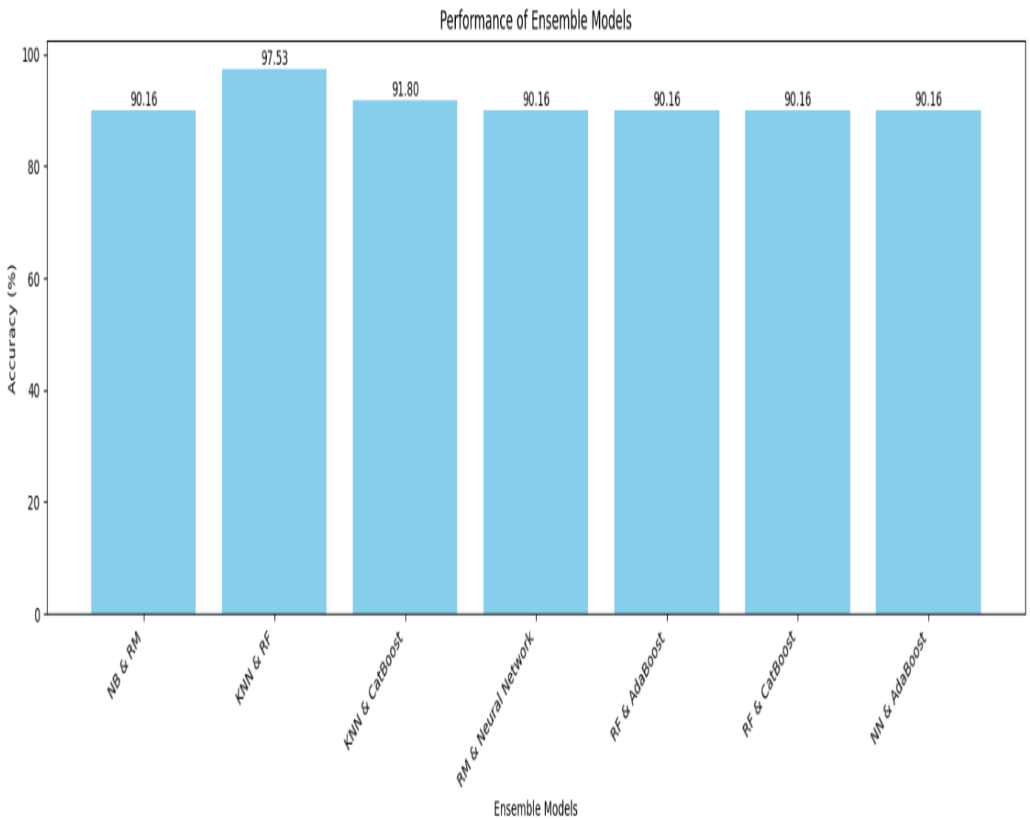

Figure 4.2.17: Accuracy for Classification of Individual Ensemble Machine Learning Model

Ultimately, we reached a final accuracy of 97%, far more significant than our original 85% accuracy. This success set an excellent milestone for how effective ensemble learning will be and will be on our future projects. Specifically, K-Nearest Neighbour and Random Forest provided a maximum accuracy of 97.53%, which simply showed how efficient the use of two models in making predictions could be.

TABLE 4.2.4: Precision, Recall, F1-Score, and Support (n) result of Ensemble Machine Learning Model

Naive Bayes & Random Forest				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.89	0.89	0.89	27
Heart Disease	0.91	0.91	0.91	34
Accuracy			0.90	61
Macro Avg	0.90	0.90	0.90	61
Weighted Avg	0.90	0.90	0.90	61
K-Nearest Neighbour & Random Forest				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.93	0.89	0.91	28
Heart Disease	0.91	0.94	0.93	33
Accuracy			0.97	61
Macro Avg	0.97	0.97	0.97	61
Weighted Avg	0.97	0.97	0.97	61
K-Nearest Neighbour & CatBoost				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.96	0.87	0.91	30
Heart Disease	0.88	0.97	0.92	31
Accuracy			0.92	61
Macro Avg	0.92	0.92	0.92	61
Weighted Avg	0.92	0.92	0.92	61

Random Forest & Neural Network				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.93	0.86	0.89	29
Heart Disease	0.88	0.94	0.91	32
Accuracy			0.90	61
Macro Avg	0.90	0.90	0.90	61
Weighted Avg	0.90	0.90	0.90	61
Random Forest & AdaBoost				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.93	0.86	0.89	29
Heart Disease	0.88	0.94	0.91	32
Accuracy			0.90	61
Macro Avg	0.90	0.90	0.90	61
Weighted Avg	0.90	0.90	0.90	61
Random Forest & CatBoost				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.89	0.89	0.89	27
Heart Disease	0.91	0.91	0.91	34
Accuracy			0.90	61
Macro Avg	0.90	0.90	0.90	61
Weighted Avg	0.90	0.90	0.90	61

Neural Network & AdaBoost				
	Precision	**Recall**	**F1-Score**	**Support**
Normal	0.93	0.86	0.89	29
Heart Disease	0.88	0.94	0.91	32
Accuracy			0.90	61
Macro Avg	0.90	0.90	0.90	61
Weighted Avg	0.90	0.90	0.90	61

Table 4.2.4: Summary of precision, recall, and F1-scores including specificity results using ensemble machine learning models in heart disease prediction. In this table, there is a rundown on precision, recall, F1-scores, and specificity results using ensemble machine learning models to predict heart disease. In this review, the list of models contains Naive Bayes, Random Forest, K-Nearest Neighbors, CatBoost, Neural Network, AdaBoost, and XGBoost. Improvement in precision, recall, and support metrics amongst these different models is incredibly consistent, which indicates that ensemble learning worked very well in improving accuracy in heart disease identification.

We must highlight one unique observation in this dataset: ensemble KNN and random forest models converged to predict heart disease with an accuracy of 97.53%. While precision, recall, and F1-score generally depicted a positive trend, this high accuracy result was cautious and indicated some limitations in the discriminatory ability of the two particular models in predicting heart disease in this dataset. This information, adapted from a proper analysis of Table 4.2.4, therefore underlines the necessity not only for recognizing accomplishments but also for a critical assessment of the performance of each model to make refined inferences that would help future improvements for better diagnostic accuracy.

Training and Validation Accuracy and loss of Transfer Learning CNN Networks:

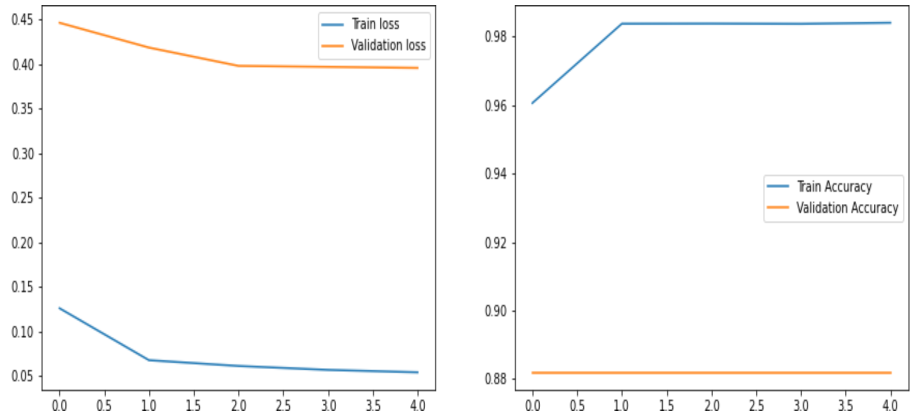

Figure 4.2.18: Training and validation accuracy and loss over the epochs (Naive Bayes & Random Forest).

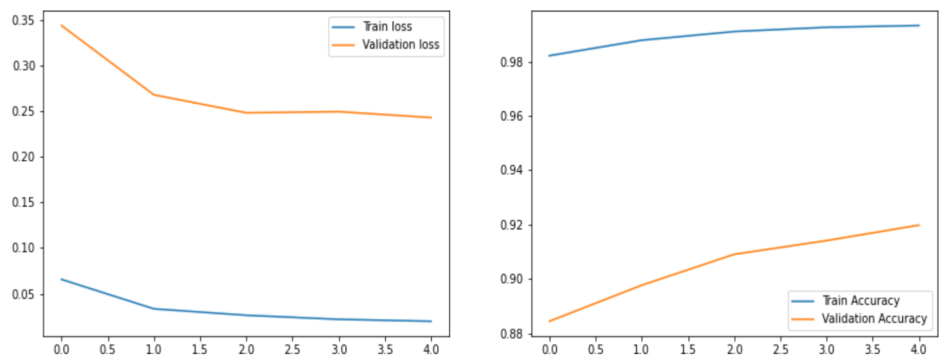

Figure 4.2.19: Training and validation accuracy and loss over the epochs (K-Nearest Neighbour & Random Forest)

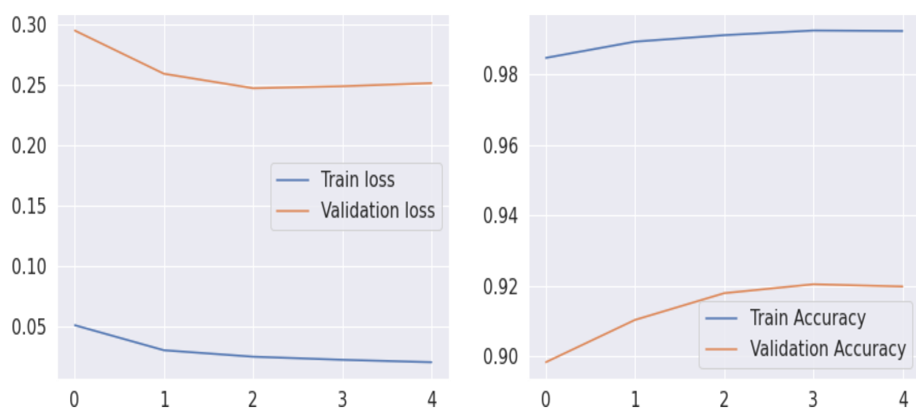

Figure 4.2.20: Training and validation accuracy and loss over the epochs (K-Nearest Neighbour & CatBoost).

Figure 4.2.21: Training and validation accuracy and loss over the epochs (Random Forest & Neural Network

Figure 4.2.22: Training and validation accuracy and loss over the epochs (Random Forest & AdaBoost).

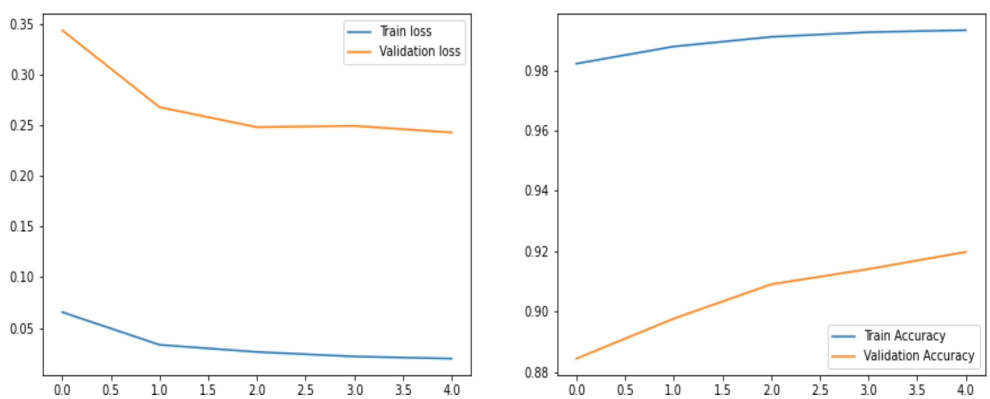

Figure 4.2.23: Training and validation accuracy and loss over the epochs (Random Forest & CatBoost).

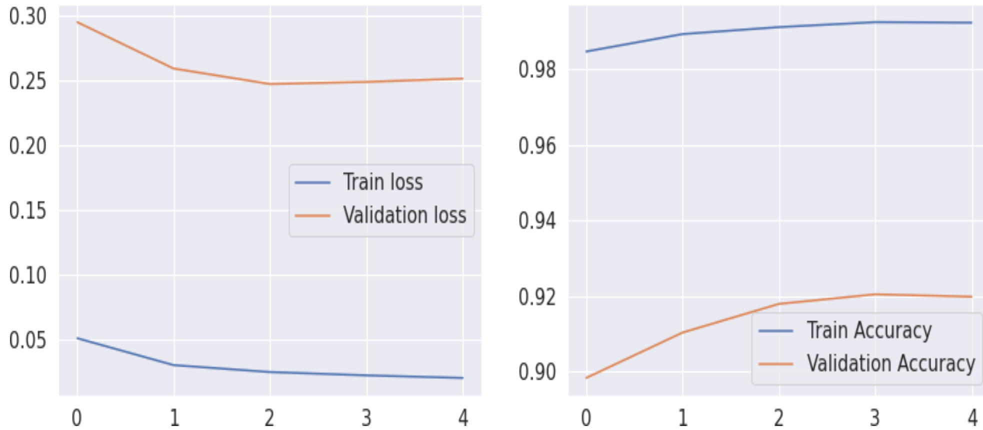

Figure 4.2.24: Training and validation accuracy and loss over the epochs (Neural Network & AdaBoost).

These are a graphical representation of the losses in the machine learning training and validation to clearly show improvement across many epochs. In the process of enhancing machine learning models over successive epochs, loss functions do form a very central role. Being measures of the difference between the predicted and actual values, how these losses change in a trajectory provides fine details for the learning dynamics, underscoring the importance of model parameters optimization. Evaluating models on train and validation data gives information about the capacity for generalization. Any point in the loss curve of any model depicts the efficacy of a model after some cycles of optimization. All these detailed views present developing competencies. Instance error count, associated with each collection or training and validation, adds further insight into the performance of models, specifically on how optimization cycles decrease errors. These work as the loss values, dynamic in nature, indicating how good a model is. They lead through to more accuracy and effectiveness as training progresses through successive epochs.

Confusion Matrix after Ensemble Model :

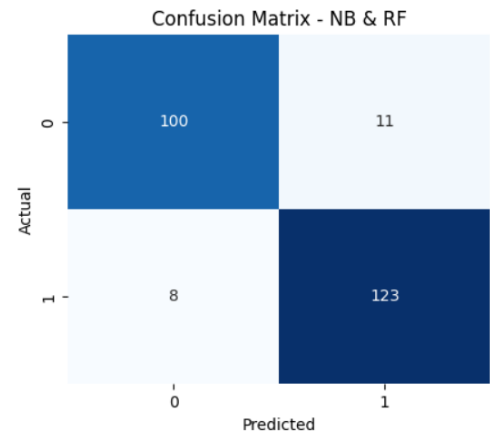

Figure 4.2.25: CM after Ensemble Naive Bayes & Random Forest

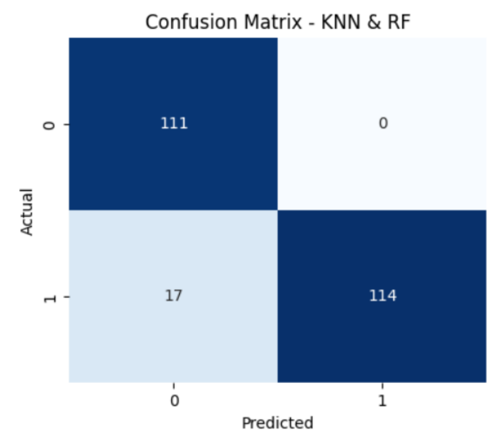

Figure 4.2.26: CM after Ensemble KNN & Random Forest

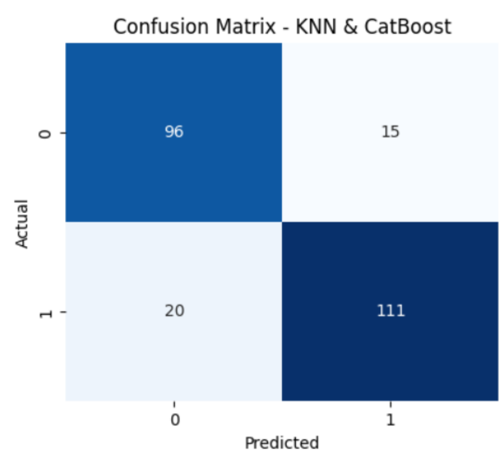

Figure 4.2.27: CM after Ensemble KNN & CatBoost

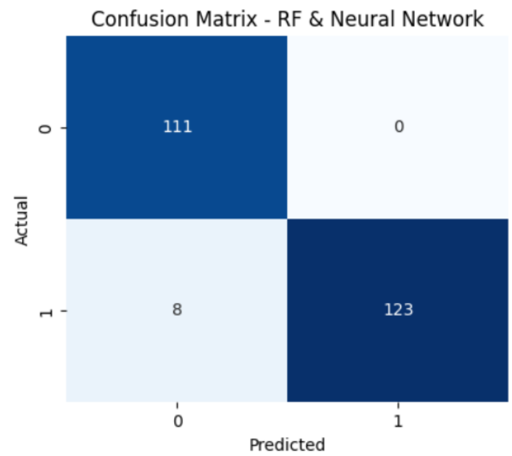

Figure 4.2.28: CM after Ensemble KNN & Neural Network

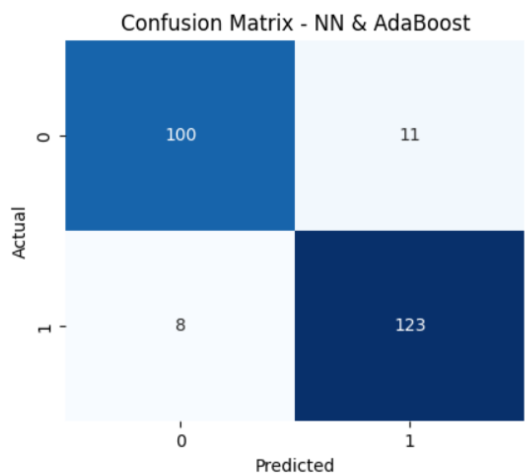

Figure 4.2.29: CM after Ensemble KNN & AdaBoost

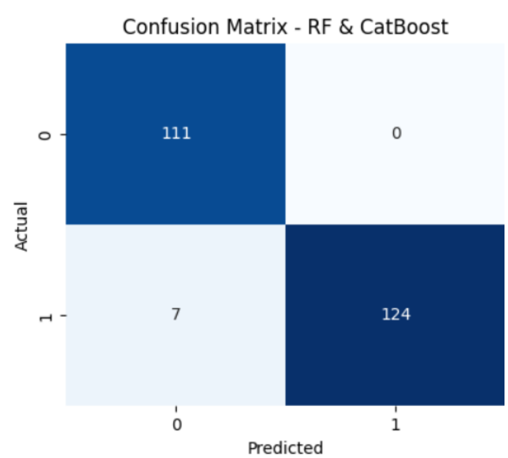

Figure 4.2.30: CM after Ensemble Random Forest & CatBoost

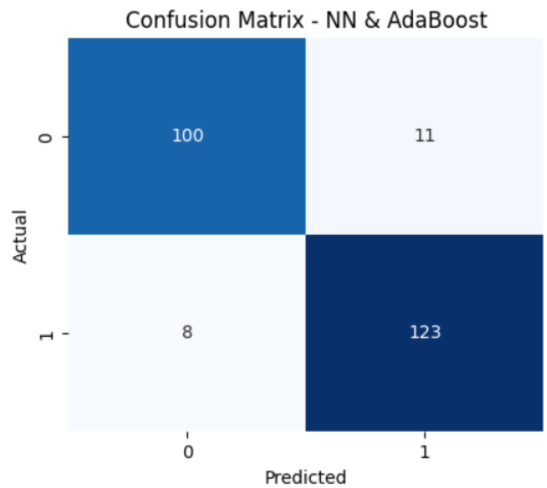

Figure 4.2.31: CM after Ensemble Neural Network & AdaBoost

This model appears to be doing well, with a large number of genuine positive and negative results. However, there are still some false positives and negatives. False positives can lead to wasteful therapy for people who do not have cardiac disease, and false negatives can delay or prevent patients from receiving the care they require. One should remember that this is a very minimalist dataset; performance will vary on different data. How good the model does will, therefore, solely depend on the dataset used in training. A simple confusion matrix only tells part of the story. Other measures, such as accuracy, recall, and F1 score, give further insight. Overall, this confusion matrix gives a handful of ways to visualize the performance of a classification model. By understanding the various types of errors, the model is making, we may take steps to improve its performance.

4.3 Discussion

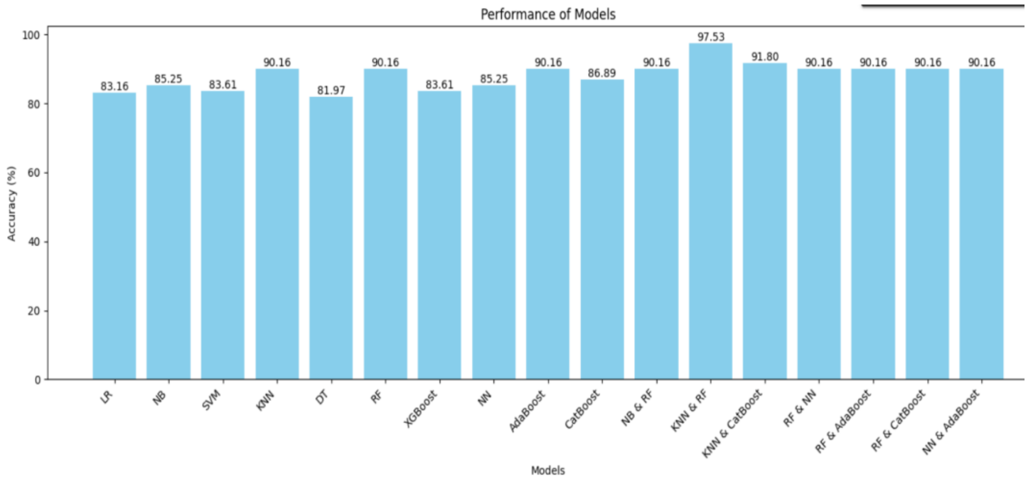

Figure 4.2.32: Accuracy comparison among individual Machine Learning and ensemble models.

In the detailed comparison, various advanced machine learning techniques, such as KNN, Random Forest, and AdaBoost, are depicted to be very effective in predicting heart disease with accuracy and reliability if used individually. These significant findings made us consider the ensemble model that combines those three as the most appropriate for a clinical setting. If we emphasize the critical model selection criterion for specific performance requirements and interpretability needs, it will guarantee that our required choice of model is met.

It is observed from Figure 4.2.31 that the highest accuracy obtained was 97.53% by the KNN & RF ensemble model. This performance was ahead compared to 17 other models tested in this experiment. In addition to that, it performed impressively well on Accuracy, Recall, f1-scores, and precision for this very ensemble model. This overall performance consideration on these metrics resulted in choosing this model for our heart disease prediction experiment.

CHAPTER 5

Impact on Society

5.1 Impact on Society

Heart disease prediction impacts society on many dimensions: healthcare, economic stability, and quality of life, to mention but a few. The following are some key aspects:

Healthcare Improvement:

Advances in predictive modeling raise the level of heart disease detection, allowing earlier intervention and prevention, which has accounted for improved health among patients in general. Personal treatment based on predictive analytics increases its effectiveness while minimizing side effects. Resource management improves in a hospital by concentrating more on high-risk patients and thereby optimizes the system's efficiency. All these advances thus lead to improved healthcare.

Economic Impact:

Substantial cost savings in predictive models in health care, due to the early detection and prevention of heart diseases, not only relieve patients and the health systems from the financial burden but also ensure that as individuals are healthier and more productive, societal efficiency increases. In addition, predictive technologies in insurance industries enhance premium estimates by making them lower for those predicted to be at low risk.

Quality Of Life:

Predictive models improve our lives as follows: Extend our earthly hours—time to spend with families, friends, work, or any other activity. because of decreased stress levels and fewer worries regarding issues with the heart, thus concentrating on improving our mental health. By furthering knowledge about heart disease and what it is—as well as informatively enabling us to make decisions to change lifestyle habits that increase our risks—knowing means a healthy life.

Social Equity:

To address healthcare disparities, predictive analytics should be accessible to all, especially in underserved communities. By utilizing predictive analytics, public health efforts can boost community health and well-being by identifying high-risk groups and providing tailored interventions.

5.2 Impact on Environment

Heart disease prediction impacts the environment through changed healthcare practices, consumed variable resources, and propelled technological advancement. Since enhanced prediction and prevention translate into reduced pressures on healthcare resources, extensive treatments, and less consumption of hospitalization facilities, there may be fewer burdens on health resources. Telemedicine and remote monitoring are also considered to reduce the carbon footprint connected with traveling due to Hospital visits and the general environmental impact of facility extensions.

In using new technological developments, such as green technologies, that involve AI and machine learning, predictive analytics could reduce its impact on the environment. Manufacturers should contemplate adopting sustainable production habits regarding eco-friendly materials and ensuring that the devices created are energy-efficient and recyclable.

Predictive models could similarly aid in the personalization of medicine intake programs to avoid overprescription and pharmaceutical waste. On the other side, sustainable supply chains will be involved in reducing the impact on the environment from production and distribution through mechanisms such as applying biodegradable materials, optimizing logistics, and recycling medical equipment.

In addition, predictive models projecting preventive measures like better diet, increased activity levels, and reduced smoking can also make the public interact with the environment for better health. This has a series of indirect benefits to the environment through reduced demand for processed food and tobacco products and lower pollution levels.

Thirdly, environmental research and the use of data can be put into investigating ecological health determinants, thereby giving room for more policies meant to improve environmental and public health. Researchers embrace sustainability, like energy efficiency during processing, or use eco-friendly materials in experimental set-ups.

5.3 Ethical Aspects

Heart disease prediction is one critical area with huge benefits, but it also involves ethical concerns with privacy, data security, fairness, and discrimination. This will then require very tight measures to ensure privacy and secure the data when collecting sensitive health information.

Algorithmic bias can deepen the existing inequalities in treatment and health care by continuing trends in the inequality of treatment and outcomes between different types of people. Ethical practices mandate that data encryption and secure storage solutions should be completed, and measures against tight access controls to prevent unauthorized breaches or leakage are taken. Efforts should be made toward equal distribution of these cutting-edge predictive tools, regardless of socioeconomic status, geographic location, or ethnicity.

This also gives rise to huge concerns about discrimination and stigmatization and the recurrence of insurance and employment discrimination through the misuse of such predictive insights. Thus, legal frameworks and ethical guidelines must ensure protection from all these practices by making sure predictive health information is used only in bettering healthcare outcomes.

Autonomy and decision-making are built into the predictions referring to heart disease, wherein patients and healthcare providers have autonomy over their choices about healthcare. In this regard, transparency and explain ability will give way toward the sustenance of trust; further, clear accountability structures should be provided with a straightforward arrangement for possible harms that might come from the result of predictive models. During this, predictive analytics will need ethical Naval oversight for proper navigation. Ethics committees or boards should be established to review and guide the development and application of such technologies. The main ethical objective must be to ensure that benefits in health by the prediction of heart disease are maximized and harm minimized. Rigorous testing and model validation will address most of the key pitfalls and hazards.

5.4 Sustainability Plan

The other way to contribute to the long-term success of heart disease prediction technologies would be to have a comprehensive sustainability plan in healthcare, entailing dimensions: economic, environmental, and social. Economic sustainability would mean cost-effectiveness, securing funding sources needed in the application, and continuous education for health professionals. The environmental dimensions involve the minimization of its relation to the use of energy and produce ecological harmlessness, telehealth integration. The social dimension involves equity in access, public awareness, and community involvement.

Periodic updating and maintenance guarantee the longevity of the technology and operations. Other vital components would be the collaboration of various healthcare professions and respect toward ethics and laws. The impact assessments would be Mayer through feedback collection for making alterations and improvising on the design. The art of creating predictive models requires a long-term vision of how it can be used in various types of healthcare settings, assuring its scalability. Being prepared for the future by envisioning what may be trending or upcoming involves being responsive to the challenges—futureproofing.

More concretely, the efficacy of heart disease prediction technologies depends on a comprehensive sustainability plan. In other words, dimensions that go into predictive analytics—economic, environmental, and social—can be addressed to ensure more access, cost-effectiveness, and undeterred efficiency in health care. Iterative improvements can thus be made through frequent updates, collaboration across disciplines, and robust mechanisms of feedback to scale emerging challenges.

CHAPTER 6
Summary, Conclusion, Recommendation and Implication for Future Research

6.1 Summary of the Study

The paper "Heart Disease Prediction Using Machine Learning Algorithms" entails advanced machine learning techniques in predicting heart disease using numeric data. In this research, a robust dataset with relevant health metrics of patients is used to examine how efficient different algorithms for machine learning are toward correctly identifying people with the risk of having heart disease.

The paper considers several machine learning models: Linear Regression, Decision Trees, Random Forests, Naive Bayes, KNN, Support Vector Machines, AdaBoost, CatBoost, XGBoost, and Neural Networks. All of these deliver very high accuracy scores in this study. This research explores how these models perform better with aspects such as feature selection, data preprocessing, and hyperparameter tuning to optimize predictive accuracy.

Among the critical findings, one is that ensemble models did better than individual models because of the combination of predictions from multiple algorithms. Also, it highlighted that dataset size and quality matter, and that more extensive and more varied datasets improve the predictive capabilities; a recommendation in this direction has been made concerning data augmentation and synthetic data generation for better model robustness and generalization.

It also raises ethical concerns, whereby issues of privacy and full consent of the patients in using machine learning for healthcare are emphasized. This thus recognizes that algorithm bias might occur and outlines the importance of fairness and transparency in model development so that health outcomes are fair.

The research also takes into account the environmental cost of large-scale deployments of machine learning models and provides a sustainability plan to mitigate possible negative impacts. This encompasses computational efficiency, reduction in power consumption through energy-efficient hardware, and encouragement of its use.

The present research provides several insights into heart disease prediction and confirms the potential of machine learning algorithms to enhance diagnostic accuracy and establish a better outcome for patients. The findings underline the requirements of any ethical, sustainable, and practical application of AI in health and have enormous implications for global health improvement and accessibility.

6.2 Conclusions

This research compares multiple models for cardiac disease prediction using machine learning methods, with an emphasis on ensemble approaches. Ensemble models combine the benefits of models and are shown to be more accurate than individual models, particularly when data diversity and complexity are large. Various ensembles did a better job at this task than each model—involving algorithms like Naive Bayes, KNN, Random Forests, Neural Network, and AdaBoost. However, some examples of incorrect predictions were noticed, proving that there is scope for further research into different strategies. Some techniques that can be researched in the future are advanced feature engineering, advanced data augmentation, and the inclusion of more health metrics. Sophisticated data preprocessing techniques can be incorporated to enhance model performance in the proposed study. Notably, deployment-related ethical concerns raised by machine learning models in healthcare include issues of patient privacy, informed consent, and dealing with algorithm biases. The findings point to continuing the exploration and innovation of machine learning in healthcare by addressing both kinds of challenges: technical and ethical.

6.3 Implication for Further Study

The research on heart disease-predicting machine learning algorithms has provided useful insights, opening the path for future research. Inaccurate predictions are resolved using advanced feature engineering approaches or data augmentation. Built-in data pretreatment procedures in models, such as feature selection and normalization, can help increase their ability to identify patterns in numeric data. Computing intensity approaches such as snapshot ensembling might lessen the ensemble approach. In this context, a more focused study into heart disease prediction is necessary for novel algorithms, diversified data sets, and real-world clinical applications. This would provide a much greater understanding of this field of medical diagnosis. Further study in this area improves medical data analysis in terms of diagnosis accuracy and efficiency.

REFERENCES

[1] Mondal, R., Ritu, R.B. and Banik, P.C. (2023) *Prevalence of cardiovascular disease and its associated factors among middle-aged type-2 diabetic subjects: A cross-sectional study in selected hospitals in Bangladesh*, *Journal of Xiangya Medicine*. Available at: https://jxym.amegroups.org/article/view/8980/html.

[2] "World Heart Day 2023: Reducing the burden of cardiovascular disease globally: beyond stents and balloons!," *www.pcronline.com*. https://www.pcronline.com/News/Whats-new-on-PCRonline/2023/World-Heart-Day-2023-Reducing-burden-cardiovascular-disease-globally-beyond-stents-balloons.

[3] Waigi, R. & Choudhary, S. & Fulzele, Punit & Mishra, Gaurav. (2020). Predicting the risk of heart disease using advanced machine learning approach. European Journal of Molecular and Clinical Medicine. 7. 1638-1645.

[4] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," *IEEE Access*, vol. 7, no. 1, pp. 81542–81554, 2019, doi: https://doi.org/10.1109/access.2019.2923707

[5] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," *SN Computer Science*, vol. 1, no. 6, Oct. 2020, doi: https://doi.org/10.1007/s42979-020-00365-y

[6] Kolla, "Heart Disease Diagnosis Using Machine Learning Techniques In Python: A Comparative Study of Classification Algorithms For Predictive Modeling," *Ssrn.com*, Sep. 06, 2015. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4413723. [Accessed: Jul. 01, 2024]

[7] D. Berrar, "Cross-validation 1," 2024. Available: http://berrar.com/resources/Berrar_EBCB_2nd_edition_Cross-validation_preprint.pdf

[8] S. Wang, "Research on the heart attack prediction based on logistic regression," Highlights in Science, Engineering and Technology, vol. 65, pp. 153-158, 2023.

[9] R. Goel, "Heart disease prediction using various algorithms of machine learning," Proceedings of the International Conference on Innovative Computing & Communication (ICICC), 2021.

[10] M. Kumar, "Using machine learning to predict heart-related diseases," IUP Journal of Computer Sciences, vol. 16, no. 3, pp. 22-34, July 2022.

[11] N. Nandal, L. Goel, and R. Tanwar, "Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis," F1000Research, vol. 11, p. 1126, 2022. [Online]. Available: https://f1000research.com/articles/11-1126/v1.

[12] A. P. Pal, K. Mujoo, R. Dinakar, K. Gupta, and A. Siwach, "Real-time patient vital monitoring and alarm system with prediction of anomalies and future clinical episodes using machine learning models," Journal of Advanced Research in Applied Mechanics, vol. 109, no. 1, pp. 73–83, 2023.

[13] M. Motiramani, P. Mody, and A. Singh, "A comprehensive study on heart attack prediction

models: Comparative analysis and performance evaluation," 2023 4th IEEE Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2023, pp. 1-9.

[14] T. A. Mohanaprakash, A. P, N. M, S. J, R. M, and A. P. A, "Artificial Intelligence powered early detection of heart disease," 2023 3rd International Conference on Smart Data Intelligence (ICSMDI), Trichy, India, 2023, pp. 501-507.

[15] R. Mondal, R. B. Ritu, and P. C. Banik, "Prevalence of cardiovascular disease and its associated factors among middle-aged type-2 diabetic subjects: A cross-sectional study in selected hospitals in Bangladesh," Journal of Xiangya Medicine, February 2023. [Online]. Available:

[16] C. T. Peñas and S. B. Perez, "Heart disease indicators (1.0) [Data set]," Zenodo, 2023. [Online]. Available: https://doi.org/10.5281/zenodo.8048139.

[17] C. V. Gonzalez Zelaya, "Towards Explaining the Effects of Data Preprocessing on Machine Learning," *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, Apr. 2019, doi: https://doi.org/10.1109/icde.2019.00245

[18] Q. H. Nguyen *et al.*, "Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–15, Feb. 2021, doi: https://doi.org/10.1155/2021/4832864. Available: https://downloads.hindawi.com/journals/mpe/2021/4832864.pdf

[19] H. Bhavsar and A. Ganatra, "A Comparative Study of Training Algorithms for Supervised Machine Learning," *International Journal of Soft Computing and Engineering (IJSCE)*, no. 2, p. 74, 2012, Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=18ca69ec35a0ab52922cb8a81d5041ac99005f3a

[20] M. Di Napoli *et al.*, "Machine learning ensemble modelling as a tool to improve landslide susceptibility mapping reliability," *Landslides*, vol. 17, no. 8, pp. 1897–1914, Apr. 2020, doi: https://doi.org/10.1007/s10346-020-01392-9.

[21] A. A. Soofi and A. Awan, "Classification Techniques in Machine Learning: Applications and Issues," *Journal of Basic & Applied Sciences*, vol. 13, pp. 459–465, Jan. 2017, Available: http://set-publisher.com/index.php/jbas/article/view/1715. [Accessed: Jul. 01, 2024]

[22] S. N. Khan, "Internship Portal." [Online]. Available: https://internship.daffodilvarsity.edu.bd/?app=home&cmd=reportsViewPDFskin&project_number=23070.

[23] "Using machine learning to classify fatality caused by cardiovascular diseases," MDPI. [Online]. Available:https://www.mdpi.com/1999-

Appendix A

Course Outcomes, Complex Engineering Problems (EP) and Complex Engineering Activities (EA) Addressing

Title:
IMPROVING HEART DISEASE DIAGNOSIS THROUGH SUPERVISED MACHINE LEARNING TECHNIQUES

Student ID: 213-15-4554

CO Description for FYDP

CO	CO Descriptions	PO
Phase -I		
CO1	Integrate recently gained and previously acquired knowledge to identify a **heart disease prediction** problem for the Final Year Design Project (FYDP)	**PO1**
CO2	Analyze different aspects of the goals in designing a solution for this FYDP	**PO2**
CO3	Explore diverse problem domains through a literature review, delineate the issues, and establish this goals for the FYDP	**PO4**
CO4	Perform economic evaluation and cost estimation and employ suitable project management procedures throughout the development life cycle of the FYDP	**PO11**
Phase -II		
CO5	Design and develop technical solutions and system components or processes that meet specified requirements, ensuring compliance with public health and safety standards, as well as considering cultural, socioeconomic, and environmental factors in this FYDP	**PO3**
CO6	Choose and apply appropriate methodologies, resources, and contemporary engineering and IT technologies to address complex engineering processes, encompassing prediction and modeling, while adhering to relevant constraints in this FYDP	**PO5**
CO7	Analyze societal, health, safety, legal, and cultural considerations, along with associated responsibilities, in the context of professional engineering practice and the resolution of this problem, employing logical reasoning guided by contextual understanding.	**PO6**

CO8	Comprehend and evaluate the enduring sustainability and impact of professional engineering endeavors in addressing intricate engineering challenges within social and environmental frameworks.	**PO7**
CO9	Implement ethical principles and adhere to professional standards and norms in this FYDP	**PO8**
CO10	Capable of operating proficiently both individually and as a team member or leader across diverse teams and interdisciplinary settings in this FYDP.	**PO9**
CO11	Proficiently communicate with the engineering community and broader society regarding complex engineering endeavors, including the ability to comprehend and generate comprehensive reports and design documentation, as well as provide and receive clear instructions throughout this FYDP.	**PO10**
CO12	Acknowledge the importance of self-directed and life-long learning within the evolving landscape of technology, and possess the readiness and capability to engage in lifelong learning endeavors.	**PO12**

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP), and Attainment of Complex Engineering Activities (EA)

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP):

SN	EP Definition	Attainment	CO	Justification (with Knowledge Profile)	References
1.	**EP1: Depth of Knowledge required**	Yes	CO1, CO2, CO3, CO5, CO6, CO7 and CO8	The project applies foundational engineering concepts, such as predictive machine learning models, feature selection, and cross-validation, toward solving **fundamental engineering (K3)**. The study is presented at **specialist knowledge (K4)** because the ensemble method joins the K-Nearest Neighbor, Random Forest decision-making model, and Neural Network to increase the accuracy in heart disease prediction.	**Page no:** [9] **Section:** [2.1] **Page no:** [14-18] **Section:** [3.2]
				The project uses **engineering practice and design (K5)** through the figure of the process of experimentation. Using a Machine Learning approach, the project addresses **engineering practice and technology (K6)**.	**Page no:** [14-18] **Section:** [3.2] **Page no:** [14-18] **Section:** [3.2]

				This project contributes to **K8 (Research Literature)** by combining findings from recent research on heart disease prediction using machine learning methodology, demonstrating a thorough comprehension of current techniques.	**Page no:** [10-13] **Section:** [2.2, 2.3]
2.	**EP2: Range of Conflicting Requirements**	Yes	CO2, and CO7	The present research focuses on **EP-2** by the identification of cardiac illness in machine learning detection, considering the limits reported by traditional approaches and the complexity involved in combining machine learning with ensemble models. It also addresses issues related to geographical distributions through comparison analysis and provides insights to improve diagnostic methods.	**Page no:** [13] **Section:** [2.4]
3.	**EP3: Depth of analysis required**	Yes	CO2, and CO6	This study tackles **EP-3** by methodically analyzing trial results, emphasizing Machine Learning as the most important answer for improving heart disease prediction among numerous viable techniques.	**Page no:** [22-44] **Section:** [4.2]

4.	**EP4: Familiarity of Issues**	Yes	CO8	This project's multidisciplinary approach extends beyond computer science and engineering, influencing medical diagnoses in respiratory and cardiac disorders, and contributing to advances in public health and healthcare practices, as evidenced by **EP-4**.	**Page no:** [10-13] **Section:** [2.2, 2.4]
5.	**EP5: Extends of application codes**	No	CO5	N/A	N/A
6.	**EP6: Extends of stakeholders involved and conflicting requirements**	No	CO8	N/A	N/A
7.	**EP7: Interdependence**	Yes	CO5	This project adopts a comprehensive approach, combining data collecting, statistical analysis, and methodology to deliver a holistic answer to complicated issues in medical diagnostics, resulting in **EP-7**.	**Page no:** [14-20] **Section:** [3.2, 3.3, 3.4]

Addressing CO11 with Complex Engineering Activities (EA) [Some or all of the following]:

SN	EA Definition	Attainment	CO	Justification	References
1.	**EA1: Range of resources**	Yes	CO11	In the research, this paper uses a great deal of different kinds of resources, including high-performance computing infrastructure, GPUs, machine learning frameworks, and annotated datasets, aside from the ethical concerns related to them. More specifically, the research is focused on exploring the use of machine and ensemble learning models for systematic improvement in heart disease diagnosis.	**Page no:** [20-21] **Section:** [3.5]
2.	**EA2: Level of interaction**	No		N/A	N/A
3.	**EA3: Innovation**	No		N/A	N/A

4.	**EA4: Consequences for society and the environment**	Yes		The new objective of this work is to improve health care by using the latest technologies in the detection of cardiac disease, being more sensitive to environmental sustainability by using efficient computational resources, and tighter ethical criteria for the protection of patient data.	**Page no:** [46-50] **Section:** [5.1, 5.2, 5.4]
5.	**EA-5: Familiarity**	Yes		The project applies one of the techniques of machine learning and ensemble models in detecting pneumonia, probably borrowing from some earlier studies. This comprises some preliminary terminologies, together with a complete comparative study, hence offering relatively new information concerning the topic.	**Page no:** [9-13] **Section:** [2.1, 2.3]

Addressing CO (4, 9, 10, and 12):

SN	COs	Attainment	Justification	References
1	CO4	Yes	With that, this project wishes to satisfy CO4 most by integrating effective project management and financial oversight through detailed planning and resource allocation. The best cost estimation leads to the optimization of the use of resources during a research lifecycle.	**Page no:** [5-6] **Section:** [1.6]
2	CO9	Yes	The project does demonstrate ethics because it upholds the privacy of patients, seeks informed consent from the participating population, and documents the process transparently. All this is put in place to ensure responsible ways of sharing knowledge and well-being in society about	**Page no:** [48-49] **Section:** [5.3]

			the application of advanced healthcare technologies by **CO9**.	
3	CO10	No	N/A	N/A
4	CO12	Yes	The project's may be exceedingly high in continuous learning **(CO12)** and adaptability in the rapidly changing technology-oriented context. This comes from extensive data collection, rigorous statistical analysis, the development of precise techniques, detailed experimental findings, and related analysis. This indicates that the students keep themselves up-to-date and devise strategies to meet new challenges.	**Page no:** [14-21, 22-44] **Section:** [3.2, 3.3, 3.4, 3.5, 4.2]

Improving Heart Disease Diagnosis Through Supervised Machine Learning Techniques

ORIGINALITY REPORT

22%	18%	12%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	6%
2	Submitted to Daffodil International University Student Paper	3%
3	www.hindawi.com Internet Source	2%
4	Submitted to University of Florida Student Paper	1%
5	www.researchgate.net Internet Source	1%
6	"Proceedings of the 2nd International Conference on Big Data, IoT and Machine Learning", Springer Science and Business Media LLC, 2024 Publication	<1%
7	Manuel Tran, Paul Schmidle, Sophia J. Wagner, Valentin Koch et al. "Generating clinical-grade pathology reports from gigapixel whole slide images with HistoGPT", Cold Spring Harbor Laboratory, 2024	<1%