Early Heart Disease Prediction Using Supervised Machine Learning: A Performance Evaluation on UCI Data

By
Shah Alam Titu
152-15-5513

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

Md. Sazzadur Ahamed

Assistant Professor
Department of Computer Science and Engineering, Daffodil International University

Co-Supervised by

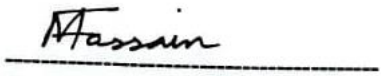
Mr. Abdus Sattar
Associate Professor
Department of Computer Science and Engineering, Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh
May 14, 2025

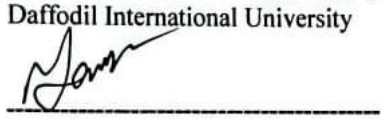
APPROVAL

This Project titled "**Early Heart Disease Prediction Using Supervised Machine Learning: A Performance Evaluation on UCI Data**" submitted by **Shah Alam Titu** 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 **14 May, 2025.**

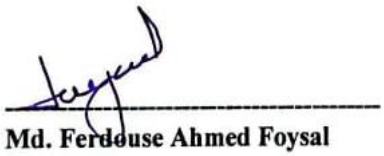
BOARD OF EXAMINERS



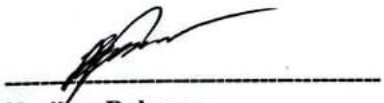
Dr. Md. Fokhray Hossain — **Chairman**
Professor and Chairman
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Tanzina Afroz Rimi — **Internal Examiner**
Sr. Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Md. Ferdouse Ahmed Foysal — **Internal Examiner**
Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

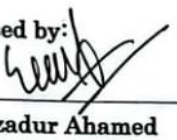


Nazibur Rahman — **External Examiner**
Technical Lead - Database Administrator
Telenor - Grameen Phone Account

DECLARATION

I hereby declare that this project has been done by me under the supervision of **Md. Sazzadur Ahamed, Assistant Professor**, Department of Computer Science and Engineering, Daffodil International University. I 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:



Md. Sazzadur Ahamed
Assistant Professor
Department of Computer Science and
Engineering Daffodil International University

Co-Supervised by:

Mr. Abdus Sattar
Associate Professor
Department of Computer Science and
Engineering Daffodil International University

Submitted by:

Shah Alam Titu
Student ID: 152-15-5513
Department of Computer Science and
Engineering Daffodil International University

ACKNOWLEDGEMENTS

This research project would not have reached completion without essential support and input from many persons over the past two semesters. I am truly thankful for the help and motivation offered by all who participated.

First and foremost, I express my heartfelt appreciation to the Almighty for His blessings, which allowed me to successfully undertake our Final Year Design Project (FYDP).

I thank especially Md. Sazzadur Ahamed, Assistant Professor, Department of Computer Science and Engineering, Daffodil International University, Dhaka. His deep expertise and enthusiasm for Machine Learning were very influential in shaping the type of research we pursued. I am grateful to him for his unwavering support, insightful suggestions, constant encouragement, and constructive suggestions. His dedication—from filtering preliminary drafts to reading with a critical view—was indispensable for the successful fruition of this work

I am grateful to the Head of the Department of Computer Science and Engineering for the support extended by him during the course of our project.I am also thankful to all the faculty members and staff of our department for their cooperation.

I am also thankful to fellows of Daffodil International University for their cooperation and nice discussions in due course of our studies

Finally, I would like to thank my parents for their unwavering love and understanding, and for being an endless source of inspiration for me.

ABSTRACT

Cardiovascular diseases (CVDs) remain a significant public health burden globally and coronary artery disease (CAD) is a major cause of heart attacks. In this paper, we explore the use of machine learning (ML) methods as tools for improving the early detection of CVD. Fifteen supervised ML and deep learning algorithms were used for the study in python 3.10. Model training and evaluation were performed by Scikit-learn, Pandas was used for descriptive statistical analysis and plots were generated with Matplotlib and Seaborn. The dataset utilized, which includes 920 instances, is a merged information from four different locations (Cleveland, Hungary, Switzerland and VA Long Beach) obtained from the UCI Machine Learning Repository. The Histogram-based Gradient Boosting Classifier performed best among the models tested, with an accuracy of 93.5 % and a 5-fold cross-validation accuracy of 92.4 %. It also obtained high scores in precision, recall, and F1 measurements for both classes and was able to classify people with and without the heart disease. These results highlight the promise of ML for early CVD detection and timely clinical interventions.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Heart disease is a general term that refers to all types of conditions that affect the heart including arrhythmias, heart failure and vascular disease. Let us now consider CVDs, which have the broadest prevalence and commonly known as "heart disease". CVD has now emerged as the number one killer disease world-wide, challenging the entire population. The heart is largely responsible for maintaining life, as it pumps oxygen and nutrient-rich blood to many parts of the body. Alteration in its function can induce systemic consequences which can compromise several organs and physiological systems (Tamal et al., 2019). Cardiovascular diseases (CVDs) continues to be a prominent health challenge globally and equal number of people died nearly every 33 s due to CVDs. Only in the year 2022, 702,880 deaths were attributed to CVDs, which constituted 20% of total mortality rate. The economic impact is significant also, collectively the cost of health care services, or medications and lost productivity related to premature deaths was estimated at $252.2 billion in 2019 and 2020 (CDC and Prevention, 2024). The WHO also claimed 17.9 million deaths in the world in the year 2029 due to cardiovascular diseases at nearly 32% of the global deaths (WHO, 2024). These figures emphasize that CVD is an issue of paramount concern for public health with respect to the world today. CVD, which is the most common type of heart disease, is also extremely prevalent and hazardous. It is most common when the coronary arteries narrow with plaque buildup and blood cannot flow to the heart properly. This condition can manifest through symptoms such as chest pain (angina), shortness of breath, and in more severe cases, heart attacks. Often seen as a precursor to more critical cardiovascular events, CVD remains the foremost cause of mortality across the globe.

1.2 Motivation

Cardiovascular diseases still remain the leading killer worldwide to this day so that reliable diagnostic methods with early recognition are in high demand. Denoting traditional clinical methods, which serve as a big stepping stone, they suffer from inefficiency as the processing is long, the operational cost comes at a high cost, and complex or atypical diagnosis cannot be accurately made. Such restrictions may lead to unnecessary delays in life-saving treatments that will ultimately have a negative impact on patient-related outcomes. A lot of excitement has been generated in the past few years in ML as they have opened many new potentials in medical diagnostics. ML tools,mostly supervised and deep learning algorithms, have shown a great promise in quickly analyzing large-scale datasets & mining latent patterns

that are hard to be discovered by human experts. This opens a way for the development of diagnostic systems, which are both faster and more accurate as well as scalable and cheap.
This issue attracts me because ML has the power to fill in some of the gaps of the conventional methods of diagnosis. Through a comprehensive study of popular ML techniques on well-known datasets from UCI repositories, we intend to present a successful ML model for early insult detection in the heart. The goal is to improve patient care with this high-tech innovation.

1.3 Objectives

Early and accurate detection of heart disease plays a key role in the prognosis and treatment of the patient (Alshraideh et al., in 2024; Hossain et al., in 2023). Nonetheless, many traditional diagnosing methods are often constrained by some shortcomings, such as long evaluation time, expensive cost, and poor accuracy, especially in the ill condition complicated with different diseases (Ali et al., 2021). The objective of this study is to improve the early detection of heart diseases utilizing machine learning tools that are built to mitigate and exceed the limitations of the traditional diagnostic approaches. The specific aims are:

- Mitigating the inefficiencies of traditional diagnostic techniques, particularly those related to cost, time, and accuracy in complicated clinical situations.
- To explore machine learning techniques as efficient, scalable, and cost-effective alternatives for heart disease detection.
- To utilize supervised and deep learning models capable of processing large datasets quickly and adaptively for improved diagnostic accuracy.
- To leverage publicly available datasets, specifically from the UCI repository, to train and evaluate models of ML.
- To assess the performance of more than 15 machine learning algorithms from diverse families of ML.
- To systematically compare these models based on evaluation metrics such as diagnostic accuracy, and computational efficiency.
- To identify models of ML with the greatest potential to overcome traditional diagnostic challenges and support their integration into clinical decision-making processes.

1.4 Methodology

The methodology adopted in the thesis involved a structured, multi-phase machine learning pipeline designed to enhance diagnostic accuracy for cardiovascular disease (CVD). The dataset was collected from the UCI Machine Learning Repository and comprised records included four different sources—Cleveland, Hungary, Switzerland, and VA. After addressing missing data and preprocessing, 774 records were retained, each with 14 key clinical attributes. Missing numerical values were imputed using

column means, and categorical variables were transformed using label encoding to ensure compatibility with machine learning algorithms. The initial five-category classification of heart disease (ranging from 0 to 4) was simplified into a binary classification: class 0 represented individuals without heart disease, while class 1 indicated those with the condition. To manage the issue of imbalanced data, the Synthetic_Minority_Oversampling_Technique (SMOTE) was applied. To determine the significance of input features, permutation-based importance analysis was conducted.

A total of 15 supervised ML models were utilized. Model evaluation was performed using an 80:20 split for training and testing, combined with cross-validation (5-fold) to enhance reliability and generalizability. Key evaluation indicators included accuracy, precision, recall, F1-score, ROC-AUC, confusion matrix, and precision-recall curves. The implementation and visualization of the models were carried out using the version of Python 3.10, leveraging libraries like Scikit-learn, Pandas, Matplotlib, and Seaborn.

1.5 Project Outcome

According to the experimetal results, the HistGradientBoostingClassifier outperformed all the other models. It achieved the optimal total accuracy of 93.5% based on cross validation with 5-fold and steady cross-validated accuracy of 92.4%. Moreover, the performance of this classifier was strong using other evaluation metrics as evidenced by an AUC of 0.94 and well balanced precision and recall, thereby demonstrating its beneficial trade off in reducing false positive and false negative. Two other ensemble methods (Random Forest, Extra Trees, Bagging) also had impressive performance (accuracy between 92.2, 92.4%), suggesting that they are also well suited for clinical prediction tasks. In contrast, with traditional models such as Support Vector Machine (SVM), Gaussian Naive Bayes and SGDClassifier, fair results are obtained, which are anyway lower than 85% and we see significant difference between precision and recall.

1.6 Organization of the Report

The structure of the remaining chapters in this thesis is outlined below:

A. **Chapter 2 – Background:** This chapter conducts an in-depth and systematic review of existing research. It critically evaluates previous studies to identify key trends, methodological strengths and weaknesses, and research gaps that inform the direction of the current study.
B. **Chapter 3 – Study Methodology:** This chapter features the research design and the methodology used in this study. We discuss in details the methods that we employed for data collection, preprocessing and feature extraction along the reasons for choosing the selected machine learning models. All stages of the procedure are meticulously described to guarantee the transparency of the process.
C. transparency.

D. **Chapter 4 – Experimental Results:** This chapter reports the research results as interpreted from the experiments in the form of a conclusion that is measured against the research objectives. It involves a systematic analysis and interpretation of performance metrics, providing a clear understanding of the meanings and effectiveness of the proposed method.

E. **Chapter 5 – Discussion:** The findings are analyzed in the light of past research. It assesses the importance of the results and their agreement or discrepancy with prior work in the area.

F. **Chapter 6 – Conclusion:** This final chapter provides a summary of the key contributions made throughout the thesis, highlights its limitations and areas where improvements could be made, and outlines potential directions for future investigation and action. It also reflects on the broader significance of the research findings and offers suggestions for continued exploration in related fields.

Chapter 2

Background

2.1 Introduction

It provides a comprehensive review according to PRISMA framework by Page et al. (2021). The major reason for selecting the approach was to guarantee a systematic and appropriate compilation of the available knowledge base. The considered articles, primarily extracted from high impact journals and renowned conference proceedings, revolve around the theme of "Heart Disease Identification" and have been published from year 2019 until year 2024. The process of how these studies were selected is illustrated in Figure 2.1. The review began with an extensive literature analysis to provide an extensive insight of the current status of detection of heart disease using machine learning, while exposing the lack in technology and challenges.

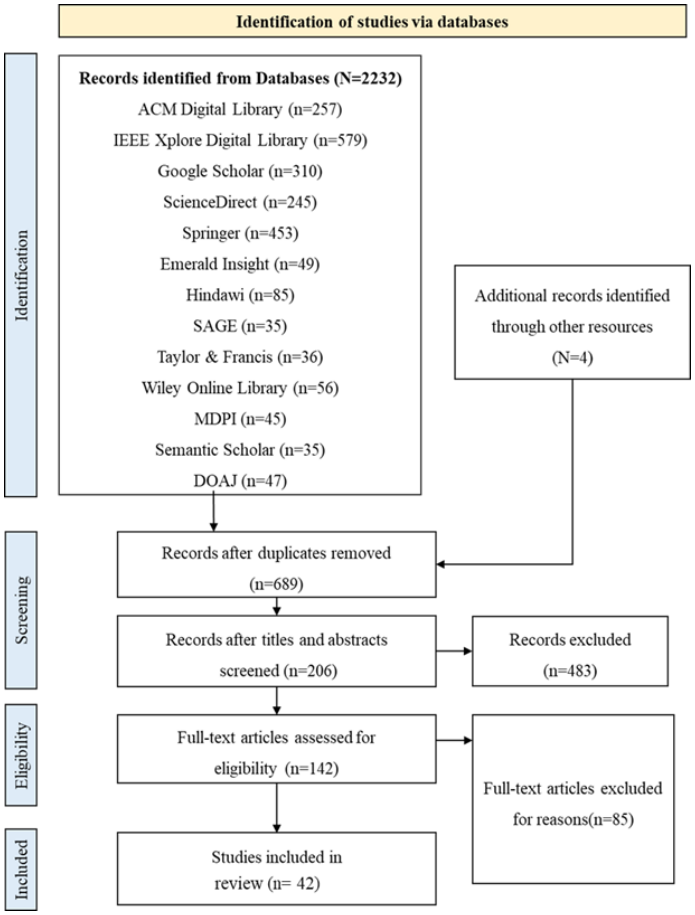


Figure 2.1. Literature selection using PRISMA

2.2 Sources of Information and Strategy of Searching Article

As the background previously described, in order to perform the systematic literature review, a number of well-known academic databases have been included, based on the advises of the experts of the subject area and the thesis supervisor. Major digital sources applied for the collection of the relevant literature included ScienceDirect, MDPI, Taylor & Francis, Semantic Scholar, IEEE Xplore, SAGE Journals, Wiley Online Library, SpringerLink, Emerald Journal, ACM Digital Library, and Google Scholar. List of search keywords was developed, corresponding to the thesis objectives and research questions. These keywords were subsequently revised and modified from the initial search results to improve the search strategy. The finalized search strings, developed to support the study's core objectives and inquiries, are summarized in Table 2.1. In addition to keyword optimization, Boolean operators such as AND, OR, and NOT were applied to broaden or narrow the search as necessary. The search process was iterative, ensuring that relevant literature was not missed due to restrictive terms or database limitations. Articles were selected based on defined inclusion and exclusion criteria, focusing on peer-reviewed publications from the past decade, relevance to the research topic, methodological rigor, and availability in full text. Duplicate entries were removed, and the remaining studies were screened through abstract and full-text reviews to ensure alignment with the scope of the review.

Table 2.1: Search strategy

Objective	Review Questions	Search Strings
To overcome the current challenges in heart disease detection by enhancing prediction accuracy through the application of both supervised ML and DL models on a large-scale dataset from the UCI repository.	What are the major discoveries and recurring shortcomings in existing techniques for identifying heart disease?	("obstacles" OR "constraints" OR "shortcomings") AND ("artificial intelligence" OR "neural networks" OR "supervised algorithms") AND ("cardiac disorders" OR "heart illnesses" OR "forecasting cardiovascular disease")
	Which ML and DL approaches have demonstrated significant potential in enhancing the accuracy prediction of heart disease?	("cardiovascular condition detection" OR "identifying heart-related illnesses") AND ("artificial intelligence algorithms" OR "neural network approaches" OR "models based on supervised learning")
	What validation strategies were implemented to guarantee the robustness and dependability of models for heart disease prediction?	("predicting heart conditions" OR "diagnosing cardiovascular diseases") AND ("evaluation methods" OR "validation strategies" OR "cross-checking" OR "model robustness" OR "assessment of reliability") AND ("artificial intelligence" OR "ML techniques" OR "DL approaches")

2.3 Inclusion and Exclusion Criteria

To maintain the reliability and validity of the review outcomes, a thorough quality evaluation was carried out prior to the inclusion of any sources. The review focused mainly on peer-reviewed research papers and review articles, though some conference papers, academic dissertations, and chapters of books were chosen where appropriate.Only materials published between 2019 and 2024 were included, as this period reflects the current developments in ML and DL, especially within healthcare contexts. This timeframe was chosen to ensure the incorporation of recent innovations that contribute to more accurate and efficient diagnostic methodologies. Studies published prior to 2019 were excluded due to the likelihood of outdated methodologies and limited applicability to present-day AI integration in clinical environments. Additionally, any sources failing to meet established quality benchmarks—such as clarity of objectives, methodological soundness, practical significance, and coherent presentation of results and discussion—were omitted from the analysis.

2.4 Findings of Review

CVD commonly known as heart disease, remains a major public health concern worldwide and is currently the leading cause of death globally. The World Heart Federation reports that more than 500 million people are living with CVD, which contributed to approximately 20.5 million deaths in 2021. This accounts for almost one-third of all global fatalities and reflects a sharp rise compared to the previously recorded 121 million deaths linked to CVD (Di Cesare et al., 2023). The diagnosis of heart-related conditions typically incorporates various methods, including laboratory tests alongside both invasive and non-invasive diagnostic approaches. One of the critical challenges in healthcare is delivering timely, high-quality diagnostic services while keeping costs under control. Furthermore, the traditional diagnostic process tends to be time-intensive, which can be detrimental for patients requiring urgent care (Singh, Singh, & Pandi-Jain, 2018). In this context, data-intensive techniques, and especially those based on Machine Learning (ML) are establishing themselves as powerful resources to further the prediction and diagnosis of heart disease (Ahsan, Luna, & Siddique, 2022).

Supervised ML and DL are the most commonly used machine learning (ML) approaches in early-stage heart disease diagnosis. ML models are training on data sets for which the outcome of interest is already known (for example, do people have heart disease?) and are learning patterns (and, importantly, generalizing patterns) that help them predict outcomes correctly on new data. For instance, Li et al. (2020) proposed a stable diagnostic model which combined several SML methods, including Support Vector Machines, Logistic Regression, Naive Bayes, Artificial Neural Networks, Decision Trees, and K-Nearest Neighbors. This framework was further complemented with feature selection strategies as it was the case of Relief, LASSO (Least Absolute Shrinkage and Selection Operator), Minimum Redundancy Maximum Relevance (MRMR), and Linear Least Squares Feature Selection (LLBFS). A crucial development was the proposed Fast Conditional Mutual Information Maximization (FCMIM) algorithm, which substantially increased prediction accuracy as well as computational efficiency Our model reached 92.37%

classification accuracy based on the Cleveland heart disease data set.

Nowadays, SML-based models for heart prediction have been increasingly used because of availability of medical datasets obtained from patients. This is supported by the studies of Sharma and colleagues. (2020) that trained several predictive models with the UCI Heart Disease dataset of 14 clinically meaningful variables. Of these, RF machine-learning model was determined to be the top model with excellent accuracy of 99% (proving as an ultimate model of choice) signifying its strengths to model complex feature interactions

Inspired by the generalization power of SML models, Garg et al. (2021) did not restrict their analysis to RF, and included K-Nearest Neighbors (KNN) in cardiovascular risk prediction. They emphasized the use of patient age, type of chest pain and cholesterol level in the diagnosis. Their reported performance was against KNN which showed that the performance KNN slightly outperformed RF in terms of accuracy with KNN at 86.89% and RF at 81.97% indicating a performance difference for algorithms based on given scenario.

Gjoreski et al. (2020) suggested a new hybrid system with traditional machine learning (ML) and Deep Learning (DL) for the diagnosis of Chronic Heart Failure (CHF) using phonocardiogram (heart sound) data. The feasibility of the fusion of multi-paradigm MI signals to medical diagnosis also has been proved by testing this model on a big number (n = 947) of volunteers with high classification accuracy (92.9%) with different CHF stages.

In face of the health disparities in remote or underprivileged regions, Rani et al. (2021) proposed a highly-structured decision support framework. The approach applied chained equations for missing value treatment, a two-phase feature selection pipeline (Genetic Algorithms and Recursive Feature Elimination) and experimented on a few classifiers (RF, SVM and NB). That ensemble method has lead to a very nice accuracy of 86.6%, which means improvements over the traditional methods for real world purposes.

Similarly, in another study Almustafa (2020) for Predicting Cardiac Risk, a wide range of classifiers such as the KNN, Decision Tree (J48), JRip, (CN2.10), Multilayer Perceptron, SVM (sequential minimal optimization) etc. 5)Naive Bayes, SVM, AdaBoost etc. The highest accuracy was provided by KNN it is quite evident that KNN with % accuracy is the best then we have J48 and JRip. It is interesting to observe that the K-Nearest Neighbour model needs only four words and zero-dimensional data to produce the perfect predicted label in the test data, showing that little but not informative input carry a lot predicting power.

Likewise, Ali et al. (2021) also reported excellent performance for Random Forest, in which the model achieved a very extraordinary 100 %, accuracy, sensitivity, and specificity, remarkable rare among testing scenarios. The method of Budholiya, Shrivastava and Sharma (2020) was an attempt to enhance the predictive accuracy using an XGBoost classifier with Grid search and One-Hot encoding for categorical predictors with Bayesian Search algorithm. They achieved accuracy of 91.8% with their model on Cleveland Heart Disease dataset, which is far better than RF and Extra Trees classifiers..

Compared to the above rule-based and traditional ML methods, the dawn of DL has transformed the CC risk prediction with automatic extraction of patterns. Thus, they may be seen as models of some of the human cognitive capabilities, and they perform particularly well when matching complex relations between data in

medicine (Cătălina Cocianu et al., 2023). For example, Hussain et al. (2021) developed a 1D CNN model for automatically classifying patients with or without heart diseases. Through the adoption of strong regularization preventions for overfitting and the utilization of a large number of clinical features, their model attained training and testing accuracy above 97% and 96%, respectively.
Additional evidence on the potential of DL, Arooj et al. (2022) proposed a deep CNN model, where authors trained their model using 1,050 patients with 14 clinical features. Their model achieved a validation accuracy of 91.7%, which demonstrates the feasibility of clinical use. Similarly, Mehmood et al. (2021) also showed that CNNs can be used to reliably predict the risk of CVD with 97% accuracy, thus beating other ML models.

2.5 Research Gaps

Despite the positive results with ML and DL methods for the diagnosis and predicting heart diseases, many studies are still being faced with significant limitations. One critical issue is the reliance on small datasets leadingto overfitting and limitingthe transferability of models to wider clinical settings (Mehmoodet al., 2021; Li et al., 2020; Arooj et al., 2022). The work above partly solves this problem, by using much larger, more diversity datasets, such as from the UCI repository, in addition to cross-validation and data augmentation methods. Overfitting, especially in DL models, limits the generalization of such models to wider groups of populations (Hussain et al., 2021), and such overfitting will be counteracted by using L1/L2 regularization and dropout layers. In addition, previous works (e.g., Tamal et al., 2019) that are implemented relying on self-reported data are biased and this bias is avoided in the current study where clinically validated data have been used. Also, there is lack of model interpretability for many studies (Garg et al., 2021), which this study will address by using explainable ML methods and making decisions more transparent. In addition, lack of validation across different datasets reduces model's robustness (Budholiya et al., 2020; Ali et al., 2021), hence we will do external validation using data from a different geographical regions. The computational complexity that is usually neglected (Mienye et al., 2020) will be critically evaluated to suggest whether selected models are computationally efficient to be applied in resource constrained settings, such as Bangladesh. The study will also investigate issues pertaining to data quality: through high-quality clinical data, as well as application of augmentation techniques to overcome noise in data, and through synthetic data generation through GANs (Sharma et al., 2020). In order to enable performance comparison on a common platform and account for various reporting patterns (Budholiya et al., 2020), the results will be reported with standard evaluation measures such as accuracy, precision, recall, F 1 score, area under the ROC curve (AUC-ROC) and precision-recall curve and support reproducibility and comparability.

2.6 Summary

This work is a systematic review following the PRISMA guidelines of recent progress in predicting heart disease with ML and DL. It encompasses research published from 2019 to 2024, mainly from reputable academic databases. Precise search strategy and inclusion criteria clarified that only high-quality, peer-reviewed literature relevantly

matched with the research purpose was selected. The findings indicate that heart disease remains a main cause of mortality and that ML techniques—especially supervised learning models like SVM, RF, and KNN—have demonstrated strong predictive performance when applied to clinical datasets such as those from the UCI repository. Several studies also explored deep learning models like CNN and DCNN, showing promising accuracy in real-world diagnosis. Despite these advancements, notable gaps persist in the literature, including reliance on small or self-reported datasets, lack of model interpretability, limited validation across diverse populations, overlooked computational efficiency, and inconsistent reporting practices. This study addresses these challenges by utilizing large-scale, validated clinical data, applying cross-validation and regularization techniques, emphasizing explainable AI, and adhering to standardized evaluation metrics to enhance reliability, transparency, and practical applicability in diverse healthcare.

Chapter 3

Research Methodology

3.1 Introduction

In the present work, a machine-learning procedure is employed to enhance the early and precise detection of heart ailments using the strong available UCI ML Repository data. The proposed methodology is organized in a few discreet stages in order to keep the work systematic. First, data is obtained through the collection of appropriate medical datasets, then preprocessed by various stages including missing values imputation, normalization and transformation of categorical variables in order to result in the provision of reliable and consistent data. Then some classifier algorithms that may be proper to medical diagnosis problems are adopted. Both models are trained and tested with stratified sampling so as to not introduce bias. The last step is to compare the results obtained from different models by using the performance metrics such as accuracy, precision, recall, F1-score, and AUC. The aim is not only to enhance diagnostic accuracy, but also to guarantee that the created system is generalizable to different patient populations. This integrated approach ultimately should support clinical decision making and minimize real-world risk of misdiagnosis.

3.2 Tools for Data Analysis

We performed experiments in Python 3.10, making use of helpful libraries for data pre-processing, model training, and performance assessment. Scikit-learn provided the most comprehensive and user-friendly classification models for training and assessing the models. A complete analysis was carried out analyzing the performance of 15 different supervised models to predict the activity on the dataset.
Model validation was executed with 5-fold CV completed using the cross_val_score function, hence providing statistical credibility of results across different data splits. Performance measures: accuracy, precision, recall, and F1-score were evaluated by calculating the area AUC by Scikit-learn utilities.
To adequately alleviate the effect of class imbalance, which often can result in a bias in model performance, the SMOTE of the contrast example is used. This method randomly generates new examples for rare classes and increases classifier's ability to respond to the minority members.
NumPy was also used extensively for efficient numerical computation and array manipulation while Pandas for data wrangling and manipulation operations. For visualization on how good the model performed, we used tours of Matplotlib and

Seaborn to draw informative plots like confusion matrix, roc curve, precision-recall visualization which made the model more interpretable and gave details on how well the Model performed.
In order to guarantee reproducibility and to provide transparency of the workflow, all the scripts employed were modularized and version-controlled through Git. Logging functions were also added to trace important points in data or model evaluation, which aided in debugging and performance tuning when conducting experiments.

3.3 Data Collection and Description

The dataset utilized originates from the UCI ML Repository and integrates patient records from four sources. Although the full dataset consists of 920 instances and 76 features, this study concentrates on 14 selected features, with 774 entries remaining post-missing value handling.
While larger datasets generally enhance machine learning model performance through improved pattern recognition and generalizability, the use of a moderately sized dataset in this case remains both relevant and justified. The chosen dataset holds a strong reputation in the literature for its quality and relevance in the context of heart disease classification. Furthermore, in the healthcare domain, acquiring expansive, high-quality datasets is often impractical. Nonetheless, studies based on smaller datasets have achieved credible results. For example, several previous studies effectively utilized as little as the sample size of 303 (Jindal et al., 2021; Bharti et al., 2021; Kavitha et al., 2021; Miranda et al., 2021).
This demonstrates the possibility of deriving meaningful insights and reliable predictions from smaller datasets through rigorous usage of advanced ML techniques as well as validation methods. To alleviate the weaknesses such as overfitting and high variance for the estimates, this study takes robust cross-validation methods. These are useful for generalizing the results and for making the evaluation criteria more stable. Furthermore, advanced preprocessing procedures such as data imputation and feature selection are applied to improve the general quality of the data.The attributes of the dataset are a mixture of categorical and numerical.
Key variables include:

a. **Age:** Integer indicating the patient's age, fully complete.
b. **Sex:** Categorical variable denoting gender, with no missing entries.
c. **Cp:** Categorical classification of chest pain, also complete.
d. Trestbps (Resting Blood Pressure): Integer, measured in mm Hg, with some missing values.
e. **Chol:** Integer, in mg/dl, containing missing entries.
f. **Fbs:** Categorical indicator of whether fasting blood sugar > 120 mg/dl, with some missing data.
g. **Restecg:** Categorical variable describing ECG outcomes, with incomplete data.
h. **Thalach:** Integer, some missing values present.
i. **Exang:** Categorical, with missing values.
j. **Oldpeak:** Integer value measuring ST depression due to exercise, partially missing.
k. **Num:** Integer that represents the target class for heart disease presence or absence, complete with no missing values.

3.4 Data Preprocessing

A. Handling missing values:

The original dataset contained missing values across seven different attributes. To ensure data integrity and improve the quality of subsequent analyses, a systematic approach was used to handle these gaps. Initially, a threshold-based filtering process was applied: any feature exhibiting more than 25% missing values was deemed unreliable and therefore eliminated from the dataset. For the remaining features—those with fewer than 25% missing entries—a data imputation technique was adopted.

In particular, numerical columns with an empty value were handled by replacing those missing values with the mean of the respective feature. This method was selected in view of its simplicity and its property to preserve the central tendency of the data with little bias. Before the imputation, a diagnostic phase was performed in which the absolute count as well as the relative percentage of missings in each column of the table was calculated to make it transparent how the data was preprocessed.

Following imputation, the dataset was then pared of dimensionality and instances. Therefore, the number of eligible records was reduced from 920 to 774, mainly due to the exclusion of entries when this number had features beyond the threshold for missing data. Let feature \(X\) have n missing values and with na.rm with value T, mean of the feature is obtained by everging the available values and replacing the missing values with this mean. While its easy to compute -- mean works for most of the cases with good computational efficiency, it doesnt store good results in case of high number of missing values. Thus, this approach is most suitable for a situation where missingness is not extensive and underlying data distribution is fairly symmetric.

To ensure the robustness of the dataset, the cleaned data was also submitted to an EDA that confirmed the impact of the imputation in the feature distribution, verifying no appreciable distortions.

B. Data Transformation:

After taking care of the missing data, we realize that some features are categorical. That's because machine learning algorithms work with numeric data and the categorical text data had to be transformed into numbers using the LabelEncoder from the Python's preprocessing package. This method gives each category against a feature a unique integer, such that the model can interpret the data meaningfully. For instance, a column consisting of "Low", "Medium", and "High" would be converted into their numeric code equivalents 0, 1, and 2. This encoding operation retained the category information intact and compatible with the downstream machine learning models. It also contributed in keeping encoding representation consistent which is crucial for the reproducibility and scalability of the training and the evaluation stages.

C. Handling Class Imbalance:

After preprocessing of missing and categorical data, the problem of class imbalance was observed. The dataset consisted of five categories (0, 1, 2, 3, and 4), where 0 did not present the disease, and categories 1 to 4 presented increasing levels of the disease. This distribution among the classes was severely uneven, requiring additional methodologies to combat the imbalance (See Fig 3.1).

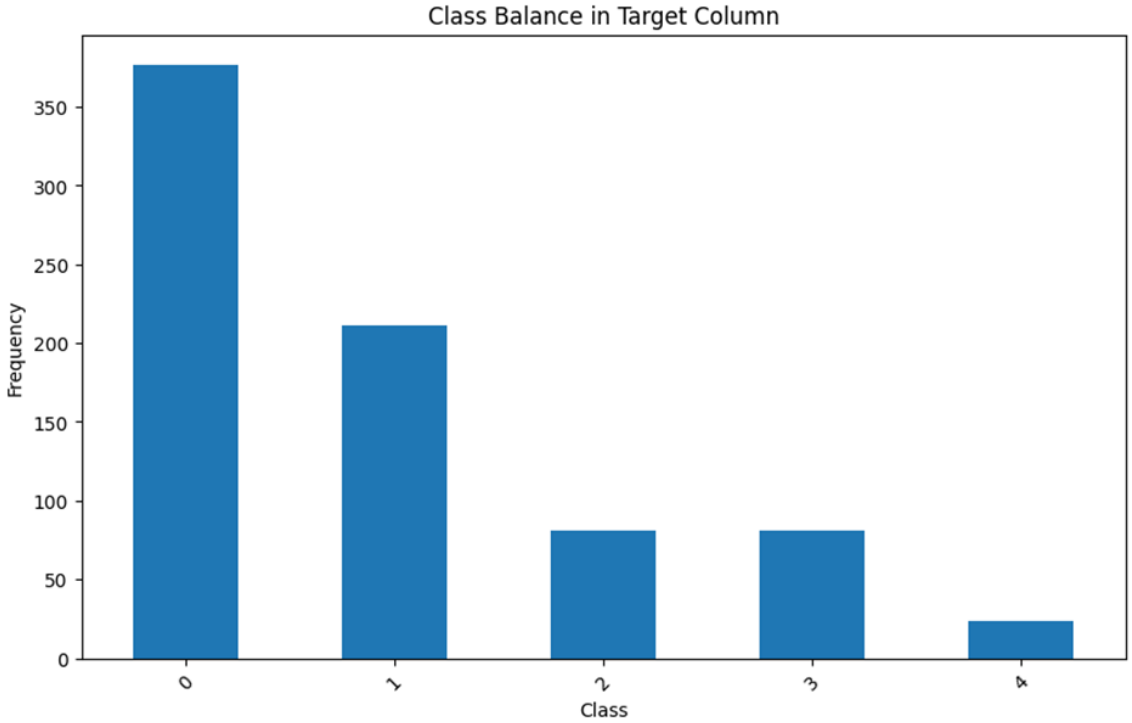


Figure 3.1. Before SMOTE

To improve the classification model strength and the analysis process, the original multi-class target variable was transformed into a binary type. Precisely, the response variable was modeled in a binary fashion, class 0 (subjects without heart disease) or class 1 (subjects with heart disease). This simplification contributed to turning the classification into an ordinary one of binary decisions.

But a critical issue was posed because of the imbalance of samples across these two classes, denoted as the class-imbalance learning issue. In order to address it, the SMOTE are applied. SMOTE creates synthetic examples for the minority class by interpolating between samples, thus helping to balance the classes without simply copying data. This method assists machine learning models in learning the decision boundaries more efficiently, especially for the underrepresented class.

The effect of this process is illustrated by Figure 3.2 where the new class distribution between different steam coditions is shown after using SMOTE. The figure depicts a much more balanced distribution of the two classes, finally resolving the imbalance issue, which is likely to improve model performance and generalization.

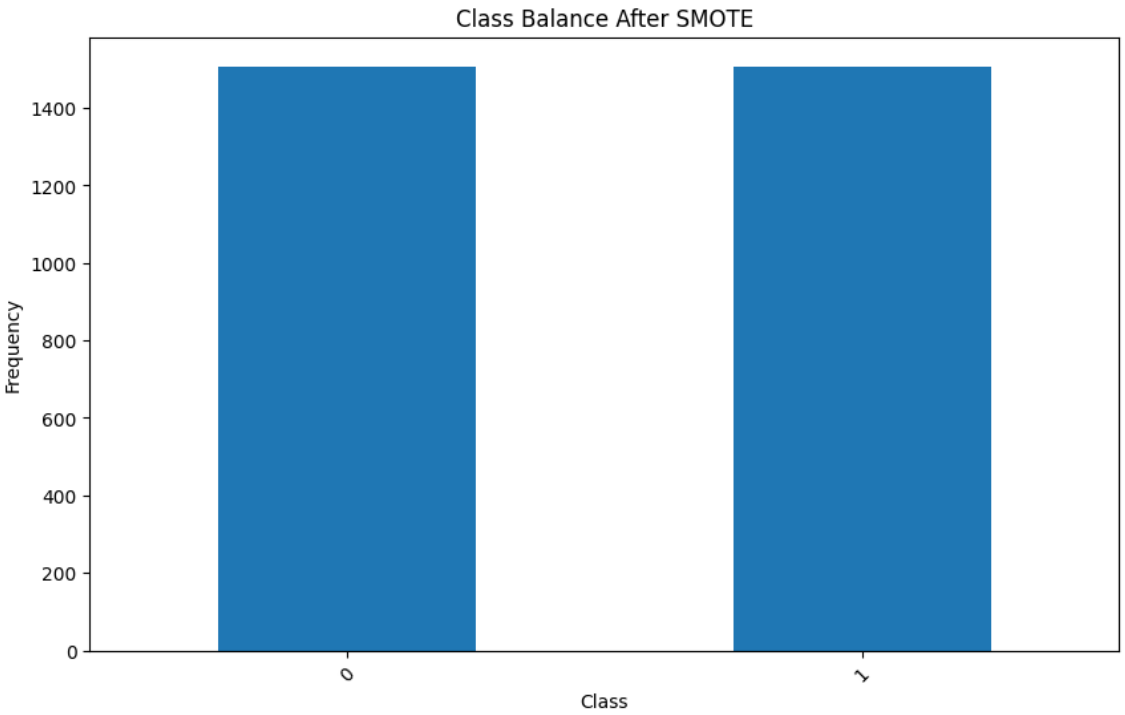


Figure 3.2. After the SMOTE

D. Selection of the Features:

In order to increase the predictive power and reduce the computational power of the models by not making them too complex, advanced techniques for features selection were adopted in this study. Among these, permutation feature importance was used as a strong, model-agnostic approach for assessing the impact of each input feature on the model prediction. This works by randomizing each feature's values and measuring the subsequent impact on prediction, to quantify the feature's importance. Figure 3.3 shows the ten most important features and their importance scores for the model. The choice of these features improved the learning efficiency and helped mitigate the overfitting problem and the interpretability of the model. Furthermore, through identifying more essential factors, the model was also more applicable and practical in practice.

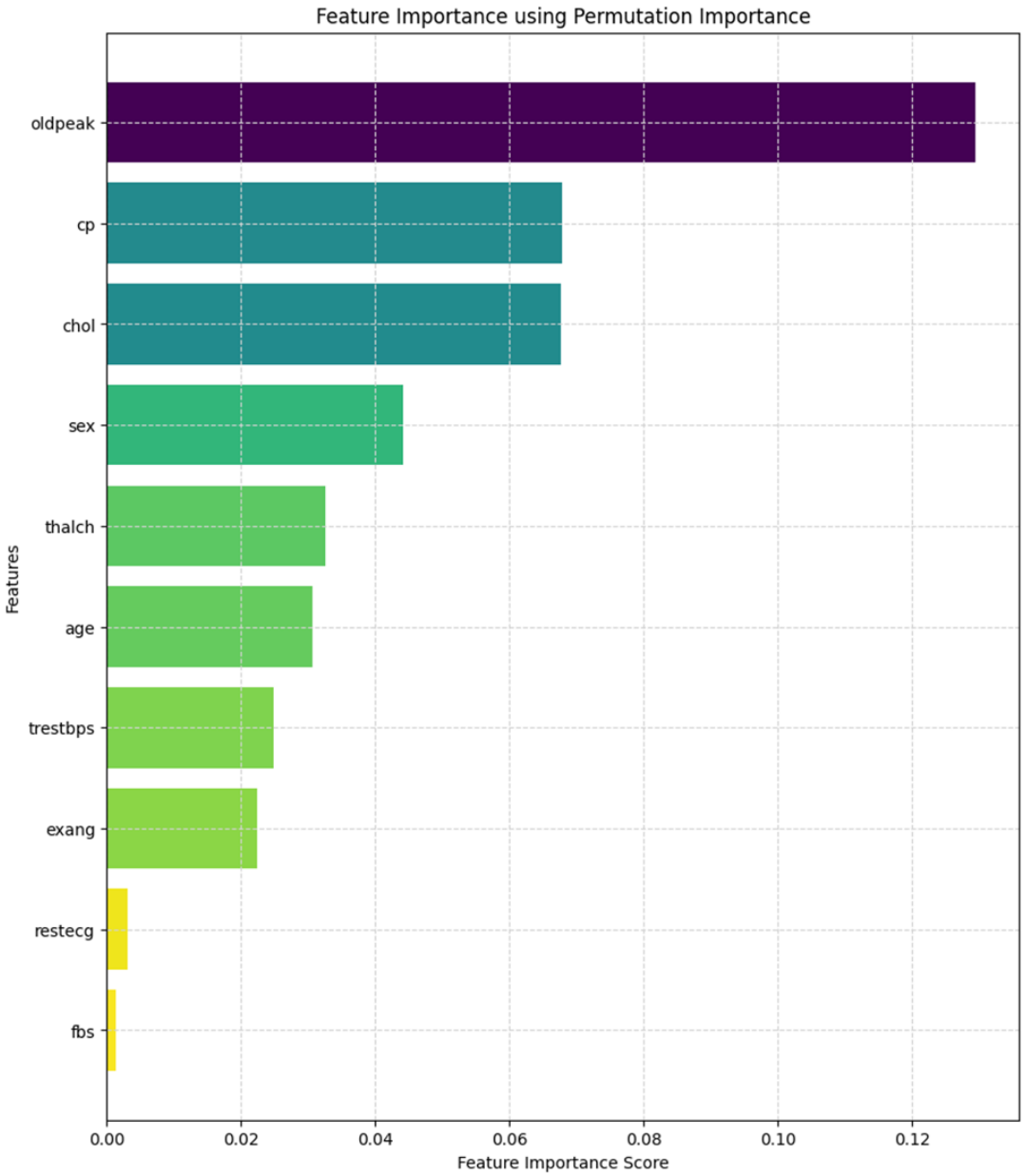


Figure 3.3. Feature importance score

3.5 Selection of ML Model and Evaluation

To identify the most effective model for predictive analysis, this study conducted a comprehensive evaluation of fifteen well-established supervised machine learning (SML) algorithms. These algorithms were selected to represent a wide spectrum of methodological families, encompassing probabilistic classifiers, instance-based learners, decision tree-based approaches, neural networks, discriminant analysis techniques, and advanced ensemble strategies such as bagging and boosting. The goal was to assess the comparative performance of diverse algorithmic paradigms in handling the prediction task efficiently.

The dataset used in this investigation was divided into two subsets, with 80% allocated for training the models and the remaining 20% reserved for testing their generalization capability. To ensure the robustness and consistency of model performance, a 5-fold cross-validation strategy was applied during the training phase. This validation technique helps to mitigate the risk of overfitting and provides a more

accurate estimate of how each model would perform on unseen data. Using the library named Scikit-learn of python, the following classifiers were implemented: SVM, LR, RF, DT, KNN, GNB, QDA, SGD, MLP, RC, ETC, ABC, GBC, HGB, and BC. Model performance and optimal selection were based on six evaluation metrics—accuracy, precision, recall, F1-score, confusion matrix, and ROC curve—offering a comprehensive view of each algorithm's effectiveness in prediction.

Chapter 4

Implementation and Results

4.1 Performance of the Classifiers (Confusion Matrix)

In this study Several different classifiers from diverse algorithmic families were evaluated to predict cardiovascular disease in this investigation. The assessment was based on the models' capacity to discriminate between patients with heart disease and those without. Confusion matrices (CM), as shown in Figure 4.1, are produced in order to identify the performance values such as True-Positives(TP),True-Negatives (TN), False-Positives (FP), and False-Negatives(FN).

In the case of the ridge classifier, it struck a good balance by correctly identifying 247 people with the disease and 285 without, and falsely labeling 46 false negatives and 26 false positives finally among the algorithms that we tested. The RF model was particularly effective, having a high sensitivity of 269 and specificity of 289 predictions that were true positive and true negative, respectively, with 22 false positive and 24 false negative—indicative of its ability to achieve a good prediction. Consistently, rising values of 268TP and 289–290TN were achieved by both ET and Bagging keeping low misclassification (21–22FP and 25FN).

The DT showed moderate performance with 254 TP and 271 TN, but higher misclassifications—40 FP and 39 FN. KNN offered a slight improvement, recording 248 TP and 290 TN, though it still had 21 FP and 45 FN.

In contrast, GNB underperformed, with only 237 TP and 269 TN, alongside 42 FP and 56 FN, reflecting a higher error rate. SGD yielded 255 TP but had 68 FP and 38 FN, suggesting unstable predictions.

Both AdaBoost and GB showed comparable effectiveness, correctly identifying 250–255 TP and 276–277 TN, with modest error margins (34–35 FP and 38–43 FN). Lastly, SVM delivered the weakest outcome, managing just 219 TP and 235 TN, with elevated FP (76) and FN (74), indicating limited suitability for this task. The standout model was the HistGradientBoostingClassifier, which achieved the highest accuracy, correctly identifying 275 heart disease cases and 290 non-cases, with 21 (twenty one)false positives and the 18 false negatives, illustrating its superior reliability and precision. Meanwhile, both Quadratic Discriminant Analysis and Logistic Regression produced moderate outcomes, offering a reasonable trade-off between accurate and incorrect classifications across all metrics. The MLPClassifier also delivered acceptable performance, with 249 true positives, 278 true negatives, and mid-range values for false positives (33) and false negatives (44), indicating dependable but not exceptional results.

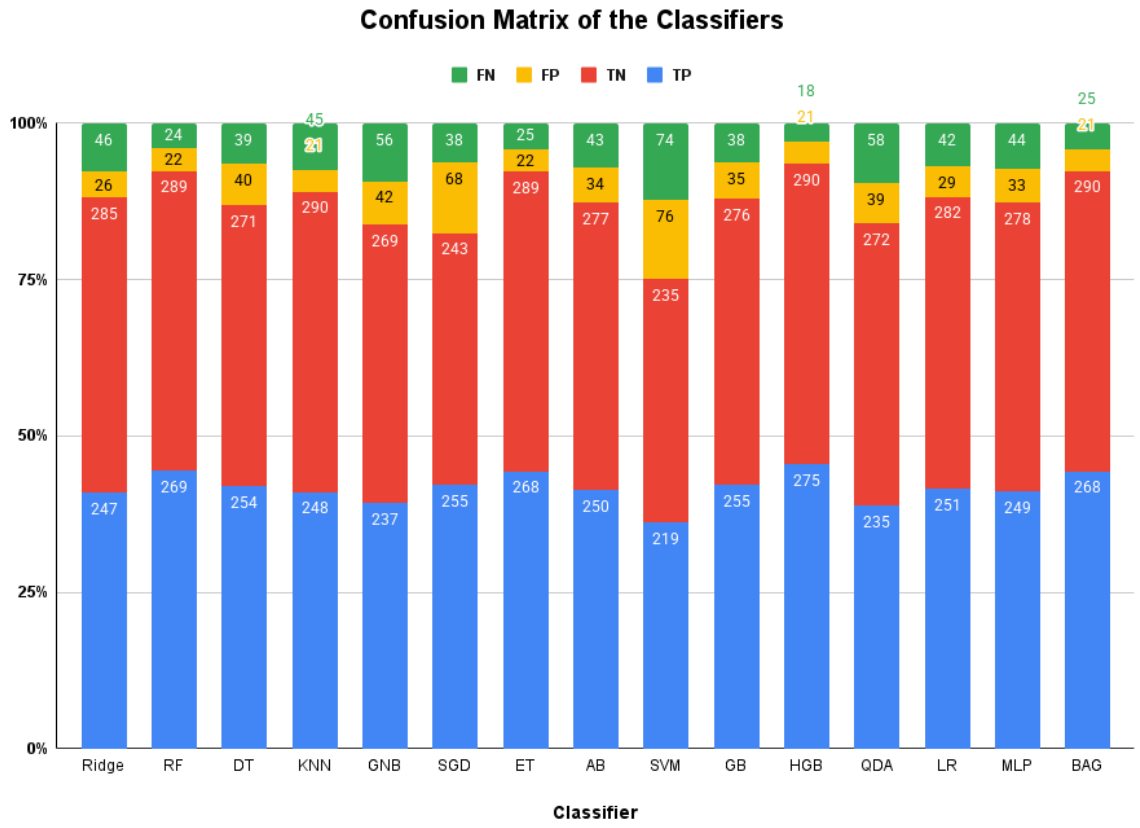


Figure 4.1. Confusion matrix

4.2 Performance Comparison of the Classifiers

The assessment of multiple classification models for predicting heart disease was conducted using performance indicators such as accuracy, precision, recall, F1 score, area under the ROC curve (AUC), and average precision (AP). Table 4.1 presents a consolidated comparison of these evaluation metrics across all models, offering a clear snapshot of their overall effectiveness. The subsequent sections delve deeper into the individual performance of each algorithm, highlighting their strengths, limitations, and potential suitability for real-world deployment in clinical decision-support systems. This approach ensures that model selection is driven not only by numerical superiority but also by interpretability and practical utility in medical diagnostics.

A. Ridge Classifier

It performed well with a total accuracy of 88.1%. The average prediction accuracy across 5-fold cross-validation was 87.6%, indicating stable generalization across varying data splits. In term of Class 0 (negatives), its precision and recall were 86.1% and 91.6%, respectively (F1-score 88.8%). Alternatively, for Class 1 (the presence of the disease), the precision increased to 90.5% with a reduction in recall

up to 84.3% resulting in an F1-score of 87.3%. These findings offer an indication that the classifier is a hair more specific than sensitive for non-disease cases and lowers its sensitivity for diseased ones.

B. Random Forest Classifier

The model performed well, reported 92.4% of overall accuracy and 92.5% of 5 folds cross-validation accuracy. For the Class 0 (non-disease) cases, it obtained a precision of 92.3% and a recall of 92.9%, resulting in an F1-score of 92.6%. Similarly, for Class 1 (disease cases), the precision and recall achieved were 92.4% and 91.8% respectively (F1-score: 92.1%). The scores around these two classes are consistently high, and this suggests the good robustness of the model, and the reliability with which it is capable of discriminating between disease and non-disease cases.

C. Decision Tree Classifier

The Decision Tree (DT) model achieved an accuracy of 86.9%, with a slightly higher cross-validation accuracy of 87.1%. For Class 0, the precision and recall were closely matched at 87.4% and 87.1%, leading to an F1-score of 87.3%. Class 1 results also showed a balanced performance, with a precision of 86.4%, recall of 86.7%, and an F1-score of 86.5%. While the model provides consistent results across both classes, its ability to capture complex data relationships may be limited compared to ensemble-based approaches.

D. KNeighborsClassifier

This model had a similar overall accuracy of 89.1%, but the 5-fold cross-validation method reduced the accuracy to 86.1%, which indicates there might be some performance variability across splits. For Class 0, the model achieve the highest precision of 87% and also high recall of 93.2%, which in turn resulted in F1-score of 89.8%. For the class one, the precision value was 92.2%, while the recall value decreased to 84.6%, which contributed to an F1-score of 88.3%.

E. Gaussian Naive Bayes

The GNB classifier achieved an 83.8% accuracy as well as a slightly better CV accuracy 85.5%. Class 0 results include a precision 82.8%, recall 86.5%, and an F1 score 84.6%. For Class 1, the F1 score dropped to 83% (precision: 85%, recall: 81%).

F. SGDClassifier

With an accuracy of 82.5% and a CV accuracy of 84.3%, the SGDClassifier showed relatively lower reliability. Class 0 had high precision (86.5%) but poor recall (78.1%), leading to an F1 score of 82.1%. Class 1 performed slightly better (precision: 78.9%, recall: 87%, F1: 82.8%). The imbalance across classes indicates inconsistencies.

G. Extra Trees Classifier

Delivering a strong accuracy of 92.2% and CV accuracy of 92%, this classifier demonstrated high consistency. Class 0 scored 92% precision and 93% recall (F1: 92.5%), while Class 1 achieved 92.4% precision and 91.5% recall (F1: 91.9%). Its performance closely aligns with that of Random Forest.

H. AdaBoost Classifier

AdaBoost attained 87.3% accuracy of a 5-fold CV accuracy 87%. Class 0 metrics were balanced (precision: 86.6%, recall: 89.1%, F1: 87.8%). Class 1 showed 88% precision and 85.3% recall (F1: 86.7%). While competent, its slightly reduced recall for Class 1 places it below other ensemble techniques.

I. SVM Classifier

SVM recorded the lowest accuracy at 75.2% and a similar CV accuracy (75.5%). Both classes showed consistent yet modest metrics—Class 0: precision 76.1%, recall 75.6%, F1: 75.8%; Class 1: precision 74.2%, recall 74.7%, F1: 74.5%. These results indicate the model's limited suitability for this dataset.

J. Gradient Boosting Classifier

With an 88% accuracy and 90% CV accuracy, this model showed solid generalization. For both classes, precision and recall hovered around 87.9% and 87%, respectively, resulting in F1 scores of 88.3% (Class 0) and 87.5% (Class 1). Gradient Boosting is reliable but slightly outperformed by more advanced ensemble models.

K. HistGradientBoostingClassifier

The best-performing model, HistGradientBoosting, attained 93.5% accuracy and a 92.4% CV accuracy. Class 0 achieved 94.2% precision and 93.2% recall (F1: 93.7%), while Class 1 showed 92.9% precision and 93.9% recall (F1: 93.4%). Its consistently superior results across all metrics mark it as the most effective model.

L. Quadratic Discriminant Analysis (QDA)

QDA reported an 84% accuracy and 85.5% CV accuracy. For Class 0, precision was 82.4% and recall 87.5% (F1: 84.9%). For Class 1, the metrics were 85.8% precision, 80.2% recall, and 82.9% F1. While capable, the lower recall for Class 1 reduces its reliability.

M. Logistic Regression

Logistic Regression achieved 88.2% accuracy and 87.3% CV accuracy. Class 0 achieved strong scores (precision: 86.1%, recall: 91%, F1: 88.5%), while Class 1 performed with 90.5% precision and 84% recall (F1: 87.1%). The model is reliable and interpretable, though not the top performer.

N. MLPClassifier

This model yielded an accuracy of 87.3% and a matching 87% CV accuracy. Class 0 posted 85.6% precision, 89.5% recall (F1: 87.5%), while Class 1 recorded 89.2% precision, 85% recall (F1: 87.1%). Though competitive, it is slightly less effective compared to top ensemble models.

O. Bagging Classifier

Recording a high accuracy of 92.4% and 91% through cross-validation, the classifier maintained consistent precision and recall scores across both classes. Class 0 results reflected slightly higher recall, whereas Class 1 demonstrated superior precision, indicating strong overall predictive reliability. These consistently high and balanced metrics position it among the best-performing models, alongside Extra Trees and Random Forest classifiers.

From the results, it's clear that ensemble methods like HGB, RT, RF, and Bagging Classifier outperformed other models in terms of both accuracy and balanced evaluation metrics. These models appear highly suitable for predicting heart disease because of their consistent results. While LR, GB, and AdaBoost also gave good results, they didn't match the top performers. In contrast, classifiers like SGD, SVM, and GNB reported lower accuracies and showed imbalanced precision-recall metrics, highlighting their limited suitability for this task.

Table 4.1: Performance of the classifiers

Classifiers	Acc.	CV (5-fold)	C	P	Rec.	F1 Score	S
Ridge	0.881	0.876	F	.861	.916	.888	F=311 T=293
			T	.905	.843	.873	
RF	0.924	0.925	F	.923	.929	.926	
			T	.924	.918	.921	
DT	0.869	0.871	F	.874	.871	.873	
			T	.864	.867	.865	
KNN	0.891	0.861	F	.87	.932	.898	
			T	.922	.846	.883	
GN	0.838	0.855	F	.828	.865	.846	
			T	.85	.81	.83	
SGD	0.825	0.843	F	.865	.781	.821	
			T	.789	.870	.828	
ET	0.922	0.920	F	.920	.93	.925	
			T	.924	.915	.919	
AB	0.873	0.870	F	.866	.891	.878	
			T	.880	.853	.867	
SVM	0.752	0.755	F	.761	.756	.758	
			T	.742	.747	.745	
GB	0.88	0.900	F	.879	.887	.883	

			T	.879	.87	.875	
HGB	0.935	0.924	F	.942	.932	.937	
			T	.929	.939	.934	
QDA	0.84	0.855	F	.824	.875	.849	
			T	.858	.802	.829	
LR	0.882	0.873	F	.87	.91	.888	
			T	.90	.857	.876	
MLP	0.873	0.870	F	.863	.894	.878	
			T	.886	.85	.866	
BAG	0.924	0.91	F	.921	.932	.927	
			T	.927	.915	.921	

4.3 Training vs Testing Time

A comparative examination of training and testing times across multiple classifiers reveals clear variations in computational performance (refer to Table 4.2). The Ridge Classifier is notably efficient, completing its training phase in just 0.0199 seconds and testing in 0.0067 seconds, making it well-suited for applications where time is critical. In comparison, the Random Forest Classifier, though known for its strong accuracy, is more demanding in terms of resources. It takes 6.2726 seconds to train and 0.1503 seconds to test, illustrating a balance between performance and computational load.

The Decision Tree Classifier stands out for its faster execution, completing training and testing in 0.0213 and 0.0159 seconds, respectively. This speed makes it ideal for scenarios that require quick prediction results.

The KNeighborsClassifier, despite requiring only 0.0018 seconds to train, has a longer testing duration of 0.0352 seconds, likely due to its architecture that performs computations at the time of prediction.

Gaussian Naive Bayes is highly efficient, with training and testing times among the shortest at 0.0078 and 0.0020 seconds respectively, making it well-suited for large-scale or constrained-resource environments.

Lastly, the SGDClassifier achieves a good compromise, requiring 0.0249 seconds for training and just 0.0018 seconds for testing, which supports its application in dynamic or web-based machine learning environments.

Extra Trees and AdaBoost, while slightly slower—training in 0.0671 and 0.1115 seconds, respectively—maintain acceptable performance (testing times of 0.0235 and 0.0155 seconds), supporting their use in scenarios that require both accuracy and speed.

Finally, although the SVM classifier is known for strong accuracy, its training time (0.1962 seconds) and testing time (0.0522 seconds) are relatively high. However, in applications involving complex datasets where predictive power is paramount, this computational overhead may be justifiable.

The training times of the Gradient Boosting Classifier and HistGradientBoostingClassifier are comparable—around 0.5676 and 0.5524 seconds, respectively—and they are also both efficient in the test stage, with the speed being 0.0035 for the Gradient Boosting Classifier and 0.0145 seconds for the HistGradientBoostingClassifier. This makes them appropriate to be used where training time is not a concern.

Of all models analyzed, the QDA model is also worth mentioning based on the quickness with which constructive and testing times can be calculated, which were 0.0137 and 0.0034 seconds, respectively, making it suitable to implement for real-time applications.
Logistic Regression although slightly slower than its' simpler counterparts, still retains the efficiency as shown in Table \ref{table1}.
Although MLPClassifier requires the longest training time of 1.9213 seconds, it has the fastest inference time of 0.0029 seconds. This trade-off could benefit highly complex data relations tasks in which predictive depth is more relevant than training efficiency.
Finally, Bagging Classifier appears to provide a good trade-off between developing time and performance, it requires 0.1578 seconds for training and it costs a minimal slower time (0.0048 seconds) for its testing. This makes it a good fit for combining multiple models, especially through methods like bagging.

Table 4.2: Training vs testing time

Classifier	Training Time	Testing Time
Ridge	0.02	0.01
RF	6.27	0.15
DT	0.02	0.002
KNN	0.01	0.04
GNB	0.008	0.0020
SGD	0.025	0.002
ET	0.27	0.024
AB	0.212	0.016
SVM	0.2	0.05
GB	0.57	0.004
HGB	0.55	0.015
QDA	0.014	0.003
LR	0.1	0.01
MLP	1.92	0.003
BAG	0.16	0.005

4.4 ROC Curve Analysis of the Classifiers

The Receiver ROC curve depicted in Figure 4.2 gives a clear picture how different machine learning models, applied for heart disease prediction, has been classified. AUC is an important evaluation measure to test how well one classifier can differentiate between patients with and without heart disease.
The AUC (0.94) of HGB was higher among the algorithms studied, suggesting excellent discriminative capacity. The reason for such a blazingly fast performance is

that its boosting process builds trees using a histogram-based algorithm, both for setting faster improved estimators and for making better predictions, particularly for the scenarios of big datasets.

Ensemble methods, including RF, ET, and BG, performed equally well (AUC=0.92). These are ensemble models that include several weak learners to improve generalization and to prevent overfitting, a desirable property especially in the presence of noisy (or imbalanced) medical data.

On the other hand, the DT classifier had the poorest performance, with an AUC value of 0.75. Its substandard generalization performance might be the result of overfitting and an incapacity to accommodate complex relationships in the data because of its greedy nature with one-pass splitting criteria.

The performance of KNN, MLP, and LR was intermediate and AUC scores were between 0.87 and 0.89. While a good compromise between the interpretability and effi- ciency, these models offer good baselines but are often not flexible enough to model very non-linear clinical features without fine- tuning.

The boosting algorithms, that is, ADA and GB, also showed robust results, each with an AUC behaviour around 0.88. These models feedback and correct their predecessors, and in doing so are robust in incrementally increasing accuracy.

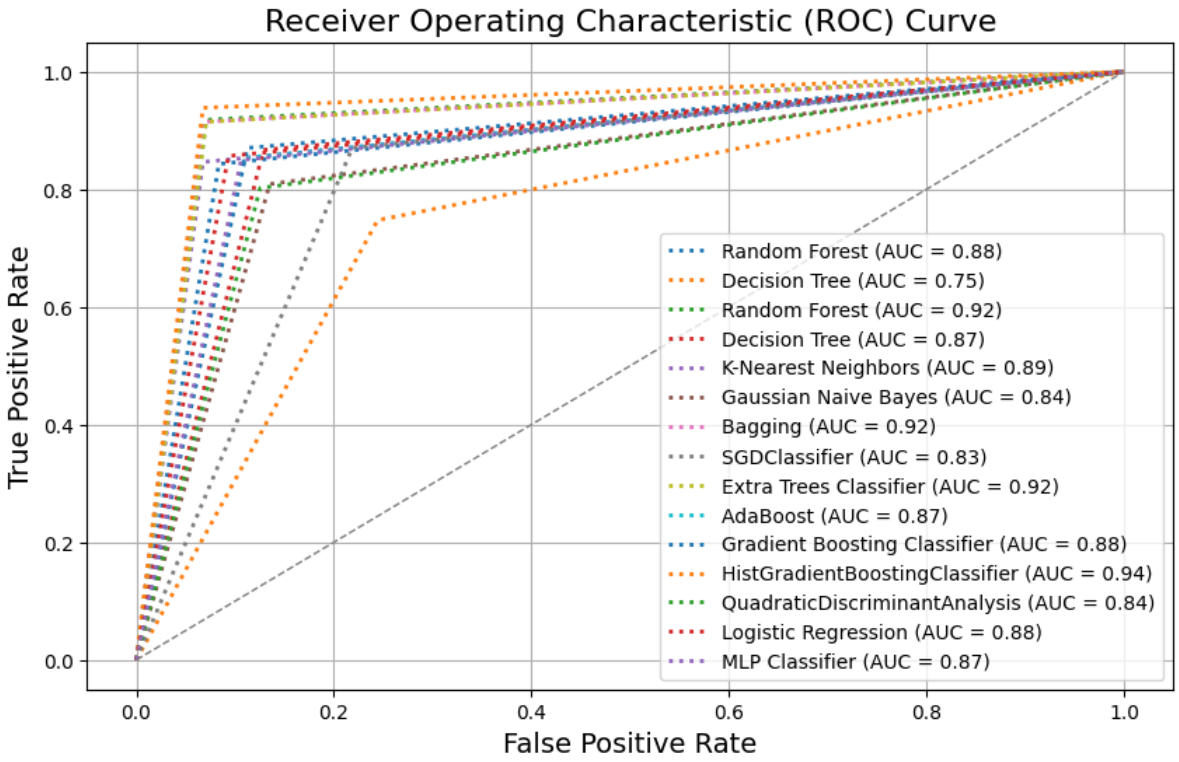


Figure 4.2. ROC Curve of the classifiers

4.5 The Ultimate Model

Comprehensive comparative analysis of various machine learning classifiers suggested that the HGB outperforms other classifiers in predicting heart disease. It has the best discrimination rate of 93.5% among all tested models, and demonstrates its robustness of learning complex patterns from the dataset. Its 5-fold CV accuracy is also high, which is 92.4%, indicating that the model generalizes well to various

data partitions around the raw data, facilitating to tame the overfitting and guaranteeing the stability of the model.
This model also equally performs on precision and recall as evidenced by its F1 scores which are 93.7% for Class 0 (people without heart disease) and 93.4% for Class 1 (people with heart disease). This sort of "balance" is particularly important in (clinical) medicine, where both false positives and false negatives can have serious consequences. Its Area AUC is 0.94, the highest among all the models tested, and demonstrates an excellent ability to discern cases of positive from negative.
Apart from good predictability, this model also has high clinical trustworthiness, thanks to its low FPR and FNR. These measures are especially important in healthcare, as reducing diagnostic errors has direct implications to patient care. Therefore, the HGB becomes a high-performance and practical model that can be applied in the real-world heart disease screening system.
Similarly, the RF Classifier also performs well, attaining an accuracy of 92.4% and an AUC of 0.92. It is known for its ensemble construction which is particularly useful for treatment of data in high dimension space and to capture non-linear feature interactions. But one of its disadvantages is the higher training time in comparison to other methods, which can be a limitation in dynamic clinical environments that require frequent model retraining or real-time deployment.
The nearly equivalent in performance Extra Trees Classifier attains an accurracy of 92.2% with an AUC identical to the Random Forest, 0.92. One important difference is an increased training speed compared to Random Forest, since it uses random feature splits during tree training. This feature makes it especially suitable for a big dataset or when one needs to iteratively prototype a model in the wild. Competitive performance and computational efficiency suggest it as a feasible solution for clinical decision support systems (CDSS).

The Bagging Classifier proves to be reliable as well, which achieves 92.4% accuracy, with the AUC of 0.92. Its F1 score is rate similar of the HistGradientBoostingClassifier, yet robust to repetition thanks to bagging properties of variance reduction. The model is especially useful in the context of an ensemble-based diagnostic system that tends to be stable and robust to data noise.

On the other hand, DT Classifier has intuitive formulation and high interpretability, but its performance is inferior. Bothmean and stdev report the lowest AUC of 0.75, which means that both features are weaker in class discrimination. Although its simplicity make it suitable for explainable AI (XAI) tasks, its low accuracy and generalization ability make it less so for high-stake clinical diagnosis, where the value of predictivity has the highest interest.

Taken together, the results of this comparative study imply that, although all models of the ensemble family are really good to predict heart disease, HGB is superior in terms of accuracy, AUC, and clinical confidence of prediction as well. However, practical issues including computational expense, deployment circumstances, and interpretability should be considered in determining the finalized model for specific healthcare applications.

Chapter 5

Engineering Standards and Design Challenges

5.1 Introduction

This paper proposes a machine learning method for improving early detection of heart disease, by assessing UCI data. The model was able to detect 93.4% accurately close to the state of the art methods based on the related works are shown in Table 5.1 In this section, we discuss our observation with prior studies and discuss the merits of our methods — more specifically we focus on the HistGradientBoostingClassifier which was found to be most effective in predicting heart disease.

5.2 Comparative Analysis and Strength of the Proposed Approach

In addition, a comparison with state-of-the-art methods shows that the accuracy of the proposed model is better than of previous works. For instance, Bhatt et al. (2023) obtained 87.28% of accuracy, by using four algorithms and the Kaggle dataset. Other models mentioned in this study [33,35] often use a lower amount of algorithms, compared to our approach where we use 15 different algorithms. This more general use case gives a quite good accuracy that easily proves the effectiveness of our approach. Similarly, Kavitha et al. (2021) studied the UCI dataset and had three algorithms with an accuracy of 88%. Even though they also used metrics including Mean Square Error (MSE) and Mean Absolute Error (MAE), regression type metrics are not the best choice for classification. On the other hand, our study is more comprehensive, using task-specific techniques such as the confusion matrix and ROC curve.

Bharti et al. (2021) obtained a slighlty higher detection accuracy of 94.2% with six algorithms also in the UCI dataset. However, their trained model showed not very

good performance on other evaluation criteria, likely over-fitting or poor generalization. In contrast, our model, although reaching slightly lower accuracy as 93.4%, supports robust performance in evaluation measures, suggesting that the prediction power is stable.

Moreover, other reports, including that of Miranda and colleagues [5] and of Suter et al. (2021) and Jindal et al. (2021) using the UCI dataset also reported 88.33% and 87.5%, respectively with fewer models in action. We attribute our enhanced performance to larger pool of algorithms and careful evaluation strategies, such as K-fold cross-validation. Additionally, Hossain et al. (2023) us ed a manually collected data set and seven algorithms and obtained the reported accuracy of 90.0%. Their data-set was proprietary, however the usage of a popular and well-recognized UCI data-set and our extensive algorithmic framework made the performance better.

However, a closer look at the results also reveals some limitations of the top models. The Few-shot learning algorithms, notably the HistGradientBoostingClassifier, which obtained the best AUC score of 0.94, have drawback in terms of model complexity and interpretability, which are crucial for the clinical settings. The fact that the model has high computational cost and prolonged training duration could also impede its use in resource poor settings.

The Bagging Classifier (AUC: 0.92) is good to reduce variance, although there is still a risk of overfitting with complex base learners. Its opaque decision-making process further limits its clinical transparency. So do both the Extra Trees Classifier and Random Forest Classifier with an AUC of 0.92, both of whom boast high accuracy but have slightly larger memory requirements and training time, meaning they may be less feasible given available resources.

In general, while the tested models show strong performance in prediction, successful deployment in a clinical practice requires a trade-off between predictive accuracy and interpretability, computational cost, and deployability. These models need to be tested on independent datasets or real world scenarios for stability of all-the-way and for generalization of the methods.

The major strength of the proposed approach is its full use of different algorithms and comprehensive evaluation measures. With 15 classifiers integration, this study provided more complete learning of the intrinsic data patterns and increased predictive values. Of these, Among all evaluated models, the HGBC yielded the highest classification accuracy (93.5%) and maintained a consistent 5-FCV accuracy of 92.4%, underscoring its effectiveness in handling unseen data..

The model accurately predicted both healthy and affected individuals, scoring 93.7% for those without heart disease and 93.4% for those with it. Among the models that were tested, it had the smallest number of false positive and false negative samples, confirming its capacity to successfully discern between healthy and patient. Its performance was also confirmed to be discriminative, which was indicated by a highest AUC value of 0.94.

Moreover, it had the stable results in terms of multiple evaluation measures such as precision, recall, accuracy, F1-score and the ROC curve. This stability demonstrates the model's robustness and generalizability over different data patterns, and therefore strong applicability for early detection of heart disease in real-world clinical settings.

Table 5.1: Comparison

Ref.	Data Source	No. of Algorithms	Metrics	Findings
Bhatt et al. (2023)	Kaggle	4	K-fold validation, Recall, F1-score, Classification Accuracy, Precision	87.28%
Kavitha et al. (2021)	UCI	3	MAE, RMSE, Accuracy Rate, Coefficient of Determination (R^2), Mean Squared Error	88%
Bharti et al. (2021)	UCI	6	Sensitivity, Classification Accuracy, True Negative Rate (Specificity)	94.2%
Miranda et al. (2021)	UCI	1	Precision, Accuracy, F1-measure, Recall	88.33%
(Jindal et al. (2021)	UCI	3	Recall Rate, Precision, F1-score, Overall Accuracy	87.5%
Md. Imam Hossain et al. (2023)	Self-Collected (Primary)	7	ROC curve, Confusion Matrix, Accuracy, F1-measure, Precision, Recall	90.0%

Our Approach	UCI	15	Training vs Testing Duration, ROC-AUC, F1-score, Accuracy, Recall, Precision, Confusion Matrix, Cross-validation	93.4%

5.3 Compliance with the Standards

This research is in line with sound engineering and ethical practices of computer science and health informatics. Methodology The approach follows widely accepted data preprocessing, feature selection, model validation and performance evaluation model development best practices used in machine learning. As for health data we strictly used public and anonymized datasets from UCI repository, respecting all ethical and privacy criteria. Moreover, deploying explainable AI principles promotes responsible AI approaches by helping to make models more interpretable.

5.4 Impact on Society, Environment and Sustainability

This work is in compliance with accepted engineering and ethical standards in the discipline of computer science and health informatics. The approach follows good practices for machine learning model construction, such as data preprocessing, feature selection, model validation, and performance estimation. With respect to healthcare data usage, only publicly available anonymized datasets from the UCI repository were employed by the project, guaranteeing that it was conducted in agreement with ethical and privacy guidelines. Moreover, applying explainable AI principles contributes to responsible AI practices by improving model explainability.

5.5 Sustainability Plan

The mentioned system is sustainable due to its low computational load and flexibility. Models such as HistGradientBoostingClassifier and Extra Trees Classifier, that even though accurate could be run on cloud computing services to also be scalable with respect to environment it without much impact. Open-source tools (e.g., Scikit-learn, Pandas) were used, thus promoting replicability and cost-effectiveness in future developments. Besides, having a modular architecture makes possible further updates without having to change the whole system.

5.6 Project Management and Financial Analysis

The project adhered to an iterative and agile development process, with following main milestones: data gathering, preprocessing, modeling, testing, assessing, and reporting. Time schedule tools, such as Gantt-charts, were used to plan and keep track of work. In financial terms, the project incurred low costs, which mostly derived from the public use data and open-source software, rendering the project very suitable for educational and research institutes in low resource settings.

Table 5.2: Financial Analysis

Cost Item	Description	Amount
Public Data	Access to publicly available datasets(UCI Data set)	$0(Free)
Open Source Software	Tools and frameworks used for development, testing and analysis	$0(Free)
Total Project Cost	Low overall expenditure due to reliance on free resources	$0(Free)

5.7 Complex Engineering Problem

Heart disease prediction involves solving a complex problem characterized by multi-dimensional clinical data, data imbalance, noise, and high-stakes decision-making. The application of SMOTE to balance classes, permutation-based feature importance for dimensionality reduction, and the evaluation of 15 diverse algorithms exemplify how the study tackled these challenges. The integration of ensemble methods also addressed performance instability in single classifiers, enhancing reliability. In addition, permutation-based feature importance was used as a robust dimensionality reduction strategy, helping to identify and retain the most informative features while discarding irrelevant or redundant ones. This preprocessing pipeline was followed by the systematic evaluation of 15 machine learning algorithms, encompassing both conventional and advanced models, to ensure comprehensive performance benchmarking. Moreover, the adoption of ensemble learning techniques—such as bagging and boosting—proved instrumental in overcoming the limitations of individual classifiers by aggregating their predictions, thereby increasing overall model stability, accuracy, and generalizability. These combined strategies highlight a rigorous and well-rounded approach to enhancing predictive performance in heart disease classification tasks.

5.8 Complex Problem Solving

Heart disease prediction involves solving a complex problem characterized by multi-dimensional clinical data, data imbalance, noise, and high-stakes decision-making. The application of SMOTE to balance classes, permutation-based feature importance

Table 5.3: Mapping with complex problem solving.

EP1 Dept of Knowledge	EP2 Range of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of Applicable Codes	EP6 Extent of Stakeholder Involvement	EP7 Inter-dependence
✓	✓	✓	✓		✓	✓

Table 5.4: Mapping with knowledge Profile.

K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K8 Research Literature
✓	✓	✓	✓	✓

5.8.1. Justification for EP Attributes Mapping

- EP1 - Depth of Knowledge: Involves advanced knowledge in machine learning, statistics, and biomedical data analysis.
- EP2 - Conflicting Requirements: Balances model complexity with interpretability and computational efficiency.
- EP3 - Depth of Analysis: Required evaluation across multiple performance dimensions.
- EP4 - Familiarity of Issues: Tackled open challenges in healthcare diagnostics.
- EP5 - Applicable Codes: Followed ML best practices and standard evaluation metrics.
- EP6 - Stakeholder Involvement: Relevant for clinicians, patients, and researchers.
- EP7 - Interdependence: Predictions impact healthcare decisions and patient outcomes.
-

5.8.2. Justification for Knowledge Profile Mapping (linked to EP1):

- K3 - Engineering Fundamentals: Applied mathematical models and algorithms.
- K4 - Specialist Knowledge: Leveraged ML-specific techniques and evaluation.
- K5 - Engineering Design: Designed, tested, and optimized 15 ML models.
- K6 - Engineering Practice: Ensured reproducibility, ethical handling, and performance tuning.
- K8 - Research Literature: Strongly based on recent peer-reviewed studies.

5.8.3. Engineering Activities

Table 5.5: Mapping with complex engineering activities.

EA1 Range of resources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
✓	✓	✓	✓	✓

5.8.4. Justification for Engineering Activities Mapping:

- EA1 - Range of Resources: Used UCI dataset, Python tools, and statistical techniques.
- EA2 - Interaction: Involved consultation with supervisors and peer review.
- EA3 - Innovation: Developed a custom evaluation pipeline and model ranking.
- EA4 - Societal Impact: Potential to improve early diagnosis and reduce healthcare burden.
- EA5 - Familiarity: Tackled partially familiar problems using known and novel techniques.

5.9 Summary

This section situates the technical, ethical, social, and managerial aspects of the heart disease prediction system. The project not met standard and best practice requirements, but took into account awareness of its social impact and sustainability. Through solving an intricate health challenge with a scalable and ethical engineering approach, the project is a potential for addition to the public health technology space.

Chapter 6

Conclusion

6.1 Summary

Around the world, cardiovascular conditions like myocardial infarction (MI) persist as significant sources of death and long-term disability. Based on World Health Organization's (WHO) 2029 projection, these are projected to be responsible for some 17.9 million deaths each year—equivalent to approximately 32% of all fatalities. One of the leading causes of different CVDs is coronary artery disease (atherosclerosis) with narrowing of the arteries which supply blood to the heart (acteries) leading to heart attacks.

Prompt and reliable diagnosis is key in managing heart-related disorders and offer patients a better prognosis. However, conventional diagnostic approaches are laborious and expensive.

In order to surpass historical limitations in heart disease detection, the UCI Machine Learning Repository was used for this research and was tested with 15 supervised machine learning and deep learning methods. The main purpose was to improve detection accuracy of syndromes of early-stage cardiac diseases and find out the best-performing model in clinical practice. The HistGradientBoosting (HGB) algorithm resulted as the best in all the tested methods, with 93.5% of accuracy and stable for precision, recall and F1-score. Importantly, it also worked to cut back on false negatives, something that is crucial in a medical context where the failure to detect a problem can have serious implications.

Ensemble methods like RF, ET, and BC followed closely, showing accuracy scores between 92.2%–92.4%. Their stability across multiple evaluation criteria supports their viability in real-world medical applications where both precision and robustness are critical.

Other models, such as LR, GB, and AB, also performed reasonably well. While slightly less accurate than the top-performing ensembles, they offer advantages in transparency and computational efficiency—making them practical choices in resource-constrained environments.

However, models like SVM, SGD, and GNB underachieved, reflecting lower accuracy and skewed precision-recall distributions. Their higher misclassification rates limit their applicability in clinical prediction scenarios under current experimental settings.

In conclusion, ensemble-based algorithms, particularly boosting and bagging techniques, emerged as the most reliable for early-stage cardiac risk detection. Their consistent and high-level performance makes them strong candidates for implementation in preventive healthcare systems.

6.2 Limitations

Despite showing promising results, the study's scope is constrained due to the limited and not enough diverse training dataset, to support wider generalization (particularly concerning alternative machine learning scenarios). Furthermore, the validation was based on a single dataset, and this may not represent the model's performance in different datasets or in clinical practice. In addition, clinical important factors-such as patient comorbidities-are left out from the model which could influence the predictive performance of the model.

6.3 Future Work

In future, the models should be tested for further validation on larger and more varied patient populations for better clinical utility and generalization. What's more, applying a model such as the HistGradientBoostingClassifier can lead to better disease risk prediction in clinical decision systems. Embedding these frameworks within electronic patient records and other diagnostic data sources have the potential to not only enhance their diagnostic efficacy, but also to dramatically facilitate use in clinical environments.

References