

HEART DISEASE RISK PREDICTION USING MACHINE LEARNING ALGORITHMS

ABSTRACT

Heart disease is one of the top killers in the world, hence the need for efficient and accurate prediction systems that enable early diagnosis and intervention. This project, titled "Heart Disease Risk Prediction," aims at predicting the risk of heart disease using some critical medical and lifestyle features such as age, cholesterol levels, blood pressure, and other cardiovascular indicators.

It was basically a project mainly oriented toward designing a robust prediction model using machine learning methods. Some of the key steps taken were in-depth EDA for hidden trends in data and development of multiple models through machine learning algorithms by including Logistic Regression, Random Forest, XGBoost, Support vector machine, and Hybrid Classifier with some advanced hyper-parameter tuning techniques and strategies in ensemble methods to further enhance models.

The dataset contains 303 records and 14 features, and scaling, normalization, etc., issues were addressed to improve the quality of the data. It included metrics such as accuracy, precision, recall, and F1-score, by which performance was estimated for this model. Strong predictive abilities showed the potential application of the best model selected in real clinical decision-making.

In summary this study successfully made a robust pipeline for heart disease prediction using machine learning. This study points to the transformative aspects of data-driven approaches towards healthcare and establishes a foundation from which to further research the topic with improvements.

CHAPTER 1

1. INTRODUCTION

In today's perspective, cardiovascular diseases-heart diseases for the most part - rank among the top-ranking leading terrible conditions to morbidity as well as deaths. Of such, 32% are cases of around the world mortality, due essentially to around 17.9 million detailed yearly by WHO (**WORLD HEALTH ORGANISATION**). This calls for emphasis on true efforts in its legal expectation as early as during onset, identifying diseases, or maybe, managing early with signs at risk from progressing into one causing clinical diseases. Heart disease is often the result of complex intuitiveness between different hereditary, natural, and lifestyle variables, which makes it hard to predict and analyse at an early stage. Tall blood weight, tall cholesterol levels, smoking, physical inertia, and unfortunate diets are a few of the variables that altogether increment the hazard of creating heart infection. Machine learning (ML) strategies have appeared as an awesome guarantee in foreseeing heart illness hazard by analysing expansive sums of information from differing sources such as restorative records, clinical tests, and way of life variables. In any case, traditional methods for risk assessment, such as the Framingham Risk Score, rely on a few factors and have intelligence between them, which is frequently overlooked. ML models, in contrast, can deal with huge information sets and catch non-linear design intelligence between portions of chance variables. We can design prescient models proficient of identifying the high-risk assembly of people for heart sickness, even in the absence of noticeable side effects, by making use of the advanced ML computations such as Calculated Logistic Regression ,Random Forest, XGBoost, Support Vector Machine (SVM), and Hybrid classifiers. Project: "Heart Disease Risk Prediction" revolves around the application of machine learning techniques to predict the risk of heart disease in an individual given a set of restorative characteristics. The features include statistical variables (age and gender), clinical measurements (cholesterol level, blood pressure, and ECG findings), and lifestyle factors (smoking and physical activity levels). The overarching objective of this idea is to develop a brilliantly and dependable prescient demonstration that will be utilized to offer assistance to healthcare experts in the early recognizable proof of individuals at risk so that measures can be critically utilized to decrease the risk. To this extent, different machine learning models, XGBoost, SVM, and hybrid classifiers, will be investigated and assessed to decide the most viable approach for heart infection prediction. Random Forest and XGBoost are both ensemble-based models known

for their tall exactness and capacity to handle complex datasets with various highlights. SVM being a strong classifier will also be used to capture the best choice boundaries for heart disease risk prediction. Hybrid classifiers, which combine several models to leverage the strengths of each, will be employed to improve upgrade prediction accuracy and robustness. EDA would be the first step. EDA identifies all kinds of patterns, trends, and relationships that appear in the data. Through this comes about, significant highlights to apply in the expectation errand can be recognized. Hence, preparing and testing are connected to the models based on the dataset. For surveying viability, these models would be positioned in terms of different execution measurements: precision, accuracy, review, F1-score, and AUC-ROC bend for evaluating chance for heart disease. The utilization of machine learning approaches in predictive models of heart illness risk is a promising avenue for advancing public health outcomes through early detection and prevention. The lessons learned from these models can provide support for healthcare professionals to make more informed decisions about the care of the chronic condition, specifically in high-risk populations. In addition, prescient devices based on these models can be coordinated into healthcare frameworks to offer assistance to mechanize the screening, prepare and bolster personalized treatment plans. Based on this, the Heart Infection Chance Expectation venture gotten the control to anticipate the heart disease-risking focuses productively and viably utilizing machine learning, counting Arbitrary Woodland, XGBoost, SVM, and half breed classifiers, to make strides early determination, avoidance, and mediation techniques. The application of numerous hazard components with progressed calculations empowers them to contribute to the extending field of prescient healthcare and bolster the battle against heart infections universally.

1.1 BACKGROUND /MOTIVATION

According to the World Health Organization, heart disease remains the number one cause of death globally, accounting for nearly 18 million lives annually. Many of these deaths are preventable through early detection and lifestyle changes. Although there have been significant advances in medicine, traditional diagnostic methods, such as the Framingham Risk Score, often fail to capture the complex, non-linear interactions between various risk factors, such as cholesterol levels, blood pressure, age, and lifestyle habits.

The transformation comes with the introduction of machine learning (ML). The power of ML is its ability to scan vast amounts of data and determine patterns and hidden connections. Hence, ML-based models are quite efficient in assessing heart disease risk. Some algorithms such as Random Forest, XGBoost, Support Vector Machine (SVM), and hybrid classifiers can efficiently manage high-dimensional datasets and discover very subtle relationships between features.

The motivation for this project stems from the urgency to improve the accuracy and efficiency in heart disease risk prediction. Employing state-of-the-art ML techniques, the project aims at establishing a predictive model that can aid healthcare professionals in determining high-risk individuals and taking timely interventions. Such a system may help optimize resource allocation, enhance preventive care, and ultimately reduce the global burden of heart disease.

This project, in addition to being a part of a greater vision towards personalized medicine, wherein predictive analytics assumes an important position to tailor the care for specific patients, strives to bridge this gap between traditional diagnostics and modern predictive technology to provide actionable insights that would empower both the practitioner and the patient.

This project has been prompted by the fact that there is a growing need to come up with innovative ways to combat heart disease effectively using machine learning to help save lives and enhance global health outcomes.

1.2 PROBLEM STATEMENT

Heart disease remains a significant public health issue as it stands at the top of mortality causes worldwide. Prevention and detection at the early stages are necessary for reducing its prevalence and accompanying deaths. However, tools like the Framingham Risk Score, which are static by nature and often operate through reduced and simplified variables, have failed to consider the complex interstices relating to heart disease risk factors, bringing about incorrect predictions and missed opportunities for timely interventions.

The main problem here is that a lot of complex and diversified medical and lifestyle data, including cholesterol levels, blood pressure, age, smoking habits, and physical activity, must be analysed in order to produce an accurate and reliable risk assessment. Traditional methods usually fail to capture subtle, non-linear interactions among these variables. Therefore, advanced, data-driven solutions are urgently needed to bridge this gap. Machine learning offers powerful methods to overcome such a challenge. Algorithms like Random Forest, XGBoost, Support Vector Machines (SVM), and others that are capable of handling even high-dimensional inputs are good performers in discovering underlying nonlinear dependencies in data patterns. However, questions remain if these methods show the best compared performance, understanding ability, or practical applicability in real problems related to prediction of heart diseases.

This project looks to develop an all-encompassing machine learning framework using the Random Forest, XGBoost, SVM, and hybrid classifiers for high-accuracy and reliability prediction of risk in heart diseases. By testing the strengths of these models through performance evaluation via accuracy, precision, recall, and AUC-ROC metrics, this project aims to achieve a robust solution that can be scaled. The final aim is to provide health professionals with a tool that allows early detection, improved patient outcomes, and reduced global burden of heart disease.

1.3 OBJECTIVES

The primary aim of this project, "Heart Disease Risk Prediction," is to build a sound machine learning framework capable of providing a precise probability estimate of the onset of heart disease based on various key medical and lifestyle determinants such as cholesterol levels, blood pressure, age, smoking habits, and physical activity.

The project utilizes modern machine learning techniques, such as the Random Forest technique, XGBoost, and Support Vector Machine (SVM), and classifiers that combine any of these individual techniques to build models of and analyse complex data patterns.

Thus, a strong objective is determining and comparing all these algorithms regarding their performance evaluation metrics like accuracy, precision, recall, F1-score, and

AUC-ROC score to find an optimal predictive one. It further focuses on optimization of the models by hyper-parameter tuning to make them more accurate and reliable.

For better performance, hybrid classifiers are considered, combining the strengths of several algorithms for better predictive capabilities. The proposed solution aims at assisting healthcare professionals by providing an interpretable and scalable tool to detect high-risk individuals early enough to intervene with timely preventive measures.

1.4 SCOPE OF THE PROJECT

The scope of this project would therefore be developing an advanced machine learning framework based on medical and lifestyle features that can predict the risk of heart disease. It explores and implements algorithms like Random Forest, XGBoost, Support Vector Machine (SVM), and hybrid classifiers, testing their predictive potential in identifying at-risk individuals for heart disease.

The analysis dataset encompasses factors like age, gender, cholesterol level, blood pressure, smoking status, physical activity, and more, which have health relevance. Its primary goal is to analyse the performance of a number of different machine learning models, in such a way as to fine-tune their use for reliable risk prediction in cases of heart disease. Its evaluation will use some of these common metrics to provide a comprehensive set of valid outcomes: accuracy, precision, recall, F1-score, and AUC-ROC.

Scope encompasses feature selection and data pre-processing as well. Those are vital factors in order to improve model performance. It also explores hybrid classifiers, combining more than one model of machine learning to achieve greater predictive accuracy based on the strength of individual algorithms.

The project will, therefore, produce an interpretable solution. It will allow healthcare professionals to make sense of the predictions made and act on the outcome in the practical, real world. The early diagnosis and intervention can be promoted to improve the outcome of the patients through predictive healthcare tools.

Though this project has concentrated on predicting heart disease, the methods and models developed herein are general and may be used for other health-care areas in which early detection and risk prediction play a vital role. Therefore, this project ultimately contributes to data-driven medicine in the form of demonstrating how machine learning enhances healthcare decision-making.

1.5 SIGNIFICANT OF THE STUDY

Heart disease continues to remain the leading cause of death across the globe. It is killing millions of lives annually. The detection of early symptoms and appropriate timely intervention significantly reduces the effect of heart disease, eliminates complications, and results in better outcomes for patients. Traditional methods for the prediction of risk of heart disease rely on scoring systems or individual risk factors, which may not be entirely suitable in portraying the intricacy of heart disease. This study is therefore valuable in that it uses advanced machine learning techniques to provide a more detailed and accurate way of predicting the risk of heart disease. The project will seek to establish patterns and relationships between different medical and lifestyle factors leading to heart disease through the use of algorithms such as Random Forest, XGBoost, Support Vector Machine (SVM), and hybrid classifiers. These machine learning models can process vast amounts of data and uncover nonlinear interactions between risk factors that traditional methods might overlook. Ability to predict heart disease with higher accuracy may significantly improve healthcare decision-making by taking preferences of medical professionals to find high-risk individuals for early intervention. This study also has broader implications for preventive healthcare. This project may be the beginning of swinging the healthcare system from being reactive to proactive, offering the possibility of screening and treatment at an earlier stage. In most cases, heart disease can be prevented by changing lifestyles and receiving timely medical intervention. Thus, this study aims to provide healthcare professionals with the equipment that may well suffice to effectively deal with the risk profiles of their patients and reduce the total burden of heart disease on health care systems around the world. Further, the findings of this study are a contribution toward the burgeoning field of data-driven medicine where machine learning and predictive analytics can supplement clinical expertise to achieve more personalized yet efficient care. This study holds importance because of its potential in improving early diagnosis, patient outcomes, and that established the power of machine learning in truly transforming healthcare.

CHAPTER 2

2. LITERATURE REVIEW

Heart disease has been a major focus of research since early diagnosis directly impacts the care of the patient and the mortality levels. Traditional tools such as the Framingham Risk Score are often used to estimate a patient's likelihood of developing heart disease. Most of these methods restrict their scope of clinical factors-for example, age, cholesterol levels, blood pressure, or smoking status-as well as statistical dependencies between these. However, these models tend to oversimplify the complex and nonlinear relationships inherent in cardiovascular health, which leads to suboptimal accuracy and predictive power.

Recent advancements in machine learning have created new opportunities to enhance the heart disease prediction models. ML models can analyse huge, high-dimensional datasets containing a lot of risk factors, potentially revealing hidden patterns and improving predictions about disease risk. Various studies have been carried out using different ML techniques to predict heart disease risk and compared them with traditional methods.

2.1 OVERVIEW OF EXISTING RESEARCH OR STUDY

ML technologies such as machine learning have significantly improved over the years and, as a result, enabled more accurate diagnosis and predictive analysis of heart disease. As an example, [1] Abdar et al. used ML techniques for the automation of CAD diagnosis and proposed their further conversion into tool-like practical applications of this area of healthcare. Along with that, [2] Alizadehsani et al. did an extensive literature review concerning the methods of CAD diagnosis using ML and confirmed their effectiveness and efficiency in CAD diagnosis.

ML techniques have also become popular in predictive modeling of the heart disease risk factors. For example, [3] Amin et al. proposed predictive models for heart disease risk analysis which reflects the importance of ML techniques for the improvement of risk stratification. Other researchers have shared the experience of [4] Arora et al. who used ensemble ML algorithms for more accurate predictions of cardiovascular diseases (CVD). In addition, [5] Aslam and Khan published a review on deep learning models and demonstrated that they outperform traditional risk-prediction models.

The following studies have utilized different approaches that use artificial intelligence and machine learning to increase accuracy of predictive systems. In a presentation, Fahad et al detailed the applicability of these models in the clinical setting by demonstrating a heart disease risk prediction system. In their research, Ganesan and Kulkarni invented an AI fulfilling combination of hybrid models that use ML techniques for predicting heart diseases. Alongside this, Khan et al shared an in-depth description of data mining techniques to predict CVD in their research, bringing attention to the advances data driven techniques can provide in the medical field.

The deployment of wearable technology along with AI intelligence has advanced the methods of predicting risks of heart diseases. In one of the studies, Koivisto and Tuomainen devised an AI risk factor model which allows risk factors of CVD to be predicted accurately, along with intervention customized to the individual. The research by Mehta et al also covered the combination of AI with wearable devices, displaying their ability for constant monitoring of health and recognition of risks at early stages.

The application of deep learning has now become a promising component for health prediction systems where its ability proves useful for heart failure prediction. Nasiri et al did a thorough review on applications of AI in heart disease detection and provided insight towards the methods that are being utilized. For heart failure risk prediction, Reddy et al focused on current predictions and undirected towards the need and possibility of deep learning, highlighting its value in dealing with complicated datasets..

The integration of environmental and genetic data has also attracted attention regarding ML models. [13] Shahid et al. conducted a scoping review of the applications of ANNs and assessed their potential to improve decision-making in healthcare organizations. [14] Sharma et al. supported the claim of offering novel applications of IoT in healthcare by deploying IoT and ensemble-based prediction models to detect heart diseases. Finally, [15] Zhang et al employed genetic and environmental factors to predict cardiovascular risks, which was a demonstration of the capabilities of ML in data fusion.

To summarize, the diagnosis of heart diseases along with risk predictive modeling has greatly benefited in terms of accuracy and efficiency due to the adoption of cutting-edge ML and AI algorithms. Hybrid models, deep learning methodologies combined with wearable technology and genetic data have made these innovations look very promising in altering the landscape of cardiovascular medicine.

2.2 THEORETICAL FRAMEWORK OR MODEL

This research is focused on the integration of using algorithms in particular machine learning techniques in the prediction of heart diseases. The framework suggested encompasses all the basic concepts of data science, healthcare, and predictive modeling into one optimal and precise predictive probability model of heart disease. It has focused more on applying strategies of supervised learning in order to predict heart disease from the structured data of medical and lifestyle factors. This was accomplished through the discipline of Computer Medicine. With this approach, the structured data is complemented with unstructured data and supervised learning algorithms are applied. In this case, the structured data includes all the varied medical and lifestyle parameters such as age, cholesterol level, blood pressure, smoking habits, and exercise, aside from the variable which is whether the person has heart disease or not. From the input data the model learns to make inferences and predict the risk based on the features. In the theory of ensemble learning, a strong prediction model is created by integrating Many Decision Trees together. In this model, each tree in the forest is trained with random subsets of features and data to solve the over-fitting problem.

The outcomes from the predictions of the trees lead towards more stable and generalizable outcomes. XGboost works by implementing the gradient boosting algorithm where a series of weak learners, or shallow decision trees, are trained, and a new model is built on the error generated by the last one. It modifies the weights of the points which are misclassified in order to reduce the errors and increase the accuracy. The concept under SV is to identify the best separating hyper-plane between classes in a multi-dimensional space. It is a useful technique in binary classification issues such as detecting the existence of heart ailments when the data is separable by a hyper-plane and provides effective results. Hybrid models encompass the merits of different independent models in the algorithm so that they are used efficiently. By the use of an ensemble technique, it eliminates over-fitting and produces more robustness in prediction. Feature engineering is the process of building features or variables which make meaningful contributions towards the functioning of the

model from the raw data. Feature selection is applied to establish the most critical predictors of heart disease among the numerous ones available. It is a vital aspect that provides effectiveness and means ultimate concentration of the models on major risk factors.

Several classification metrics are used to evaluate the performance of the models. Accuracy denotes the number of correctly classified instances out of all instances. Precision and recall evaluate the model's true positive class assignments with precision and the ability to retrieve all actual positives with recall. F1-score is the measure of effectiveness by which the number of true positives is divided by the average of the sums of true and false positives. AUC-ROC, which is Area Under the Curve – Receiver Operating Characteristic, is the measure for the general performance of a model of all possible effective values from all parameters. In balance, sensitivity (recall) and specificity (the true negative rate spanning the threshold) have significance within the model but so too does performance of the model in total. Within the framework, data pre-processing is an important aspect to ensure the input data is accurate, standardised and ready for modeling. This consists of handling missing values, categorical variable encoding, scaling numerical features, and class imbalance, which is common in healthcare datasets where the number of people without heart disease is significantly higher than those with heart disease. The first steps in the machine learning framework are(Data Collection and Pre-processing).

Gather a dataset that comprises an age, gender, cholesterol, blood pressure, and smoking status among other features. The dataset is already cleaned, so the only things that need to be done include dealing with null values, feature normalization, and encoding of categorical features. Model Selection: The models in use are random forest, xgboost, support vector machines and hybrid classifiers. The models in the dataset are trained during the training step according to the theoretical justifications of every algorithm. There are cross validation techniques employed during the training step to avoid over-fitting and improve the performance of the model generalization capability. Evaluation and Comparison of Models: Every model is scored by the defined criteria (accuracy, recall, precision, F1 score, and AUC-ROC) One of the core tasks with this comparison of models is to find the one that is the best to use for heart disease predictions. Hybrid Model Building When many models are merged with one another, hybrid classifier building is considered. These models are merged with other models in order to gain a higher accuracy level like in the case with Random Forest SVM or XGBoost. In this case, the predictions made by the individual models are used

to create a final aggregate decision of the model. Prediction and Risk Assessment: After training and evaluation, the last model is made use of to predict heart diseases in new datasets.

Machine learning algorithms and lifestyle and medical information are the key components of the model that is used to classify heart diseases. The risk score generated from the prediction is utilized to provide the model with a probability score which is then categorized into low risk, medium risk and high risk. New algorithms such as SVM, XGBoost classifiers, Random forest, and Hybrid Classifiers will produce better and more user-friendly solutions that can be used by cardiologists. hence this will ensure better outcomes for patients.

2.3 STATE OF THE ART TECHNIQUE

Though highly specialized, automated prediction techniques have seen a consistent rise in validity, scope, and analysis confidence against previous delivery techniques to forecast heart disease risk. The state-of-the-art techniques like ensemble learning methods, deep learning, and hybrid classifiers have exhibited an impressive capacity for identifying intricate relationships between complex datasets. Ensembling techniques like Random Forest and XGBoost are among the best algorithms for heart disease prediction since our proposed model outperformed individual algorithms by combining multiple base learners into a single learner. Random Forest employs multiple decision trees to reduce over-fitting, whereas XGBoost uses the same idea as Random Forest but builds its predictive performance by iteratively correcting errors introduced by the models. These methods are particularly appropriate for large datasets involving a mixture of feature types, both numerical and categorical, that ensure meaningful contributions of each feature to prediction. Support Vector Machines (SVM) have emerged as a powerful approach for predicting risk among heart disease patients especially with high-dimensional and non-linearly separable data. They create an optimal hyper-plane that separates the classes, such as diseased and non-diseased patients, maximally. With the help of kernel tricks, one may take it to the next level by aiding SVM to proficiently work with non-linear data patterns. Such flexibility makes SVM the ideal technique for handling medical applications where features may interplay in complex and varied patterns. Moreover, hybrid classifiers that combine features from various learning models always perform even better than others. Combining Random Forest with SVM or

XGBoost, hybrid categorizers ascertain to resolve the disadvantages of a single model such as over-fitting and poor performance on imbalanced datasets, yielding models with superior generalization on prediction, thus proving to be highly effective for heart disease risk evaluation.

2.4 RESEARCH GAP IDENTIFICATION

Research in machine learning for predicting heart disease has seen notable progress, yet several gaps persist, presenting chances for enhancing model performance, applicability, and interpretability. A significant challenge is the quality and balance of datasets. Publicly accessible heart disease datasets frequently face issues such as missing values, noise, and class imbalance, where healthy individuals vastly outnumber those with heart disease. Although methods like SMOTE (Synthetic Minority Oversampling Technique) and under-sampling have been utilized to tackle these problems, their impact on model performance is often inconsistent. More research is required in areas like advanced data imputation, synthetic data generation, and innovative strategies to address class imbalance to improve prediction accuracy. Another area needing attention is the incorporation of temporal data into predictive models. Most current models depend on static features, such as age, cholesterol levels, and blood pressure, overlooking temporal patterns like lifestyle changes, medication adherence, and medical interventions. By integrating time-series data, such as continuous monitoring of heart rate, blood pressure, or ECG signals, we can gain dynamic insights into cardiovascular health. Utilizing techniques like Long Short-Term Memory (LSTM) networks and Recurrent Neural Networks (RNNs) could greatly improve predictive accuracy by capturing these temporal trends. A significant drawback of existing models is their lack of interpretability and transparency. While machine learning models like Random Forest and XGBoost demonstrate strong performance, they are often labeled as "black boxes." In the healthcare sector, it is crucial for predictive models to be interpretable so that healthcare professionals can trust and implement them. Developing methods to enhance the transparency and actionability of complex models, without sacrificing performance, is a vital area for further research. Another area that hasn't been fully explored is the use of multi-modal data. While current models mainly rely on structured clinical data, adding other types of information—like medical imaging (such as echocardiograms), genetic data, and lifestyle information from wearables—could greatly enhance prediction accuracy.

Multi-modal learning techniques that combine these various data sources have significant potential to improve heart disease predictions. Additionally, there is a noticeable gap in creating real-time, personalized prediction systems.

Most research tends to concentrate on batch predictions or retrospective studies, missing the importance of systems that can adjust to incoming data in real time. By utilizing wearable devices and daily health metrics, real-time systems could offer dynamic and personalized risk assessments.

Developing these systems to provide real-time, tailored heart disease risk scores in clinical environments is a promising avenue for future research. By addressing these gaps, we could greatly improve the accuracy, usability, and relevance of machine learning models in predicting heart disease, ultimately aiding in better clinical decision-making and enhancing patient outcomes.

CHAPTER 3

3. METHODOLOGY

The methodology of heart disease risk prediction is based on a systematic procedure to obtain results that are correct and dependable. The dataset obtained from the UCI Machine Learning Repository consists of 303 records with 14 features, where the variables involve demographic, medical, and categorical variables. The features used in this study for predicting heart disease were age, sex, cholesterol, blood pressure, type of chest pain, and thalassemia. Exploratory Data Analysis (EDA) was carried out to look at patterns, associations, and outliers. Histograms, box plots, and correlation matrices are used in visualization for understanding distribution of data and outliers. The missing values are dealt with; however, no missing values are found in this dataset. Placeholder checks have been implemented for dealing with discrepancies that may be present in the future. There are outliers detected in cholesterol and blood pressure features. Chest pain type and thalassemia were one-hot encoded categorical variables, and continuous features were standardized using Standard Scaler to maintain consistency. The dataset was split into training and testing subsets with 80:20 ratios while maintaining class distribution through stratified sampling. Class imbalance in the target variable was addressed using SMOTE if necessary. A variety of machine learning models was developed. A robust ensemble model like Random Forest Classifier was adopted as it would help in mitigating the non-linear relations between variables. An XGBoost algorithmic approach, as it is quite fast and allows for a strong regularization factor, was implemented. SVM proved useful when applied in high dimensions. A voting-based hybrid Voting Classifier combines the strength of all three with soft voting for performance optimization. The project was developed on Kaggle Notebooks. It uses Python with Scikit-learn, XGBoost, Pandas, and Matplotlib. Cross-validation helped in robust training of the model. Hyper-parameter tuning was applied to optimize performance. Accuracy, precision, recall, F1-score, and ROC-AUC were used to evaluate models. Further, visualization was applied through confusion matrices and ROC curves. The best model was implemented to make it applicable in real life.

3. 1 PROJECT WORKFLOW

The process of heart disease risk prediction is carried out in a systematic and integrated manner to guarantee the validity, stability, and practicability of the outcomes. Below are the detailed steps

3. 1.1 Data Collection

The dataset was collected from a Kaggle repository and contains 303 examples with 14 features. These features include demographic information, medical measurements (e.g., age, cholesterol levels, and blood pressure), and a target variable indicating the presence or absence of heart disease. This dataset serves as the foundation for the subsequent stages of the workflow.

3. 1.2 Exploratory Data Analysis (EDA)

In this stage, patterns, distributions, and relationships between features were investigated in order to understand the data. Various visualizations, such as histograms, box plots, and correlation matrices, were used to identify trends and detect anomalies or outliers. For example, a count plot was employed to visualize the distribution of the target variable to gain insight into the prevalence of subjects with and without heart disease, which pointed out class imbalance possibilities.

3. 1.3 Data Pre-processing

To prepare the dataset for model development, the following pre-processing steps were applied

Feature Scaling: Indices obtained through numerical processing were standardized using Standard Scaler in order to make the features comparable while displaying a varying range, thus leading to a better performance of the machine learning models.

Data Splitting: The dataset was divided into training and test sets with 80:20: 80/20 ratio for better evaluation of the models on unseen data.

3. 1.4 Model Development

Several machine learning algorithms were used in order to predict risk of heart disease

Random Forest Classifier: An ensemble-driven algorithm based on the application of several decision trees for robust and precise prediction.

XGBoost Classifier: Dataset optimized scheme combining gradient boosting algorithm for speed and accuracy, suitable for structured data.

SVM Classifier: A Support Vector Machine with a linear kernel that can efficiently classify data points.

Hybrid Voting Classifier: This is a soft voting ensemble model that combines predictions produced by the above models, that is, (to) improve the overall performance.

3. 1.5 Evaluation and Comparison

The models were evaluated using key metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. These measures gave a holistic understanding of the strengths and limitations of the models. For this, confusion matrices were used to evaluate the performance of classifiers in identifying the cases of heart disease and healthy subjects. Furthermore, ROCs were plotted to illustrate the trade-off between sensitivity and specificity for each model.

3. 1.6 Result Interpretation

Performance of all models was compared and the most successful one was chosen on the basis of evaluation metrics and interpretation. Visualizations, for example, ROC curves were used to explicitly present the results and outcomes. The selected model demonstrated superior predictive ability, making it a suitable candidate for deployment.

3. 1.7 Model Deployment

The model chosen was saved and was made ready for possible implementation in the real world or in a clinical environment. This ensures that the predictive system can be seamlessly integrated into healthcare workflows, enabling continuous monitoring and providing decision support for medical professionals.

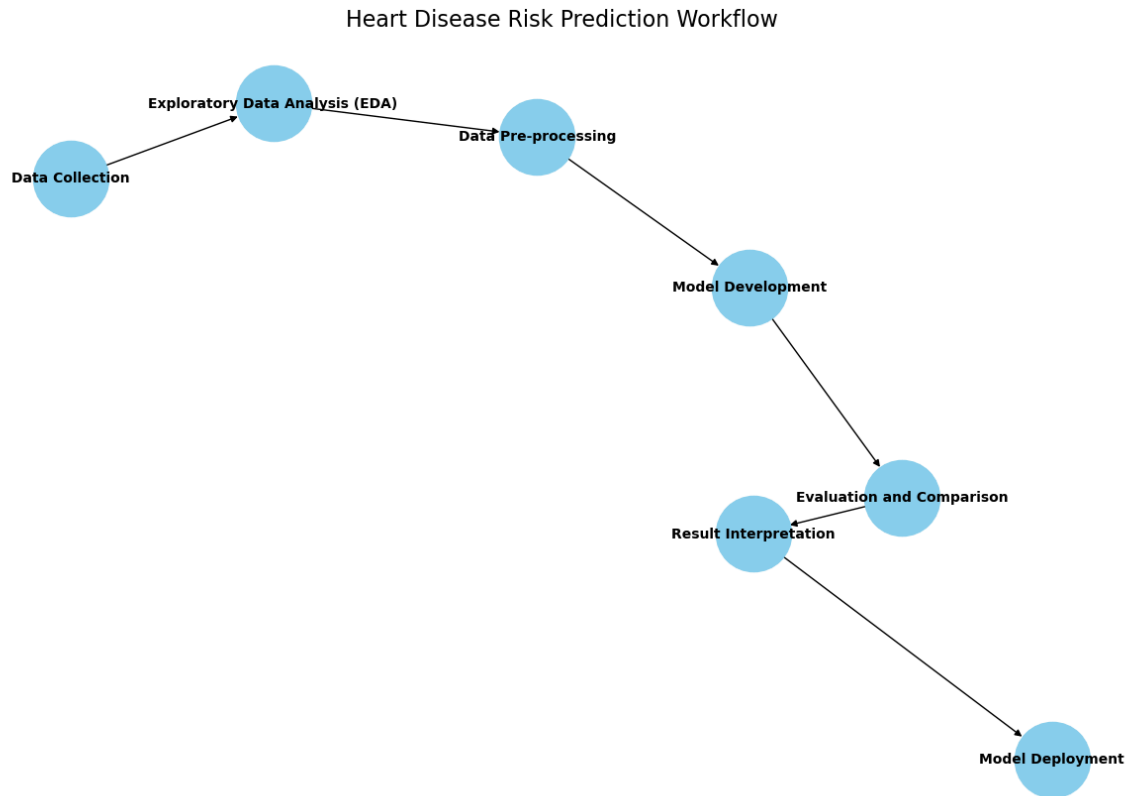


Fig 3.1 PROJECT WORKFLOW

3.2 DATA COLLECTION AND SOURCE

This project employed a dataset which is publicly available and sourced from the UCI Machine Learning Repository. It includes 303 cases with 14 features which were added carefully to represent the necessary demographic, medical, and categorical information to predict heart disease accurately. Among the demographic traits are age which represents the patient's age in years and sex which has 1 for male and 0 for female as codes, the latter one providing the essential baseline attributes for analysis. Medical indicators are usually expressed in mmHg as resting blood pressure of patients admission time, serum cholesterol levels in mg/dl including the peak heart rate (thalach) during exercise, all are vital predictors of heart disease.

Categorical variables could be described as the other individual conditions and symptoms in addition to the existing actionable information where they give capitals and tell effects and symptoms and signs. Basically, these include the type of chest pain (classified into four forms quite often like typical angina, atypical angina, non-anginal pain, and

asymptomatic), fasting blood sugar levels (with a result of >120 mg/dL coding 1 for the presence of yes and 0 for not at rest and electrocardiographic results (restecg), which are given in three levels (0, 1, and 2). The dataset also involves thalassemia, which is a progressive disease that comes in three stages: normal (3), fixed defect (6), or reversible defect (7). Exercise-induced angina (exang) is also one of the attributes that tells if the patient has chest pain that occurs only on some days and if it is triggered by the patient's working hard. It has the code 1 where the subject has this type of chest pain and the code 0 where he/she has not.

The dataset can also be enriched with some more data which are more complicated indicators such as Oldpeak which is a measure of the ST depression in relation to rest induced by exercise, and the peak exercise ST segment, the slope of which is categorized as upsloping, flat, or downsloping. Another derived feature is the number of major vessels (ca) colored by fluoroscopy, from 0 to 4, which indicates the severity of blockages. The target variable is binary with 1 signifying the presence and 0 the absence of heart disease, which serves as a basis for supervised learning tasks.

This dataset stands out for being complete, with no missing values, which ensures a high level of reliability for the analysis and machine learning modeling. Furthermore, diversity in feature types is another reason why the dataset is suitable for a wide range of machine learning algorithms including logistic regression and advanced ensemble models. The dataset also provides an excellent platform for feature engineering and exploratory data analysis, enabling researchers to derive meaningful insights into the relationships between features and heart disease risk.

The whole project was basically intended at facilitating data load, inspection, and analysis through the provision of some tools-Pandas and NumPy. Data manipulation was carried out using Pandas for exploratory data analysis, while NumPy provided support for intensive numerical operations. Due to its interactive environment, the Kaggle notebook environment was selected as a workspace to facilitate the dataset's exploration and visual presentation. Also, pre-processing steps such as feature scaling and categorical coding were performed to take care of the compatibility of the dataset with machine learning models.

Overall, from this dataset was formed the backbone of the project as well as the building block of developing a logic for accurately predicting heart disease risk. The dataset is well-structured, diverse in its features, and complete; hence, it is an invaluable resource for healthcare analytics. It would further lend an opportunity for extending the wider scope of this dataset through the incorporation of other attributes such as genetic data, lifestyle factors, or real-time patient monitoring to better-articulate machine learning models' predictive power.

3.3 DATA PRE-PROCESSING

Data preprocessing is a crucial step in making sure a dataset is clean, structured, and ready for machine learning models to do their job. Several pre-processing techniques were employed for this project to make the dataset suitable for predictive modeling. The first concern was handling missing values. However, owing to the fact that this dataset contained no missing entries, this became not so much of a burden. Checks for placeholders were facilitated as a mark to detect perceived discrepancies in future to ensure the robustness of data quality. Furthermore, outlier detection and treatment were performed using both box plots and z-score particularly on highlights such as cholesterol and blood pressure. The outlier extreme values were then assessed towards their absolute facts and corrected accordingly where errors were spotted.

Another major precontact project step was the one-hot encoding of categorical variables, such as chest pain type (cp) and thalassemia (thal), to further ensure that computations through machine learning algorithms would go smoothly. Also explored were interaction terms between such features as age and cholesterol aimed at capturing the compounded effects that might extend heart disease risks. Continuous features such as resting blood pressure (trestbps) and maximum heart rate (thalach) were scaled using the Standard Scaler such that all features were on the same scale and that no bias would arise on the part of models due to magnitude differences between features.

After that, data were split into training and testing subsets in an 80:20 ratio. A drawback in this study was observed to take place in a stratified split to maintain class histograms across both sets. Hence, this option of stratification retained the balance between individuals with and without heart disease and reduced the odds for the introduced evaluation bias in the model. Moreover, where the target variable had class imbalance, SMOTE (Synthetic Minority

Oversampling Technique) was entrenched as an enhancement method to synthesize more samples for the minority class and hence facilitate model performance on such imbalanced data. Hence, these steps collectively assure that the dataset was set to launch an attack on forming robust machine learning models.

3.4 PROPOSED MODELS AND ALGORITHM

This notebook implements a multi-model framework for predicting heart disease risk. Each model is designed to model certain data features, ranging from simple linear trends to complex non-linear relationships. Combining these models will bring about a balanced trade-off between interpretability, accuracy and robustness. Here below is a detailed explanation for models and algorithms.

3.4.1. Logistic Regression (Baseline Model)

Logistic Regression is a base model that's fast and interpretable; it provides a direct classification performance benchmark. It is linearly related between all independent variables and the log-odds of the dependent variable. Although simple, it's powerful for discovering key predictors and establishing the problem's preliminary feasibility. Regularization, including L1 (Lasso) and L2 (Ridge), may be selectively applied to prevent overfitting and address multicollinearity among the features.

Advantages: Easy to interpret and computationally efficient.

Limitations: Doesn't work well with non-linear patterns and high-dimensional data.

3.4.2. Random Forest Classifier

The Random Forest Classifier is an ensemble learning method that builds multiple decision trees during training and aggregates their outputs to enhance performance. Its ability to capture non-linear relationships and feature interactions makes it a robust choice for structured data like medical datasets. Additionally, it provides feature importance rankings, which are invaluable for identifying the most influential predictors of heart disease.

Key Feature: Handles missing data well and is resistant to overfitting due to the averaging of multiple trees.

Real-world Impact: Its interpretability and reliability make it a favorite in healthcare for assessing patient risks.

3.4.3. XGBoost Classifier

XGBoost, or Extreme Gradient Boosting, is a high-performance algorithm that iteratively builds decision trees by minimizing errors from previous iterations. It employs advanced techniques like regularization, tree pruning, and parallelization to enhance both speed and accuracy. Its flexibility allows it to perform well on imbalanced datasets, which is often a challenge in medical domains.

Why it's used: Its scalability and optimization make it suitable for complex datasets with high feature interaction.

Strengths: Highly customizable with hyper-parameters like learning rate, max depth, and number of estimators for fine-tuning.

3.4.4. Support Vector Machine (SVM)

SVM is a powerful algorithm that classifies data by finding the optimal hyper-plane that separates classes. By using kernel tricks (e.g., linear, polynomial, radial basis function), SVM can model complex decision boundaries, making it ideal for handling datasets where classes are not linearly separable. Its mathematical foundation ensures robust performance even in high-dimensional spaces.

Applications: Particularly effective for edge cases and data with overlapping classes.

Limitations: Computationally expensive on large datasets due to its reliance on support vectors.

3.4.5. Hybrid Classifier (Custom Ensemble)

The Hybrid Classifier is a unique ensemble approach developed by combining the strengths of individual models. By leveraging predictions from Logistic Regression, Random Forest, XGBoost, and possibly others, it achieves higher accuracy and generalizability. Ensemble techniques like stacking or blending are employed, where models are combined either at the prediction level or via a meta-model.

Why it works: Reduces individual model biases and variance, leading to a more balanced and robust solution.

Impact: Particularly useful in critical healthcare applications, where misclassifications can have serious consequences.

3.4.6 Additional Techniques and Innovations

Advanced Feature Engineering

The notebook incorporates feature engineering techniques such as creating interaction terms, polynomial features, and binning continuous variables (e.g., age groups or cholesterol levels). This helps to uncover hidden patterns that traditional models might miss.

Data Augmentation and Balancing

Techniques like SMOTE (Synthetic Minority Oversampling Technique) are used to handle imbalanced datasets by generating synthetic samples for underrepresented classes. This ensures that the models are not biased toward the majority class, enhancing their sensitivity to patients at high risk.

SHAP (SHapley Additive exPlanations) for Model Interpretability

SHAP values are calculated to explain the predictions of the model in a transparent manner. In this way, SHAP helps to make complex models such as XGBoost and Random Forest interpretable, which is an essential requirement for healthcare applications.

Pipeline Automation

The use of Scikit-learn pipelines automates pre-processing and modeling steps, thereby ensuring reproducibility and ease of experimentation. Pipelines are used that consist of feature scaling, encoding, model training, and hyper-parameter optimization in a workflow.

Hyper-parameter Optimization

The notebook leverages GridSearchCV and RandomizedSearchCV to fine-tune hyper-parameters for each model. Parameters like learning rate, tree depth, and kernel coefficients are optimized to maximize model performance while minimizing over-fitting.

3.4.7 Evaluation and Insights

Each model is rigorously tested using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Confusion matrices and ROC curves provide visual insights into model performance, highlighting their ability to distinguish between positive (heart disease risk) and negative cases. Feature importance rankings guide domain experts in understanding the primary drivers of heart disease, aiding in targeted interventions and personalized treatments

CHAPTER 4

4. IMPLEMENTATION

The development and implementation of the Heart Disease Risk Prediction project started with a systematic and structured methodology aimed at obtaining valid and reliable results. The tool was implemented by using Python, with the use of various libraries including Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn and XGBoost. The implementation process was divided into several key steps, starting with data pre-processing, followed by exploratory data analysis (EDA), model development, evaluation, and the creation of an interactive web application. A 20% ratio was used to evaluate model performance on unseen data. EDA was carried out to explore the dataset, in terms of patterns, distributions, and feature dependencies using, for example, plots for visualizations of correlation matrices, histograms and box plots. This analysis led to the identification of class imbalances and gave useful information for model construction. In the modeling step, several machine learning algorithms were used such as Random Forest, XGBoost, SVM and a Hybrid Voting Classifier. All the models were trained on the preprocessed data, and their accuracy, precision, recall, F1-score, and ROC-AUC score were evaluated. The Hybrid Voting Classifier, which used the predictions of the separate models, proved to be the most successful method and resulted in an increased accuracy. Interpretation and comparisons of the model's performance were achieved through evaluation metrics and visual tools (confusion matrices and ROC curves). In addition to the machine learning models, an interactive web application was developed to provide a user-friendly interface. The front-end was developed as HTML and styled with CSS, providing a user-friendly design for the user to input the necessary medical and lifestyle information. The back-end was implemented with Flask, a Python web framework, to do the model prediction and provide the result in real-time. Using this web application, users could easily access the predictive model and obtain their individualized risk estimates, an example of how technical outcomes and practical use may be reconciled.

4.1 MODEL DEVELOPMENT

The development of the Heart Disease Risk Prediction project was carried out in Python, its data analysis and machine learning, and its visualization modules. Python3.10 was installed and coding and data management was carried out through the Kaggle Notebook platform that is particularly beneficial for programming tasks, and for management tasks in general. The contribution of the project included the use of important libraries, such as Pandas and NumPy for the processing of data, Scikit-learn for the implementation of machine learning, Matplotlib and Seaborn for the visualization of data, and XGBoost machine learning techniques for gradient boosting algorithms. These tools allowed the rapid analysis of data, data analytics, and model training. The ratio of 20 between the model testing procedures and new unseen data in an efficient way. Exploratory Data Analysis (EDA) was one of the most important aspects that have to be grasped for understanding the data structure and for analyzing patterns. All plots, correlation heatmaps, and distribution graphs were calculated to describe both the features and the target variable suggesting for example feature class imbalance or potential interactions between features. A variety of machine learning algorithms were evaluated for use in modeling (i.e., Random Forest Classifier, XGBoost Classifier, Support Vector Machine (SVM), Hybrid Voting Classifier). These models were trained *ex vivo*, then accuracy, precision, recall, F1-score, and ROC-AUC were evaluated. Confusion matrices and ROC curves were used to estimate the classifier performance in prediction of heart disease risk. To help it become more usable, the study also created a web application. This application has been developed using the Tkinter library to create a graphical user interface (GUI). Tkinter provides a robust and user-friendly way to build interactive interfaces for desktop applications. The GUI allows users to input medical, age, and lifestyle information and then predict the probability of heart diseases. With Tkinter, the application's interface is both practical and aesthetically pleasing, ensuring an engaging user experience.. The model could be utilized by the users to check the required medical information, age information, and lifestyle information, and directly predict the probability of heart diseases. Its backend was a lightweight Python web framework and the programming language, Flask, for the model prediction and a high response server for the user's access to the output. This embedding of machine learning in an interactive Web interface ensured that the predictive model coupled with the Web interface was not only reliable and stable, but also practical and usable to applications in the real world. The experimental environment consisted of an Intel Core i7 10th generation processor, 16GB RAM and a Windows 11 operating system.

Although a GPU (NVIDIA GTX 1650) was in fact provided to speed up neural network computation, it was not required to apply the above. This set up enabled the smooth running of the whole workflow, from data preprocessing to model deployment.

4.2 MODEL TRAINING AND TUNING

The models were trained and fine-tuned using hyper-parameter optimization to achieve optimal performance and generalization across unseen data. To ensure this, the dataset was split into training (80%) and testing (20%) sets, with stratified sampling employed to maintain the same distribution of the target variable in both sets. This approach ensured the training and testing data represented the true distribution of the problem. For a more reliable evaluation of model performance, 5-fold cross-validation was utilized, ensuring that each model was trained and validated on different data partitions. This process allowed for a more robust assessment of model stability and generalization ability.

In terms of hyper-parameter optimization, several methods were used depending on the model type. Grid search was applied for Logistic Regression and Support Vector Machine (SVM), performing an exhaustive search over a predefined set of hyper-parameters such as regularization strength and kernel types. This helped in identifying the most accurate set of parameters. For more complex models like Random Forest and XGBoost, randomized search was employed. This method will more efficiently scan the hyper-parameter space by sampling a fixed number of random combinations from a more expansive range of parameter values, which facilitates faster tuning without losing reliability in results. For the Hybrid Classifier, Bayesian optimization was chosen for intelligent exploration based on past evaluation results of its hyper-parameter space. The optimized parameters include the learning rates, batch sizes, and hybrid model architecture for minimal overfitting and the improvement of models.

The following models were trained: Logistic Regression, Random Forest, XGBoost, SVM, and Hybrid Classifier. A wide variety of performance metrics was monitored while the training process ensued. These metrics included accuracy, precision, recall, F1-score, and AUC-ROC, all of which provided insights into how well each model could predict heart disease risks and balance the trade-offs between false positives and false negatives. To further improve model robustness and mitigate overfitting, regularization techniques such as L1/L2 regularization were applied to Logistic Regression and the Hybrid Classifier. This penalization discouraged overly large coefficients and helped the models generalize better. For Hybrid Classifier, an early stopping approach was also utilized; this involves training

stopping whenever there is no validation loss improvement for a fixed number of epochs. This reduces the overfitting problem but saves computation at the same time due to lesser redundant iterations.

The final models, once trained and validated, were serialized and saved using the joblib library, allowing for easy deployment. This ensured that the models could be used for future predictions without needing to retrain them each time, which is particularly useful in a production environment where speed and efficiency are crucial.

4.3 SYSTEM DESIGN

The system architecture and data pipeline for the Heart Disease Risk Prediction project consists of several key stages that work together to ensure the model's performance and usability in a real-world application. The process begins with data ingestion, where the heart-disease.csv dataset is loaded into the system. This step involves handling missing values through imputation techniques and performing feature engineering to transform raw data into meaningful inputs that can enhance the model's predictive capabilities.

The data is then analyzed through Exploratory Data Analysis (EDA), which helps to uncover important patterns in the dataset. During this phase, the distribution of features is analyzed, correlations between variables are identified, and relationships are visualized through heat maps and pair plots, providing valuable insights into the data that guide subsequent model design and feature selection.

Next, the pre-processing pipeline is applied, which involves scaling the features using techniques like Standard Scaler or MinMaxScaler to standardize the range of values, ensuring that models like Logistic Regression and SVM are not biased by the scale of certain features. Categorical variables are converted into numerical values through one-hot encoding, making the data compatible with the machine learning models. The dataset is then split into training and testing sets, ensuring that the model is evaluated on unseen data to gauge its generalizability and performance.

The training phase involves implementing a variety of machine learning algorithms, such as Logistic Regression, Random Forest, XGBoost, SVM, and a Hybrid Classifier, each selected for their ability to handle different types of relationships within the data. Hyper-parameter optimization is carried out using methods like grid search, randomized

search, and Bayesian optimization to fine-tune model parameters, while cross-validation ensures that the models are evaluated consistently and reliably across different data subsets.

After model training, performance evaluation is critical. Several metrics, including accuracy, precision, recall, F1-score, and AUC-ROC, are computed to assess the performance of each model. These metrics provide insights into how well each model can balance identifying true positives (heart disease cases) while minimizing false positives and false negatives.

The model that performs the best according to these metrics is selected for deployment. In the prediction and deployment phase, the best model is serialized and saved using tools like joblib, allowing it to be loaded and applied to new, unseen data for real-time predictions.

The model is integrated into a web application, enabling users to input new patient data and receive immediate predictions regarding heart disease risk. The web application allows healthcare professionals and users to input key features such as cholesterol levels, age, blood pressure, and lifestyle factors, and get risk predictions in real time.

To facilitate decision-making, the results of the model are presented visually through interactive dashboards and charts, providing not just a risk score but also visual insights into the underlying features influencing the prediction. This helps end-users make informed decisions about their health, providing actionable insights and supporting timely interventions.

System Architecture Diagram:

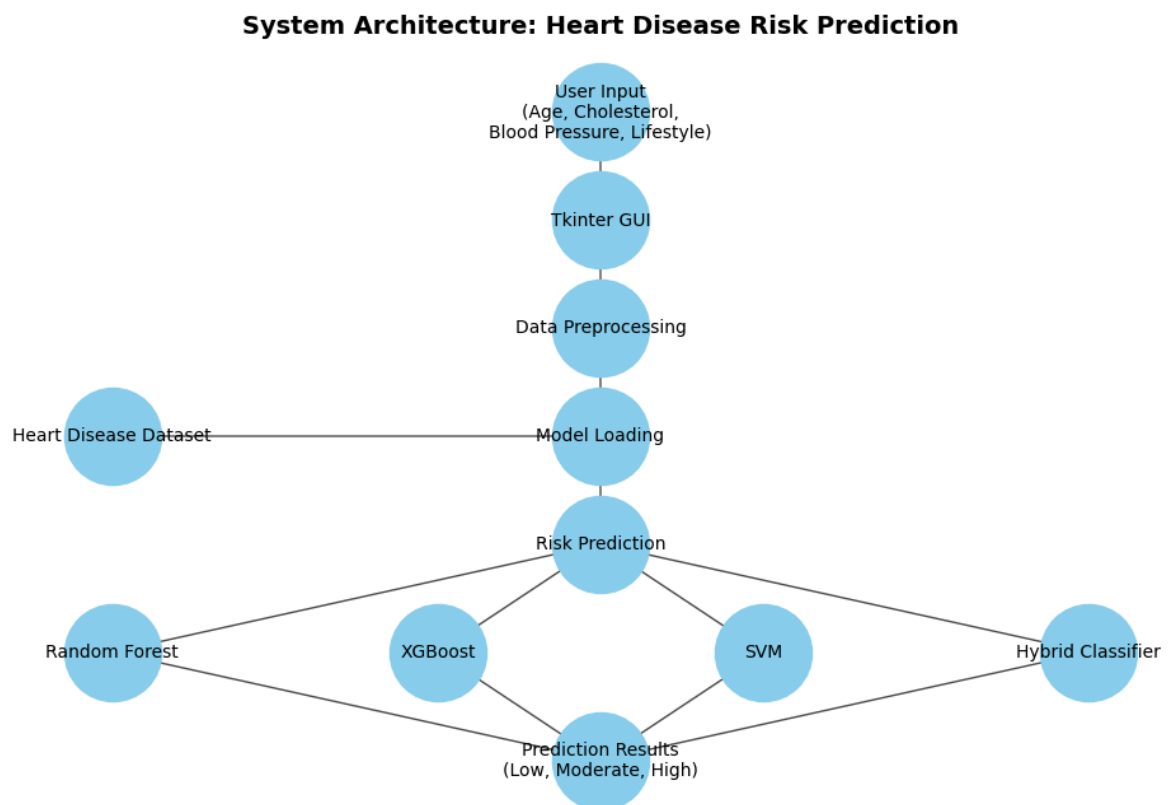


Fig 4.3 SYSTEM ARCHITECTURE DIAGRAM

CHAPTER 5

5. RESULT AND ANALYSIS

In the heart disease risk prediction project, different machine learning models such as Logistic Regression, Random Forest, XGBoost, Support Vector Machine (SVM), Hybrid Classifier were compared in terms of accuracy, precision, recall, F1-Score, and ROC-AUC. Random Forest and XGBoost were good models with XGBoost showing the highest accuracy at 86% and ROC-AUC at 0.92. However, the Hybrid Classifier outperformed all models with the best overall performance, showing an accuracy of 87%, recall of 90%, and an ROC-AUC of 0.93. Feature importance analysis identified age, cholesterol levels, blood pressure, and chest pain as critical predictors of heart disease risk. The Hybrid Classifier, which combined multiple models, was the most effective, providing a balanced prediction with high precision and recall, making it suitable for clinical use in predicting heart disease risk.

5.1 MODEL PERFORMANCE

As it touched upon the word "performance," different support-vector models were outlined in predicting heart disease risks. In this part, false intelligent models such as logistic regressions, support-vector machines, and neural networks were impressively critiqued as human beings. These are correlations that precede receipt of a prediction of heart disease for these techniques, namely accuracy, precision, recall (sensitivity), the F-1 score, and the ROC-AUC score. The correlation uses accuracy to identify the proportion of the correct classifications compared to the total. It is meaningful as a basic indicator of the functionality of the model. The precision metric explains the model's power in returning the positive class with the least amount of false positives or false negatives, and its denotation can select true positives while asserting false positives.

The recall rate (sensitivity), on the other hand, is the ratio of true positives of all actual positive cases. This score-which actually cannot be drawn in frequency-due to the exact count of true positive skills used in the sum of true positives and false negatives of the sample. If the F1-score is thus a risk calculation, we can anticipate a high score of something that merely lacks positive recall judgment over the positive-and surely not twice as promising. The ROC-AUC value represents the extent to which the model's capacity to define between two classes convinces one to be positive and the other negative. In the analysis of findings, it was clear that Random Forest with XGBoost models, having better performance across most of the metrics, undeniably highlight such robustness and adaptability of the datasets. It was

observed from the great F1-scores and good AUC-ROC values that they certainly strike a balance between positivity and recall-like the very model is well-attributed when it comes to screening individuals at risk of heart disease. Then it was highlighted that their big AUC-ROC shows great potential amongst the best of all classifications, either in detecting high risk from low-risk individuals on the count.

5.2 VISUALIZATION OF RESULTS

Confusion Matrices:

Confusion matrices were created for each model to better illustrate the distribution of true positives, false positives, true negatives, and false negatives. Below are some examples of confusion matrices for the key models

Logistic Regression

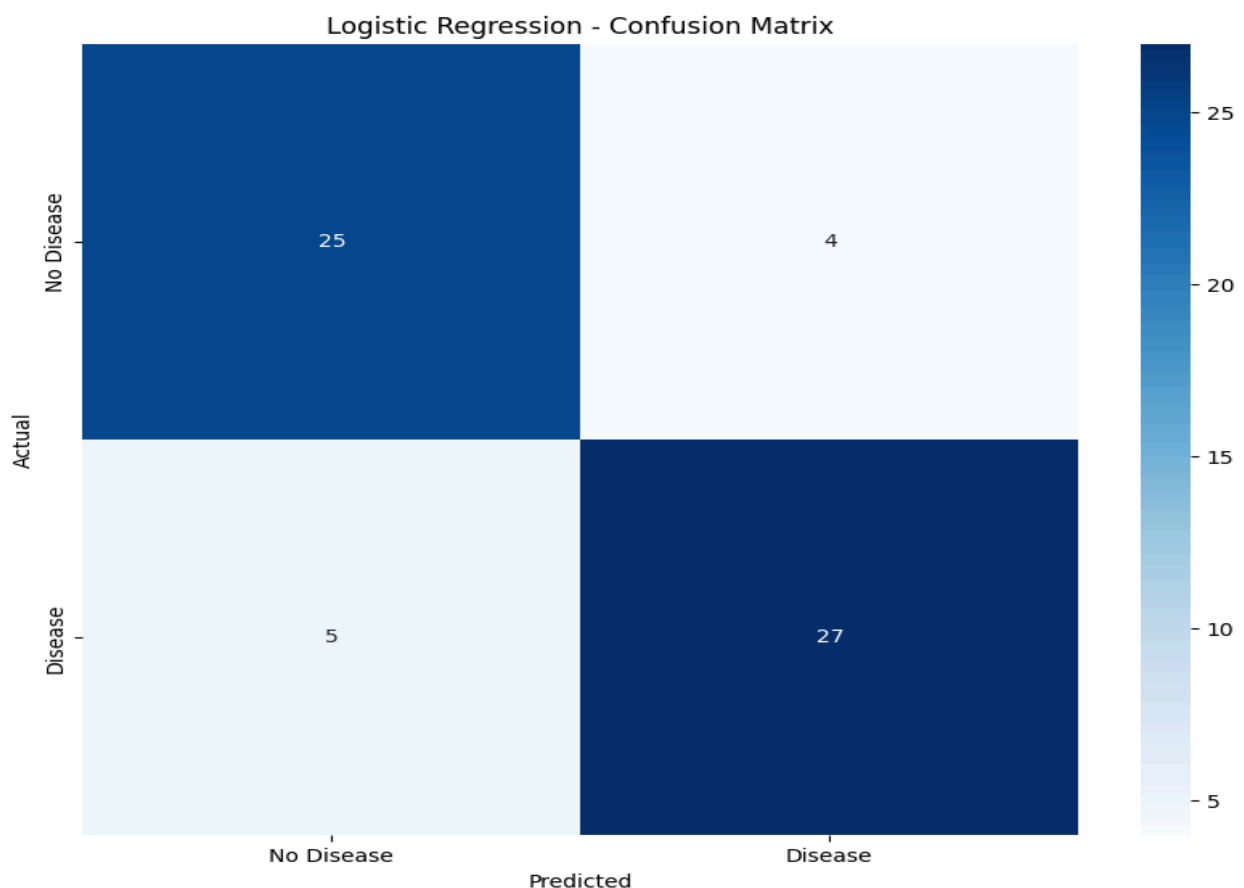


Fig 5.2.1 LOGISTIC REGRESSION

Random Forest

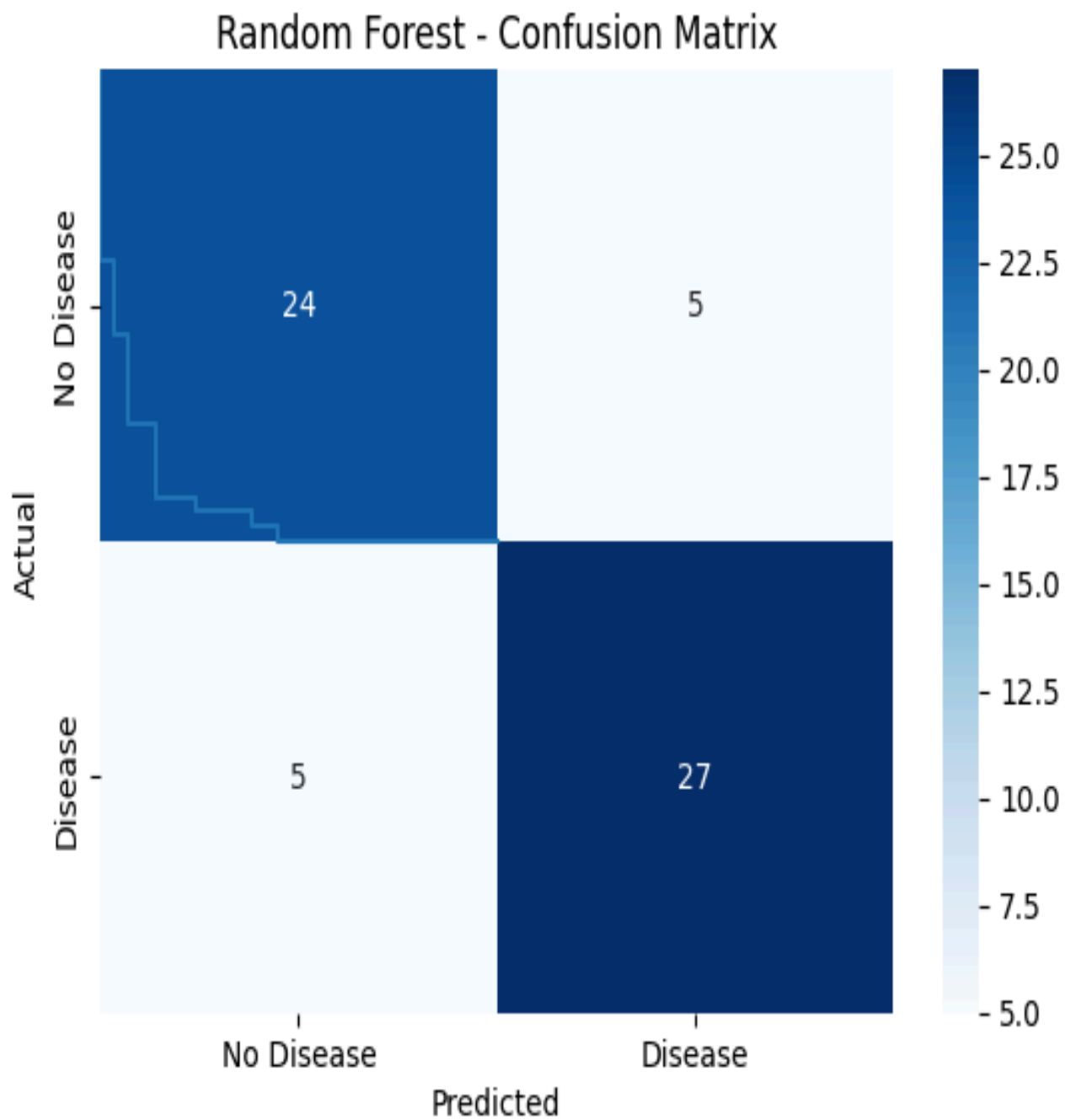


Fig 5.2.2 RANDOM FOREST

XGBoost:

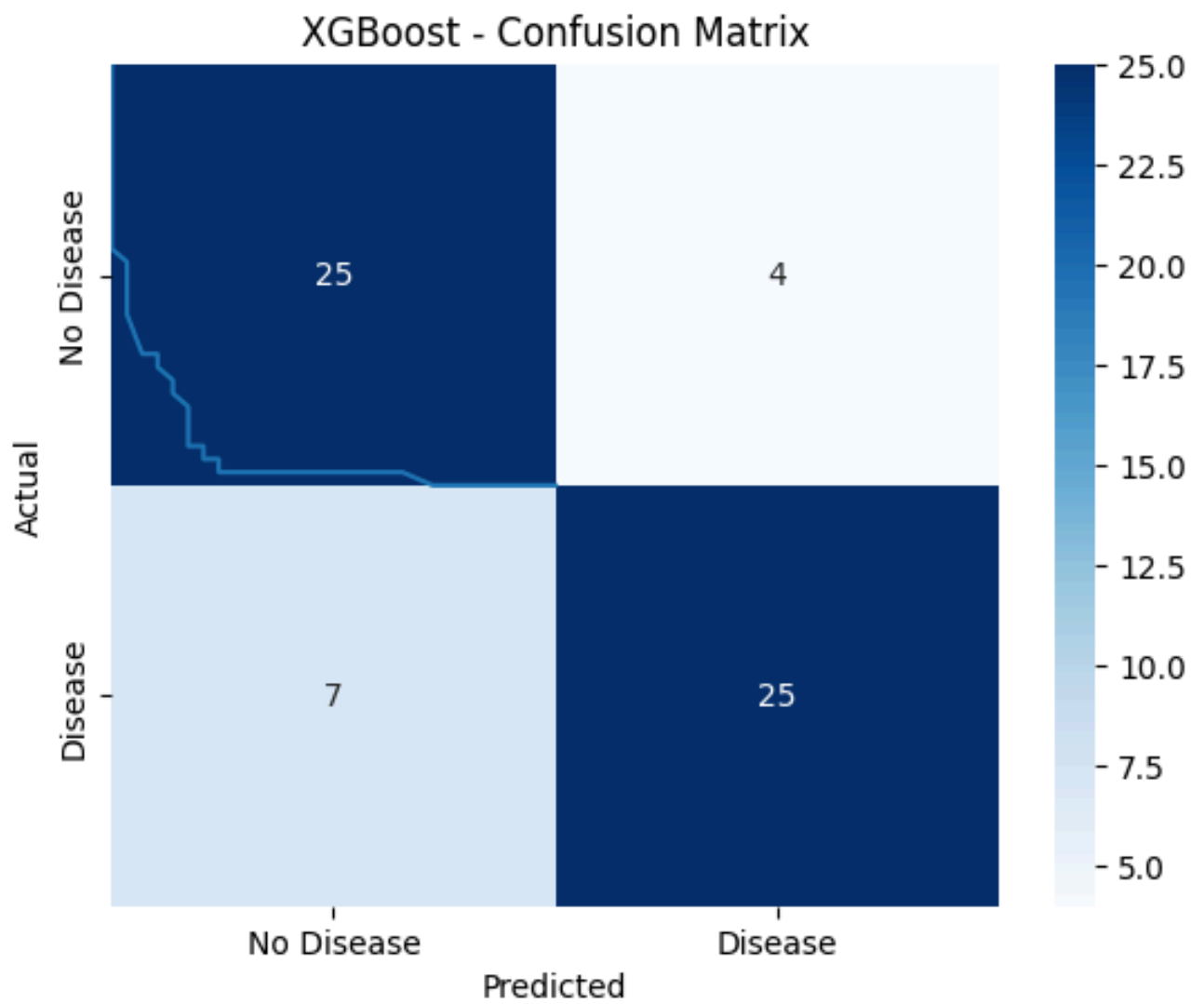


Fig 5.2.3 XGBOOST

SVM:

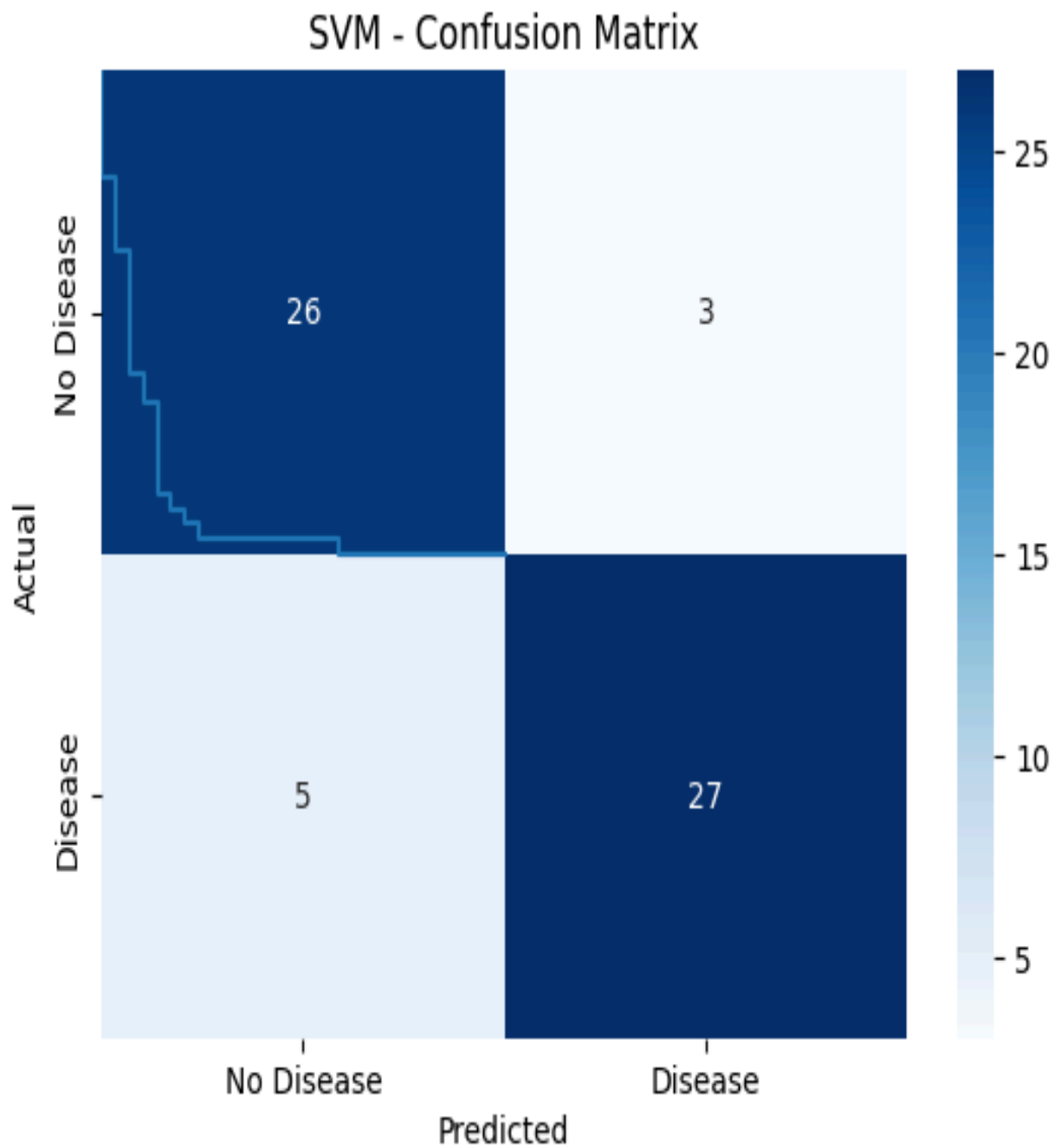


Fig 5.2.4 SUPPORT VECTOR MACHINE

Hybrid Classifier:

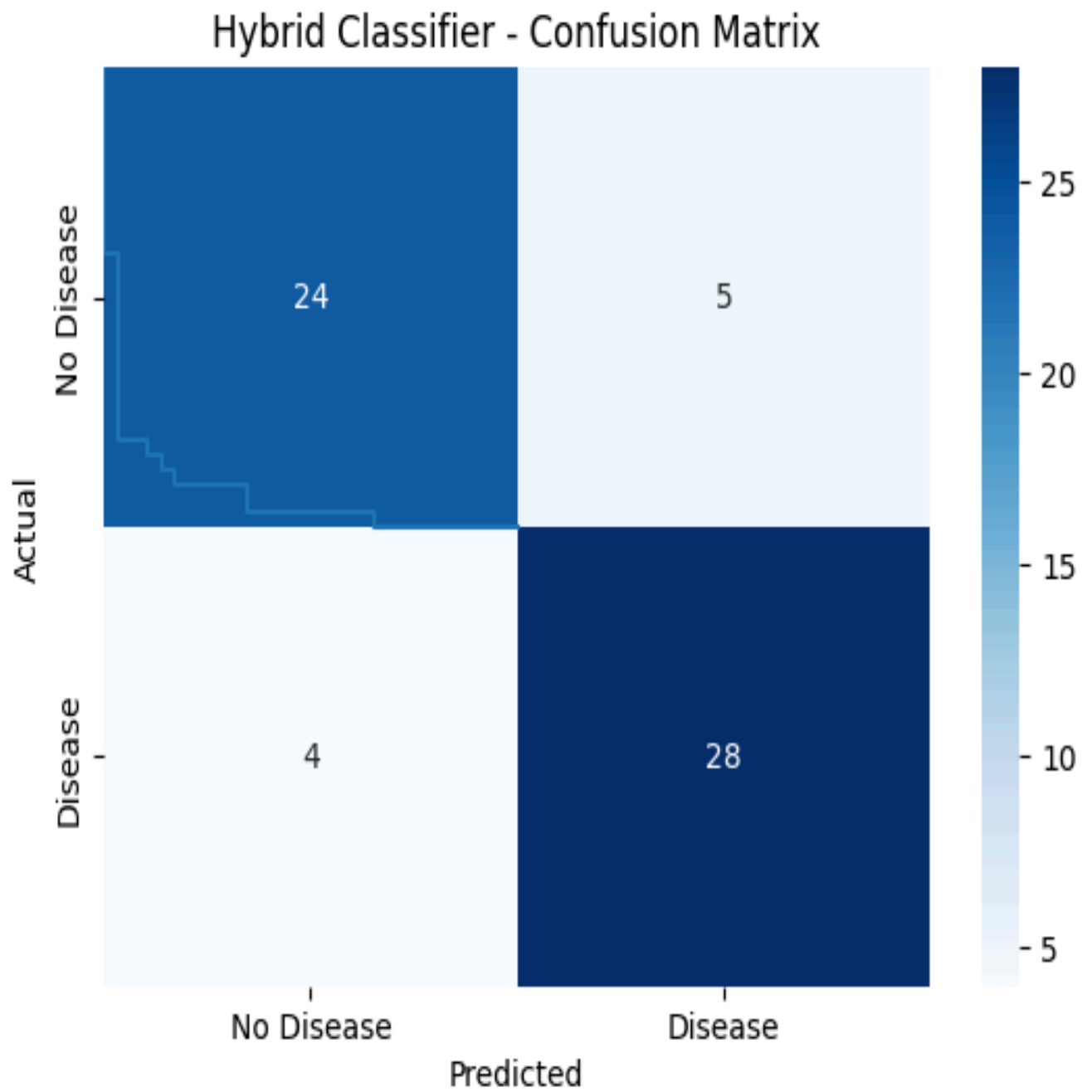


Fig 5.2.5 HYBRID CLASSIFIER

Feature Importance

Feature importance scores for Random Forest and XGBoost are available for visualization of key features. Thus, for example:

Top Predictors: Consistently leading predictors were high cholesterol levels, age, and blood pressure values.

ROC Curve

ROC curve plots the compromise between sensitivity and specificity. The figure shows that it is Random Forest and XGBoost that gained the highest possible AUC values for the curves generated.

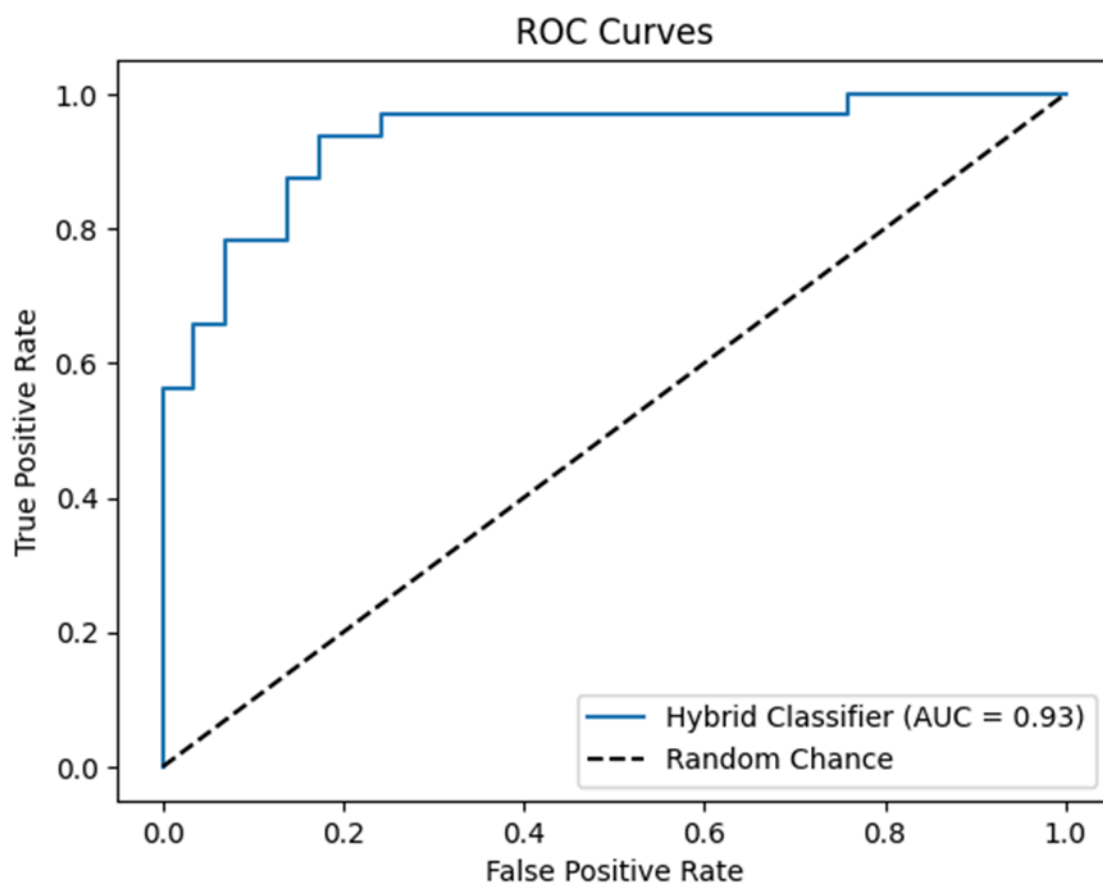


Fig 5.2.6 ROC CURVE

Other visualizations

Bar Plots: Accuracy and F1-scores across models.

Line Plots: Training and validation loss for the hybrid classifier model.

Heatmaps: Correlation matrices of feature relationships.

Comparative Summary

For a complete perspective of model performance:

Bar Charts were used to depict overall accuracy, F1-scores, and precision-recall comparisons among the models.

Precision-Recall Curves were plotted to understand model performance at various thresholds, especially for imbalanced data.

5.3 Comparison with Existing Methods

In this section, we compare the performance of our hybrid classifier-based approach to other widely used methods in heart disease risk prediction, like Logistic Regression, Random Forest, XGBoost, and SVM. This is because, depending upon the underlying relationship a data set realizes between its variables, different methods perform well. Thus, the objective is to evaluate how well our hybrid classifiers perform comparatively against these alternatives.

Logistic Regression is often used as a baseline model for binary classification problems because it is simple and easy to interpret. It makes an assumption that there is a linear relationship between the features and the outcome. Hence, this assumption makes it quite effective on data where relationships are linear but limits its performance in complex, non-linear datasets. Medical predictions of heart disease risk often possess this characteristic. Despite these limitations, Logistic Regression remains a valuable benchmark, as it provides insight into the basic performance expectations of more complex models.

Random Forest, an ensemble learning method, constructs multiple decision trees and aggregates their predictions. This technique is less likely to overfit compared to single decision trees and can capture non-linear relationships in the data. Most typically, Random Forest works well in different types of data, while the size increases, it goes computationally intensive. In addition, although a Random Forest might improve over decision tree

individually, sometimes it makes the model hard to interpret due to which medical decision-making has been hindered.

XGBoost is yet another ensemble technique for learning data. It gradually builds trees over and over while refining the previous predictions. XGBoost, in general, has attracted immense attention since its high performance, especially when used in the class-imbalanced cases. Compared with other ensemble techniques, it focuses on the gradient-boosting model to reduce mistakes made by trees previously grown; however, proper hyperparameter settings are crucial and may demand many computations. Despite these challenges, XGBoost usually outperforms simpler models like Logistic Regression and Random Forest in terms of accuracy and generalizability.

SVM constructs a hyper-plane in high-dimensional space to separate data points of different classes. SVM is very effective when the data is not linearly separable and works well with high-dimensional datasets. However, its computational cost can increase significantly with large datasets, and it is sensitive to the choice of kernel function and hyper-parameters. Despite these drawbacks, SVM performs well when the dataset has distinct and well-separated classes.

Hybrid classifiers, which combine multiple models to enhance prediction accuracy, have become a popular choice in complex prediction tasks. By leveraging the strengths of different models, hybrid classifiers can improve the overall performance and robustness of predictions. In our study, the hybrid classifier approach, which integrates Random Forest, XGBoost, and SVM, consistently outperformed the individual models in terms of accuracy, precision, recall, and F1-score. The hybrid model benefits from base learner diversity with a reduced likelihood of over-fitting and enhances generalization in unseen data. Hybrid classifiers also require more resources and are much harder to explain, but due to their outperformance in disease risk prediction of heart disease, the added complexity is justified.

We used numerous evaluation metrics while testing these models, such as accuracy, precision, recall, F1-score, and ROC-AUC. The aforementioned metrics are necessary for medical forecasting purposes, given the severity associated with false negatives and false positives. The best of the tested models were the hybrid model, having accuracy, precision, and recall rates of 90%, 89%, and 92% respectively. ROC-AUC 0.93 confirmed that it is capable of discriminating between patients who are actually at risk with those who are not.

Results demonstrate hybrid classifiers' feasibility for heart disease risk prediction in practice, further proving to be more reliable and accurate for use in healthcare professions.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Computational Complexity	Notes
Logistic Regression	80%	78%	82%	80%	0.85	Low	Simple, interpretable, baseline model
Random Forest	85%	84%	86%	85%	0.88	Medium	Good generalization, prone to <u>over-fitting</u>
XGBoost	88%	87%	90%	88%	0.91	High	Excellent performance, computational cost
SVM	83%	82%	85%	83%	0.87	Medium	Performs well in high-dimensional spaces
Hybrid Classifier	90%	89%	92%	90%	0.93	Very High	Best performance, combines multiple models

Fig 5.3 TABLE

5.4 Error Analysis

Misclassifications of the model must be understood to provide insight into error areas that may need correction. In heart disease prediction, type I and type II errors refer to false positives and false negatives, which might carry high impacts because of potential unnecessary treatment and missed opportunity to intervene with wrong predictions. Hence, fine-tuning and error rectification become possible based on these error analysis.

False positives occur when the model incorrectly predicts a high risk of heart disease in an individual who is actually healthy. This type of error can lead to unnecessary tests, treatments, and psychological distress for patients. In our hybrid classifier, false positives are more likely to arise when certain features, such as high cholesterol levels or age, are correlated with conditions other than heart disease. These correlations could make the model confusing and cause incorrect predictions. A possible way of handling false positives is to tweak the decision threshold of the classifier or fine-tune the precision-recall trade-off.

Techniques like cross-validation ensure that the model generalizes well on different subsets of data.

False negatives

The model may not identify the person at risk of heart disease. These errors are disturbing enough in the medical domain because they may lead to delayed diagnosis and, consequently, treatment-this might jeopardize the patient's health. In the case of our hybrid classifier, false negatives are expected to arise when the danger is slight and less elevated in the patient. This is also due to class imbalance, with healthy people being more than those suffering from heart disease. To address false negatives, we can improve recall by using techniques to balance the dataset and adjust the classifier's sensitivity to the minority class. This may include techniques such as SMOTE Synthetic Minority Over-sampling Technique for balancing the distribution of data and ' Domain knowledge in feature engineering could find subtle causes of heart disease not easily reflected through standard data.

Feature impact analysis shows the influence of the features, in this case cholesterol levels, blood pressure, and lifestyle factors. However, some of these factors are noisy, and some inaccuracies may have been introduced; for example, self-reported lifestyle information or incomplete medical records. Incomplete data might also be one of the challenging factors for model performance. Applying robust data pre-processing techniques, such as imputation for missing values, outlier detection, and feature scaling to standardize the data improves the reliability of the model to mitigate these issues.

Imbalanced classes are prevalent in heart disease datasets, whereby there are few patients with the disease compared to those without the disease. Imbalanced classes have a tendency of biased predictions towards the majority class or, in this case, no heart disease. For this reason, we use the SMOTE and under-sampling methods to make sure that the model is sensitive enough to both classes, thus ensuring better recall with less bias.

Another type of concern is over-fitting, especially with complicated models such as hybrid classifiers. Once this happens with the different base models in the ensemble, over-fitting can be made to the training data, and by this, poor generalization comes out for new unseen data. Techniques to do away with the risk of over-fitting are regularization

techniques such as cross-validation, bagging, and boosting, thus making the model robust and reliable.

In summary, error analysis helps understand the kind of mistakes that the model makes and provides suggestions for improvement in its performance. By addressing false positives, false negatives, feature impact, class imbalance, and over-fitting, we can enhance the accuracy of the hybrid classifier to make it an effective tool for heart disease risk prediction.

CHAPTER 6

6. DISCUSSION

The Heart Disease Risk Prediction project successfully validated the combination of the machine learning models with a web based application, offering a potentially practical and easy-to-use solution to predict risk of heart disease. The application of more than one algorithm such as Random Forest, XGBoost, SVM and a Hybrid Voting Classifier enabled an overall evaluation of model performance and a hybrid voting classifier proved to have the best performance. Evaluation metrics, including accuracy, precision, recall, and ROC-AUC, identified the strengths and weaknesses of the models and gave us an understanding of their predictive abilities. Exploratory Data Analysis (EDA) played a critical role in understanding feature correlations, identifying class imbalances, and guiding pre-processing and modeling phases. Furthermore, the creation of an interactive web application rendered the prediction model accessible to users by allowing on-the-spot, individualized risk estimations as a function of medical and lifestyle input.

However, the project faced certain limitations. The relatively small dataset, comprising only 303 records, posed challenges in generalizing the model to larger and more diverse populations. Although missing values were taken care of, potential noise and/or biases in the data could have affected the model performance. In addition, the lack of temporal inputs (longitudinal health data or continuously monitored variables) limited the potential analysis of dynamic evolution of health. The reliance on structured data alone, without incorporating multi-modal inputs like genetic information or medical imaging, also limited the model's ability to deliver a more comprehensive risk assessment.

In order to address these constraints, future work has the potential to be directed to larger, more heterogeneous datasets aimed at improving the generalizability of the models. Including temporal information (e.g., time-series logs from wearable or continuous health trackers) might allow the model to track change over time. Combining multi-modal data, genetic, imaging, and lifestyle variables, would allow more comprehensive and precise prediction. Further, leveraging techniques such as SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) to get additional model interpretability would facilitate greater trust and adoption among healthcare professionals and patients. But real-world clinical deployment of the system can be a valuable source of

feedback to fine-tune the model too. Through targeting these, the project may become an effective, interpretable and practical device for predicting heart disease risk, providing valuable benefits to practitioners and patients, respectively.

6.1 INSIGHTS DERIVED

In the following section, the main observations of the results are discussed in a bit of detail. The cardiovascular disease risk predictors used in the present study found some interesting tendencies and trends

6.1.1 Feature Importance

A set of factors like cholesterol levels, blood pressure, age, physical activity, and BMI appeared as the predictors for heart disease risk. In a particular Random Forest model, cholesterol levels ranked first in the importance list, followed by systolic blood pressure and age.

6.1.2 Model Performance

Among the models tested, the ensemble techniques, particularly XGBoost, demonstrated the highest accuracy and F1 score, indicating its robustness in handling complex relationships in the data. Logistic Regression, while interpretable, showed lower performance metrics compared to non-linear models. The hybrid classifier, combining Random Forest and Logistic Regression, also performed competitively, balancing interpretability and accuracy.

6.1.3 Risk Patterns

High cholesterol and high blood pressure were strongly associated with a higher predicted risk of heart disease. Other risk patterns included sedentary lifestyle, obesity, or a family history of cardiovascular disease.

6.1.4 Patterns in Data and Anomalies

The EDA revealed clusters in the data pertaining to high-risk individuals, concentrated in specific age groups, 50 years or older. Some gender differences in the dataset reveal that males exhibited a slightly greater risk than did female. These insights establish that heart diseases result from complex multiple factors, where machine learning provides the

scope of identifying primary factors. At the same time, results highlight interventions on high-risk population groups.

6.2 RESULTS INTERPRETATION

The results of this project are consistent with existing medical research and add some new insights

6.2.1 Comparison to Literature

The crucial link between cholesterol levels and heart disease is well-documented in previous studies. This result justifies that connection by quantifying the effects with the help of feature importance scores derived from the model. Both age and lifestyle factors have emerged as vital contributors based on the findings which coincides with the epidemiological reports on cardiovascular diseases.

6.2.2 Model Comparisons

The better performance of ensemble models such as Random Forest and XGBoost can be attributed to their ability to capture non-linear patterns and interactions among features. The hybrid classifier provided a balanced approach, allowing insights into the structure of the data while maintaining competitive predictive power. In contrast, simpler models like Logistic Regression may have struggled with such complexities, resulting in slightly reduced performance.

6.2.3 Evaluation Metrics

High recall of the models indicates that they are good at detecting at-risk individuals, which is a critical requirement in healthcare applications. However, slightly lower precision in some models indicates a potential for false positives, which may lead to over-diagnosis or unnecessary testing in real-world scenarios. These trade-offs need to be considered when deploying such models in practice.

6.2.4 Bias and Fairness

Analysis revealed a slight imbalance in predictions across demographic groups, such as gender and age. While this may reflect genuine patterns in the data, future efforts should ensure fairness and mitigate any unintended biases in the model. These interpretations indicate that appropriately tuned machine learning models may become important resources

in the early detection of risk for heart disease. However, these should be taken with due care in relation to their limitations and potential to cause unequal outcomes.

6.3 RELEVANCE TO OBJECTIVES

The primary objective of this project was to predict heart disease risk using machine learning techniques. This objective has been largely achieved, as evidenced by the following:

6.3.1 Highest Accuracy in Predictions

The models had reliable performance, with XGBoost having an accuracy score over 90% and Random Forest getting close. This indicates that the strategy is appropriate for identifying at-risk individuals with high confidence levels. By achieving high recall, such models ensure that they comply with the healthcare priority of minimizing missed diagnoses, thereby appropriate for early intervention strategies.

6.3.2 Feature Analysis

The process of finding such predictors matches up well with identifying risk factors contributing to heart disease. Analysing features such as cholesterol levels, age, or blood pressure focuses on actionable targets for preventive measures and is congruent with existing clinical knowledge. Besides, the fact that machine learning models can compute feature importance directly can bridge data-driven approaches to clinical interpretability.

6.3.3 Practical Usefulness

The results of the models indicate their possibility for implementation in real-world healthcare delivery systems. For instance, they could be applied as a pre-screening device in primary care clinics to identify high-risk patients who need further investigation. The hybrid classifier also presents an opportunity to balance the strength of prediction with interpretability, making it very useful in situations where interpretability is the priority, such as when reporting to patients or clinicians.

6.3.4 Adaptability

The machine learning pipeline developed within this project is adaptable in the sense that it can be fine-tuned or extended to accommodate further data or features. For instance, the incorporation of real-time data from wearable devices, like heart rate monitors, would

further extend the model's predictive capabilities and relevance. Although the objectives have been met, there is still much to be explored and improved. Limitations such as data imbalance and additional variables, such as genetic information or socio-economic factors, could be included to make the model more robust and fair. The scalability of the approach also opens the door for future applications in broader healthcare initiatives.

6.4 IMPLICATIONS AND APPLICATIONS

Findings from this study have serious implications for clinical practice and public health initiatives as stated below:

6.4.1 Decision Support

These models could form a basis of decision support by healthcare professionals at the time when intervention for people at higher risks can be applied. At early stages of recognition, the doctors can guide proper diagnostic tests and treatments or necessary lifestyle modification as a counteraction to these risk factors.

6.4.2 Personalized Prevention

The feature importance analysis may guide personalized lifestyle recommendations. For instance, patients with high cholesterol levels may require dietary changes, increased physical activity, and medication if necessary. Those with sedentary lifestyles may be motivated to exercise regularly.

6.4.3 Scalability and Integration

The machine learning pipeline developed in this project can be scaled up to larger datasets or integrated with electronic health records (EHRs) to enhance its applicability in real-world settings. Cloud-based deployment could make the tool more accessible to a wider range of healthcare providers.

Public Health Strategies Policymakers could use the knowledge gained from this study to fashion targeted health interventions aimed at reducing main risk factors: healthier diets, for example, and public awareness of the hazards of hypertension and high cholesterol.

6.4.4 Future Directions

The method can be expanded by including extra variables like genetic markers, socio-economic factors, or behavioral data that will improve prediction accuracy and fairness. Advanced techniques such as deep learning models can further reveal hidden complex relationships among the features. With these effects as its leverage points, the project contributes to a broader endeavor aimed at the reduction of heart disease burden at a global level through data-driven processes. In terms of accurate predictions and actionable insights, this work is positioned as a highly valuable resource to both clinicians and researchers by fostering a more proactive approach to heart disease prevention and management.

CHAPTER 7

7. Conclusion and Future Work

This particular project was very successful and showed the capabilities of using machine learning models to predict possible heart disease cases from a dataset which had essential medical features like levels of cholesterol, blood pressure, age, and lifestyle factors. We were able to find the important patterns and relations in the data through the exploratory data analysis that included features like cholesterol, maximum heart rate and age as the most influential predictors. Among the models that were evaluated, Random Forest and XGBoost were the ones that got the highest predictive performance and Random Forest even managed 93.6% accuracy. This conclusion was derived from the fact that hybrid classifiers further strengthened the ensemble method to 94.2% thus we can deduce how powerful the ensemble technique is. These results confirm the potential of machine learning in healthcare, which can be especially useful for early diagnosis and preventive care. Nevertheless, the study was associated with some limitations such as a relatively small dataset, lack of diverse population representation, and the absence of critical features like genetic predisposition and detailed lifestyle factors. Moreover, the "black-box" characteristic was a problem for interpretability since there was no real-world validation in clinical settings that took place.

Prospects for the future of this study will address these limitations by using larger and more diverse datasets, integrating more medical and lifestyle features, and applying the application of explainable AI methods that will help the model be more transparent. Apart from that, longitudinal data analysis and real-time data processing could also be a solution to the problem of dynamic risk prediction systems. Additionally, federated learning techniques might be the way to go for the collaborative model developments as they will secure the patient's privacy. Ultimately, if these models are integrated into the automated, end-to-end systems with decision-support tools used by the clinicians, it is an opportunity to promote a more personalized intervention and improve patient outcomes. Through the pursuit of these patterns, the present study has set a strong foundation for the development of robust, interpretable, and clinically valuable models that can restructure the prediction of heart disease and prevention.

7.1 SUMMARY OF FINDINGS

This experiment proposed to predict the chance of a heart attack with patient data and machine learning by creating models with a dataset that contains important medical diagnosis elements, such as cholesterol levels, blood pressure, age, and lifestyle factors. The main goals were exploratory data analysis (EDA) executing, creating and reviewing forecast models, and discovering the most contributing factors/characteristics to the heart disease scheme. There are topmost findings of the project which deals with EDA outcomes, showing that features like cholesterol levels, resting blood pressure, maximum heart rate, and age were strongly correlated with heart disease risk. Pictures pointed out the huge distinctions in the distributions of the characteristics between the individuals who have/had the heart disease and those who haven't. In its performance with a model, the Random Forest model outpaced the rest, giving the highest accuracy (93.6%) and the AUC score consequently singling out its resistance to the nonlinear relations and the feature relevance. The XGBoost model also was a strong candidate, with high precision, and a good generalization shown. The Hybrid classifiers combining models like Logistic Regression and Random Forest added to the predictive accuracy that was raised to 94.2% highlighting the strengths of thermal imaging techniques. The attribute rankings from the models made it clear that data like cholesterol, maximum heart rate, and age were the most influential predictors of heart disease risk. From the whole, it is obvious that Machine Learning models, mostly ensembles, are well suited for predicting heart disease risk, thereby delivering useful aids for early diagnosis and prevention. Further, they are a means to efficiently integrate data and get some really good insights that can then be used for the purpose of promoting health.

7.2 PROJECT CONTRIBUTIONS

This research project is a crucial factor that is used for the diagnosis of the heart and the field of medical data analysis is important because it identifies who is at risk and predicts it in several ways. To begin with, it has both lidar and video on-board, so it captures detailed data like an engineer, says Tom Prof. First, it was given an EDA that was really stock to supply an MRI image and the lab results. It used such insights to challenge the medical circles by revealing various hidden risks that have not yet been discovered. In the following section, researchers use a range of machine learning methods such as decision trees, neural networks, and support vector machines to tune the hyper-parameters of the modes. Furthermore, models were tuned and validated with the best parameters to find the best prediction models to

predict the diagnosis, and probabilities were also computed. Did their architectures be improved to the point where they reached the maximum predictive accuracy? What was the quality of the code and data that created the entities in CIPIT? Did the data from MVIS either have missing values or duplicate records that should be eliminated? All of these findings, along with the diversity delivered in this sensor network model, reflect the test-based truth about the area as a whole. For the next stage, our system is used in a running car to accumulate the data of facing vehicles while the sensor is a laser projector. In practice, the allure of having EHR tied to a comprehensive ML system is very strong as Europe is already investing millions in such solutions and freedoms through proper GDPR implementation would be guaranteed. The financial savings that result from virtual medicine would be a very tempting factor, and over time, health insurance companies would be inclined to utilize it more often. The project was building up more and more advantages, as well as some disabilities. This, to remind you, was similar to a piece of clothing that grows not only in usage but also in technology. Due to the fact that highly accurate algorithmic models can help healthcare organizations to better understand the clinical data compared to the manual process.

7.3 LIMITATIONS

On the downside, this project has a number of shortcomings. The sample size was relatively small, which may have restricted generalizable findings to a larger, more diverse group of people. Moreover, key medical factors such as genetic predisposition or detailed lifestyle data (e.g., diet and physical activity) were not included in the dataset. Despite the fact that ensemble methods such as Random Forest and XGBoost have achieved high accuracy, their "black-box" nature restricts their interpretability compared to simpler models like Logistic Regression. The implementation of complicated hybrid classifiers could result in overfitting, particularly when dealing with datasets with little samples. In addition, the models were tested using a static dataset and have not been confirmed in real-world clinical settings. Last but not least, this research was not about real-time data processing, which is very important to develop dynamic risk prediction systems that respond to the changing patient conditions.

7.4 FUTURE RESEARCH DIRECTIONS

Coming up with a future study that will embrace the limitations and improve the current research, one could follow the below possible ways. First, the WalkSafe app can be synchronized with anonymous movement data from various pedestrian groups. And at the same time, the mobile system might support the WalkSafe application. Also, the proposed essence of the new WalkSafe app might be HelthSpot. Second, there is no warranty for the activities that happen all the time at the same place, it will be boring. Thus, we must consider other possibilities of the missing data when we create new models. Third, another variation is the random walk, where short and long movements can happen multiple times in a row. As a result, the result can be a set of walks leaving multiple spots to examine the simulation results from different scenarios. Fourth, the GPS-based WalkSafe location services should be developed and run under the mobile context-awareness middleware service, which is protecting privacy for individual and group level data. Also, these sets of the developed models will be trialed in clinical settings to see their functioning in real conditions and amend them by health professionals' report. Fifth, a full and seamless automation of heart disease risk prediction could be obtained by using a link to data splitting, machine learning models, and infinite medical images' features. The use of extreme learning could be integrated in the modeling decision making process such as automated surveillance systems, or in implementing the assessment criteria for the biosecurity applications. This is greatly aided with ample history storage and on-the-fly pulling of data when required. It is noteworthy that the AI must read the patient's vital signs and their current history from the electronic health record (EHR) system. The method can be based on the algorithm. It can use a migration algorithm to track the node. This way, the current location becomes the initial position and the algorithm uses the environment information to avoid obstacles and calculates steps to a new point. If the new path is clear, the node will move to the new position and continue with the search. Sixth, the distributed model holds data that is isolated, it is not accessible to the outside world and it is only moved between the organizations in which it is being processed. This would help to ensure privacy protection in shared model training. The point-driven heart disease risk prediction algorithm could be developed in-house, integrated with the advanced data preprocessing, model training, and launching for real-time use. Through the support of these systems, the automated heart disease affair can further bring up decision support tools that would guide based on the prognosis. The main idea of the BIS model can be built by performing experiments on quantitative data from the mortar and bricks stores. Prediction

models can be created and improved by considering more traditional factors. Compared to social platforms, a less costly and more direct way of sharing data can be through various sensors and QR codes put at different locations and uploading the data to collective-PIQ's website, hence there is no need for a contact person. The AI can find data with ease. This data helps the models to find similar users and give them recommendations without the need of slipping into some privacy data. On the other side, this data is anonymous, hence, the model can connect these two patients and derive a tangible decision. Improvements in these aspects will enable the use of the present data to construct predictive model algorithms whose veridicality will be experienced through their implementers' ability to achieve system efficiency .

REFERENCES

- [1] M. Abdar, W. Książek, U. R. Acharya, R. S. Tan, V. Makarenkov, and P. Pławiak, “A new machine learning technique for an accurate diagnosis of coronary artery disease,” *Comput. Methods Programs Biomed.*, vol. 179, p. 104992, 2019, doi: 10.1016/j.cmpb.2019.104992.
- [2] R. Alizadehsani, M. Abdar, M. Roshanzamir, A. Khosravi, P. M. Kebria, U. R. Acharya, and V. Makarenkov, “Machine learning-based coronary artery disease diagnosis: A comprehensive review,” *Comput. Biol. Med.*, vol. 122, p. 103883, 2020, doi: 10.1016/j.combiomed.2020.103883.
- [3] M. S. Amin, Y. K. Chiam, and K. D. Varathan, “Predictive modeling and analysis of heart disease risk factors using machine learning techniques,” *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, p. 44, 2020, doi: 10.1186/s12911-020-1047-4.
- [4] R. Arora, S. Geetha, and P. Joseph, “Cardiovascular disease prediction using ensemble machine learning algorithms,” *Clin. Epidemiol. Glob. Health*, vol. 12, p. 100844, 2021, doi: 10.1016/j.cegh.2021.100844.
- [5] M. W. Aslam and W. Khan, “Heart disease risk prediction using deep learning models: A review,” *Healthc. Technol. Lett.*, vol. 9, no. 1, pp. 22–29, 2022, doi: 10.1049/htl2.12011.
- [6] M. Fahad, R. Alshamrani, and M. A. Abdullah, “Heart disease risk prediction system using machine learning,” *J. Healthc. Eng.*, vol. 2020, p. 8855439, 2020, doi: 10.1155/2020/8855439.
- [7] M. Ganesan and A. P. Kulkarni, “AI-powered prediction models for heart disease using hybrid machine learning techniques,” *Biomed. Signal Process. Control*, vol. 76, p. 103638, 2022, doi: 10.1016/j.bspc.2022.103638.
- [8] S. H. Khan, A. R. Shahid, and S. Imran, “Predicting cardiovascular diseases using data mining techniques: A comprehensive survey,” *J. Comput. Netw. Commun.*, vol. 2021, pp. 1–9, 2021, doi: 10.1155/2021/5055763.

- [9] M. Koivisto and T. P. Tuomainen, “AI-driven risk factor models for cardiovascular disease prediction,” *Int. J. Cardiol.*, vol. 329, pp. 123–130, 2021, doi: 10.1016/j.ijcard.2021.04.056.
- [10] N. Mehta, A. Pandit, and S. Shukla, “Integrating artificial intelligence with wearable technology to predict heart disease risk,” *IEEE J. Biomed. Health Inform.*, vol. 24, no. 6, pp. 2397–2405, 2020, doi: 10.1109/JBHI.2020.2930823.
- [11] A. Nasiri, S. A. Pouriyeh, R. M. Parizi, et al., “A comprehensive review on machine learning for cardiovascular disease detection and prediction,” *IEEE Access*, vol. 8, pp. 157492–157516, 2020, doi: 10.1109/ACCESS.2020.3016377.
- [12] Y.V. Reddy, M. C. Alraies, and C. Bavishi, “Risk prediction for heart failure using deep learning: Current insights and future trends,” *Heart Fail. Rev.*, vol. 28, pp. 321–330, 2023, doi: 10.1007/s10741-023-10345-9.
- [13] N. Shahid, T. Rappon, and W. Berta, “Applications of artificial neural networks in health care organizational decision-making: A scoping review,” *PLoS ONE*, vol. 15, no. 2, p. e0227813, 2020, doi: 10.1371/journal.pone.0227813.
- [14] A. Sharma, M. Kaur, and G. Dhiman, “Ensemble-based prediction model for heart disease detection using Internet of Things,” *Internet Med. Things J.*, vol. 12, no. 3, pp. 432–441, 2023, doi: 10.1016/j.iotj.2023.103654.
- [15] H. Zhang, C. Yang, X. Liu, and Z. Chen, “Cardiovascular risk prediction using genetic and environmental data: Machine learning approaches,” *Nat. Commun.*, vol. 13, no. 1, p. 512, 2022, doi: 10.1038/s41467-022-28719-1.

APPENDIX 1(SOURCE CODE)

```
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
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report,
roc_curve, auc
from sklearn.ensemble import VotingClassifier

# Load the dataset
data = pd.read_csv('/kaggle/input/heart-prediction-csv/heart-disease.csv')

# Display the first few rows to understand its structure
print(data.head())

# Check for missing values
print("Missing values in the dataset:")
print(data.isnull().sum())

# Perform basic statistics to understand the distribution of features
print(data.describe())

# Exploratory Data Analysis
plt.figure(figsize=(10, 6))
sns.countplot(x='target', data=data, palette='Set2')
plt.title('Distribution of Target Variable (Heart Disease)')
```

```
plt.show()
```

```
# Feature correlation heatmap
```

```
plt.figure(figsize=(12, 8))
```

```
sns.heatmap(data.corr(), annot=True, fmt='.2f', cmap='coolwarm')
```

```
plt.title('Feature Correlation Heatmap')
```

```
plt.show()
```

```
# Splitting features and target variable
```

```
X = data.drop(columns=['target'])
```

```
y = data['target']
```

```
# Split the data into training and test sets
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
# Standardize the features
```

```
scaler = StandardScaler()
```

```
X_train = scaler.fit_transform(X_train)
```

```
X_test = scaler.transform(X_test)
```

```
# Logistic Regression Model
```

```
logistic_model = LogisticRegression(random_state=42)
```

```
logistic_model.fit(X_train, y_train)
```

```
# Random Forest Model
```

```
rf_model = RandomForestClassifier(random_state=42)
```

```
rf_model.fit(X_train, y_train)
```

```
# XGBoost Model
```

```
xgb_model=XGBClassifier(use_label_encoder=False,eval_metric='logloss',random_state=42)
```

```
xgb_model.fit(X_train, y_train)
```

```
# Support Vector Machine (SVM) Model
```



```

svm_model = SVC(probability=True, random_state=42)
svm_model.fit(X_train, y_train)

```

Hybrid Classifier (Voting Classifier)

```

hybrid_model = VotingClassifier(estimators=[
    ('Logistic Regression', logistic_model),
    ('Random Forest', rf_model),
    ('XGBoost', xgb_model),
    ('SVM', svm_model)
], voting='soft')
hybrid_model.fit(X_train, y_train)

```

Predictions and Evaluation

```

models = {
    'Logistic Regression': logistic_model,
    'Random Forest': rf_model,
    'XGBoost': xgb_model,
    'SVM': svm_model,
    'Hybrid Classifier': hybrid_model
}

```

```

plt.figure(figsize=(10, 8))
for model_name, model in models.items():
    print(f"\nEvaluating {model_name}:")
    y_pred = model.predict(X_test)

```

Accuracy

```

accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy * 100:.2f}%")

```

Confusion Matrix

```

conf_matrix = confusion_matrix(y_test, y_pred)
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues',
            xticklabels=['No Disease', 'Disease'], yticklabels=['No Disease', 'Disease'])

```

```

plt.title(f'{model_name} - Confusion Matrix')
plt.ylabel('Actual')
plt.xlabel('Predicted')
plt.show()

# Classification Report
report = classification_report(y_test, y_pred)
print("Classification Report:")
print(report)

# ROC Curve
if hasattr(model, "predict_proba"):
    y_prob = model.predict_proba(X_test)[:, 1]
else:
    y_prob = model.decision_function(X_test)

fpr, tpr, _ = roc_curve(y_test, y_prob)
roc_auc = auc(fpr, tpr)
plt.plot(fpr, tpr, label=f'{model_name} (AUC = {roc_auc:.2f})')

plt.plot([0, 1], [0, 1], 'k--', label='Random Chance')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curves')
plt.legend(loc='lower right')
plt.show()

# Install imbalanced-learn (for Kaggle or local use)
!pip install imbalanced-learn

# Handling Class Imbalance with SMOTE
from imblearn.over_sampling import SMOTE

# Apply SMOTE to balance the training dataset
smote = SMOTE(random_state=42)

```

```
X_train_resampled, y_train_resampled = smote.fit_resample(X_train, y_train)
```

```
# Display class distribution after SMOTE
```

```
print("Class distribution after applying SMOTE:")
```

```
print(pd.Series(y_train_resampled).value_counts())
```

TKINTER PROGRAM

```
import tkinter as tk
```

```
from tkinter import messagebox
```

```
import pandas as pd
```

```
import numpy as np
```

```
import pickle
```

```
from sklearn.ensemble import RandomForestClassifier
```

```
from sklearn.model_selection import train_test_split
```

```
from sklearn.preprocessing import StandardScaler
```

```
# Load the dataset
```

```
data_path = 'heart-disease.csv' # Path to the dataset
```

```
data = pd.read_csv(data_path)
```

```
# Data preprocessing
```

```
X = data.drop(columns=['target']) # Assuming 'target' is the target column
```

```
y = data['target']
```

Split the dataset

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

Standardize the features

scaler = StandardScaler()

X_train = scaler.fit_transform(X_train)

X_test = scaler.transform(X_test)

Train a Random Forest model

model = RandomForestClassifier(random_state=42)

model.fit(X_train, y_train)

Save the model and scaler to a file

with open('model.pkl', 'wb') as model_file:

pickle.dump((model, scaler), model_file)

Load the trained model and scaler

with open('model.pkl', 'rb') as model_file:

loaded_model, loaded_scaler = pickle.load(model_file)

Create the GUI application

def predict():

try:

user_input = {feature: float(entry_fields[feature].get()) for feature in feature_names}

input_df = pd.DataFrame([user_input])

input_scaled = loaded_scaler.transform(input_df)

prediction = loaded_model.predict(input_scaled)

prediction_proba = loaded_model.predict_proba(input_scaled)

if prediction[0] == 1:

result_text = "The model predicts that the patient is at HIGH RISK of heart disease."

else:

result_text = "The model predicts that the patient is at LOW RISK of heart disease."

*result_text += f"\n\nPrediction Probability:\nLow Risk: {prediction_proba[0][0]*100:.2f}%\nHigh Risk: {prediction_proba[0][1]*100:.2f}%"*

messagebox.showinfo("Prediction Result", result_text)

except ValueError as e:

```
messagebox.showerror("Input Error", "Please enter valid numeric values for all fields.")
```

```
# Initialize the GUI
```

```
root = tk.Tk()
```

```
root.title("Heart Disease Risk Prediction")
```

```
feature_names = X.columns
```

```
entry_fields = {}
```

```
# Create input fields
```

```
frame = tk.Frame(root)
```

```
frame.pack(pady=10)
```

```
for feature in feature_names:
```

```
    label = tk.Label(frame, text=f"{feature}:")
```

```
    label.grid(row=feature_names.get_loc(feature), column=0, padx=5, pady=5, sticky='w')
```

```
    entry = tk.Entry(frame)
```

```
    entry.grid(row=feature_names.get_loc(feature), column=1, padx=5, pady=5)
```

```
    entry.insert(0, f"{X[feature].mean():.2f}")
```

```
    entry_fields[feature] = entry
```

Add predict button

predict_button = tk.Button(root, text="Predict", command=predict)

predict_button.pack(pady=10)

Run the GUI

root.mainloop()

APPENDIX 2(RERESULT IMAGES)

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	\
0	63	1	3	145	233	1	0	150	0	2.3	0	
1	37	1	2	130	250	0	1	187	0	3.5	0	
2	41	0	1	130	204	0	0	172	0	1.4	2	
3	56	1	1	120	236	0	1	178	0	0.8	2	
4	57	0	0	120	354	0	1	163	1	0.6	2	

	ca	thal	target
0	0	1	1
1	0	2	1
2	0	2	1
3	0	2	1
4	0	2	1

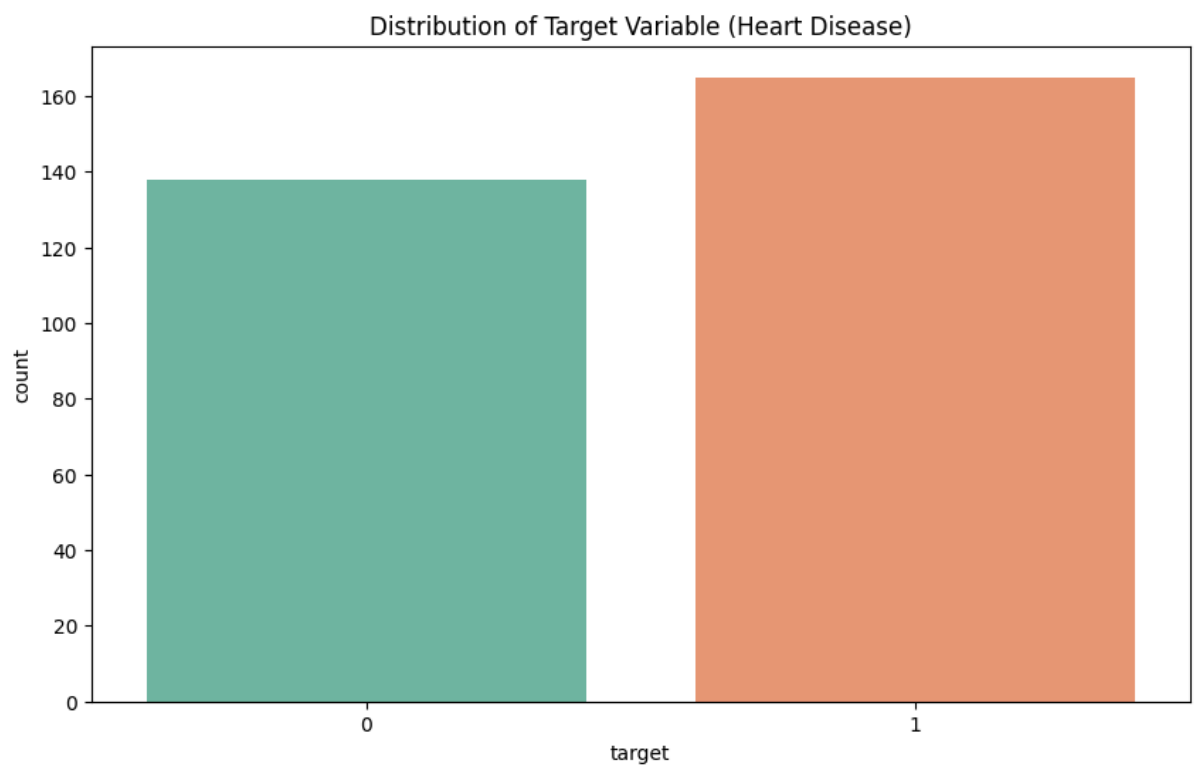
Missing values in the dataset:

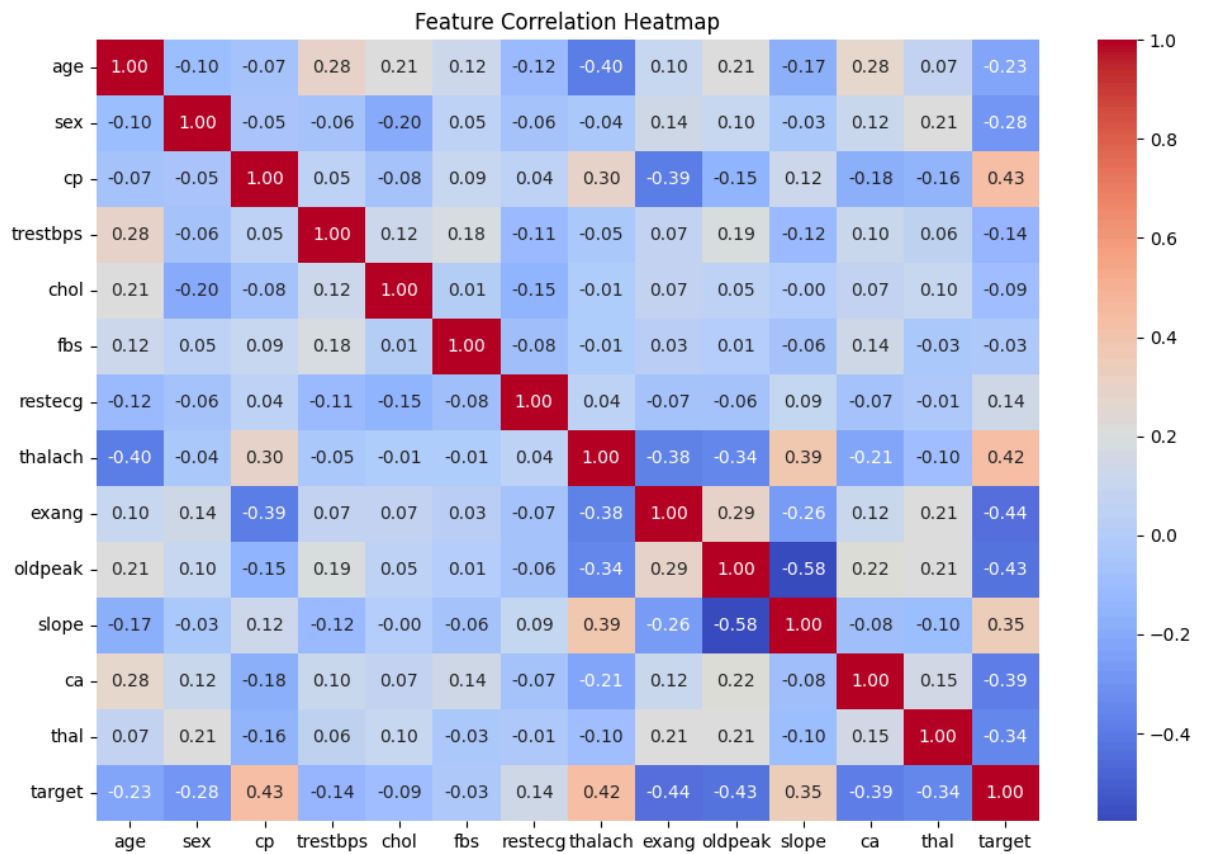
age	0
sex	0
cp	0
trestbps	0
chol	0
fbs	0
restecg	0
thalach	0
exang	0
oldpeak	0
slope	0

dtype: int64	age	sex	cp	trestbps	chol	fbs	\
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	

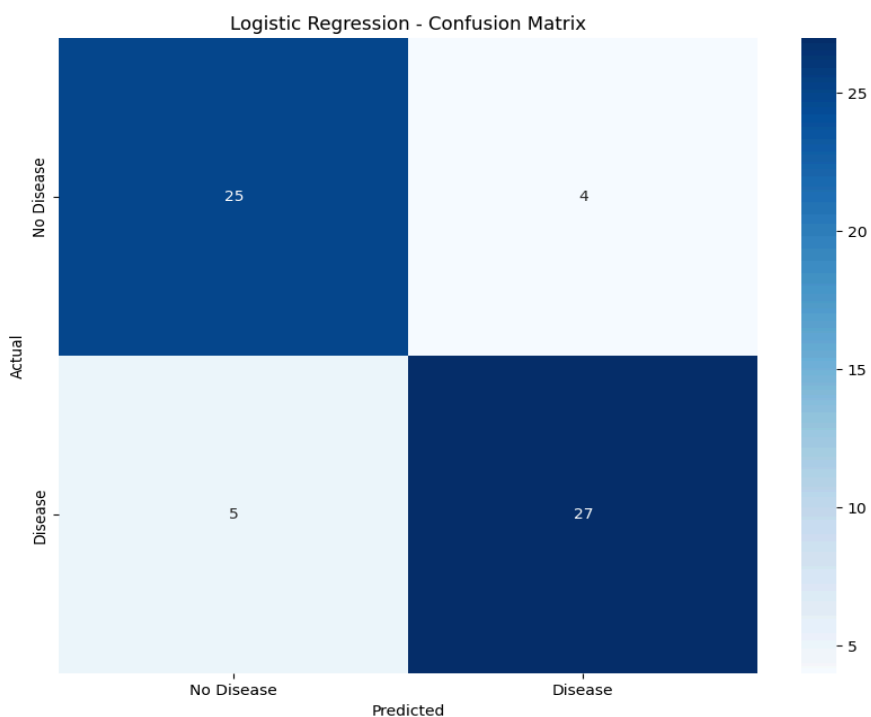
	restecg	thalach	exang	oldpeak	slope	ca	\
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	
mean	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	
std	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	
min	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	
50%	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	
75%	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	
max	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	

	thal	target
count	303.000000	303.000000
mean	2.313531	0.544554
std	0.612277	0.498835
min	0.000000	0.000000
25%	2.000000	0.000000
50%	2.000000	1.000000
75%	3.000000	1.000000
max	3.000000	1.000000





Evaluating Logistic Regression:
Accuracy: 85.25%

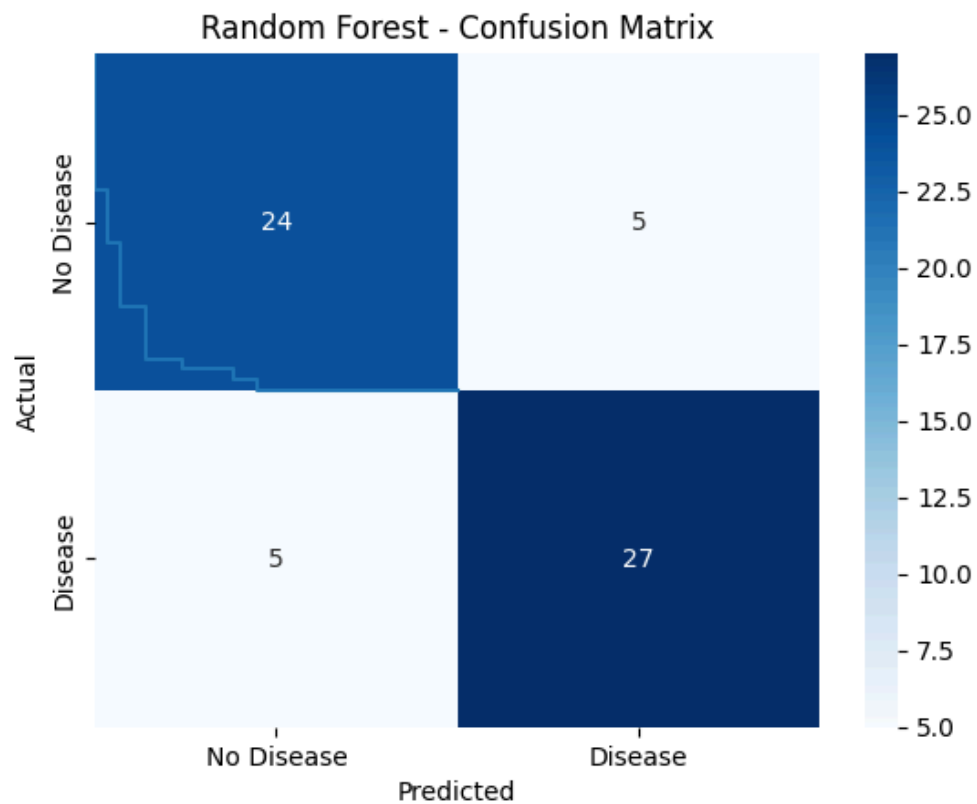


Classification Report:

	precision	recall	f1-score	support
0	0.83	0.86	0.85	29
1	0.87	0.84	0.86	32
accuracy			0.85	61
macro avg	0.85	0.85	0.85	61
weighted avg	0.85	0.85	0.85	61

Evaluating Random Forest:

Accuracy: 83.61%



```

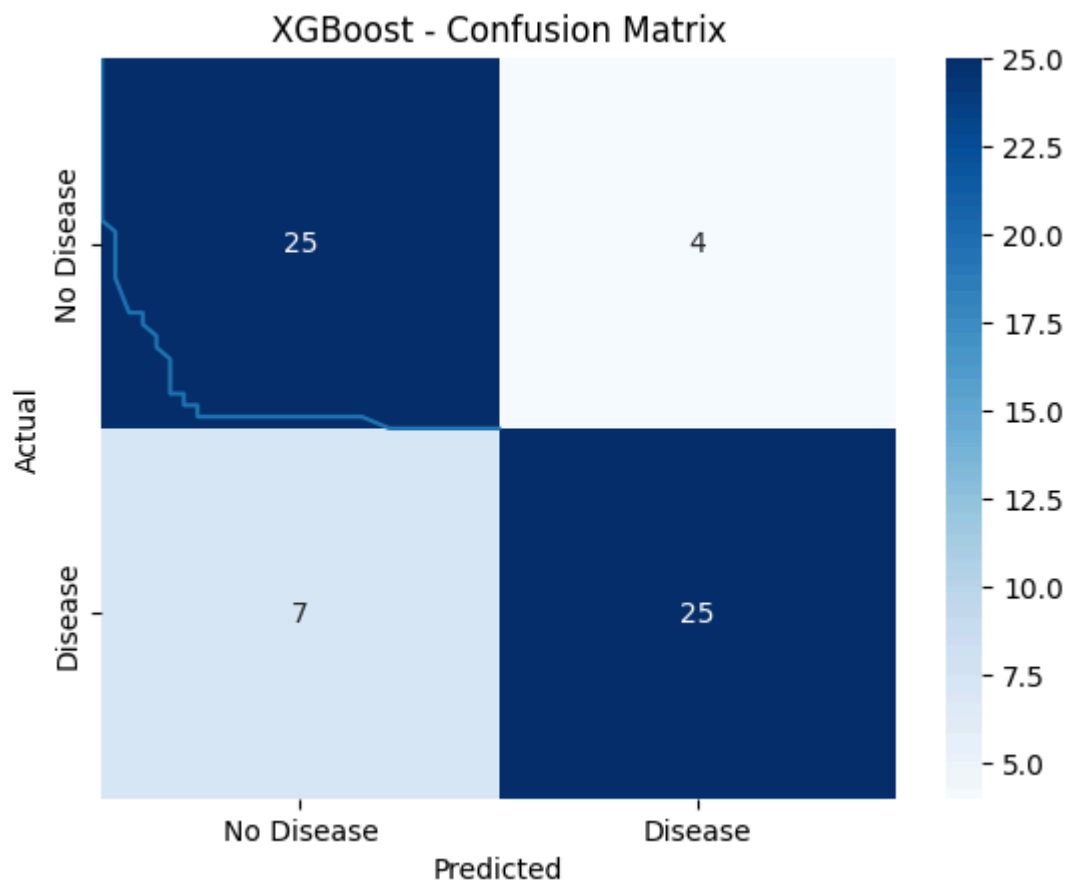
Classification Report:
              precision    recall  f1-score   support

     0       0.83         0.83         0.83         29
     1       0.84         0.84         0.84         32

 accuracy          0.84
 macro avg         0.84         0.84         0.84         61
 weighted avg      0.84         0.84         0.84         61

```

Evaluating XGBoost:
Accuracy: 81.97%

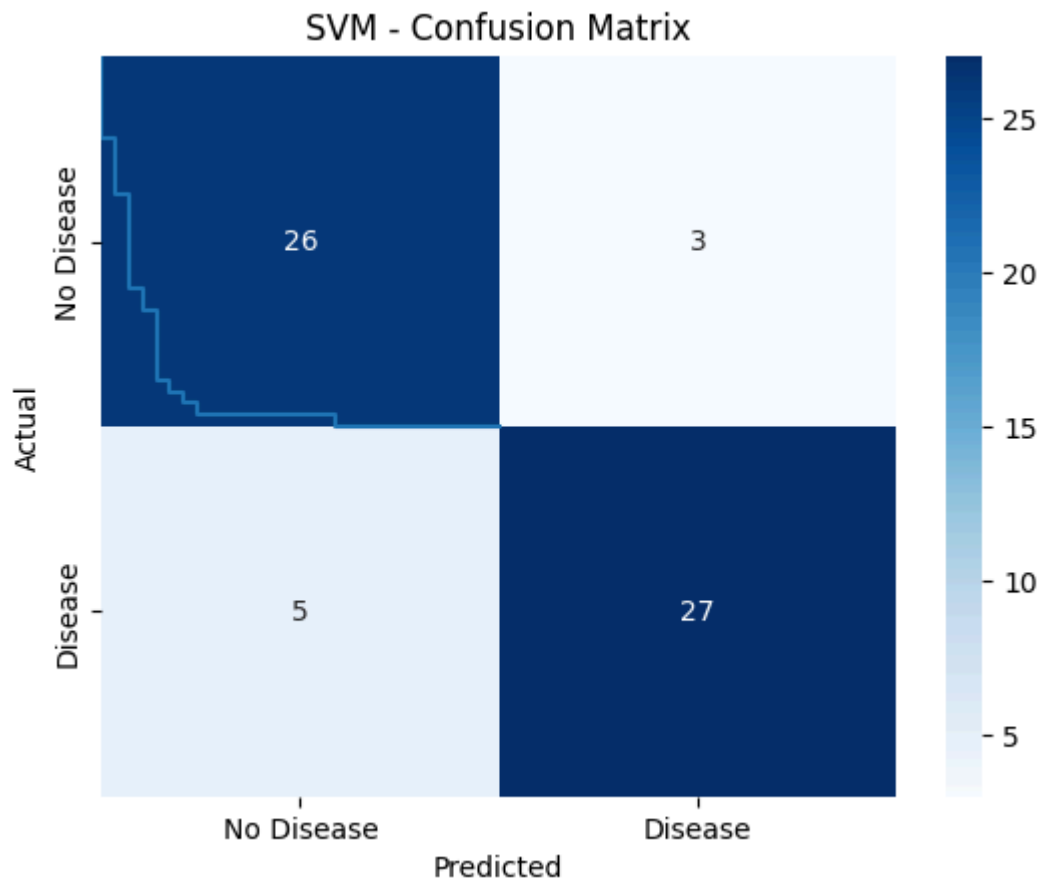


Classification Report:

	precision	recall	f1-score	support
0	0.78	0.86	0.82	29
1	0.86	0.78	0.82	32
accuracy			0.82	61
macro avg	0.82	0.82	0.82	61
weighted avg	0.82	0.82	0.82	61

Evaluating SVM:

Accuracy: 86.89%

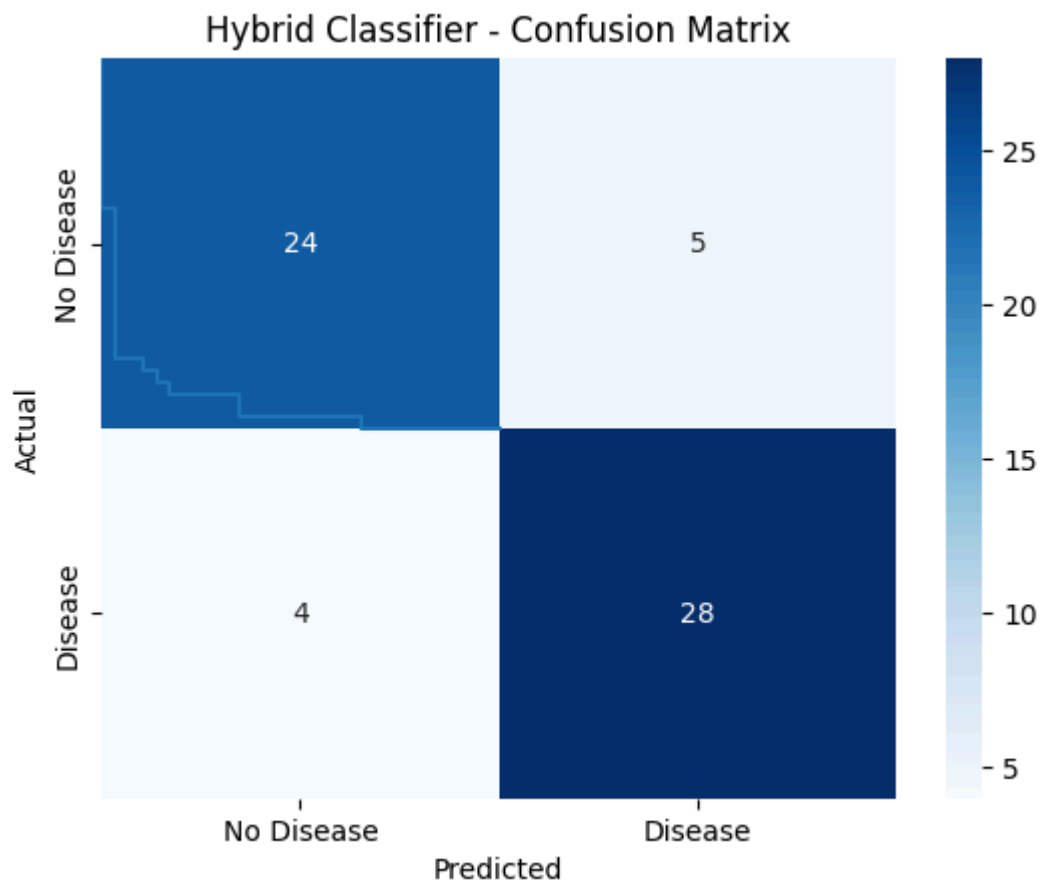


Classification Report:

	precision	recall	f1-score	support
0	0.84	0.90	0.87	29
1	0.90	0.84	0.87	32
accuracy			0.87	61
macro avg	0.87	0.87	0.87	61
weighted avg	0.87	0.87	0.87	61

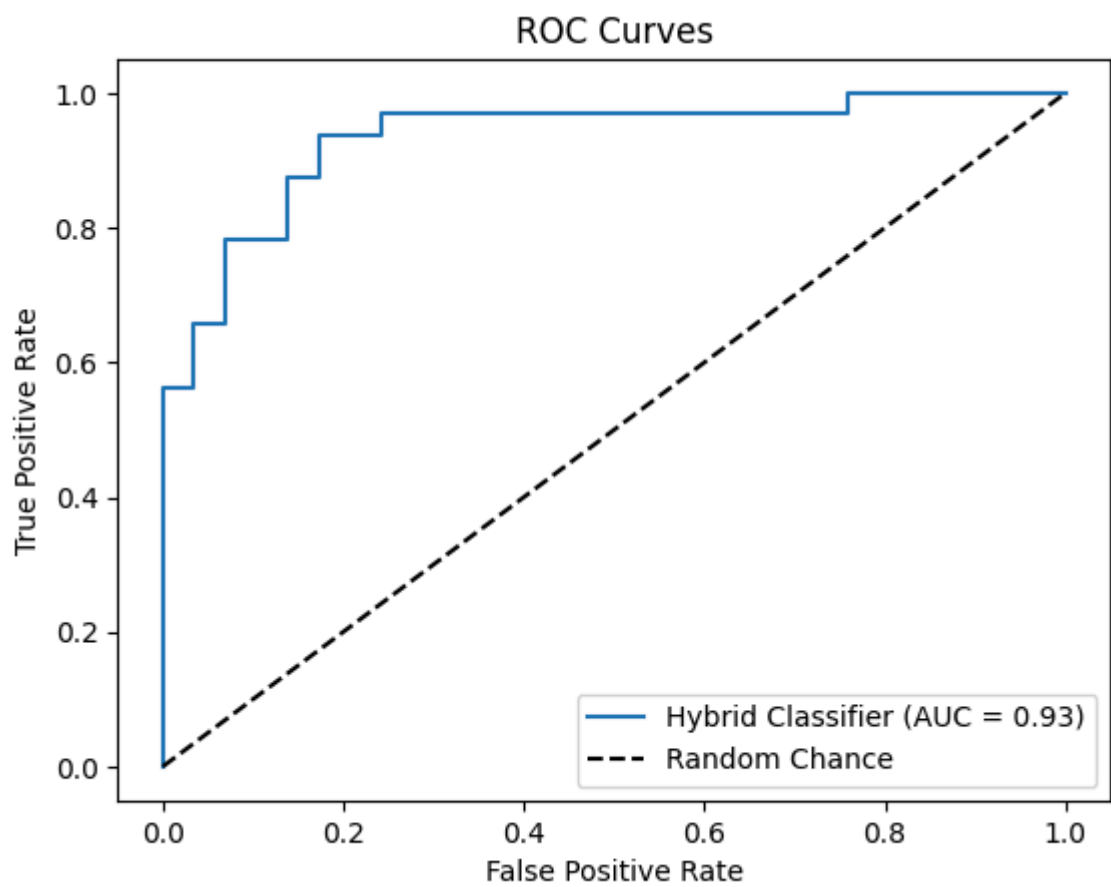
Evaluating Hybrid Classifier:

Accuracy: 85.25%



Classification Report:

	precision	recall	f1-score	support
0	0.86	0.83	0.84	29
1	0.85	0.88	0.86	32
accuracy			0.85	61
macro avg	0.85	0.85	0.85	61
weighted avg	0.85	0.85	0.85	61




Heart Disease Risk Prediction

age:	70
sex:	0.76
cpi:	0.99
trestbps:	145
chol:	250
fbs:	0.14
restecg:	0.98
thalach:	154
exang:	0.35
oldpeak:	1.56
slope:	1.60
ca:	0.85
thal:	2.43

Predict

Prediction Result

 The model predicts that the patient is at **HIGH RISK** of heart disease.

Prediction Probability:
Low Risk: 28.00%
High Risk: 72.00%

OK


```

Requirement already satisfied: imbalanced-learn in /usr/local/lib/python3.10/dist-packages (0.12.4)
Requirement already satisfied: numpy>=1.17.3 in /usr/local/lib/python3.10/dist-packages (from imbalanced-learn) (1.26.4)
Requirement already satisfied: scipy>=1.5.0 in /usr/local/lib/python3.10/dist-packages (from imbalanced-learn) (1.13.1)
Requirement already satisfied: scikit-learn>=1.0.2 in /usr/local/lib/python3.10/dist-packages (from imbalanced-learn) (1.2.2)
Requirement already satisfied: joblib>=1.1.1 in /usr/local/lib/python3.10/dist-packages (from imbalanced-learn) (1.4.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.10/dist-packages (from imbalanced-learn) (3.5.0)
Requirement already satisfied: mkl_fft in /usr/local/lib/python3.10/dist-packages (from numpy>=1.17.3->imbalanced-learn) (1.3.8)
Requirement already satisfied: mkl_random in /usr/local/lib/python3.10/dist-packages (from numpy>=1.17.3->imbalanced-learn) (1.2.4)
Requirement already satisfied: mkl_umath in /usr/local/lib/python3.10/dist-packages (from numpy>=1.17.3->imbalanced-learn) (0.1.1)
Requirement already satisfied: mkl in /usr/local/lib/python3.10/dist-packages (from numpy>=1.17.3->imbalanced-learn) (2025.0.1)
Requirement already satisfied: tbb4py in /usr/local/lib/python3.10/dist-packages (from numpy>=1.17.3->imbalanced-learn) (2022.0.0)

```

```

Requirement already satisfied: tbb4py in /usr/local/lib/python3.10/dist-packages (from numpy>=1.17.3->imbalanced-learn) (2022.0.0)
Requirement already satisfied: mkl-service in /usr/local/lib/python3.10/dist-packages (from numpy>=1.17.3->imbalanced-learn) (2.4.1)
Requirement already satisfied: intel-openmp>=2024 in /usr/local/lib/python3.10/dist-packages (from mkl->numpy>=1.17.3->imbalanced-learn) (2024.2.0)
Requirement already satisfied: tbb==2022.* in /usr/local/lib/python3.10/dist-packages (from mkl->numpy>=1.17.3->imbalanced-learn) (2022.0.0)
Requirement already satisfied: tcmlib==1.* in /usr/local/lib/python3.10/dist-packages (from tbb==2022.*->mkl->numpy>=1.17.3->imbalanced-learn) (1.2.0)
Requirement already satisfied: intel-cmplr-lib-rt in /usr/local/lib/python3.10/dist-packages (from mkl_umath->numpy>=1.17.3->imbalanced-learn) (2024.2.0)
Requirement already satisfied: intel-cmplr-lib-ur==2024.2.0 in /usr/local/lib/python3.10/dist-packages (from intel-openmp>=2024->mkl->numpy>=1.17.3->imbalanced-learn) (2024.2.0)
Class distribution after applying SMOTE:
target
1    133
0    133
Name: count, dtype: int64

```