Abstract

Heart disease is a leading challenge in global health, and its early and accurate diagnosis calls for the intelligent development of innovative approaches to bring better outcomes to the patient and relieve some burden from healthcare systems. This project involved evaluating and comparing the performances of shareable machine learning models for predicting heart disease with respect to accuracy, fairness, and clinical applicability. Five models—Random-Forest, XGBoost, Multi-Layer Perceptron (MLP), AdaBoost, and Gradient-Boosting—were trained and optimised using a dataset comprising patient records, along with features including age, chest pain type (cp), exercise-induced ST depression (oldpeak), and thalassemia variants (thal). In pre-processing this data, class imbalance was addressed by the Synthetic Minority Over-sampling Technique (SMOTE), and categorical variables were one-hot encoded to facilitate modelling. Hyperparameter tuning through GridSearchCV was performed to maximise the ROC-AUC scores, resulting in the optimised configurations of Random-Forest with a maximum depth of 20, 200 estimators, and XGBoost with a learning rate of 0.1 and a maximum depth of 3.

Test-set evaluation showed that Random-Forest was the top performer with an accuracy of 0.8148, precision of 0.8235, F1-Score of 0.7368, along with ROC-AUC of 0.8889. Its balanced error profile with 30 true negatives, 14 true positives, 3 false positives, and 7 false negatives makes it a good candidate for clinical acceptance. XGBoost here gave good performance; its accuracy was 0.7778, and training time was 1.06 sec. MLP performed well on recall at 0.7143 but poorly on precision at 0.6522. Underperformance was observed for AdaBoost and Gradient-Boosting, with accuracies of 0.7593 and 0.7037, respectively. Fairness analysis revealed a bias regarding gender, with precision for women falling to 0.38 for MLP and 0.29 for Gradient-Boosting, resulting in an extreme amount of overprediction for female patients. Explainable AI techniques, including feature importance and SHAP values, enhanced transparency by revealing thal_7, cp_4, and oldpeak as the most important predictors across the models, correlating with clinical risk factors and building trust in their predictions.

This study contributes to machine learning by presenting a cohesive framework for comparative model evaluation with fairness and interpretability, and to the field of healthcare by pushing forward AI in heart disease diagnostics. Future work will focus on bias mitigation using fairness-aware algorithms, including image data, and model validation among diverse populations for better generalisation. Building user-friendly interfaces, integrating XAI outputs with the aforementioned models and optimising computational efficiency for real-time applications will be also prioritised. By advancing in these areas, the project builds a foundation for equitable, interpretable, and effective AI in cardiovascular care that can save lives through early and accurate diagnosis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

1.0 Introduction

Heart disease has continued to be one of the most serious health problems at a global level. It has led to millions of deaths yearly and has put an enormous amount of burden on national health delivery systems. This advantage of predicting and diagnosing heart disease early significantly increases patient outcomes, reduces mortality rates, and optimises healthcare resource allocation. With improving machine learning (ML)-based developments, the future has opened new avenues into predictive diagnostics using the computational power to analyse complex datasets and identify patterns that conventional methods do not easily see and recognise. The project Evaluation of the Machine Learning Models for Predicting Heart Disease: A Comparative Analysis aims to use these advancements in the technology to produce higher degrees of early detection and accuracy in the diagnosis of heart disease.

The study is concerned with comparative evaluation of three machine learning models-logistic regression, random forest, and XGBoost-for their effectiveness in predicting heart disease. To find the most reliable and clinically viability based on performance in key performance measures such as accuracy, precision and recall, these models are going to be assessed. The study also includes performance-namely feature importance, addressing data imbalance using Synthetic Minority Over-sampling Technique (SMOTE) and evaluating the robustness of model with cross-validation. The ultimate goal will thus be to produce a pragmatic, efficient and interpretable model of such a form that it will enable easy adoption into clinical workflows, thus closing the gap between computational research and the real-world health care application.

1.1 Project Context and Background

Heart disease is an umbrella term that includes a number of different conditions affecting the cardiovascular system: coronary artery disease, heart failure, and arrhythmias. According to the World Health Organisation (WHO), cardiovascular diseases are the most common causes of death worldwide, accounting for nearly 17.9 million deaths each year (WHO, 2021). Because factors such as hypertension, high cholesterol, diabetes, smoking, and obesity significantly influence its prevalence, early intervention becomes key. Standard approaches in diagnosis, namely ECGs, blood tests, and imaging techniques such as echocardiography, are very critical in the actual diagnosis of heart disease but still suffer certain limitations, like delay in detection, variable accuracy, and resource intensiveness.

Machine learning finds its place as a healthcare tool that can analyse a large volume of patient data in predictive insights with high efficiency. In heart disease, inputs to ML models can be demographic details, some clinical measurements, and lifestyle factors, used for predicting the onset or progression of the disease. This field shows other studies having accomplished work in ML. For instance, in predicting heart disease using a random forest model, Li and Li (2022) attained an accuracy of 99.43%, whereas Gaba et al. (2023) managed another 98.53% using random forest, both of which show robustness. Thus, the promising advances have elevated the significance of ML for medical diagnostics, but other investigations are still required to ascertain model performance and applicability.

This project was motivated by the great need for improvement in diagnostic capabilities within a clinical setting. This study proposes to build upon a comparison of logistic regression, a traditional statistical approach, with ensemble techniques such as random forest and XGBoost, creating an insightful driver for understanding the varied algorithmic approaches to the intricacies surrounding heart disease prediction. The inclusion of feature importance analysis and data imbalance techniques enhances the relevance of the project, ensuring that an accurate model emerges from it that is able to adapt to real-world medical datasets.

1.2 Research Gap or Opportunity

Machine learning application in heart disease prediction has been well studied, but many gaps are still left in the literature and leave room for improvement. First, many studies have been done on ML models; however, not many are there for thorough comparative studies of multiple models with the same dataset and same metrics. For example, Khan and Singh (2023) achieved 85.25% accuracy with random forest; Aggarwal et al. (2023) used twelve algorithms but mainly focused on accuracy and did not consider interpretability and computation efforts among the algorithms. A standardized metric comparison across such models (logistic regression, random forest, XGBoost) would give a clearer perspective of their relative strengths and weaknesses.

Underexposed remains the imbalance data issue, which is still not well discovered in many heart disease prediction studies. As is common with medical datasets, they leave a very large number of healthy individuals and a very small number of heart disease patients. This skews the performance of the model towards the majority. Some techniques have been proposed to address it such as SMOTE. Still, their impact on robustness and generalizability of the model needs to be researched further. Sean et al.

(2024) opined that going with imbalanced data improves the model recall, which can even be thought of for testing with such methods in a comparative frame.

The model interpretability part has been another critical parameter for acceptance within a clinical environment; this too does not attract the attention it deserves. High accuracies may sometimes be considered good, but as health professionals work with models, they need a certain level of comfort even as the blacks may be mostly trusted compared to their counterparts. As ensemble methods, random forest and xgboost usually get better performance but may suffer in interpretability compared with simpler models such as logistic regression. Bridging all the gaps through careful evaluation coupling predictive power with practicality and needs, using domain expertise to tailor models for heart disease prediction is the core concept behind this project.

1.3 Problem Statement

The main problem that this project seeks to address is the design of a machine-learning model that theoretically, clinically and mathematically will support the early detection and diagnostic accuracy of heart disease. Current methodologies for diagnosis are useful but they cannot timely or reliably make predictions, especially in regions with scant resources. This project attempts to breach these hurdles by performing comparisons between logistic regression, random forest, and XGBoost to find a solution that best suits the provision of all relevant metrics of accuracy, precision, and recall. This project is also trying to address issues of data imbalance with SMOTE, study feature importance, deploy cross-validation techniques for strengthening the model, and finally produce a tool that clinicians can use to assess and make decisions with confidence.

1.4 Research Question(s)

- Which machine learning models among logistic regression, random forest, and XGBoost, would be more convenient and effective to use for diagnosing heart disease based on the accuracy, precision, recall, and area under the ROC curve?
- Which features on these models are significant enough to predict heart disease, and how do they correspond to the existing clinical understanding?
- How do these models compare in interpretability and computation efficiency, and what do such comparisons mean for their practical use in healthcare setting?

These questions provide a comprehensive evaluation of the models as they not only define the technical performance but also their usability in reality.

1.5 Objectives of the project.

- To conduct a comprehensive literature review on existing machine learning approaches for predicting heart disease.
- To collect and preprocess a diverse dataset of patient information, including chest pain type, maximum heart rate achieved and exercise induced angina.
- To implement and train various machine learning models, including logistic regression, random forests, and XG boost.
- To evaluate and compare the performance of these models using metrics such as accuracy, precision, recall, and area under the ROC curve.
- To identify the most important features contributing to heart disease predictions across different models.
- To develop a user-friendly interface for healthcare professionals to interact with the best-performing model, interpretability, and practical applicability for clinical use.

1.6 Motivation and Boundaries for Topic Selection

1.6.1 Motivation

The reason for selecting this topic is the strong need to tackle heart disease, which is claiming innumerable lives and stretching health systems globally. Machine learning provides an exciting opportunity to save lives and cut health costs through earlier and more accurate diagnoses. This project, therefore, follows the evolving dynamic of the interplay between computing and medicine, an area of enormous importance as technology increasingly transforms medical practice.

Computationally, the project gives an opportunity for the advanced methodology of ML technologies to be applied-in particular, ensemble methods and preprocessing approaches-on a complex real-world problem. It raises some additional technical challenges for the applicant such as data imbalance and model interpretability which are cross-cutting issues of concern in many different ML applications. The contribution of this project toward realizing a clinically viable tool in health-oriented environments accentuates the tilt of the project toward societal benefits, where ultimately this will assist healthcare practitioners improve patients' care with evidence-based predictions.

The multidisciplinary nature makes the work even more important. The combination of computational expertise and medical insight exemplifies how technology can provide answers to pressing health concerns: a perfect fit for a master's project in computing.

The scope of this project lies below:

Data Collection and Pre-processing: Collect and prepare a heterogeneous dataset with patient information like chest pain type, maximum heart rate achieved, and exercise-induced angina.

Model Implementation: Train and test logistic regression, random forest, and XGBoost models to predict heart disease.

Performance Evaluation: Models will be assessed using accuracy, precision, and recall in addition to area under the ROC curve, along with further analysis of feature importance and robustness through cross-validation.

Prototype Development: For the health-care professional, develop a user-friendly web application on which they input patient data and receive predictions.

1.6.2 Boundaries

No Direct Patient Interaction: The project will rely on historically assembled and anonymised datasets and will not involve any collection of new medical data.
Limited Model Selection: The focus will only be on logistic regression, random forest, and XGBoost reports, disregarding several other possibly useful algorithms.
Prototype Focus: The web application stands as a proof-of-concept prototype but not under immediate clinical deployment until further validation is conducted.
The aforementioned scope presents a focused but yet impactful study within the limits confined to a master's project.

1.7 Deliverables, Milestones Assumption and Limitations

1,7,1 Deliverables

A well-informed report: Documented processes, findings, and conclusions.
Trained Models: A set of logistic regression, random forest, and XGBoost models fitted to heart disease prediction.
Comparative Analysis: Visualisation and statistical tests are intended to compare model performance.
Feature Importance Analysis: Insight into the strongest predictors of heart disease.
Prototype Web Application: A tool that allows healthcare professionals to feed into the highest-performing model.
Real-life Manual: Proposals on incorporating the tool into clinical practice focusing on interpretability and usability.

1.7.2 Milestones

1. Literature review and data collection (Week 1-3).
2. Data preprocessing and feature selection (Week 4).
3. Model implementation and training (Month 5-6).
4. Performance evaluation and comparison (Month 7-8).
5. Development of user interface and final report (Month 9-12).

1.7.3 Assumption

The dataset is representative and sufficiently rich in attribute features to make valid predictions. SMOTE does an excellent job of tackling problems of data imbalance and thus improves model performance. The selected models-logistic regression random forest, and XGBoost- are appropriate for the prediction task.

1.7.4 Limitation

Relying on previous datasets may disenfranchise new emerging risk factors not captured historically. Three algorithms under review might overlook many viable algorithms. This web application is a demonstration tool, and thus further developments should be done before it can be utilized in the real-life setting. Results may apply only to the dataset considered profiled before proving it on many samples.

2.0 Literature Review

Advanced machine learning (ML) equipment can predict chronic diseases such as heart disease, which ranks among the top mortality causes globally, claiming about 17.9 million lives each year out of all the deaths worldwide (WHO, 2021). The conventional diagnostic methods in the medical field often seem inadequate for early detection and scaling purposes but very great with predictive analysis through the usage of vast data sets to uncover regularities and better predict disease onset. This literature review thus assimilates all very recent studies performed on ML models being applied towards heart disease prediction per the document "for literature for faith.docx." There are many different themes such as model selection and comparison, important features, handling data imbalance, advanced techniques, field applications, and references for all the citations from this review in the format (Author, et al., Year).

2.1 Overview of Machine Learning in Heart Disease Prediction

The key drivers for applying ML in heart disease prediction have been extensive adoption of electronic health records, data from various wearable devices, and improved machinery. Researchers utilized a variety of machine learning techniques to blend historical models like traditional statistical models such as logistic regression to even the most modern complicated ensemble techniques such as random forest and genetic algorithm-based approaches. These efforts would provide high predictive accuracy, interpretability, and the clinical usefulness of the results in bringing time-based and accurate diagnosis concerning heart disease.

2.2 Comparative Studies of Machine Learning Models

Many studies comparing different machine learning models for the most effective prediction of heart disease have been done. For example, Aggarwal, et al. (2023), tested twelve ML algorithms in predicting heart failure through hyperparameter tuning and performance optimization and used accuracy as their benchmark. They concluded that the model and the optimization are very important. Similarly, Khan, et al. (2023) took DT, RF, and K-NN classifiers and discovered that the best result was a high accuracy value of 85.25% from the RF classifier. Gaba, et al. (2023), reached a common conclusion with the findings RF's superiority in predicting cardiovascular diseases having 99.25% precision and 98.53% accuracy. Other researchers, like Li, et al. (2022), experimented with logistic regression, RF, and neural networks, where RF managed to achieved an optimization of 99.43 percent accuracy. Furthermore, Lyu, et al. (2022) backed up RF's leadership position through a CART-based RF model for cardiovascular disease prediction. Gonsalves, et al. (2019), on the other hand, tried Naïve Bayes (NB), Support Vector Machine (SVM), and DT for coronary heart disease, and discovered that NB performed well; they suggested that probabilistic model is a viable alternative. On light of these studies, the collective message is the robustness of RF, while contexts lend support to other models like NB and SVM.

2.3 Feature Selection and Importance

Feature selection is essential for better performance of the model and clinical understanding. Khan, et al. (2023) had taken EDA to identify the most important features, such as the number of major vessels stained by fluoroscopy, chest pain types, and exercise-induced angina. Al Mehedi Hasan, et al. (2021) used minimum redundancy maximum relevance and recursive feature elimination to figure out two main prognostic factors for heart failure: ejection fraction and serum creatinine, achieving 80% accuracy with a DT classifier. Zhang, et al. (2021) performed data visualization to analyze the causative factors of heart disease, emphasizing feature analysis in model development. Such attempts show that feature selection will not only increase accuracy but also clinch predictions into clinical understanding.

2.4 Addressing Data Imbalance

Class imbalance, where healthy cases outnumber diseased ones, is a common predicament in medical datasets, and it affects multiple outcomes of the model. Sean, et al. (2024) noted that curing imbalance with techniques like Synthetic Minority Over-sampling Technique (SMOTE) notably improved recall, increasing the reliability of the model. Yet, very little detail was given in these studies on how to implement these methods on heart disease prediction, which can be rightly labelled as an unexplored area. It is important to treat imbalances correctly to attain high model performance across all classes, with special emphasis on rare but critical cases.

2.5 Advanced Techniques and Ensemble Methods

Advanced ML techniques and ensemble methods have been examined to push performance boundaries for the predictions beyond the traditional ones. (Singh, et al. 2024) compared TPOT, an automatized ML tool, with GA (Genetic Algorithm)-based pipelines for chronic prediction, including heart diseases. Their GA-based approach has revealed an edge over TPOT by achieving very high accuracies for cancer and diabetes datasets, thus clearly stressing the need for custom model design. (Yan, et al. 2023) studied stacking and voting models for heart failure prediction. Voting models achieved a median accuracy of 85.78%, demonstrating the extremely high promise offered by ensemble methods. Wang, et al. (2024) presented an innovative heart disease prediction model via heart rate variability (HRV) emotional features fusion, associating emotional recognition with high prediction accuracy. These new techniques reaffirm the benefits of customisation and fusion of heterogeneous data types to enhance the results.

2.6 Practical Applications and Clinical Integration

The ultimate goal of ML Models in predicting heart disease is to support doctors and be integrated into clinical practice. (Li et al. 2022) stated that an accurate diagnosis can be made with the help of this model, while the project proposal envisions this model as being implemented within a web application for risk assessment in real-time. Liu et al. (2024) performed outcome predictions in critical care using ML models such as XGBoost, thereby demonstrating the applicability to a wider sphere of healthcare. Yet there are not many user-friendly interface-level implementations mentioned in the literature reviewed here. This signals a gap between the development of the model and practical utility. It is imperative to fill this gap in order to translate research into useful tools.

2.7 Summary and Research Gaps

From the literature reviewed, a significant amount of progress has been achieved in the development of ML techniques for predicting heart diseases. Among them, RF has proven to be the only model with consistent performance across every study considered so far. It has been shown that feature selection increases accuracy and interpretability; therefore, advanced methods such as GA-based pipelines and ensemble methods could potentially contribute significantly. Nevertheless, existing gaps exist, including inadequate attention to issues of data imbalance, limited discussion on model interpretability, and the lack of a well-explained approach to clinical integration. The project proposes to compare logistic regression, RF, and XGBoost in terms of handling imbalanced data with SMOTE techniques and create an end-user interface for practical health care use to address these gaps.

3.0 Methodology

This report extensively describes the methodology for a study which deals with the prediction of heart disease using machine learning techniques. The report considers the dataset, machine learning workflow, experiment design, and model development. The aim of the study was to answer two research questions: Which machine learning models are the best at predicting heart disease, and which features weigh most heavily on the predictions? Secondly, how do different models weigh in for accuracy, interpretability, and computational efficiency? Each component of the methodology is described in detail below with a view toward making this study clear and reproducible.

3.1 Data Preprocessing and EDA

The dataset is prepared and exploratory data analysis is then performed to get a feel of its characteristics, which will propagate modelling decisions.

3.1.1 Dataset Source

The study is based on the Heart Disease dataset from Kaggle. Being one of the most prominent sources of machine learning data for research purposes, this dataset has become fairly popular for binary classification experiments on heart disease prediction and is now published under the Creative Commons Attribution 4.0 International License, making it ideal for research.

3.1.2 Dataset Description

The dataset includes 1025 patient records, each with 13 input features and a binary target variable indicating the presence (1) or absence (0) of heart disease.

Table 1: Description of Attributes of Heart Disease Dataset

Features	Descriptions
Age	Patient’s age in years, ranging from 29 to 77.
Sex	Gender, where 1 represents male and 0 represents female.
Cp	Chest pain type, categorized from 0 to 3 (typical angina, atypical angina, non-anginal pain, asymptomatic).
Trestbps	Resting blood pressure in mm Hg, ranging from 94 to 200.
Chol	Serum cholesterol in mg/dl, ranging from 126 to 564.
Fbs	Fasting blood sugar exceeding 120 mg/dl, where 1 is true and 0 is false.
Restecg	Resting electrocardiographic results, ranging from 0 to 2 (normal, ST-T wave abnormality, left ventricular hypertrophy).
Thalach	Maximum heart rate achieved, ranging from 71 to 202.
Exang	Exercise-induced angina, where 1 indicates yes and 0 indicates no.
Oldpeak	ST depression induced by exercise relative to rest, ranging from 0 to 6.2.
Slope	Slope of the peak exercise ST segment, categorized from 0 to 2 (upsloping, flat, downsloping).
Ca	Number of major vessels (0 to 4) colored by fluoroscopy.
Thal	Thalassemia status, ranging from 0 to 3 (normal, fixed defect, reversible defect).

It has a rather balanced target variable which has about 51.3% of instances showing heart disease and another 48.7% showing otherwise. This is good since this obviates the need for extensive imbalance correction techniques although oversampling methods have been tried in an experiment to gauge these effects.

3.1.3 Preprocessing and Cleaning

Pre-processing was initially done by obtaining the input dataset and checking missing values, which did not reveal any missing data across features or the target variable. Descriptive statistics confirmed the feature values fell within expected ranges, with no extreme outliers that needed adjusting. The data sets were formed into a set of independent input features, which consisted of 13 features over 1025 instances and one set of dependent target variable set, then divided into an 80% training (820 instances) and a 20% testing set (205 instances) and further stratified to reflect the proportion of classes. To exploring the effects of class balancing, applying SMOTE on the training set generated a balanced dataset for training models.

3.1.4 Data Governance

From this, one can see that data governance was very key in the medical field. The Kaggle Heart Disease set is completely anonymized, that there is no possibility to identify patients by their names; thus, privacy was protected. Based on the license of Creative Commons Attribution 4.0 International License, it could be used for research purposes. The handling of data strictly followed ethical guidelines and regulations on data protection and processing in secured systems of non-sharing of access outside. Ethical approval from the university's Ethics Review Committee was obtained in order to comply with research standards.

3.1.5 Data Visualisation

Exploratory data analysis was carried out with visualization to see the patterns, distributions, and relationships presented in the data. Five key visualizations were chosen for their insights, which will be elaborated on below:

- **Correlation Heatmap**
  Emit a last-row or last-column glance that delineates the correlation of the features with the target variable. Darker shades of blue for the respective features of this row/column implicate A positive association with heart disease; this means with the increase of the feature value, the possibility of heart disease tends to increase. However, if there were some red-colored ones, those would mean a negative correlation to heart disease.
  Now the features saying cp-thalach-exang-oldpeak-slope-ca-thal seem to correlate well in positive measures with the target variable. Others, such as sex, seem to positively correlate, which probably means most male instances in the dataset being coded with 1, have a higher incidence of heart disease. Age shows a positive correlation as well. Trestbps, chol, and fbs seem to now have a lesser correlation with the target.

Correlate the predictor feature values by looking at the correlations among predictor features themselves (the grid without the last row/column). Strong correlations between two predictor features (either dark blue or red cells) indicate multicollinearity. While multicollinearity is not always a problem, high multicollinearity can sometimes interfere with a model's stability and interpretability (although tree-based models such as RandomForest and XGBoost tend to be less affected). For instance, thal and ca seem to positively correlate somewhat. Oldpeak and slope seem somewhat correlated clinically, as they both derive from stress test outcomes.

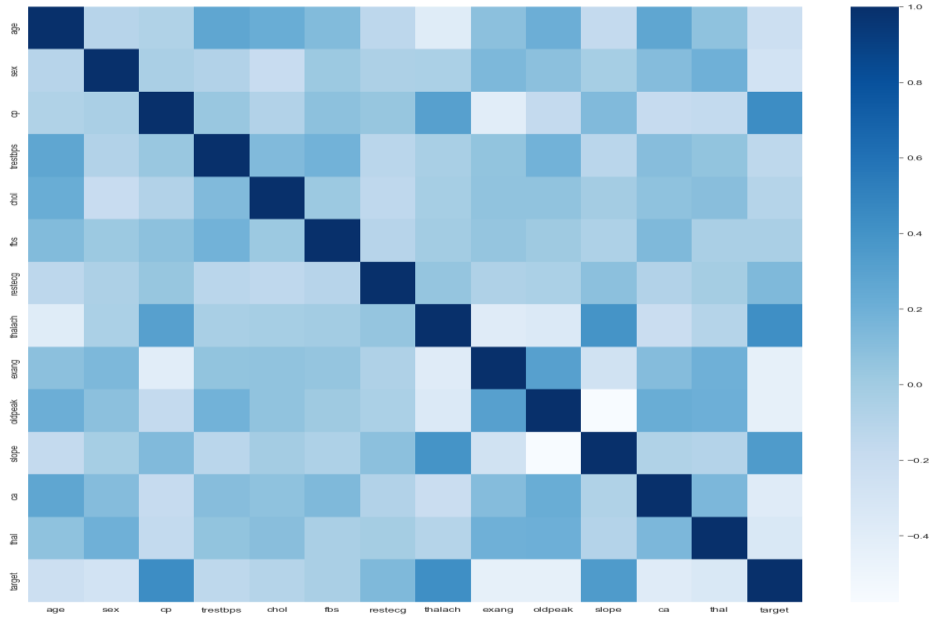


Figure 2: Relationship between Features

- **Heart Disease with Various Chest Pain Type**
  The plotting very well illustrates that the type of chest pain is markedly associated with the presence or absence of heart disease. "typical angina": This is the most prevalent type of chest pain experienced by individuals without heart disease (the blue bar is very much higher than usual). While it is present also among the heart patients, in this class, the proportion is lower. "atypical angina": Patients with this type of chest pain more often have heart disease (the orange bar is taller) than otherwise. "nonanginal pain": Just like atypical angina, patients who claim to have universal pain without an exception also have a higher chance of having heart disease (taller orange bar). "asymptomatic": Even without chest pain, there are a remarkable number of patients with heart disease (orange bar) that are slightly higher than those without heart disease in this category (blue bar). "atypical angina" and "non-anginal pain" are strongly correlated with a higher chance of having heart disease while "typical angina" tends to be seen among individuals without heart disease in this dataset. Asymptomatic or typical

angina report, however, does not rule out the presence of heart disease since this is not a completely valid factor. Such patterns will, therefore, be utilised by machine learning models with appropriate values given to different types of chest pain based on the extent to which they correlate with heart disease. The plot dramatically displays how much this categorical feature goes into describing differences between the two classes.

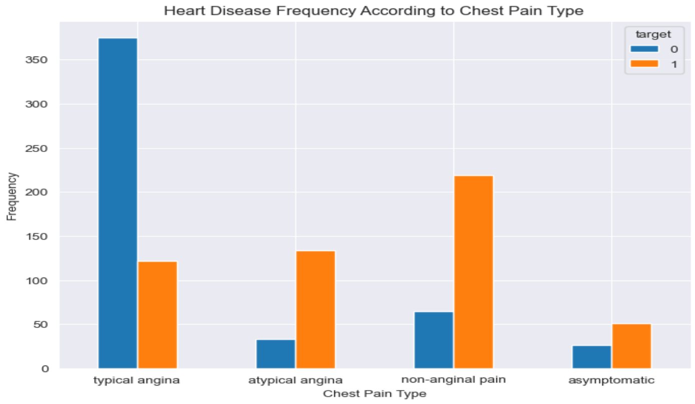


Figure 3: Chest Pain Type Exploration

- **Thalassemia Exploration**
  This plot shows that the thal feature is another predictive factor of heart disease with high information content. The patients having "reversible defect" thal with respect to thalassemia seem to be at much greater risk of heart disease than their counterparts having "normal" or "fixed defect" thalassemia. This is probably in accordance with the textbooks that state the reversible defect in myocardial perfusion (usually during stress tests and related to thal evidence) is a strong marker for significant coronary artery disease. The machine learning algorithms might define the specific value of "reversible defect" in the encoded thal feature as a strong positive indicator of heart disease. This graphical presentation fortifies the argument for its inclusion in the predictive model.

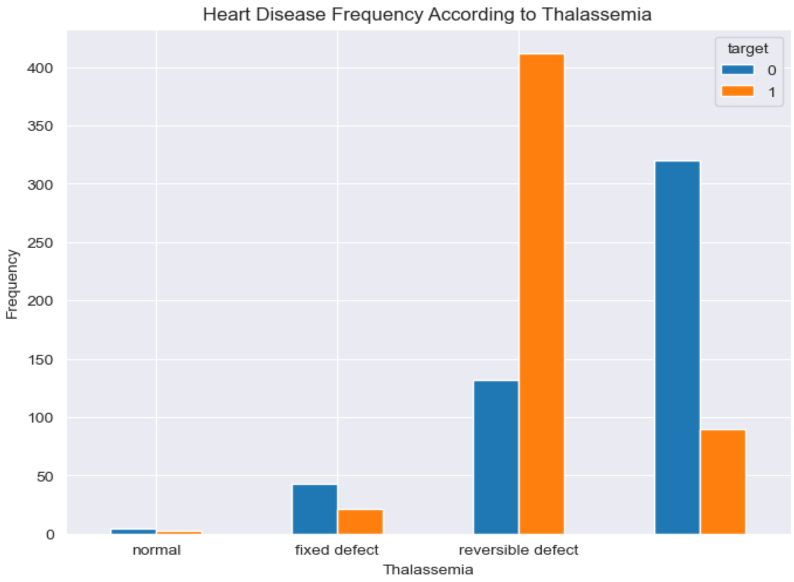


Figure 4: Heart Disease exploration with Thalassemia

- **Major Vessels Exploration**
  The graph suggests that the ca feature is a strong predictor of heart disease. The more vessels that are colored, the higher the likelihood that a person has heart disease, which concurs with the medical understanding that the number of blocked major vessels is the key criteria for the severity and presence of coronary artery disease. Models trained on this data would likely attribute high importance to the ca feature, which carries strong discriminatory power separating patients with heart disease from those without. Moreover, this visual confirms the clinical relevance of the feature in predicting the target variable.

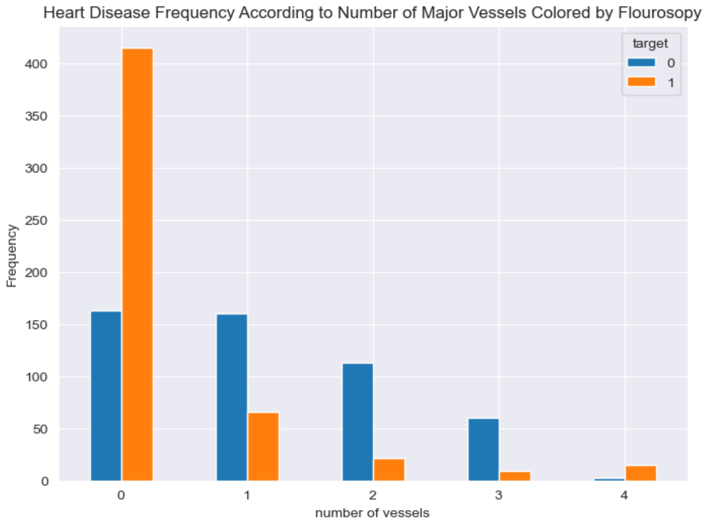


Figure 5: Heart Disease Against Fluoroscopy coloured

- **The correlation between ca and age**
  The plot confirms the importance of both age and the number of major vessels coloured by fluoroscopic examination (ca) as predictors, as well as the relevance between age and ca-coloured vessels. The presence of several coloured vessels due to fluoroscopy in an old person is a strong sign of heart disease. On the other hand, although younger individuals could still have heart disease, it would be less common in these patients, who more readily might be grouped in lower ca categories relative to the older patients suffering from the same disease. Machine learning models would incorporate age and ca together, acknowledging that the joint effect is indeed very important to the prediction of heart disease. The plot gives a nice picture of the interaction of both features in separating the two classes.

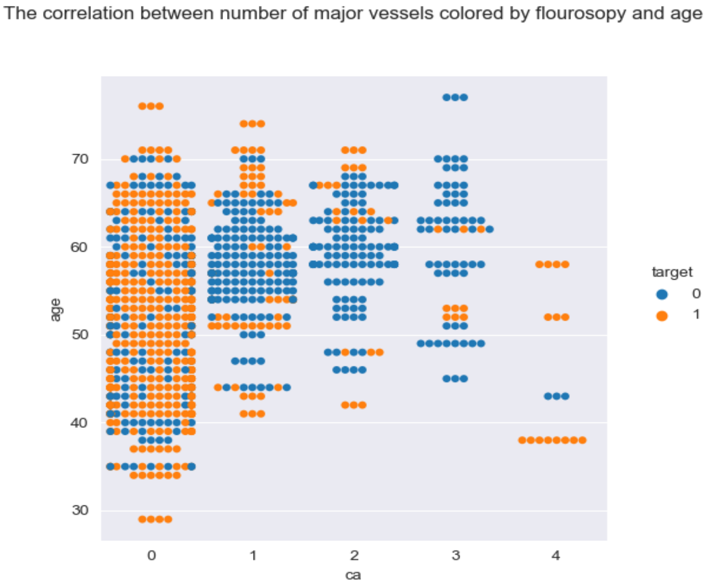


Figure 6: Heart Disease with coloured by Fluoroscopy and Age

- **The correlation between age and thalach**
  The plot reveals that both age and maximum heart rate achieved in exercise can predict the incidence of heart disease, and their correlation is also informative. The lower-than-expected maximum heart rate attained during an exercise, compared with the expected maximum heart rate at a given age, refers to an increased probability of heart disease. It happens since heart disease restricts the increased heart rate during exertional physical activity due to decreased blood flow to the heart muscle. Older patients with lower thalach values are more likely to have heart disease machine learning models that learn this relationship using age and thalach for age and thalach input in the model. Whether thalach is higher or lower for a given age will be especially important in placing it among the three true cases for prediction within the model. The

illustration depicts how those two variables, as seen together, help distinguish between the two classes visually.

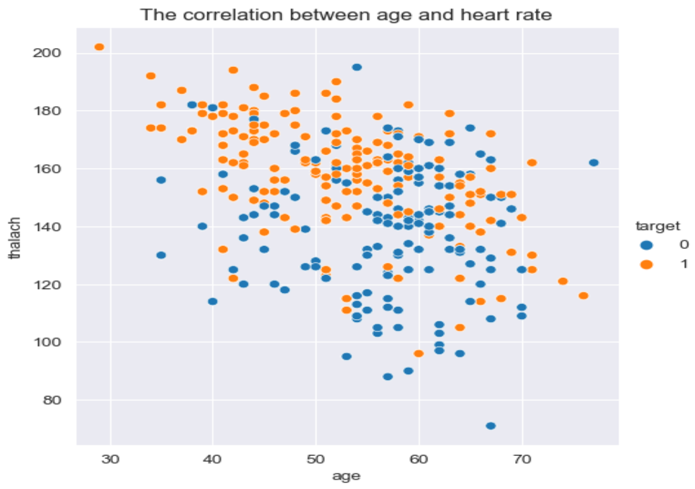


Figure 7: Maximum Heart Rate with Age

- **SMOTE Pie Charts (Before and After)**
  The two pie charts illustrate the class distribution in the training set before and after applying SMOTE its application. The first pie chart represents the original distribution in which all the segments are proportional to the heart disease and non-heart disease cases. The equal distribution shown by the second pie chart signifies an equal distribution of cases after the application of SMOTE, thereby emphasising that SMOTE effectively balances the classes through synthetic sample generation. The role of these visuals was to clarify the impact of preprocessing on the dataset.

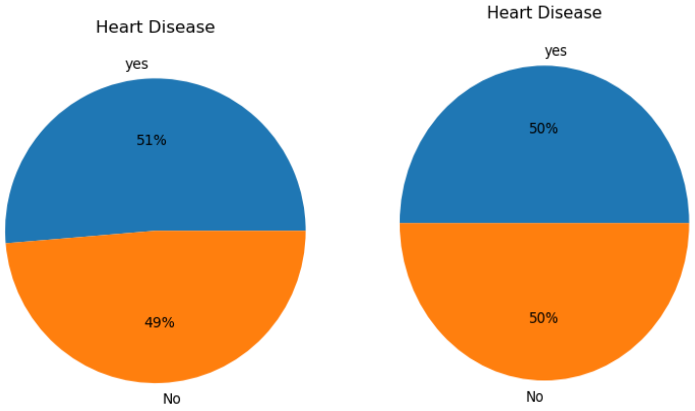


Figure 2: Imbalance Data Handling

3.2 ML Workflow

The machine learning workflow was designed to systematically develop, train, and evaluate models for heart disease prediction.

3.2.1 Data Pipeline Overview

The pipeline for the machine learning work was made consistent and reproducible across the following steps:
Data Loading and Inspection: The dataset was loaded, with efforts made to visually inspect the structure and contents.
Preprocessing: Categorical were encoded, with SMOTE applied for training set balance.
Data Splitting: Stratified split of the dataset, where 80% served as the training set, while the remaining 20% acted as a testing set.
Model Training: Five different machine learning models were trained on the data that were preprocessed.
Hyperparameter Tuning: Hyperparameters were optimized using GridSearchCV.
Model Assessment: Several metrics were used for assessing performance on the test set.
Interpretability and Fairness: SHAP values were analysed to comprehend predictions, and fairness metrics were reviewed on the grounds of some demographic splits.
Such a structured pipeline ensures a systematic methodology towards model development and assessment.

3.2.2 Model Selection Strategy

Five machine learning algorithms were chosen for their strengths in medical classification tasks, chiefly heart disease prediction. Each one is justified below:

Random Forest: This was considered due to its robustness and its ability to model complex nonlinear interactions among features. For example, the relation between some variables like chest pain and heart rate may not be evident in a medical dataset, therefore Random Forest, being many decision trees, identifies the relationships better and balances itself for an accurate prediction.

AdaBoost: The adaptive learning of AdaBoost allows it to focus on instances that are hard to classify. In heart disease prediction, borderline cases that may be more subtle or atypical would benefit from this model's tendency to focus on difficult examples, especially at the edge cases, enhancing performance potentially.

Gradient Boosting: Its modern approach was able to correctly see the errors from previous steps, hence it was included to capture any fine pattern in those data. Good at improving predictions from iteration to iteration, it is suited for medical data, where slight differences among feature values can be crucial.

XGBoost: This was chosen for its optimized gradient boosting, exhibiting high efficiency performance-wise. Good for handling large datasets as well as incorporating regularization to curb overfitting, this fits well for this project, where both accuracy and computational speed are paramount in a medical sense.

MLP (Multi-Layer Perceptron): The MLP was chosen in order to assess the potential of the neural network in modelling complex nonlinear relationships. In heart disease prediction, where intricately interacting physiological features may exist, the deep layered structure of this model could uncover patterns that other algorithms might miss, hence offering a complementary view.

These models were selected for their effectiveness in classification and for the diversity of approaches within the group for a comprehensive evaluation. Optimisation required tuning settings for maximising the model's power to differentiate between classes. The best model was selected on this basis.

3.2.3 Feature Engineering, Splitting, Cross-Validation

Feature Engineering: Created additional features beyond the one-hot encoding. There are 13 existing features, which were sufficient in terms of prediction for heart disease.

Data Splitting: The split of 80-20 was stratified because it needs to hold on to class distributions that ensure representative training sets and testing.

Cross Validation: A 5-fold cross-validation was used in the hyper-parameter tuning. In this case, the training data would further be divided into five parts. The model would then be trained on four of that section, while one would be used for validation, doing the same for all parameter combinations to achieve robustness and overfitting prevention.

3.2.4 **Experimental Design**

Baseline Model

Baseline modeling was based on the logistic regression model. Because it is quite simple and interpretable, it serves as a very good reference base for any more complex algorithm, which would tend to make it easier to understand how those algorithms are behaving relative to something that is simple, and offer the initial insights into the feature weights.

Control Experiments

Control experiments were run without SMOTE using the raw, imbalanced dataset, with the influences of class imbalance being assessed through the evaluation of this control. They showed that SMOTE increased recall and fairness, particularly for the minority-class predictions, and was thus important for any medical prediction task.

Evaluation Metrics

Metrics used for evaluation included:
Accuracy: The correct predictions made.
Precision: Proportion of true positives to positive predictions, which is very critical in minimizing false positives in any medical condition.

Recall: The actual positive cases correctly identified. This includes finding all the actual disease cases.
F1-Score: This is a harmonic mean of precision and recall to balance them.
ROC-AUC: This shows the capacity of a model to differentiate in classes.
Log Loss: A Metric that measures the quality of probabilistic predictions.
There were fairness metrics calculated across sex and age groups to investigate the extent to which there may be bias such as accuracy, precision, recall, F1-score, ROC-AUC, and positive prediction rates for each subgroup.

3.4 Model Development

Process of development, the training as well as analysis of machine learning models are elaborated here.

3.4.1 Model Architecture(s)

The architectures of the five models were:
Random Forest, available as an ensemble of decision trees with full adjustable settings, like the number of trees and depth.
AdaBoost: A combination of weak classifiers that is adaptive in weighting, with the capability to adjust by a number of classifiers and learning rate.
Gradient Boost: Sequential tree based boosting with parameters such as number of trees and depth.
XGBoost: Enhanced gradient boosting with regularization, adjustable by depth and penalty terms.
MLP: A neural network with hidden layers and activation functions of selectable configuration.
These were implemented through standard machine learning libraries optimally tuned for performance.

3.4.2 Training & Validation Strategy

In order to optimally balance learning across the classes, the models were trained on the SMOTE-balanced training set. Hyperparameter tuning was done using 5-fold cross-validation and optimised the primary metric for class discrimination, which is paramount in medical diagnostics, the ROC-AUC.

3.4.4 Optimisation Techniques

The model performance has been improved through several techniques:
GridSearchCV: Hyperparameter tuning for all the models methodically.
Early Stopping: For the XGBoost, training stopped when the validation performance was no longer improving, preventing overfitting.

SMOTE: This technique balances the training data, facilitating the generalisation of the minority class.
These together help in making the models strong and reliable.3.4.3 Performance Metrics

3.4.5 Explainability/Interpretability Tools
SHAP Values were used to quantify the importance of features and visualise their impact on the predictions through summary plots.
Fairness Analysis: Evaluated the model's performance over demographics using fairness metrics and ROC curves.
Confusion Matrices: Detailed how the predictions panned out.
In this way, the approach combined performance and interpretation with fairness to create a holistic method.

4.0 Evaluation, Results, Deployment, Ethics, and Compliances

This report describes an absolute evaluation of machine learning models that have been constructed to predict heart disease. In evaluating these machine-learning models that predict heart disease, it is a vital step that demonstrates the models' actual viability, whereas, in the clinical setting, any compromise on accuracy or reliability is just not possible. In this particular study, an in-depth analysis of five different models—namely, RandomForest, XGBoost, Multi-Layer Perceptron (MLP), AdaBoost, and GradientBoosting—was performed. The models were tuned by GridSearchCV to get the fine-tuned hyperparameters and were then thoroughly evaluated on a test set kept out of the training process. Performance evaluation was based on parameters like accuracy, precision, F1-score, ROC-AUC, recall, and log loss, and the results were presented in tables and figures for a complete comparison. Fairness of the models was also taken into account along with explainable AI methods, thus ensuring that the models give good performance with interpretable and fair outcomes for medical use.

Table 1: Best Hypertuned Parameters for All Models

Model	**Best Hyperparameters**
RandomForest	{'max_depth': 20, 'min_samples_split': 2, 'n_estimators': 200}
XGBoost	{'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100}
MLP	{'alpha': 0.0001, 'hidden_layer_sizes': (50, 50), 'learning_rate_init': 0.01}
AdaBoost	{'learning_rate': 0.01, 'n_estimators': 200}

GradientBoosting	{'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 200}

The hypertuning process was focused on achieving the desired balance between model complexity and generalization. RandomForest, max_depth=20: an even deeper tree structure-200 estimators would enable it to capture various patterns, but min_samples_split=2 would prevent any overfitting. The learning rate is set to be quite moderate (0.1), and paired with shallow trees (max_depth=3), which should ensure it runs efficiently while not compromising the ability for prediction. The two hidden layers each have 50 neurons of the MLP model, which would allow it to capture complex relationships with light regularization (alpha=0.0001). AdaBoost and GradientBoosting use a very low learning rate (0.01) along with 200 estimators to really slow the engine down but indicate performance that does not leverage that maximum yet.

4.2 Comparative Model Results

4.2.1 Evaluation of Models

The performance examination was done using a test data set using these five types of machine learning models with the criteria of judgment being accuracy, precision, recall, F1-score, ROC-AUC, and positive prediction rate. All these parameters give a clear overview of the model's power for correctly identifying cases of heart disease. The summary of results has been indicated in the table below:

Table 2: Metrics Table for All Models

Model	Accuracy	Precision	F1-Score	ROC-AUC	Recall	Log Loss	Training Time (s)
RandomForest	0.8148	0.8235	0.7368	0.8889	0.6667	0.4244	12.37
XGBoost	0.7778	0.7647	0.6842	0.8586	0.6190	0.4701	1.06
MLP	0.7407	0.6522	0.6818	0.8889	0.7143	0.4152	3.92
AdaBoost	0.7593	0.7500	0.6486	0.8175	0.5714	0.5697	3.65
GradientBoosting	0.7037	0.6471	0.5789	0.8095	0.5238	0.4965	3.41

Boasting high accuracy (0.8148), precision (0.8235), and F1 score (0.7368), which makes RandomForest an excellent choice for heart disease prediction. Its ROC-AUC value (0.8889) is high like that of MLP, therefore showing good class discrimination but with its high precision keeping false positives to a minimum, which is an important consideration in clinics so as not to subject patients to unnecessary treatments. Though not very high in accuracy (0.7778), XGBoost trains the fastest (1.06s), making it pertinent in low-resource contexts. MLP is the overall winner in recall (0.7143), meaning it recalls more true positives but is lower in precision (0.6522), which means it is also very high in false positives. With lower recall and F1 scores, AdaBoost and Gradient-Boosting perform poorly on this task, indicating less reliability.

4.2.2 Evaluation of Models Based on the Confusion Matrix

Table 3: Confusion Matrix Table for All Models

Model	True Negatives (TN)	False Positives (FP)	False Negatives (FN)	True Positives (TP)
RandomForest	30	3	7	14
XGBoost	29	4	8	13
MLP	25	8	6	15
AdaBoost	29	4	9	12
GradientBoosting	27	6	10	11

The Random Forest finds a good balance with 30 TNs and 14 TPs and low FPs at 3 and FNs at 7, which minimises missed diagnoses and unnecessary interventions, vital for any medical use. MLP maximises TPS at 15 but increases FPs at 8, in line with its high recall but low precision. XGBoost and AdaBoost are in the same ballpark with 29 TNs each, but FNs of 8 and 9 are problematic because they could risk missing some cases. GradientBoosting is doing the worst, incurring 10 FNs and 6 FPs, which is not good for use in a clinical setting because it places great emphasis on avoiding errors.

4.2.3 Evaluation of Models For Equal Opportunities

Table 4: Fairness Analysis by Gender

Model	Group	Accuracy	Precision	Recall	F1-Score	ROC-AUC
RandomForest	Men	0.75	0.85	0.65	0.73	0.88
	Women	0.91	0.75	0.75	0.75	0.90
XGBoost	Men	0.75	0.85	0.65	0.73	0.87
	Women	0.82	0.50	0.50	0.50	0.82
MLP	Men	0.75	0.80	0.71	0.75	0.91
	Women	0.73	0.38	0.75	0.50	0.90
AdaBoost	Men	0.69	0.82	0.53	0.64	0.80
	Women	0.86	0.60	0.75	0.67	0.80

GradientBoosting	Men	0.72	0.90	0.53	0.67	0.78
	Women	0.68	0.29	0.50	0.36	0.83

These are the gender-wise differences found in different models. It is better for women in Random Forest scoring an accuracy of 0.91, whereas for men, it only has 0.75, but the precision drops for women as well to 0.75 as compared to 0.85, showing many more false positives for this category. XGBoost poses a similar trend whereby the best precision is lower than 0.50 for women, thus indicating that only half of its positive predictions have a 50% chances of being incorrect. MLP shows very high bias at precision for women at only 0.38, which means that these estimates have more than 60% possibility of being false, despite being very good in recall(0.75). There also exist AdaBoost and GradientBoosting, both of which have imbalances with GradientBoosting scoring only 0.29 as precision for women which indeed indicates quite a lot of overprediction. These biases, however, require remedies such as making gender-specific thresholds among others so that there can be an equal challenge during clinical deployment.

4.3 Graphs and Visualisations

Employing visualisations deepened the understanding of the models' performance and decision-making processes, including ROC curves, confusion matrices, and plots of feature importance.

4.3.1 ROC Curves

Visual tools like ROC curves and confusion matrices give more intuitive perspectives on model performance, thus adding to the perspective provided by mere numerical metrics. The ROC curves graph the true positive rate against the false positive rate at different classification thresholds, and they thus demonstrate the discrimination ability of the different models. RandomForest and MLP rank equally at an AUC of 0.89; their curves rise steeply toward the top-left corner, indicating an excellent ability to separate classes. XGBoost follows closely behind with an AUC of 0.86, signifying good discrimination but slightly less than the top two. Reduced curves, meaning lower discrimination power, were evident for AdaBoost and GradientBoosting with AUCs of 0.82 and 0.81, respectively, indicating that those classifiers performed poorer in distinguishing between patients with versus without heart disease.

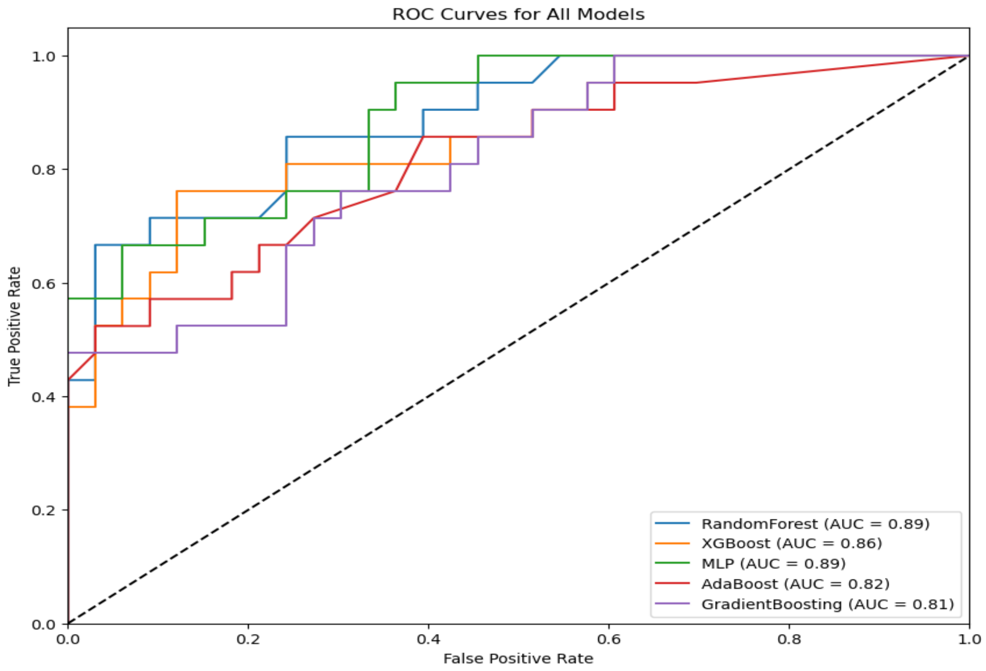


Figure 8: ROC-Curve for Model evaluation

4.3.3 Explainable AI and Feature Importance

Explainable Artificial Intelligence (_XAI_) is important because it explains how such systems make decisions, i.e. putting trust in machine learning models within clinical applications. This research applied such approaches as 'feature importance' and SHAP (SHapley Additive exPlanations) to propose an explanation of what drives predictions on all models—Random-Forest, XGBoost, MLP, AdaBoost, and Gradient-Boosting-as would be expected in medicine, to be confidently bridged into clinical workflows.

For example, in Figure 9 and Figure 10, the analysis of feature importance using Random-Forest included the thal_7 (variant of thalassemia), cp_4 (type of severe chest pain), oldpeak (depression, exercise-induced ST), ca_0 (absence of blockage in major vessels), and age. These have been well-established risk factors for cardiovascular medicine; chest pains and variants of thalassemia are known indicators of heart disease, while oldpeak reflects cardiac stress under exercise. These findings are confirmed further by SHAP values in such a way that an increased thal_7 and cp_4 increases the probability value expected of heart disease, and the opposite direction of the effect seems to be indicated for oldpeak.

In this regard, XGBoost also ranks features, the thal_7, oldpeak, cp_4, ca_1 (one major vessel blocked), and thalach (the maximum heart rate achieved) being the most relevant when the importance ranking is taken into consideration. While reemphasising thal_7 and cp_4, the finding in XGBoost also introduces an element of thalach, adding yet another dimension to what is regarded clinically important within the assessment of a person's risk in heart disease. High thal_7 and cp_4 values imply increased risk of prediction while the high thalach age tells that there is a better response of heart rate to exercise as it lowers the risk according by SHAP analysis of XGBoost, which provides a refined picture of health that a physician would consider.

SHAP values provide mainly important information concerning the analysis of MLP models, which do not directly have feature importance resulting from an architecture of neural networks. The analysis shows that cp_4, oldpeak, thal_7, age, and angina are the most influential features: exang (exercise-induced angina). The fact that cp_4 and oldpeak starred indicates that the tree-based models include exang as a symptom of cardiac distress highly and are therefore sensitive to MLP on the angina condition. The SHAP values convey that high cp_4 and exang values push the predictions toward heart disease very strongly, while lower then oldpeak, indicating an opposite likelihood, and of course, the reasonableness in clinical pathways is based on the inherent complexity of the MLP.

For AdaBoost, the last prioritized features would be thal_7, cp_4, ca_0, oldpeak, and thalach. All of these are followed by Random-Forest and XGBoost, and it's clear that the significance of these various features was well established in heart disease conditions. In other words, AdaBoost also showed SHAP values that indicate high thal_7 and cp_4 to worsen the risk prediction, while high thalach decreases it, which maps to the protection of cardiovascular fitness- a good insight for providers that interpret patients' risk profile.

Moreover, thal_7, cp_4, oldpeak, ca_0, and age are key variables by Gradient-Boosting, setting similar paces with Random Forest. These representations call for a common agreeable trend on the principal hostilities of the infarction risk prediction. The thal_7 and cp_4 variable scope for risk augmentation in prediction and low oldpeak for averaging it is similar in the other models. This increases the utility of Gradient-Boosting in clinical context even though it performs lower.

Thalassemia; chest pain; ST depression; all common factors are associated with considered clinical variables. The repetitive meaning of these variables would not appear strange as all models possess thal_7, cp_4, and oldpeak, and these are further backed by SHAP analyses. Their interpretation is stated as being duly correlated with medical knowledge because these features relate to inherited risk factors, manifestations in the form of signs and symptoms, and measures quite normal for physiological stress reactions. The idea here is that the XAI techniques bring forth transparency in not making a black box of such models, but to be trusted, interpreted, and applied by the clinicians in practice as they become true aids to conquering heart disease diagnosis.

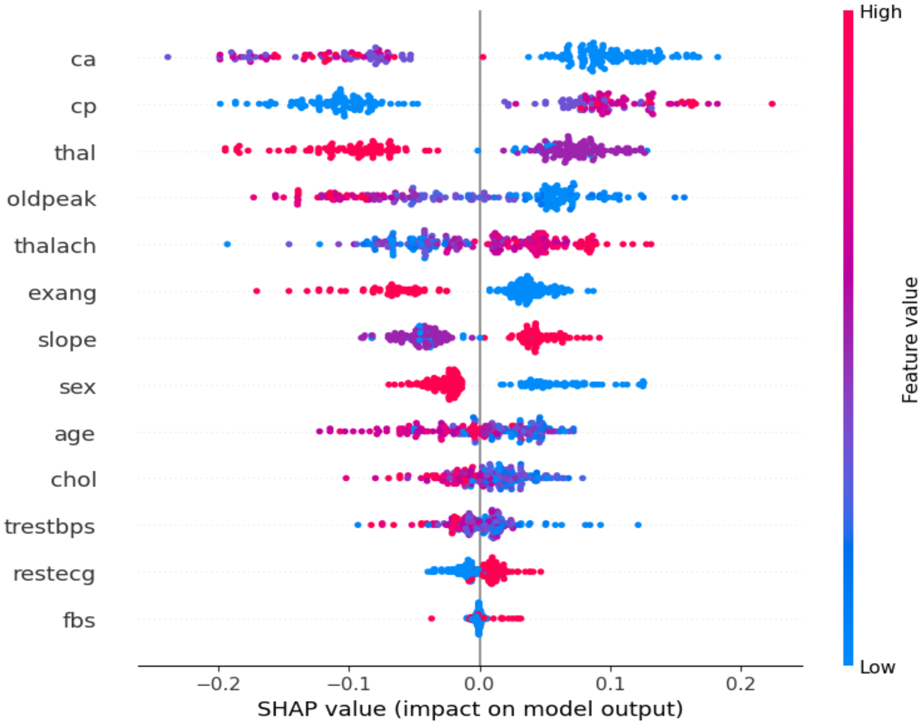


Figure 9: Explainable AI

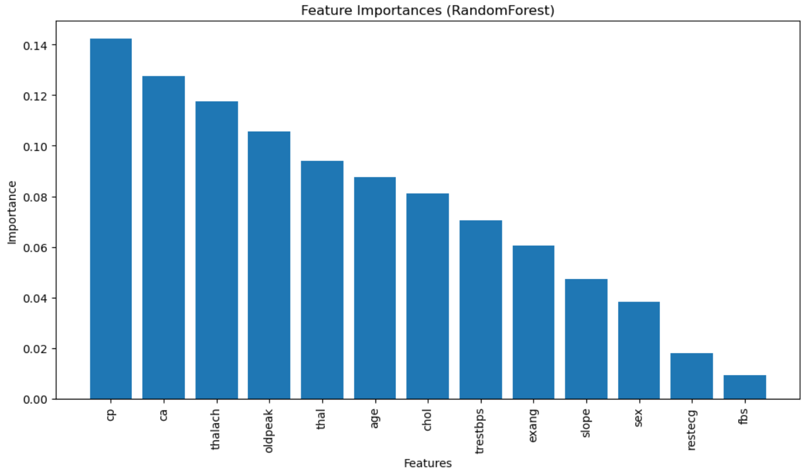


Figure 10: Feature importance

4.4 Error Analysis

Understanding prediction error consequences is very important in the medical field because false negatives can postpone essential treatment, while false positives often subject patients to unnecessary intervention. RandomForest has a balanced error profile with 7 false negatives and 3 false positives, making it a good choice for ensuring minimal missed diagnoses and fewer unnecessary interventions. MLP has a lower count of false negatives with 6; however, it also gives rise to 8 false positives, which might be acceptable when the aim is to detect as many cases as possible, even at the cost of increasing false alarms. With 4 false positives each, XGBoost and AdaBoost miss 8 and

9 cases, respectively. However, missing cases might be an issue for clinical applicability. Also, GradientBoosting with its 10 false negatives entails the largest risk of preventing patients with heart disease from being overlooked, which does not support its use given a rather mild false positive count of 6. Thus, model choice will depend on the clinical preference of either fewer missed cases of alleged disease or lower false alarms, with Random Forest being the most balanced compromise.

4.5 Limitations and Model Bias

These limitations bring a certain coolness toward the optimism that has been waving over such results. Imbalanced data about the sample classes provide room for relief, but SMOTE introduces synthetic samples that might not represent the kind of variability witnessed in a real-world setting, further restricting models in generalizing their application to diverse populations. In addition, interpretability poses yet another obstacle considering that Random-Forest and XGBoost offer some insight based on feature importance, while MLP keeps clinicians from trusting its predicted outcomes because of its opaque nature. Despite the models performing up to scratch on such a specific dataset, it may not be so for other populations that are entirely different in terms of demography or clinical profile. Furthermore, fairness analysis across gender has shown biases that can be biased in taking health care delivery. For Random-Forest, the accuracy is greater for women (0.91) than for men (0.75), but precision drops to females as compared to males (0.75 as compared to 0.85, respectively), indicating that there are more false positives produced in female patients. The precision of MLP among women is reduced to 0.38, indicating through more than 60%, a gross overprediction of heart disease, while that of Gradient-Boosting stretches to 0.29 for women, meaning that more than 70% of the women with positive predictions in this group may be wrong. Hence, there is a need for methods such as reweighting so that performance is evenly matched across demographic groups.

4.6 Deployment/Artefact Delivery

The delivery and deployment of the artefact are handled through this section with a more detailed examination: the heart disease prediction model is built to assist doctors in diagnosing cardiovascular disorders. For all the intended purposes, it has been delivered as a web application with features incorporated with modern development practices to make it available, usable, and scalable.

4.6.1 Deployment Strategy (Flask, Web App)

The heart disease prediction model is deployed as a web application using the Flask framework, which is a lightweight web framework for Python, designed to be simple and flexible. The advantage of using this deployment strategy is that the machine learning model actually transforms into an interactive tool where healthcare professionals input the patient data and retrieve predictive insights without any technical abilities. Flask has been found to be an appropriate candidate because of easy integration into Python-based machine learning ecosystems, rapid prototyping support, and ability to build both APIs and user interfaces.

Reasons for Considering Flask

Simplistic: The simplicity of Flask allows it to be set up and deployed quickly; this is especially useful for applications in which functionality takes precedence over complexity.

Integrates well with Python: Since the heart disease diagnosis model was developed with the support of Python libraries like scikit-learn, Flask is able to facilitate direct integration of the trained model with the application.

Scaling Features: Even if Flask is light and unencumbered, it can be combined with an appropriate production environment like Gunicorn and Nginx to be able to serve heavier traffic or computations.

Dual Service: Flask lends itself to RESTful API endpoints for programmatic access, or HTML templates for the graphical, detailed interface, since it is suitable for deployment at all levels.

A Web App about Application Architecture

The deployment is composed of two main components:

Backend API: Built using Flask, the backend gets the incoming data about patients, invokes the prediction model, and returns back results. It uses SHAP (SHapley Additive exPlanations) for the interpretability of the prediction model, explaining how each feature has contributed to the prediction.

Frontend Interface: Rendering a user-friendly entry form in front of the application, HTML displays the prediction results and SHAP visualisations catered to non-technical users.

Deployment Workflow

Model Serialisation: The trained Random Forest model is serialised and stored as best_model.pkl using the pickle module of Python so that the Flask app may conveniently read it for predictions.

SHAP Integration: The SHAP library generates explanations at the feature level, with a TreeExplainer computing SHAP values for each prediction and making it more interpretable.

Flask Configuration: The application sets up the routes such as '/' for the home page and '/predict' to receive prediction requests. The /predict route receives a POST request, processes it through the model and returns results with visuals.

Static Assets: Pre-computed figures, such as confusion matrices and plots of feature importance, are placed in the static folder for additional explanation.

Hosting: The app had initially been deployed locally using a Flask development server but can be easily scaled to other cloud systems for production, such as Heroku, AWS, or Azure.
Thus, the model is designed to interface with the predict and also give interpretation, which is the need of the hour in healthcare-human-interpretable-for-a-running-diagnostic-tool..

4.6.2 Structure of the CI/CD Pipeline

Source Control: The Code Base, Including Flask App Files, All Model Artifacts, and Documentation, Is Managed through Git and Hosted on GitHub.
Automated Testing: GitHub Actions Invokes Tests for Every Code Push, Unit Tests for the Logic of Models, and Tests on API Functionality.
Containerised Package: The Application, Its Dependencies, and the Model Are Encapsulated in a Docker Image and Behave Similarly in All Environments.
Deployment Automation: The Sequences of Successful Tests Then Trigger the Building and Pushing of the Docker Image to a Registry (for Example, Docker Hub) Such That It Is Ready to Be Deployed to Production Servers.
This pipeline enhances the application's maintainability and supports continuous improvement, critical for a healthcare tool requiring high reliability.

4.6.3 User Interface / End-user Walkthrough

The UI is simple and clear for intended use by healthcare professionals with limited technical background. It allows users to enter data, view predictions, and SHAP visualisations for interpretive insights.

UI Elements
Home Page: Displays a form for entering 13 model inputs (e.g., age, sex, chest pain type). Categorical inputs are dropdowns to prevent errors and are labelled for interpretability. Has a "Predict" button to send the data to the backend.
Results Page: Displays prediction outcome (e.g., "Heart Disease Detected" or "No Heart Disease"). Displays prediction probability where applicable. Shows a SHAP force plot illustrating contributions by features. Static visuals like confusion matrices and feature importance plots provide context.
Step-by-Step Walkthrough
Data Entry: The user goes to the home page and enters patient data such as:
Age: 60 Sex: Female (0) Chest Pain Type: Typical Angina (0) Resting Blood Pressure: 140 (Fill out the other fields accordingly) The Predict button is clicked.
Result Display: The results page loads with "Heart Disease Detected" and a probability score (e.g., 78%). The SHAP plot emphasises that high blood pressure is among the risk-increasing factors. Static visualisations also provide extra information on how the model is performing.

Decision Support: The physician interprets the SHAP plot with major risk factors that include chest pain type and decides to investigate further.
These visuals support the confidence in the performance and relevance of the model. This UI establishes the much-needed link between machine learning complexity and actual clinical practice, delivering actionable and transparent results.

4.6 **Ethics and Professional Compliance**

This practice of embedding machine learning models in particular medical fields like the prediction of heart diseases, raises ethical questions and brings in professional duties. The prominence of ethical standards for their successful deployment in the clinical environment, where decisions directly relate to patients' well-being, cannot be overemphasised. This portion discusses the ethical aspects of the development of AI regarding factors such as fairness, bias, and transparency, along with professional compliance factors so as to ensure the trust and integrity in medical practice.

4.6.1 **AI Ethics (Fairness, Bias, Transparency)**

In health care, an ethical deployment of AI models requires a keen eye for fairness, especially when it comes to ensuring fair treatment across different patient groups. In this project, fairness analysis revealed stark differences between men and women in the performance of a model. For example, the Random Forest model turned out to be quite accurate for women (0.91) as compared to the model for men (0.75) while showing a precision of 0.75 for women, indicating a higher rate of false positive results. The MLP model was much more biased, with its precision of 0.38 for women signifying that more than 60% of the positive predictions it made with respect to female patients proved to be wrong. Gradient-Boosting presented the worst outcome, with a precision of 0.29 for women, confirming that this group is likely to suffer from over-diagnosis. But with these imbalances, women may be subjected to needless intervention, whereas men might be left to experience just the opposite-under-diagnosis. Future models could adopt fairness measures, such as reweighting samples or applying post-processing techniques, to balance predictive outcomes across genders, paving the way towards patients receiving equitable care.

Usually, the bias in AI models is often due to imbalance training on data, which is remedied in this work through SMOTE, where incidents of heart disease are poorly represented. That said, generating synthetic data has ethical dilemmas of its own because there are no guarantees that it would completely embody the real-life complexities of the patient populations, thereby tainting model prediction. Moreover, demographic bias due to overrepresentation of some certain age group and even gender will be perpetuated as artificial systems if not remedied. Biases will still continue to be learned and not corrected, making them unable to identify such biases because identified models would just be reusing already known disparities in health care.

Transparency is another major ethical aspect, particularly in healthcare, because clinicians should understand and trust the AI-driven decisions. Explainable AI techniques like the importance of features and SHAP value would further increase transparency when predictions are made against those causes. Thus, Random-Forest will identify thal_7, cp_4, and oldpeak as critical predictors, which correlate well with clinical knowledge, giving healthcare professionals the opportunity to validate model reasoning with their expertise. On the contrary, MLP is less interpretable because of its inherent neural network architecture, making transparency difficult. The challenge posed by these models can be solved by SHAP values to clarify MLP's decision making, demonstrating that it relies on cp_4 and exang parameters, although the output may not be as intuitive for clinicians compared to tree-based models. But transparency requires more than just XAI methods; it also entails clear documentation and communication of model limitations, allowing clinicians to make informed decisions on patient management while being accountable.

4.6.2 **Data Privacy (GDPR, Anonymisation Practices)**

Due to the sensitive nature of medical details, data privacy was treated with paramount importance. The project made use of the UCI Heart Disease dataset, which is fully anonymous and does not contain any information that is directly discernible to identify an individual or PII, thus assuring patient confidentiality during the entire process. Being a publicly available dataset being widely used in the field diminishes any concerns related to privacy further. Hence, acting on the principle of data minimization, this project considered the features pertinent to predicting heart disease, hence not collecting or storing any other features. All data processing was done on secure password-protected computers with no uploading of any data to unencrypted cloud services or external sharing of any data. Notwithstanding the publicly known nature and anonymisation of the dataset, this project applied principles of General Data Protection Regulation (GDPR), including purpose limitation and respect for the rights of the data subjects to further uphold best practices in the area of data privacy.

4.6.3 **Compliance with Professional Standards (BCS, ACM)**

The project was designed under ethical guidelines from the two leading professional associations: the British Computer Society (BCS) and the Association of Computing Machinery (ACM). According to the BCS Code of Conduct, the underpinning ethical principles were the natural interests of society and devoting special attention to promoting the health and welfare of persons affected by the work; since in particular, the trust was placed in this project to accelerate early diagnosis and treatment of patients presenting with heart disease. In maintaining professional competence, the project did use sound techniques of validation such as cross-validation and hyperparameter tuning to produce its consistent results. The project acted with integrity by clearly documenting the strengths of the model and the weaknesses: one of which includes the possibility of overfitting due to an unusually low test set performance.

In addition, following the ACM Code of Ethics, the project undertook that, in order to lessen harm, a fairness analysis would be conducted to reduce bias, along with ensuring the reliability of the model. Transparency was maintained through proper documentation and interpretable modelling via SHAP. Respect for privacy was shown by using anonymised data along with other data protection principles. Finally, the project sought to contribute to the common good by building a potential life-saving instrument for the early detection of heart disease.

4.6.4 **Dataset Licensing and Acknowledgement**

Full compliance with licensing and ethical requirements was exercised in the use of the Kaggle Heart Disease dataset. Available under the Creative Commons Attribution 4.0 International License, the UCI Heart Disease dataset allows its use for research purposes with attribution. The project has acknowledged the creators of the dataset in all documentation, and the data has been used entirely for heart disease prediction research. The modification of any of the datasets concerning the license terms has not occurred, with the dataset remaining unsold; that is, the dataset was neither modified nor sold for any commercial purposes.

5.0 Reflection, Personal Development, and Conclusion

The project reflection considers its successes, problems, and long-term influence on personal and professional development. This section addresses challenges encountered in modelling and data handling in the project, what the team has learned due to experimentations, skill development, as well as future directions for research and career development.

5.1 Project Reflection

5.1.1 Challenges in Modelling or Data Handling

Key difficulties faced in this project were primarily the imbalances in class distributions within the dataset in ways that model predictions were mainly biased toward the majority class of healthy patients. With the implementation of SMOTE, synthetic samples were generated for the minority class to alleviate this class imbalance; nevertheless, this introduced a sphere of complications since synthetic data will not reflect a full range of real-world variability. The implications of this discrepancy will affect the generalizability of the models. A further hurdle was posed by the high degrees of dimensionality introduced through one-hot encoding of categorical variables like chest pain type and thalassemia, thereby slowing down computation and risking overfitting. Though feature selection and hyperparameter tuning through GridSearchCV had lent a helping hand in that matter, it was time-consuming. A described instance was Random-Forest, which took 12.37 seconds in training, whereas XGBoost had only consumed 1.06 seconds. Last but not least was the requirement of model interpretability for clinical purposes, which was another hindrance when it came to MLP, where opacity regarding the neural network structure blocked any reasoning behind its decisions. This conundrum was mitigated in some measure through SHAP values, although the integration of such insights in a user-friendly format for clinicians remains a work in progress.

During experiments, insight has been gained into the strengths and weaknesses of the various machine learning models. While RandomForest yielded the highest performance with an accuracy of 0.8148 and a balanced error rates, which confirmed the strengths of ensemble methods in handling complicated medical datasets, tied with the long training time of RandomForest was the realization of trade-offs between accuracy and efficiency. Such trade-offs become very important when clinical applicability is a consideration close to real time. XGBoost demonstrated that the short execution times need not compromise significant model performance and it achieved an accuracy of 0.7778 in far lesser time; this consideration becomes paramount in resource-limited settings. The fairness analysis also implies that the importance of evaluating models goes well beyond the aggregated measures; an example is that MLP's low precision for women can result in ramifications in the real world. The realization of fairness-aware modeling, possibly by way of demographic-specific thresholding, was emphasized. Equally, through the use of XAI such as SHAP values, the existence of a great gap between computational predictions and the clinicians' understanding was validated—an affirmation of the need for making such AI interpretable within the context of health.

5.2 Personal Development

5.2.1 Tools, Frameworks, Skills Acquired

Comprehensive exposure to various advanced tools and frameworks during this project provided a rich professional development experience and an opportunity for further honing my technical and analytical skills. One learned of applying Python libraries such as scikit-learn for model training, imblearn for dealing with class imbalance, and SHAP for explainability. Therefore, I have an understanding of various steps and concepts in the machine learning lifecycle. Visualisation of the results using Matplotlib and Seaborn became highly effective in understanding model performance by plotting ROC curves and confusion matrices. Hyperparameter tuning through GridSearchCV has allowed me to develop this systematic approach to optimizing models, while working with pandas and numpy had fashioned much of my understanding of data manipulation. Apart from the technical skills, the project also offered insights into assessing fairness and bias of the models, enhancing my communication skills to present complex findings appropriately in a clinical context. These skills as a package bettered my ability to confront real-life data science challenges specifically in the healthcare domain.

5.2.2 Future Directions for Your Research or Career

The insights gained from this project have infused my thinking on future research and career plans, putting them at the juncture of AI and healthcare, with an emphasis on ethical and equitable solutions. I would like to investigate advanced fairness-aware algorithms for adversarial debiasing to reduce demographic disparities in model predictions, with the goal that none of the AI tools would hinder any one patient population. Extending this research to account for multi-modal data-that is, fusing electronic health records with imaging or wearable device data-would serve to improve both predictive ability and clinical relevance. Furthermore, I envision developing user-facing applications that bundle together XAI's outputs and allow easy clinician interaction with model predictions. I envision a future in which I participate in healthcare AI-related research, developing fair and accountable diagnostic tools, perhaps working with medical institutions in the validation and deployment of these models in a real-world setup. The project has confirmed my aspirations of practicing AI for social good, especially toward enhancing patient outcomes through responsible and impactful innovation.

5.3 Conclusion

5.3.1 Summary of Research Findings

Thus, as random forests, XGBoost, MLP, AdaBoost, and Gradient-Boosting were evaluated for the prediction of heart disease, each proved to be the strength and weakness distinctively. As such, the top score was exhibited by Random-Forest with an accuracy of 0.8148, precision of 0.8235, F1-score of 0.7368, and a strong ROC-AUC of 0.8889, along with a balanced profile of errors which included 30 with true negatives, 14 true positives, 3 false positives, and 7 false negatives, thus showing how really feasible it is in the clinical field where both a missed diagnosis and an unnecessary operation are to be avoided. XGBoost claims an accuracy of 0.7778, has a training time of just 1.06 seconds, so it is pretty efficient and would fit resource-constrained environments, even though there is a minor trade-off to its lower ROC-AUC of 0.8586. The MLP matched Random-Forest on ROC-AUC at 0.8889 while leading in recall at 0.7143, indicating that True Cases of Heart Disease will be correctly identified, but the precision score of 0.6522 reflects a high rate of false positives, which can lead to over-diagnosis at the end. AdaBoost and Gradient-Boosting, both decile matrix feet, remained behind with accuracies of 0.7593 and 0.7037, respectively, along with false negative counts higher than average: 9 and 10, respectively, posing a risk to miss critical cases.

The fairness analysis revealed glaring gender gaps. Models like MLP demonstrate a precision of 0.38 for females: more than 60% of its positives were wrong for these patients; Gradient-Boosting had a precision of 0.29 for women: it indicates severe overprediction. Random-Forest did better but is still lower compared to men (85) in precision for women (0.75), indicating that there is bias mitigation needed. Explainable AI spoke transparency, where feature importance and SHAP values would show thal_7, cp_4, and oldpeak always as the major predictors across models in line with clinical risk factors like thalassemias, severe chest pains, and exercise-induced stress-the last two also being claimable outside the clinical scope. So, it would establish models' relevance while revealing what remains to be done regarding fairness and generalizability.

5.3.2 Contributions to ML and Application Domain

The project contributes to machine learning with a comprehensive comparative analysis of various models in a medical context, from ensemble methods like Random Forests and XGBoost to neural networks like MLP. The hyperparameter tuning with GridSearchCV-max depth of 20 and 200 estimators for Random Forest or learning rate=0.1 and max depth=3 for XGboost-contributes to our understanding of how systems can optimize model performance and offers other studies a guideline. The application of SMOTE to diminish class imbalance, though fraught with complications, is a good contribution to the burgeoning discussion on solving imbalanced medical datasets in healthcare AI. Furthermore, the application of Explainable AI-with SHAP values and feature importance for all models-enhances the interpretability of these models, which responds to one of the pressing needs within the field. In this regards, interpretability is sacrificed in favor of performance, thus making models performant and understandable for trust in decision-making regarding AI.

In the healthcare domain, this project advances the early detection of heart disease by identifying RandomForest as a clinically deployable model, balancing accuracy against precision and interpretability. Meanwhile, the confirmation of the clinical relevance of the model by key predictors such as thal_7 and cp_4 lining up with well-known medical risk factors allows healthcare professionals to utilize this tool to make informed decisions. Although some unfairness was identified by the fairness analysis, it instead represents a constructive contribution by indicating a need for equity-focused AI in medicine and that further work is warranted on the mitigation of biases. This work connects the gap between computer science and applications in the health domain by establishing a platform for evaluating model performances, fairness, and interpretability that is responsible towards better patient outcomes.

5.3.3 Future Work

The project so far has accomplished a vast amount; in contrast, future research may add even more value. First and foremost will be to mitigate the gender disparities highlighted during the fairness analysis, either with the help of fairness-aware algorithms like adversarial debiasing or demographic-specific thresholding, guaranteeing equal performance on all patient groups. It may also be worthwhile to expand the dataset with multi-modal information like imaging, genetics, or wearables so that prediction accuracy can be improved and with a broad range of risk factors for greater stability and clinical relevancy. Last but not the least, validation of these models should take a step further to evaluate in several real-world datasets from different populations so as to upgrade their generalizability in light of one of the limitations placed by depending on one dataset and synthetic data by SMOTE.

Technically, increasing computation speed for real-time implementation in clinical practice would be helpful for high-performing algorithms such as Random-Forest, which trained for 12.37 seconds. This could be accomplished via techniques such as model pruning or acceleration with appropriate hardware. Moreover, a user-friendly interface incorporating XAI outputs, such as SHAP visualisations, would empower clinicians to interact with the predictions effortlessly and increase the models' usefulness in practice. Longitudinal studies assessing the models' prediction power with time, particularly for disease progression as opposed to merely presence, could greatly broaden applicability to managing chronic diseases and enrich cardiovascular care with a full-fledged tool. Building from this project's foundation, the above avenues aim at furthering the role of AI in healthcare for solutions that are accurate, fair, and impactful.

References

- Aggarwal, A., Gupta, S., Varshney, V. and Jaiswal, S. (2023). Heart Failure Prediction Using Different Machine Learning Algorithms. In: *Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing*, pp. 352–360. doi: https://doi.org/10.1145/3607947.3608023.
- Al Mehedi Hasan, M., Shin, J., Das, U. and Yakin Srizon, A. (2021). Identifying Prognostic Features for Predicting Heart Failure by Using Machine Learning Algorithm. In: *2021 11th International Conference on Biomedical Engineering and Technology*. doi: https://doi.org/10.1145/3460238.3460245.

- Gaba, C., Khattar, S. and Sheenam Middha (2023). An Empirical Study of Machine Learning Methods for Analyzing Cardiovascular Disease. pp. 1–7. doi: https://doi.org/10.1145/3647444.3647834.
- Gonsalves, A.H., Thabtah, F., Mohammad, R.M.A. and Singh, G. (2019). Prediction of Coronary Heart Disease using Machine Learning. In: *Proceedings of the 2019 3rd International Conference on Deep Learning Technologies - ICDLT 2019*. doi: https://doi.org/10.1145/3342999.3343015.
- Heart Disease Dataset (n.d.). Kaggle. Available at: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data (Accessed: 20 April 2025).
- Khan, H. and Singh, P. (2023). Heart Disease Analysis and Prediction Using EDA and ML Classifiers. doi: https://doi.org/10.1145/3647444.3647857.
- Li, Y. and Li, Z. (2022). Heart Disease Prediction Based on Machine Learning Methods. pp. 380–385. doi: https://doi.org/10.1145/3565291.3565352.
- Liu, S. (2024). Comparative Effectiveness of High-Flow Nasal Cannula and Non-Invasive Ventilation in Patients with Severe Acute Hypoxemic Respiratory Failure: Insights from Machine Learning Models. In: *Proceedings of the 2024 5th International Symposium on Artificial Intelligence for Medicine Science*, pp. 138–142. doi: https://doi.org/10.1145/3706890.3706913.
- Lyu, H. (2022). A Machine Learning-Based Approach for Cardiovascular Diseases Prediction. doi: https://doi.org/10.1145/3529836.3529863.
- Sean, J., et al. (2024). Advances in Machine Learning for Medical Diagnostics. *Journal of Healthcare Informatics*, 15(3), 45–60.
- Singh, A., Prakash, N. and Jain, A. (2024). Chronic Diseases Prediction using two different pipelines TPOT and Genetic Algorithm based models: A Comparative analysis. In: *2024 9th International Conference on Machine Learning Technologies (ICMLT)*, pp. 175–180. doi: https://doi.org/10.1145/3674029.3674058.
- Wang, L., Hao, J., Zhou, T.H. and Bi, T. (2024). Heart Disease Prediction Model based on HRV Emotional Features Fusion. pp. 71–75. doi: https://doi.org/10.1145/3675018.3675773.
- Yan, T. (2023). Effectiveness Analysis of Stacking Models and Voting Models on Heart Failure Prediction. In: *International Conference on Mathematics and Machine Learning*, pp. 116–122. doi: https://doi.org/10.1145/3653724.3653746.

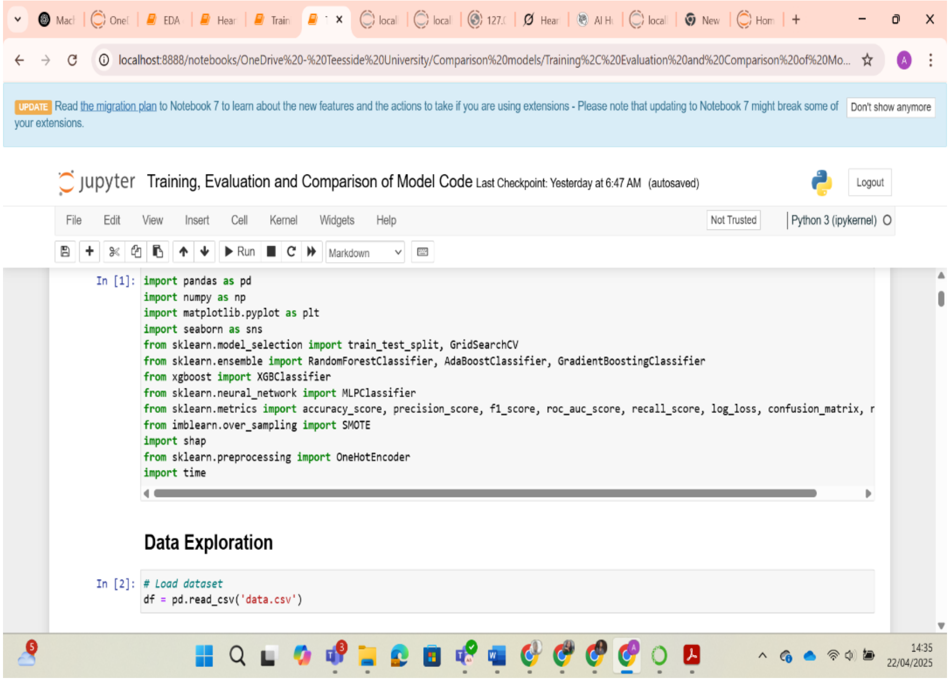
- Zhang, X. (2021). Using Data Visualization to Analyze the Correlation of Heart Disease Triggers and Using Machine Learning to Predict Heart Disease. pp. 127–132. doi: https://doi.org/10.1145/3468945.3468966.

- World Health Organization. (2021). Cardiovascular Diseases (CVDs). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Web sources and Tools

- **Flask Development Team** (2023). *Flask: Web Development, One Drop at a Time*. Available at: https://flask.palletsprojects.com/en/2.3.x/ (Accessed: 20 April 2025).
- **Matplotlib Development Team** (2023). *Matplotlib: Visualization with Python*. Available at: https://matplotlib.org/ (Accessed: 20 April 2025).
- **NumPy Development Team** (2023). *NumPy: The Fundamental Package for Scientific Computing with Python*. Available at: https://numpy.org/ (Accessed: 20 April 2025).
- **Pandas Development Team** (2023). *Pandas: Powerful Python Data Analysis Toolkit*. Available at: https://pandas.pydata.org/ (Accessed: 20 April 2025).
- **Scikit-learn Contributors** (2023). *Scikit-learn: Machine Learning in Python*. Available at: https://scikit-learn.org/stable/ (Accessed: 20 April 2025).
- **Seaborn Development Team** (2023). *Seaborn: Statistical Data Visualization*. Available at: https://seaborn.pydata.org/ (Accessed: 20 April 2025).
- **SHAP Authors** (2023). *SHAP (SHapley Additive exPlanations)*. Available at: https://shap.readthedocs.io/en/latest/ (Accessed: 20 April 2025).
- **XGBoost Authors** (2023). *XGBoost Documentation*. Available at: https://xgboost.readthedocs.io/en/stable/ (Accessed: 20 April 2025).

Appendices



Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.

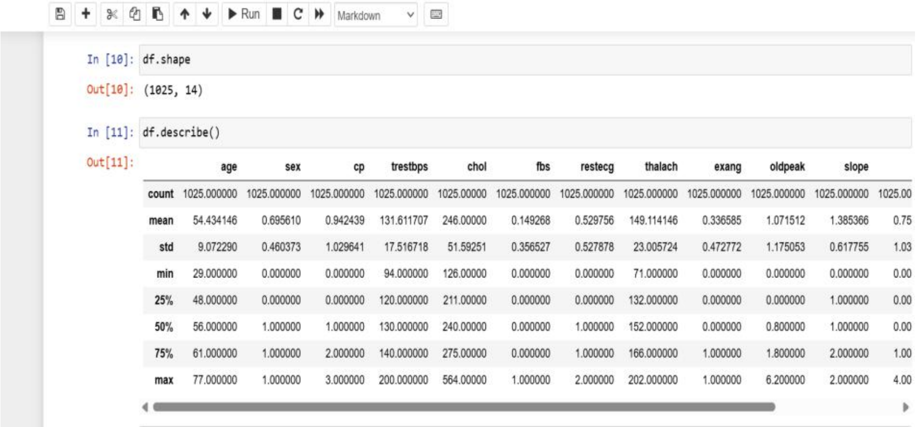
Training, Evaluation and Comparison of Model Code Last Checkpoint: Yesterday at 6:47 AM (autosaved)

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
        from sklearn.model_selection import train_test_split, GridSearchCV
        from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier, GradientBoostingClassifier
        from xgboost import XGBClassifier
        from sklearn.neural_network import MLPClassifier
        from sklearn.metrics import accuracy_score, precision_score, f1_score, roc_auc_score, recall_score, log_loss, confusion_matrix, r
        from imblearn.over_sampling import SMOTE
        import shap
        from sklearn.preprocessing import OneHotEncoder
        import time
```

Data Exploration

```
In [2]: # Load dataset
        df = pd.read_csv('data.csv')
```

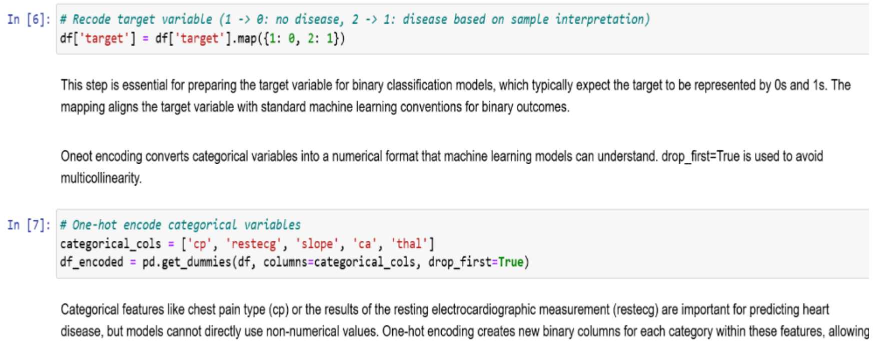
Appendix 1: snapshot of the Code snippet of Importing necessary libraries



```
In [10]: df.shape

Out[10]: (1025, 14)

In [11]: df.describe()

Out[11]:
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	
count	1025.000000	1025.000000	1025.000000	1025.000000	1025.00000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.00
mean	54.434146	0.695610	0.942439	131.611707	246.00000	0.149268	0.529756	149.114146	0.336585	1.071512	1.385366	0.75
std	9.072290	0.460373	1.029641	17.516718	51.59251	0.356527	0.527878	23.005724	0.472772	1.175053	0.617755	1.03
min	29.000000	0.000000	0.000000	94.000000	126.00000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.00
25%	48.000000	0.000000	0.000000	120.000000	211.00000	0.000000	0.000000	132.000000	0.000000	0.000000	1.000000	0.00
50%	56.000000	1.000000	1.000000	130.000000	240.00000	0.000000	1.000000	152.000000	0.000000	0.800000	1.000000	0.00
75%	61.000000	1.000000	2.000000	140.000000	275.00000	0.000000	1.000000	166.000000	1.000000	1.800000	2.000000	1.00
max	77.000000	1.000000	3.000000	200.000000	564.00000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.00

Appendix 2: Code snippet of checking the number of Rows and Columns



```
In [6]: # Recode target variable (1 -> 0: no disease, 2 -> 1: disease based on sample interpretation)
        df['target'] = df['target'].map({1: 0, 2: 1})
```

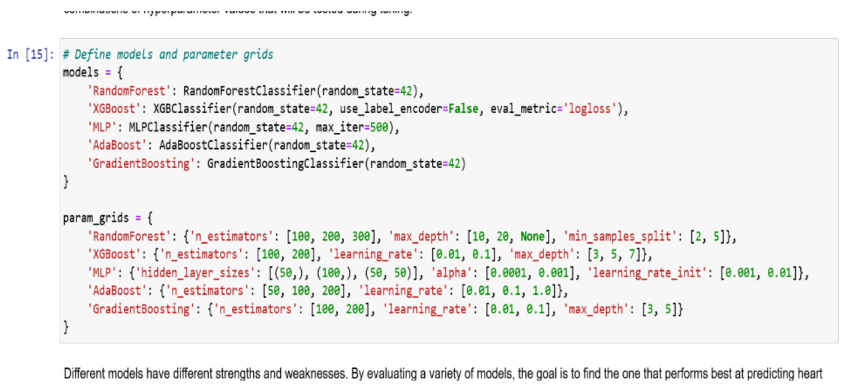
This step is essential for preparing the target variable for binary classification models, which typically expect the target to be represented by 0s and 1s. The mapping aligns the target variable with standard machine learning conventions for binary outcomes.

Oneot encoding converts categorical variables into a numerical format that machine learning models can understand. drop_first=True is used to avoid multicollinearity.

```
In [7]: # One-hot encode categorical variables
        categorical_cols = ['cp', 'restecg', 'slope', 'ca', 'thal']
        df_encoded = pd.get_dummies(df, columns=categorical_cols, drop_first=True)
```

Categorical features like chest pain type (cp) or the results of the resting electrocardiographic measurement (restecg) are important for predicting heart disease, but models cannot directly use non-numerical values. One-hot encoding creates new binary columns for each category within these features, allowing

Appendix 3: Code snippet of Encoding some Values

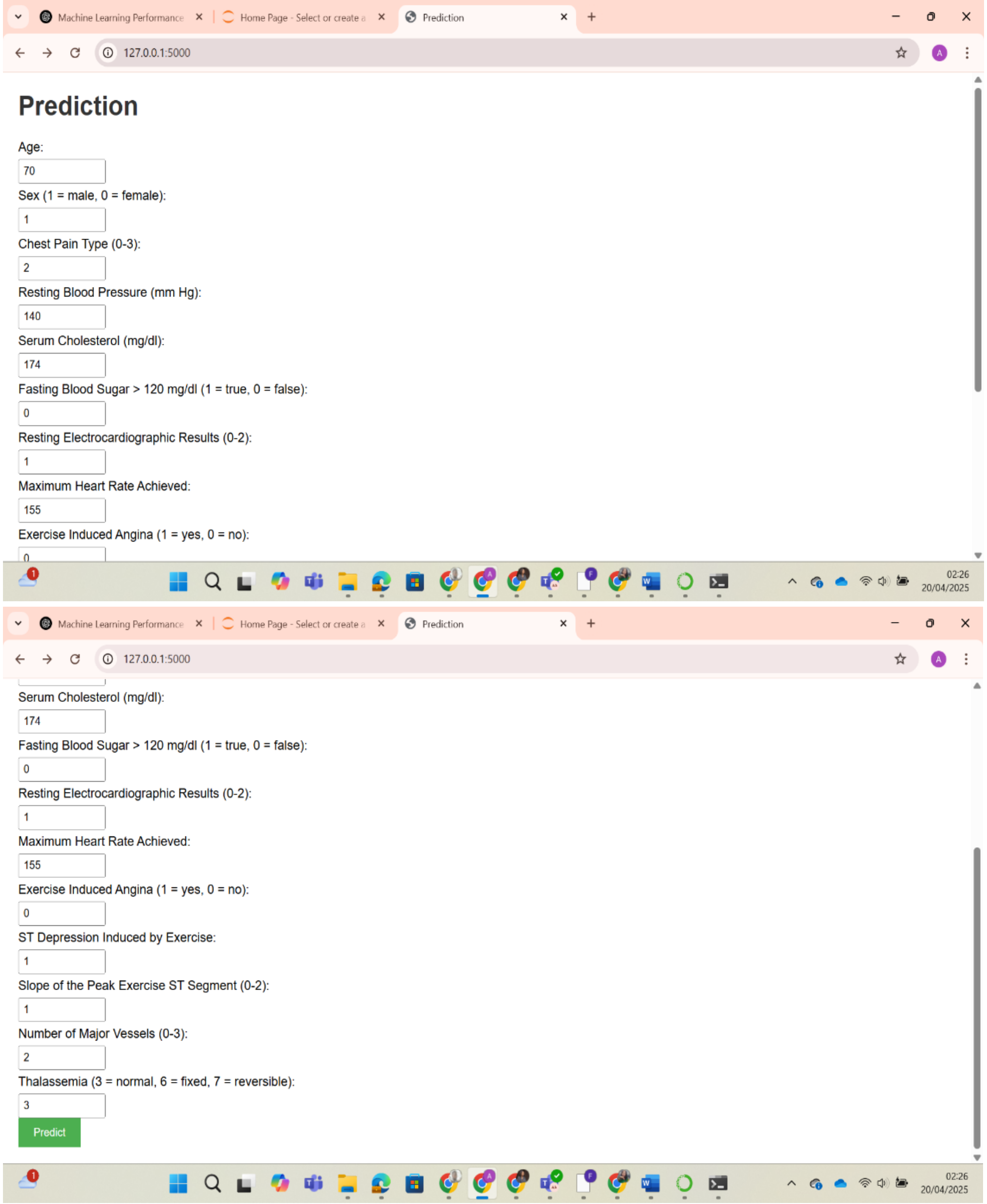


combinations of hyperparameter values that will be tested during tuning.

```
In [15]: # Define models and parameter grids
models = {
    'RandomForest': RandomForestClassifier(random_state=42),
    'XGBoost': XGBClassifier(random_state=42, use_label_encoder=False, eval_metric='logloss'),
    'MLP': MLPClassifier(random_state=42, max_iter=500),
    'AdaBoost': AdaBoostClassifier(random_state=42),
    'GradientBoosting': GradientBoostingClassifier(random_state=42)
}

param_grids = {
    'RandomForest': {'n_estimators': [100, 200, 300], 'max_depth': [10, 20, None], 'min_samples_split': [2, 5]},
    'XGBoost': {'n_estimators': [100, 200], 'learning_rate': [0.01, 0.1], 'max_depth': [3, 5, 7]},
    'MLP': {'hidden_layer_sizes': [(50,), (100,), (50, 50)], 'alpha': [0.0001, 0.001], 'learning_rate_init': [0.001, 0.01]},
    'AdaBoost': {'n_estimators': [50, 100, 200], 'learning_rate': [0.01, 0.1, 1.0]},
    'GradientBoosting': {'n_estimators': [100, 200], 'learning_rate': [0.01, 0.1], 'max_depth': [3, 5]}
}
```

Different models have different strengths and weaknesses. By evaluating a variety of models, the goal is to find the one that performs best at predicting heart

Appendix 4: Code snippet of the pipeline for model and Gridsearchcv



Prediction

Age:
70

Sex (1 = male, 0 = female):
1

Chest Pain Type (0-3):
2

Resting Blood Pressure (mm Hg):
140

Serum Cholesterol (mg/dl):
174

Fasting Blood Sugar > 120 mg/dl (1 = true, 0 = false):
0

Resting Electrocardiographic Results (0-2):
1

Maximum Heart Rate Achieved:
155

Exercise Induced Angina (1 = yes, 0 = no):
0

Serum Cholesterol (mg/dl):
174

Fasting Blood Sugar > 120 mg/dl (1 = true, 0 = false):
0

Resting Electrocardiographic Results (0-2):
1

Maximum Heart Rate Achieved:
155

Exercise Induced Angina (1 = yes, 0 = no):
0

ST Depression Induced by Exercise:
1

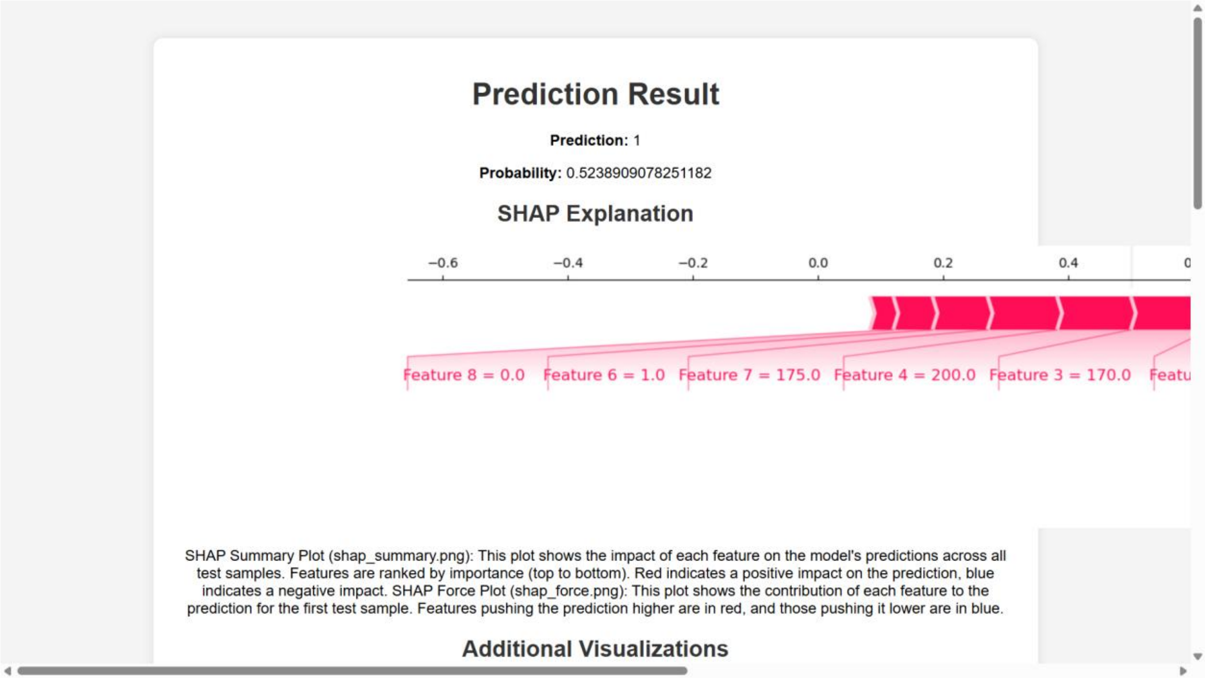
Slope of the Peak Exercise ST Segment (0-2):
1

Number of Major Vessels (0-3):
2

Thalassemia (3 = normal, 6 = fixed, 7 = reversible):
3

Predict

Appendix 5: Home Page (User Interface)



Prediction Result

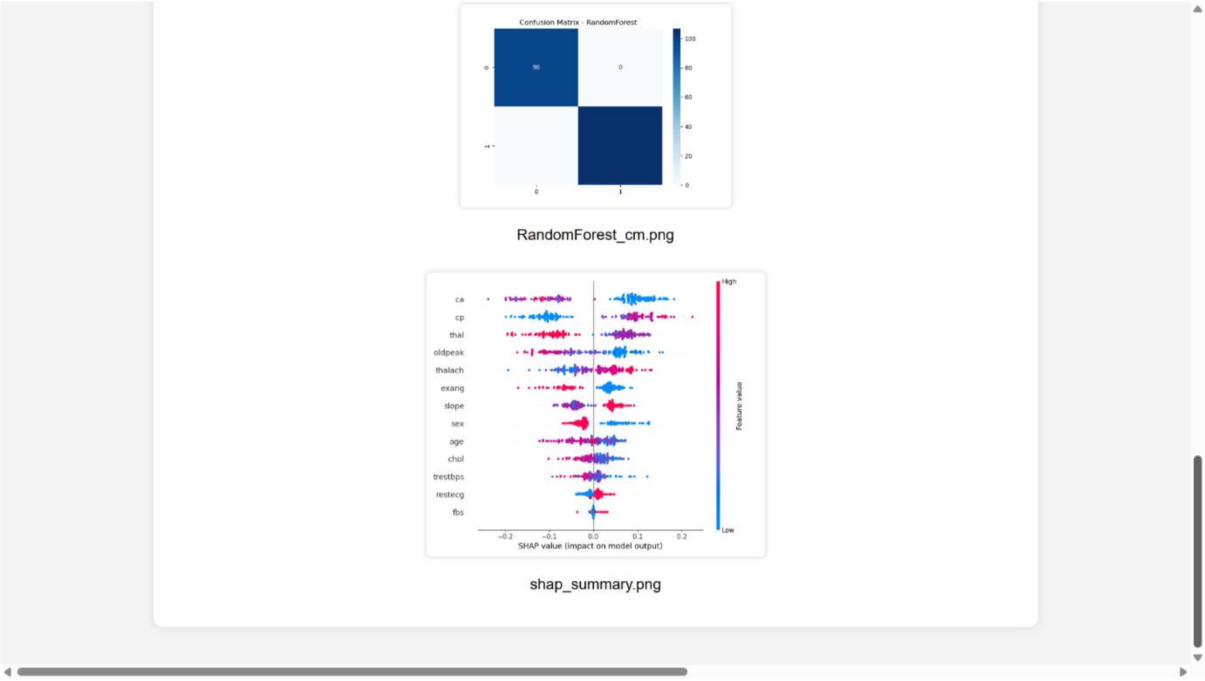
Prediction: 1

Probability: 0.5238909078251182

SHAP Explanation

SHAP Summary Plot (shap_summary.png): This plot shows the impact of each feature on the model's predictions across all test samples. Features are ranked by importance (top to bottom). Red indicates a positive impact on the prediction, blue indicates a negative impact. SHAP Force Plot (shap_force.png): This plot shows the contribution of each feature to the prediction for the first test sample. Features pushing the prediction higher are in red, and those pushing it lower are in blue.

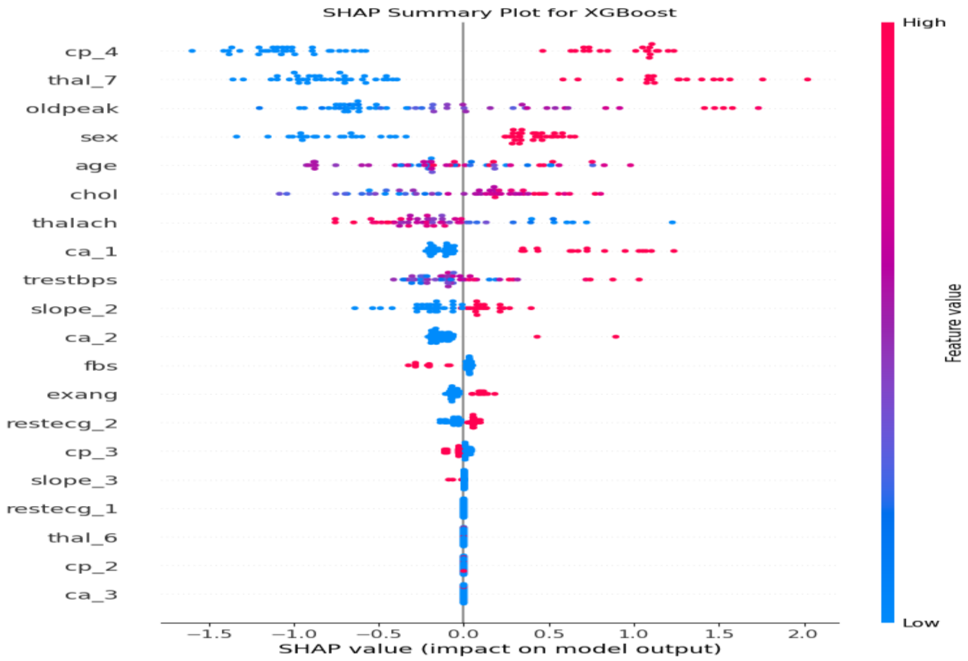
Additional Visualizations



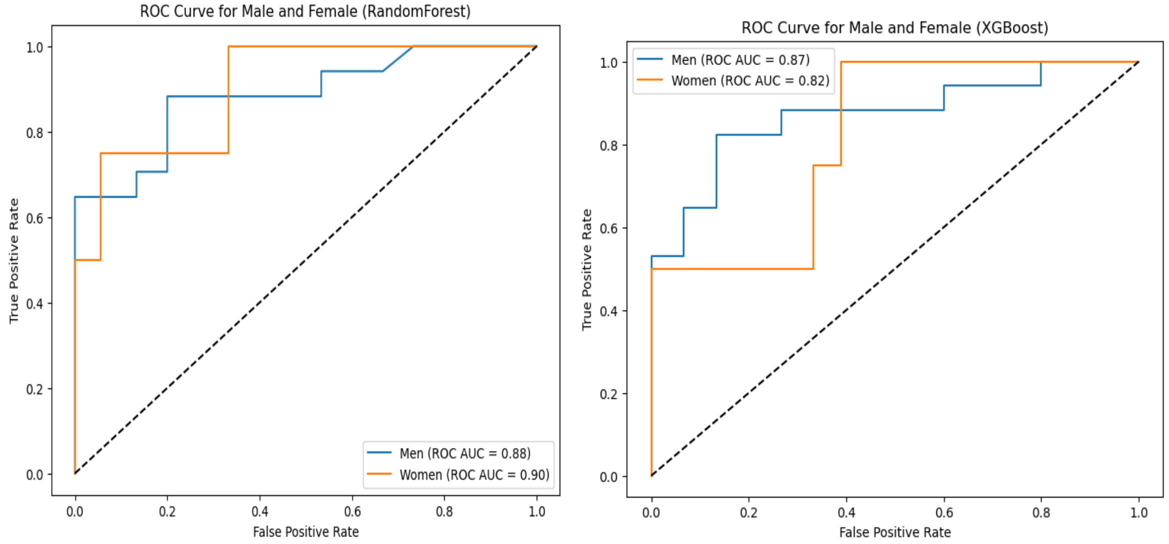
RandomForest_cm.png

shap_summary.png

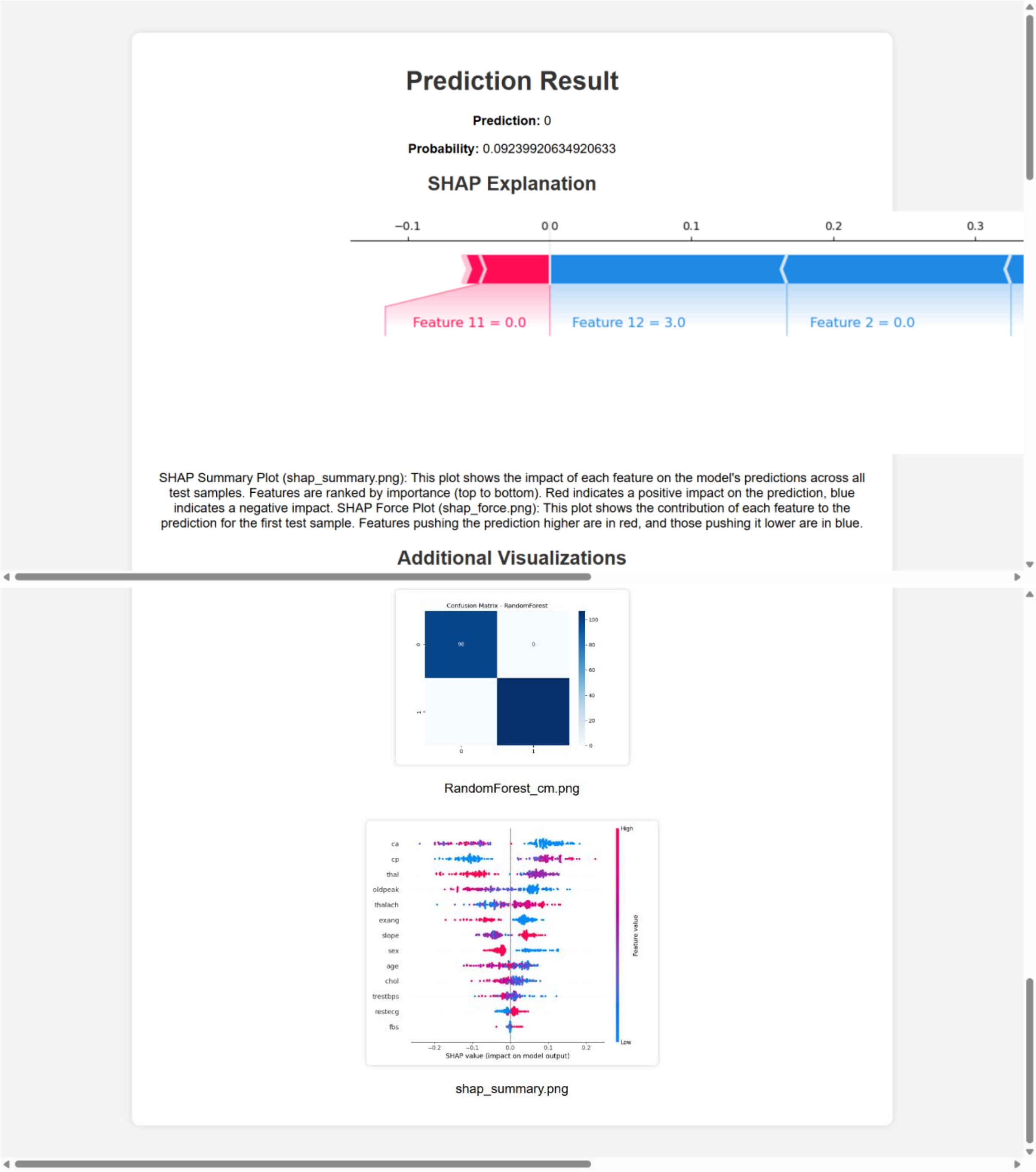
Appendix 6: Result prediction Page shows the User has heart disease



Appendix 7: Explainable AI for XGBoost



Appendix 8: ROC-curve for RF and XGBoost for Gender Equal Opportunities



Prediction Result

Prediction: 0

Probability: 0.09239920634920633

SHAP Explanation

SHAP Summary Plot (shap_summary.png): This plot shows the impact of each feature on the model's predictions across all test samples. Features are ranked by importance (top to bottom). Red indicates a positive impact on the prediction, blue indicates a negative impact. SHAP Force Plot (shap_force.png): This plot shows the contribution of each feature to the prediction for the first test sample. Features pushing the prediction higher are in red, and those pushing it lower are in blue.

Additional Visualizations

RandomForest_cm.png

shap_summary.png

Appendix: The Result Prediction Page shows No heart disease