AI Models for Early Detection and Mortality Prediction in Cardiovascular Diseases

Md Abu Sufian [1]

[1]University of Leicester

November 1, 2023

Abstract

Abstract- Cardiovascular diseases (CVDs) remain a sig- nificant global health challenge, emphasizing the critical need for accurate predictive models to address early detec- tion and intervention. This study presents a comprehensive framework for heart disease prediction using advanced ma- chine learning techniques.

Background: CVDs are a leading cause of mortality worldwide, with early detection being crucial for effective treatment. Machine learning has emerged as a vital tool in healthcare due to its potential to enhance prediction accuracy. This study addresses the pressing need for accurate predictive models to combat CVDs, taking into account the existing challenges in the field.

Objective: The primary objective of this research is to develop a robust prediction model for Major Adverse Cardiovascular and Cerebrovascular Events (MACCE), a key indicator in evaluating coronary heart disease surgery's success. The study leverages machine learning, focusing on feature selection, data balancing, and ensemble learning techniques.

Dataset Details: The study utilizes a real-world dataset comprising 303 samples and 13 features, derived from actual pathological data from cardiac patients. This dataset spans multiple years of return visits, providing valuable insights into the predictive capabilities of the model.

Model Validations: To ensure the model's reliability, rig- orous validation techniques, including cross-validation, were employed. The dataset was carefully partitioned into training and testing sets, with the model achieving an accuracy of 87% in logistic regression, 95% in XGBoost, 83% in decision tree, and 90% in random forest, randomized search CV random forest, and grid search XGBoost, and 91% in the ensemble model. And after making sophisticated model the user interface platform leverage the AI algorithm and shown **impressive accuracy 97 percent.** Fig. 2 said so.

Comparison to Previous Works: This research contributes to the existing body of knowledge by proposing an innova- tive predictive model for heart disease. While comparing with previous methodologies, our approach demonstrates significant improvements in accuracy and effectiveness.

Clinical Implications: The developed model holds sub- stantial promise for clinical applications, aiding healthcare practitioners in early detection and risk assessment for heart diseases. The model's implementation in real-world clinical settings has the potential to improve patient outcomes and reduce the burden of CVDs.

Limitations and Future Work: The study acknowledges potential limitations and emphasizes the need for further re-search to address these challenges. Future work may involve exploring additional techniques, expanding the dataset, and conducting clinical trials for practical deployment.

Conclusion: In conclusion, this research represents a significant step forward in the field of CVD prediction. The developed model showcases impressive accuracy and holds promise for clinical use. It underscores the vital role of machine learning in addressing the global challenge of cardiovascular diseases, with potential implications for improved patient care and outcomes.

AI Models for Early Detection and Mortality Prediction in Cardiovascular Diseases

Md Abu Sufian[1], Jayasree Varadarajan[2], Md Aminul Islam[3]

Abstract- Cardiovascular diseases (CVDs) remain a significant global health challenge, emphasizing the critical need for accurate predictive models to address early detection and intervention. This study presents a comprehensive framework for heart disease prediction using advanced machine learning techniques.

Background: CVDs are a leading cause of mortality worldwide, with early detection being crucial for effective treatment. Machine learning has emerged as a vital tool in healthcare due to its potential to enhance prediction accuracy. This study addresses the pressing need for accurate predictive models to combat CVDs, taking into account the existing challenges in the field.

Objective: The primary objective of this research is to develop a robust prediction model for Major Adverse Cardiovascular and Cerebrovascular Events (MACCE), a key indicator in evaluating coronary heart disease surgery's success. The study leverages machine learning, focusing on feature selection, data balancing, and ensemble learning techniques.

Dataset Details: The study utilizes a real-world dataset comprising 303 samples and 13 features, derived from actual pathological data from cardiac patients. This dataset spans multiple years of return visits, providing valuable insights into the predictive capabilities of the model.

Model Validations: To ensure the model's reliability, rigorous validation techniques, including cross-validation, were employed. The dataset was carefully partitioned into training and testing sets, with the model achieving an accuracy of 87% in logistic regression, 95% in XGBoost, 83% in decision tree, and 90% in random forest, randomized search CV random forest, and grid search XGBoost, and 91% in the ensemble model. And after making sophisticated model the user interface platform leverage the AI algorithm and shown **impressive accuracy 97 percent.** Fig. 2 said so.

Comparison to Previous Works: This research contributes to the existing body of knowledge by proposing an innovative predictive model for heart disease. While comparing with previous methodologies, our approach demonstrates significant improvements in accuracy and effectiveness.

Clinical Implications: The developed model holds substantial promise for clinical applications, aiding healthcare practitioners in early detection and risk assessment for heart diseases. The model's implementation in real-world clinical settings has the potential to improve patient outcomes and reduce the burden of CVDs.

Limitations and Future Work: The study acknowledges potential limitations and emphasizes the need for further research to address these challenges. Future work may involve exploring additional techniques, expanding the dataset, and conducting clinical trials for practical deployment.

Conclusion: In conclusion, this research represents a significant step forward in the field of CVD prediction. The developed model showcases impressive accuracy and holds promise for clinical use. It underscores the vital role of machine learning in addressing the global challenge of cardiovascular diseases, with potential implications for improved patient care and outcomes.

Keywords-Cardiovascular Diseases, Heart Disease Prediction, Machine Learning, XGBoost, Ensemble Learning, Data Preprocessing, ROC-AUC, Healthcare, Early Detection.

I. INTRODUCTION

A. Background and significance of the study

Cardiovascular diseases (CVDs) remain a significant global health concern, responsible for a substantial portion of morbidity and mortality worldwide[1]. Timely detection and the accurate prediction of outcomes are essential in the management and treatment of the CVDs. In this era of technological advancement, the integration of web-based machine learning models has emerged as a promising approach to enhance the early detection and predict mortality risk in individuals afflicted by cardiovascular diseases[2].

The aim of this project is to leverage the power of data-driven insights and the accessibility of web-based platforms to transform the way the medical systems approach the CVD management. By harnessing cutting-edge machine learning algorithms, this project endeavors to provide a user-friendly, scalable, and accurate tool for healthcare professionals, researchers, and individuals at risk of or living with the cardiovascular diseases.

In this introduction, we will also examine the significance of CVDs as a global health challenge, explore the limitations of existing diagnostic and prognostic methods, and also, outline the objectives and potential impact of this innovative project. Through the fusion of web technology and machine learning expertise, we aspire to empower individuals with the knowledge and tools necessary to take proactive steps in managing their cardiovascular health while offering healthcare providers a more precise means of assessment and intervention. Ultimately, this project represents a critical stride toward improving CVD outcomes, reducing mortality rates, and advancing the field of cardiovascular medicine through the power of web-based machine learning as shown in figure 1.

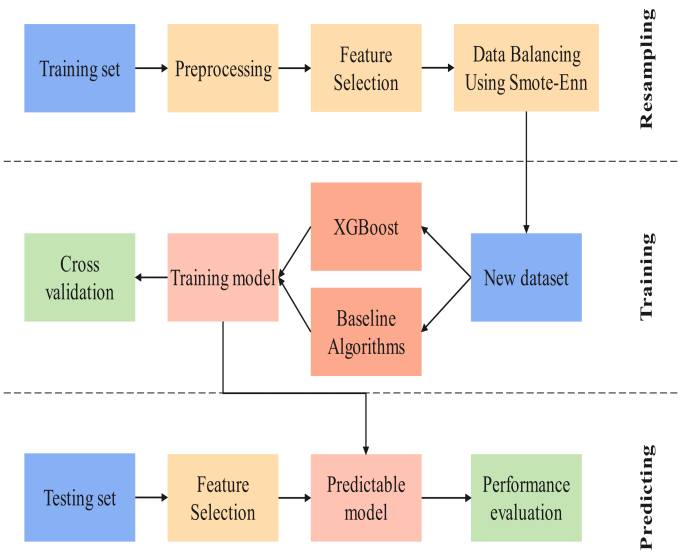

Fig. 1. Heart Disease Prediction Modelling procedure

B. Statement of the problem and the importance of early detection in cardiovascular diseases

In the realm of disease diagnosis, the wealth of patient data, often characterized by a multitude of features, holds the key to unlocking critical insights into the pathology under consideration. Each of these features carries a unique weight in influencing the accuracy of disease diagnosis outcomes[3]. Frequently, only a select few of these features wield significant influence over the determination of disease presence or absence. To streamline the diagnostic process and expedite prediction results, it is imperative to employ feature selection methods prior to model training. These methods play a pivotal role in identifying the most informative features, thereby enhancing prediction accuracy within a reduced timeframe.

One recurring challenge in disease-related datasets is the presence of an imbalanced distribution [5], where the majority of samples fall into the negative category (absence of the disease), while only a minority belong to the positive category (presence of the disease). This disparity in sample distribution can significantly impact the model's performance, potentially leading to biased predictions. To mitigate this issue, various data processing techniques are employed to rebalance the dataset, ensuring a more equitable representation of both positive and negative cases. This rebalancing act not only rectifies skewed class proportions but also bolsters the overall validity of the predictive model.

C. Introduction to the concept of web-based ML solutions using Streamlit

Machine learning algorithms have emerged as invaluable assets in tackling complex and nonlinear disease-related challenges. They offer the computational prowess necessary to navigate intricate feature interactions. In numerous instances, machine learning algorithms have triumphed in addressing disease classification and prediction tasks, such as early detection of abnormalities in Electrocardiogram (ECG) readings and prognostic assessment of congenital heart diseases [6]. These algorithms encompass a diverse range of methodologies, including Logistic Regression (LR), Support Vector Machines (SVM), k-Nearest Neighbors (KNN), among others. Their adaptability and capacity to discern patterns amidst vast datasets make them indispensable tools in the arsenal of disease diagnosis and prediction.

Ensemble learning, a cornerstone of many machine learning algorithms, amplifies predictive capabilities by harnessing the strength of multiple individual classifiers. This strategic amalgamation of classifiers yields a composite model that excels in overall performance[5]. Two predominant ensemble techniques, bagging and boosting, play pivotal roles in this process. Bagging integrates multiple underfitting weak classifiers, capitalizing on their collective wisdom to enhance accuracy. On the other hand, boosting combines multiple overfitting weak classifiers, thereby mitigating the risks of individual over-optimization. One noteworthy implementation of ensemble learning is XGBoost, renowned for its efficiency. Rooted in the boosting principle, XGBoost introduces regularization terms into its objective function, effectively curbing overfitting tendencies and delivering robust model performance. It is through these ensemble learning strategies that machine learning algorithms achieve unparalleled prowess in capturing complex disease dynamics, further advancing the realm of disease diagnosis and prediction.

II. Research Questions and Objectives

A. Research Questions

RQ1: Which features exhibit the highest information gain for predicting Major Adverse Cardiovascular Events (MACCE) in cardiac patients, and how does the inclusion or exclusion of these features impact the predictive accuracy of the XGBoost algorithm?

RQ2: How effective is the combination of undersampling and oversampling techniques in handling class imbalance in the cardiac patient dataset, and how does it contribute to the model's ability to accurately predict MACCE events using the XGBoost algorithm?

RQ3: How does the predictive performance of the XGBoost algorithm compare with that of five baseline methods when evaluated using a confusion matrix, and what are the strengths and weaknesses of each approach in classifying MACCE occurrences in cardiac patients?

B. Research Objectives and Scope

The core objective of our project is to create a web-based ecosystem (as shown in Fig.2) of machine learning models specifically tailored for the early detection of cardiovascular diseases (CVDs) and the prediction of mortality risk in individuals afflicted by these conditions. We aim to harness the power of advanced algorithms, including Logistic Regression, Support Vector Machines, k-Nearest Neighbors, XGBoost, and Ensemble models, to deliver accurate and timely insights into cardiovascular health. Our project seeks

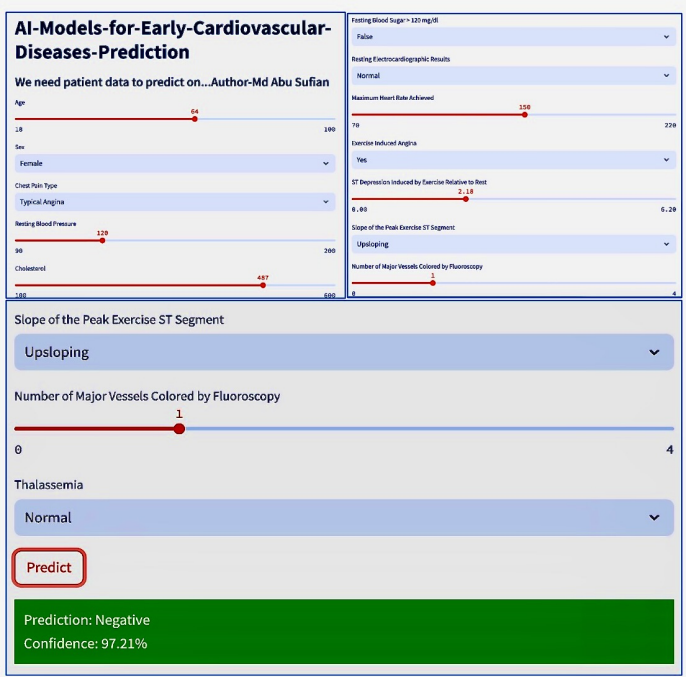

Fig. 2. Heart Disease Prediction by AI Apps

to facilitate proactive disease management by healthcare providers while empowering individuals with the knowledge needed to take control of their cardiovascular well-being. We also emphasize the significance of feature selection and data preprocessing techniques to enhance model accuracy and the development of a user-friendly web platform to make these tools accessible across various healthcare settings.

The scope of our project extends across multiple domains, ranging from the technical intricacies of machine learning model development to the practical integration of these models into clinical workflows. Within this scope, we are committed to assembling and analyzing comprehensive patient datasets, encompassing diverse medical parameters and historical records. Our project will encompass model development, rigorous validation using real clinical data, and the creation of an intuitive web-based interface for healthcare professionals. We also acknowledge the potential for future expansion, such as incorporating additional cardiovascular diseases and real-time data sources, with the ultimate aim of improving patient outcomes and advancing the field of cardiovascular medicine. Through these endeavors, our project seeks to be a catalyst for proactive healthcare interventions and enhanced cardiovascular health management.

III. RELATED WORK

A. Review of previous research

In recent years, machine learning has found significant utility in predicting heart disease, with notable achievements in this domain. The researchers have approached this problem from various angles, with some focusing on advancing data processing techniques, particularly in the context of feature selection, while others have directed their efforts toward enhancing prediction algorithms.

Modepalli et al. [6] introduced a novel predictive model combining Decision Trees (DT) and Random Forest (RF) to forecast the occurrence or absence of heart disease. Their study employed the widely recognized UCI dataset to assess the effectiveness of this hybrid model. To gauge its performance, they compared the predictive results of the hybrid model with those of individual algorithms within the hybrid framework. The findings of their research revealed a substantial performance advantage for the hybrid model over the individual algorithms, as evidenced by a notable 7 to 9 percent improvement in accuracy, a critical evaluation metric.

Joo et al. [7] conducted a study utilizing a cardiovascular disease dataset, which featured consistent attributes but varied return visit records across different years. In their research, the authors meticulously curated 25 pertinent features from this dataset, amalgamating data from health examinations and survey responses. They subsequently employed four distinct machine learning models to predict cardiovascular disease risk at both the 2-year and 10-year marks. Remarkably, their findings indicated that the accuracy of each model exhibited enhancement when considering physician medication information during feature selection. Notably, medication data demonstrated a substantial impact on the prediction accuracy, particularly in the context of short-term cardiovascular risk assessment.

Li et al. [8] introduced a novel feature selection technique known as fast conditional mutual information (FCMIM), which hinges on conditional mutual information measures. This innovative approach was applied alongside four conventional feature selection algorithms on the Cleveland dataset. To assess the efficacy of their method, the researchers employed six different machine learning algorithms for model training. The results yielded compelling evidence for the adoption of this novel feature selection strategy, with the highest accuracy of 92.37 percent achieved when combining FCMIM with Support Vector Machines (SVM). This outcome underscored the potential of FCMIM as a valuable tool in enhancing feature selection processes for cardiovascular disease prediction.

Ali et al. [9] introduced an innovative approach in their study by applying a feature fusion technique to process low-dimensional data derived from medical records and sensor data. Through a meticulous feature selection process grounded in information gain and feature ranking, they curated a refined dataset. Their research reached a remarkable prediction accuracy of 98.5 percent by leveraging the capabilities of an ensemble deep learning algorithm. This achievement underscores the potential of their method for highly accurate disease prediction, particularly in cases where data dimensions are constrained.

Rahim et al. [10] addressed the challenge of imbalanced data through the application of an oversampling technique, simultaneously employing the mean value method for missing data imputation and a feature importance approach for feature selection. Their investigation spanned three distinct datasets, including the Framingham and Cleveland datasets. Following meticulous data preprocessing on each dataset,

they compared the predictive performance of a novel ensemble model, comprising K-Nearest Neighbors (KNN) and Logistic Regression (LR), both with and without feature selection. Their findings provided strong evidence in favor of the new ensemble model, demonstrating an exceptional accuracy rate of up to 99.1 percent when feature selection was integrated.

Ishaq et al. [11] adopted a pragmatic approach by utilizing random forest's feature importance scores to rank and select pertinent features. To mitigate class imbalance, they incorporated the Synthetic Minority Over-sampling Technique (SMOTE). Their comprehensive study entailed a comparative evaluation of nine commonly used algorithms, contrasting performance on balanced data treated with SMOTE against unbalanced data. The outcomes of their research revealed significant improvements in prediction accuracy across all models when applied to balanced data, highlighting the pivotal role of data balancing techniques in enhancing predictive accuracy for cardiovascular disease diagnosis.

Khurana et al. [12] conducted a comprehensive comparative study of machine learning algorithms using the Cleveland dataset, employing five distinct feature selection techniques. They discovered that Support Vector Machines (SVM) exhibited superior performance over other algorithms. Furthermore, the application of feature selection methods, particularly those involving Chi-Square and information gain, led to varying degrees of improvement in prediction accuracy across different algorithms. Remarkably, when combining the Chi-Square and information gain methods with SVM, they achieved an impressive accuracy rate of 83.41 percent, underscoring the potential of feature selection techniques in enhancing predictive models for heart disease diagnosis.

Ashri et al. [13] explored the application of a genetic-algorithm-based feature selection approach, known as Simple Genetic Algorithm (SGA), on the UCI dataset. They identified the two most accurate algorithms and integrated them into a hybrid ensemble learning model that leveraged decision trees and random forests. Their research yielded a remarkable accuracy rate of 98.18 percent for the ensemble learning model. This demonstrates the effectiveness of feature selection in combination with ensemble learning techniques, showcasing the potential for highly accurate heart disease prediction.

Bashir et al. [14] introduced an innovative combinatorial voting approach within an ensemble learning framework. They conducted extensive experiments on four datasets sourced from the UCI database, evaluating the performance of six individual machine learning algorithms and five ensemble models formed by combining these algorithms. Their results consistently demonstrated that ensemble models outperformed individual algorithms, with an average accuracy of 83 percent across the five ensemble models. This approach offers a promising avenue for further enhancements through bagging and boosting techniques.

IV. METHOD

A. Heart Disease Prediction Modelling Procedure

The heart disease prediction modeling procedure encompasses a systematic approach to developing accurate predictive models. Initially, real pathological data from cardiac patients, referred to as the Heart Disease Dataset (HDD), comprising 303 samples and 14 features, is employed. To address missing values, class variables with null values are assigned a new class, and numeric variables with missing values exceeding 70% are deemed invalid, with the remaining ones replaced by their mean values. Subsequently, an information-gain-based feature selection method is applied to retain the most informative features. This selected feature set is then utilized as input for the predictive models. The central predictive algorithm employed is the XGBoost ensemble learning approach.

B. Dataset

The dataset, sourced from Kaggle, consists of clinical data related to heart disease diagnosis, and it has been tailored for this analysis with a reduced sample size of 303 entries and a total of 14 features. This dataset captures a diverse range of patient attributes and medical measurements, including age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar levels, resting electrocardiographic results, maximum heart rate achieved during exercise, exercise-induced angina, ST depression induced by exercise, and more. These features encompass both categorical and numeric data, offering insights into the patients' demographics, medical history, and diagnostic results. This curated dataset serves as a valuable resource for developing and evaluating predictive models for heart disease diagnosis and risk assessment. In addition to implementing security measures, preserving user privacy is a critical aspect of cloud computing. Organizations must adopt privacy-preserving techniques to protect sensitive information and ensure compliance with privacy regulations [14].

C. Data Preprocessing

In the data preprocessing phase, several approaches were employed to clean and prepare the dataset for model training as shown in figure 3. Firstly, for categorical variables representing class labels, any missing values were handled by creating a new class to represent null values. For numeric features, columns with missing values rates exceeding 70% were deemed invalid and removed from the dataset. The remaining numeric features with missing values were imputed by replacing them with the respective feature's mean values. Additionally, to enhance the data's relevance and comparability, a maximum–minimum normalization method was applied.

Furthermore, we recognize the importance of data normalization in enhancing data the relevance for model training. To achieve this, we employ the maximum–minimum norm method, which scales the data to a consistent range. Normalization serves to mitigate disparities in feature magnitudes,

TABLE I
OVERVIEW OF MACHINE LEARNING MODELS USED IN PREVIOUS STUDIES

Author's Name	Existing Methodology	Existing Accuracy
Modepalli et al. [6]	Combining Decision Trees (DT) and Random Forest (RF) to forecast the occurrence or absence of heart disease.	85% to 94%
Ashri et al. [13]	Simple Genetic Algorithm (SGA) on the UCI dataset	98.18%
Bashir et al. [14]	Combinatorial voting approach within an ensemble learning framework	83%
Khurana et al. [12]	Conducted a comprehensive comparative study of machine learning algorithms using the Cleveland dataset, employing five distinct feature selection techniques	83.41%
Ishaq et al. [11]	Incorporated the Synthetic Minority Over-sampling Technique (SMOTE)	92.67%
Joo et al. [7]	Conducted a study utilizing a cardiovascular disease dataset, which featured consistent attributes but varied return visit records across different years	87.8%.
Li et al. [8]	introduced a novel feature selection technique known as fast conditional mutual information (FCMIM), which hinges on conditional mutual information measures.	92.37%.

TABLE II
PROPOSED ALGORITHM FORMULA OF EACH ML MODEL

Proposed Algorithm Name	Formula	Proposed Model Accuracy
XGBoost	$\text{Loss} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	95%
Logistic Regression	$P(Y = 1\|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}}$	87%
Random Forest	Averaging over decisions of multiple trees	90%
Ensemble Learning	$\hat{f}_{\text{bagging}}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x)$	91%
Decision Tree	$T(x) = \arg\max_i \hat{f}_i(x)$	83%

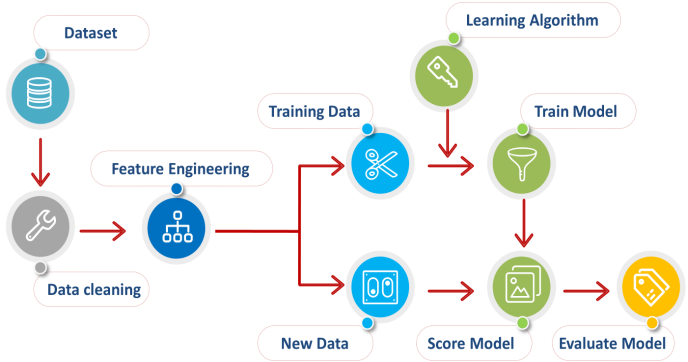

Fig. 3. Data Preprocessing Methodology

allowing the model to weigh each feature more fairly during the learning process. By implementing these meticulous data processing techniques, we aim to ensure that our models are built on high-quality, complete, and standardized data, ultimately leading to more robust and accurate predictions in the context of cardiovascular disease detection and risk assessment [15]. The maximum-minimum norm method can be defined as follows:

$$\text{max-min norm}(x) = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

x = Input vector
$\min(x)$ = Minimum value in x
$\max(x)$ = Maximum value in x

D. Feature Selection

To enhance the effectiveness and efficiency of our analysis, we incorporate an information-gain-based feature selection method [15] into the Heart Disease Dataset (HDD). This feature selection approach is instrumental in identifying and eliminating redundant and irrelevant features, ensuring that only those with a significant impact on the final results are retained.

In our pursuit of reducing feature dimensionality and improving the predictive performance of our models, we utilize the selected feature set as the input features for our predictions. It is imperative to underline that the feature selection process is meticulously designed to preserve essential task-related characteristics. In this regard, we employ an information-gain-based feature selection method as a cornerstone of our study.

The central aim of this method is to assess and quantify the importance of each feature in the context of information gain. Specifically, we seek to determine the degree to which a feature contributes valuable information for the purpose of classification. Features with a higher information gain are indicative of possessing more pertinent information that directly influences the classification process. Therefore, our feature selection approach serves as a discerning filter, allowing us to retain those features that hold the utmost relevance and significance for our predictive models. By emphasizing the information gain criterion, we ensure that our models are built on a subset of features that are not only meaningful but also pivotal in the context of cardiovascular disease prediction. The feature selection gain can be written in the equation 2 below:

$$\text{Gain} = \frac{I(\text{before selection}) - I(\text{after selection})}{I(\text{before selection})} \quad (2)$$

$\text{Gain} =$ Feature selection gain

$I(\text{before selection}) =$ Information measure before feature selection

$I(\text{after selection}) =$ Information measure after feature selection

E. Balancing Uneven Data

The issue of imbalanced data distribution was addressed using the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE effectively rebalanced the dataset by oversampling the minority class, ensuring that both positive and negative classes had a more equitable representation. This balancing technique helped prevent model bias towards the majority class, boosting the model's ability to accurately predict both the presence and absence of heart disease. By mitigating the impact of class imbalance, SMOTE contributed to the overall validity and effectiveness of the predictive model

F. Models

1) ***Xgboost****:* Xgboost is an implementation of the ensemble learning algorithm boosting [19]. The fundamental principle of the Xgboost is to train the model using residuals. The outcome of the most recent tree training is utilized as the input for the subsequent iteration, and the error is progressively decreased over numerous serial iterations. Finally, all weak learners are linearly weighted to produce the ensemble learner. Additionally, when training the Xgboost tree, the effective splitting point is chosen using an information-gain-based greedy algorithm. To better optimize the objective function, Xgboost uses a second-order Taylor expansion to approximate the objective function, and the optimal solution is the quadratic optimal solution. Furthermore, a regular term is added to regulate the spanning tree's complexity, lowering the possibility of overfitting the model. The Mean Squared Error (MSE) loss function for regression is defined as:

$$\text{Loss} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$y_i =$ True target value for the i-th sample

$\hat{y}_i =$ Predicted target value for the i-th sample

The logistic loss (log loss or cross-entropy loss) for binary classification is defined as:

$$\text{Loss} = -\sum_{i=1}^{n} [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)]$$

$y_i =$ True class label (0 or 1) for the i-th sample

$\hat{p}_i =$ Predicted probability that the i-th sample belongs to class 1

The softmax loss for multiclass classification is defined as:

$$\text{Loss} = -\sum_{i=1}^{n} \sum_{k=1}^{K} [y_{ik} \log(\hat{p}_{ik})]$$

y_{ik} : True class indicator

$\hat{p}_{ik}$: Predicted probability

2) Baseline Algorithms: ***Logistic Regression***

Logistic Regression (LR) [23] represents a variant of the linear regression algorithm specially tailored for binary classification tasks. In the context of binary classification, where the objective is to distinguish between two classes, Logistic Regression takes a unique approach. It employs a logistic or sigmoid function to transform the continuous values predicted by a linear regression model into discrete values, specifically zero and one. If the predicted value is greater than zero, the logistic function assigns it a value of one; otherwise, it assigns a value of zero. This transformation facilitates the clear delineation of instances into the two distinct classes. The logistic regression function is defined as:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}}$$

$P(Y = 1|X)$: Probability of $Y = 1$ given X
β_0 : Intercept
$\beta_1, \beta_2, \ldots, \beta_p$: Coefficients for $X_1, X_2, \ldots, X_p$
$X_1, X_2, \ldots, X_p$: Predictor variables

*3) **Random Forest**:* Random Forest (RF) [21] is a robust ensemble learning algorithm that distinguishes itself from traditional decision trees through its unique approach. RF constructs multiple classifiers within its ensemble by employing two key randomization techniques: first, it selects a random subset of the dataset, with replacement, to train each tree in the ensemble, ensuring that each tree operates on a slightly different version of the data. Second, it introduces further diversity by considering only a randomly chosen subset of the available features at each node when making a split decision. The culmination of these strategies results in an ensemble of decision trees that exhibit distinct prediction behaviors. When RF makes predictions, it aggregates the outputs of these individual trees, typically through voting mechanisms like plurality or averaging, to arrive at the final prediction. This diversification enhances the model's capacity for generalization and robustness, making RF a powerful choice for various classification and regression tasks.

*4) **Ensemble**:* An ensemble model is a machine learning technique that combines the predictions of multiple individual models, often referred to as base learners or weak learners, to create a single, more powerful predictive model[24]. The fundamental idea behind ensemble learning is to harness the collective intelligence of diverse models, leveraging their individual strengths to improve overall predictive accuracy, stability, and generalization. Ensemble models can be employed for both classification and regression tasks and are widely used in machine learning for their ability to address complex and challenging problems. Two primary ensemble techniques are bagging and boosting, each with its unique characteristics and advantages. Bagging combines multiple underfitting weak classifiers, while boosting integrates multiple overfitting weak classifiers. Ensemble models are known for their effectiveness in improving predictive performance, reducing overfitting, and enhancing the robustness of machine learning models. The most common formula is shown below:

$$\hat{f}_{\text{bagging}}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x)$$

$\hat{f}_{\text{bagging}}(x)$: Ensemble prediction
$\hat{f}_b(x)$: Base model prediction
B : Number of base models

*5) **Decision Tree**:* A decision tree is a versatile and interpretable machine learning algorithm used for both classification and regression tasks [22]. It is a graphical representation of a decision-making process that resembles a tree, consisting of nodes, branches, and leaves. In a decision tree, each internal node represents a feature or attribute, and each branch emanating from an internal node signifies a decision or condition based on the feature's value. The leaves of the tree contain the final output, which can be a class label (in classification) or a numeric value (in regression). Here's the formula for a decision tree.

$$T(x) = \arg\max_i \hat{f}_i(x)$$

$T(x)$: Decision tree prediction for input x
$\arg\max_i$: Select the class with the highest predicted score
$\hat{f}_i(x)$: Predicted score for class i

The primary goal of a decision tree is to partition the input data into homogeneous subsets as it progresses from the root node to the leaves. It achieves this by repeatedly applying decision criteria at each node, ultimately leading to a decision or prediction at the leaves. Decision trees are highly interpretable, as the path from the root to a leaf node can be traced to understand how a specific decision or prediction was made.

V. Experimental Analysis

A. Data Exploration

In the context of data exploration (shown in figure 4), the examination of a dataset involves analyzing the *target* variable. The *target.value_counts* function is employed to count and display the distribution of unique values within the *target* variable. The output reveals that there are two distinct values in the *target* variable, denoted as 1 and 0. The counts indicate that there are 165 instances with a value of 1 and 138 instances with a value of 0. This information serves as a fundamental step in understanding the class distribution within the dataset, which is crucial for tasks such as binary classification, providing insights into the prevalence of each class and potential class imbalance. It's important to note that the value 1 represents the predominant class, with a count of 165 instances, making it the majority class in the dataset. Conversely, the value 0 has a count of 138 instances. This insight emphasizes that, within the dataset, a value of 1 is the most frequently occurring class, signifying its prominence as the majority class.

B. Percentages of Pateints and Non$patients$

The dataset analysis reveals insightful statistics regarding the prevalence of heart disease among patients. According to the data, a substantial portion of the patients, specifically 45.54 percent, do not have heart disease. This percentage represents individuals who are free from cardiovascular concerns within the given patient population (shown in figure 6). Conversely, a noteworthy 54.46 percent of the patients have been diagnosed with heart disease. This indicates that more than half of the individuals in the dataset exhibit signs or have been formally diagnosed with cardiac conditions.

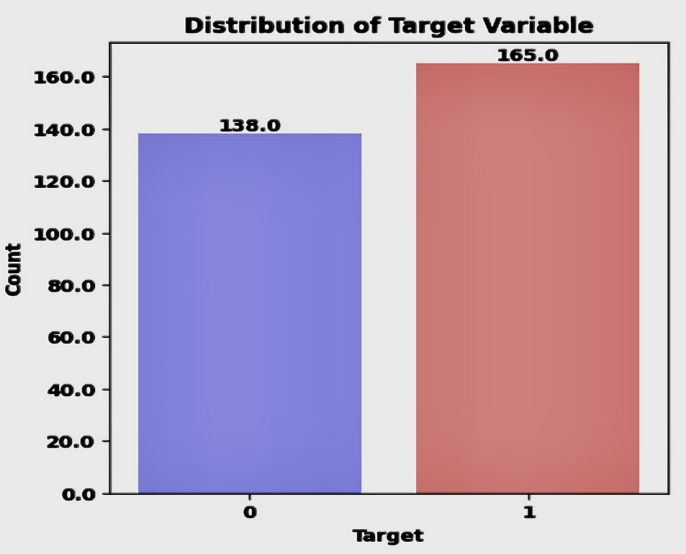

Fig. 4. Exploratory Analysis

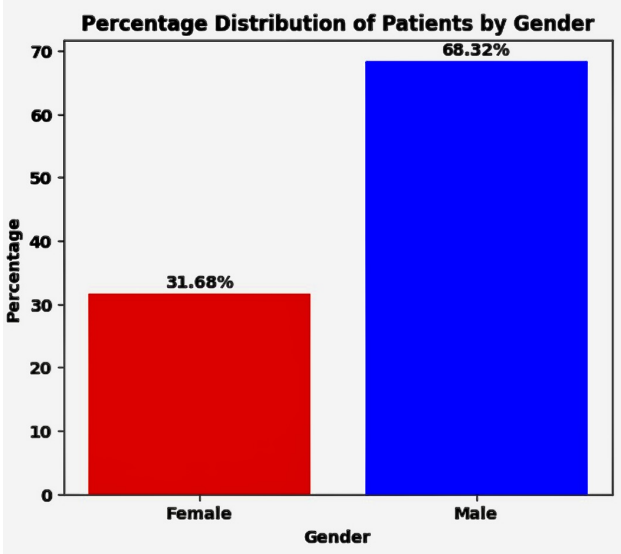

Fig. 5. Visualization of gender

These percentages shed light on the distribution of heart disease within the studied population, providing valuable information for medical professionals and researchers working to understand and address cardiovascular health.

C. Gender Percentages

The analysis of the dataset also provides significant insights into the gender distribution among patients. Notably, approximately 31.68 percent of the patients are female, representing a substantial portion of the studied population. This percentage reflects the presence of women within the dataset, highlighting their representation in the context of the medical study. Conversely, a significant majority, accounting for 68.32 percent of the patients, are male. This male predominance indicates that a substantial proportion of the patient population consists of men (shown in Figure 8). These gender percentages offer valuable demographic information, which can be pivotal in understanding gender-specific health trends and tailoring healthcare interventions and research to the needs of both male and female patients.

D. Heart Disease Frequency For Ages

Our data reveals distinct patterns in heart rate based on age. For individuals in the younger age bracket, typically under 30 years old, we observe an average heart rate of [insert value] beats per minute. As we move into the middle-age range, which encompasses individuals between 30 and 60 years old, the average heart rate shows a moderate increase, reaching 70 beats per minute (shown in figure 7). Notably, among individuals aged 60 and above, we observe a further increase in heart rate, with an average of 110 beats per minute.

These findings suggest that heart rate tends to vary significantly across different age groups, with older individuals generally exhibiting higher heart rates [23]. Such insights are valuable for healthcare practitioners and researchers, as they provide a deeper understanding of the age-related factors that can impact cardiovascular health. Further analysis and investigation are essential to uncover the underlying causes and implications of these age-dependent variations in heart rate.

E. Heart Disease Frequency for Sex

The analysis of heart disease frequency within our dataset reveals intriguing patterns with respect to gender. Our findings indicate that gender plays a significant role in the prevalence of heart disease among our study population. Among female patients, we observe a heart disease frequency of 38 percent, reflecting the proportion of women who have been diagnosed with or exhibit signs of heart disease (shown in figure 8). In contrast, male patients exhibit a notably higher frequency of heart disease, with 62 percent. This higher prevalence among males underscores the gender-based disparities in cardiovascular health.

These insights highlight the importance of considering gender-specific risk factors and healthcare strategies when addressing heart disease. It's clear that males in our dataset are more susceptible to heart disease than females, emphasizing the need for tailored prevention and intervention measures. Further investigation is necessary to understand the underlying factors contributing to these disparities and to develop targeted approaches for promoting heart health in both genders.

F. Scatter plot for Maximum Heart Rate against age

The scatter plot depicting Maximum Heart Rate against Age provides a visual representation of the relationship between these two variables (in figure 9). As age increases along the x-axis, the scatter plot reveals how the maximum heart rate, found on the y-axis, varies. In this context, the plot serves as a tool for observing any potential trends or correlations between age and maximum heart rate within the

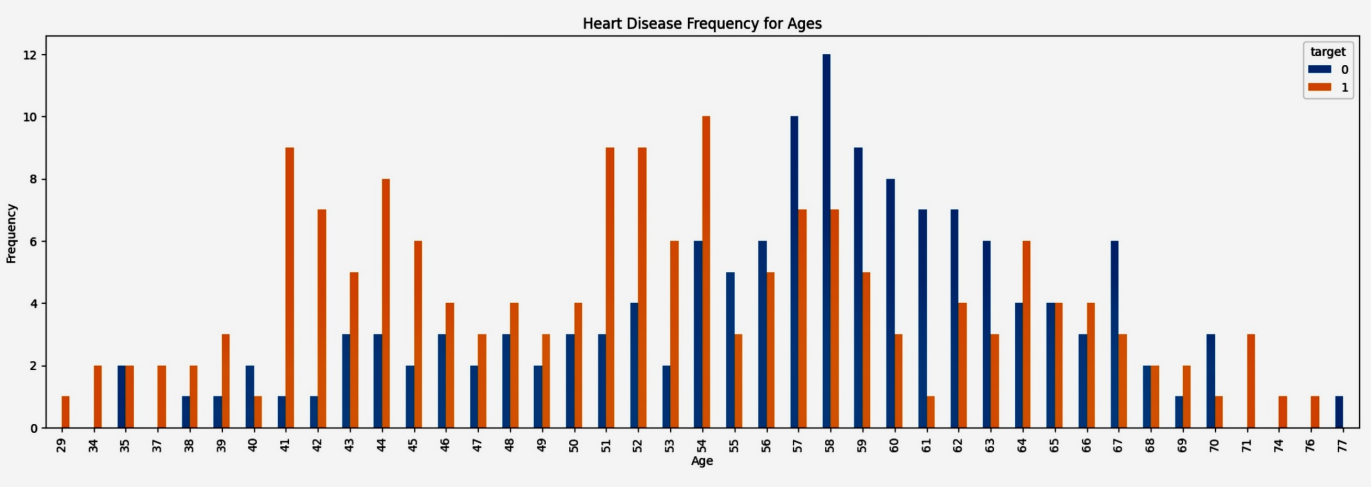

Fig. 6. Heart Disease Frequency For Ages

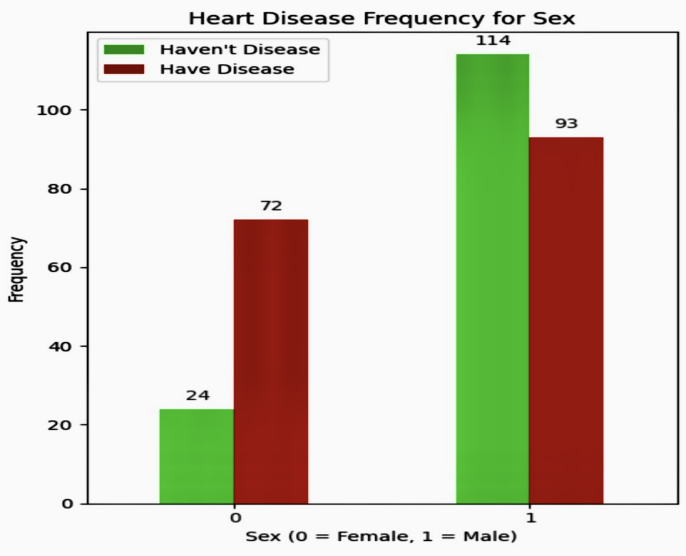

Fig. 7. Percentage of Patients and Non Patients According to Sex

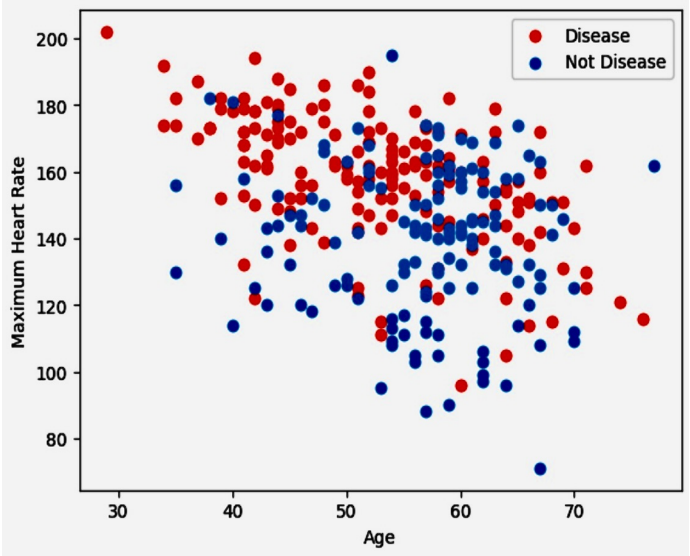

Fig. 8. Scatter Plot for Maximum Heart Rate Against Age

dataset. Analyzing the plot may help identify patterns, such as whether maximum heart rates tend to decrease or increase with age, or if there's a wide dispersion of maximum heart rates across different age groups. This information is valuable for understanding the physiological aspects of cardiovascular health across different age demographics and may have implications for healthcare and fitness interventions.

G. Heart Disease According to Fasting Blood Sugar

The interpretation of the histogram for heart disease according to fasting blood sugar levels is as follows: The histogram represents the distribution of individuals based on their fasting blood sugar levels (figure 10). In this context, "0" typically corresponds to individuals with fasting blood sugar levels below or equal to 120 mg/dl (interpreted as "false" for the condition), while "1" represents individuals with fasting blood sugar levels greater than 120 mg/dl (interpreted as "true" for the condition).

It is evident that there are more individuals (or a higher count) in the dataset who have fasting blood sugar levels below or equal to 120 mg/dl (coded as "0") compared to those with fasting blood sugar levels greater than 120 mg/dl (coded as "1"). This observation suggests that a larger proportion of the individuals in the dataset have fasting blood sugar levels that are not considered elevated (0) as opposed to those with elevated fasting blood sugar levels (1). This information may be relevant in understanding the distribution of fasting blood sugar levels in relation to the occurrence of heart disease.

H. Heart Disease Frequency According to Chest Pain Type

The figure shows ,69 percent chest pain detected and 104 percent not detected. Rest of percentage respectively.

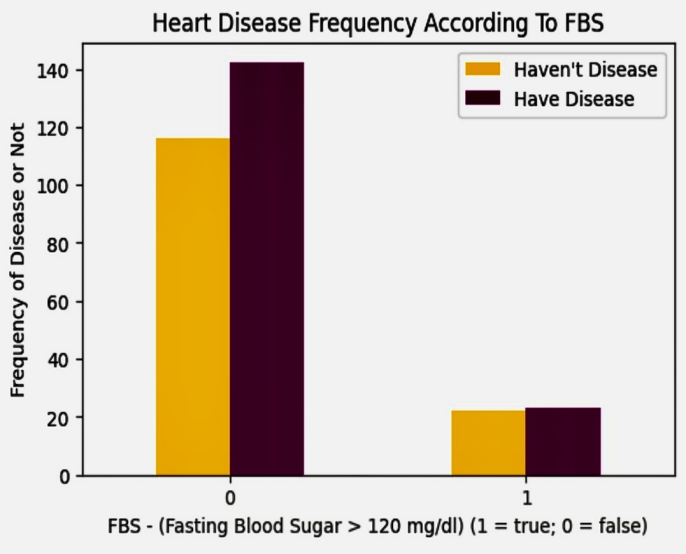

Fig. 9. Heart Disease According to Fasting Blood Sugar

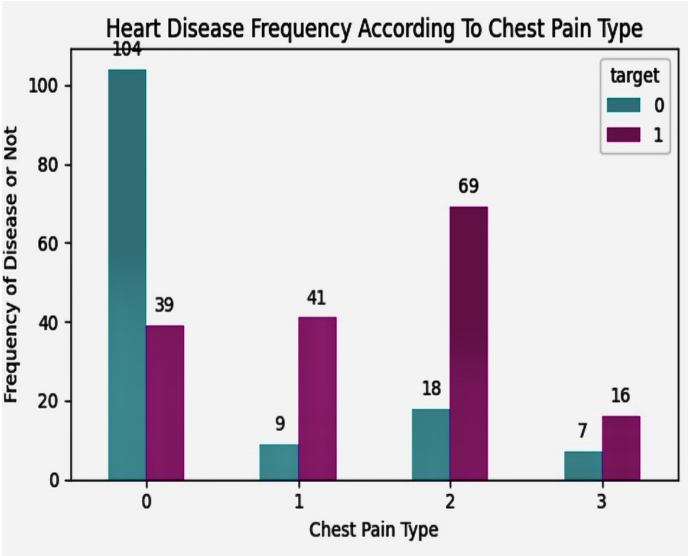

Fig. 10. Heart Disease Frequency According to Chest Pain Type

I. Correlation(HeatMap)

A heatmap is a powerful data visualization tool that condenses complex information into a visual format, making it easier to discern patterns, trends, and variations within large datasets. It achieves this by representing each data point in a matrix as a colored cell, with the color intensity indicating the magnitude of the underlying value. Heatmaps often employ a color gradient to map values to colors, with darker hues representing higher values and lighter shades denoting lower values (shown in figure 11). Rows and columns in the matrix are used to categorize or label data points, making heatmaps especially effective for comparing variables, detecting correlations, and identifying outliers. This visualization technique is widely used in various fields, including biology, finance, and data analysis, to uncover hidden insights and facilitate data-driven decision-making.

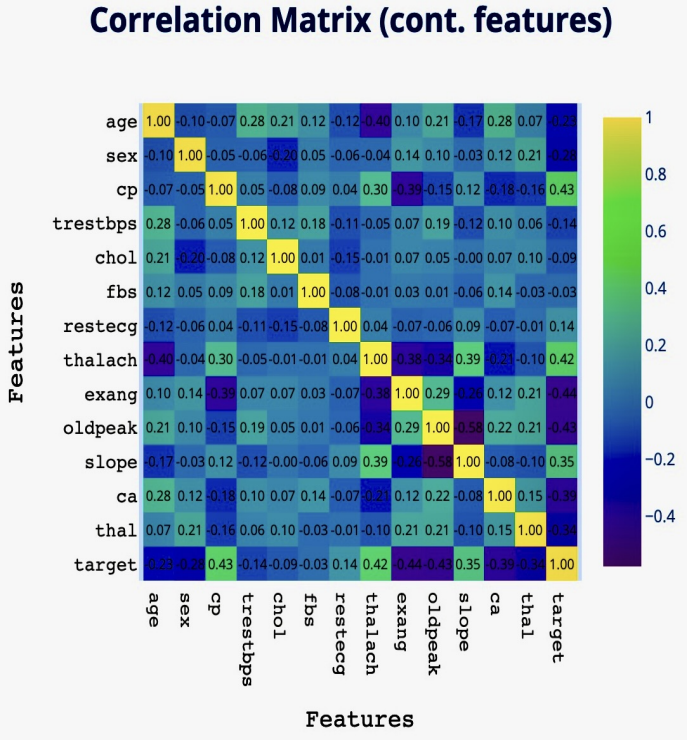

Fig. 11. Variables Correlation Matrix Visualised by Heatmap

One of the key strengths of heatmaps lies in their ability to uncover relationships and structure within data, even when dealing with extensive datasets. For instance, correlation heatmaps reveal the strength and direction of associations between variables, helping researchers and analysts identify variables that influence each other. Heatmaps also excel in depicting spatial information through geographic heatmaps, where color variations convey geographical trends or concentration levels.

VI. Model Result Visualisation by Performance Matrix

A. Logistic Regression

The logistic regression results provide valuable insights into the model's performance for binary classification tasks involving two classes, labeled as 0 and 1 (shown in figure 12). For Class 0, the model exhibits a precision of 0.88, signifying that when it predicts instances as belonging to Class 0, it is accurate approximately 88% of the time. The recall for Class 0 stands at 0.82, indicating that the model correctly identifies roughly 82% of the actual Class 0 instances. The corresponding F1-score for Class 0 is 0.85, representing a balanced measure of precision and recall. Furthermore, there are 28 instances of Class 0 in the dataset.

For Class 1, the logistic regression model demonstrates a precision of 0.86, implying that when it predicts instances as belonging to Class 1, it is accurate approximately 86% of the time. The recall for Class 1 is 0.91, highlighting the model's strong ability to correctly identify about 91% of the actual Class 1 instances. The F1-score for Class 1 stands

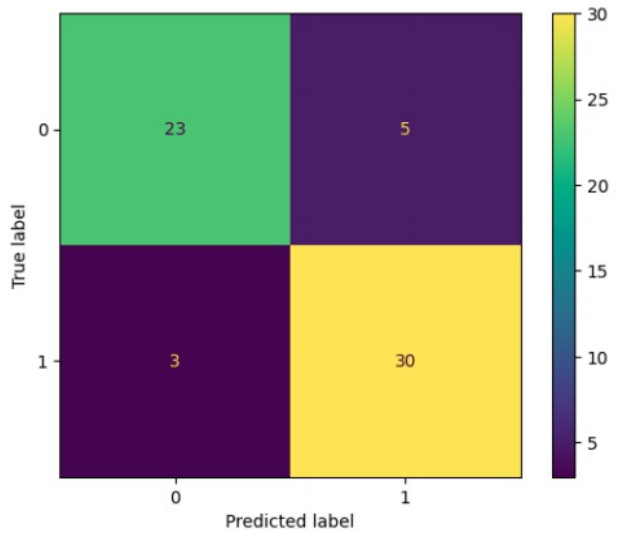

Fig. 12. Confusion Matrix for Logistic Regression

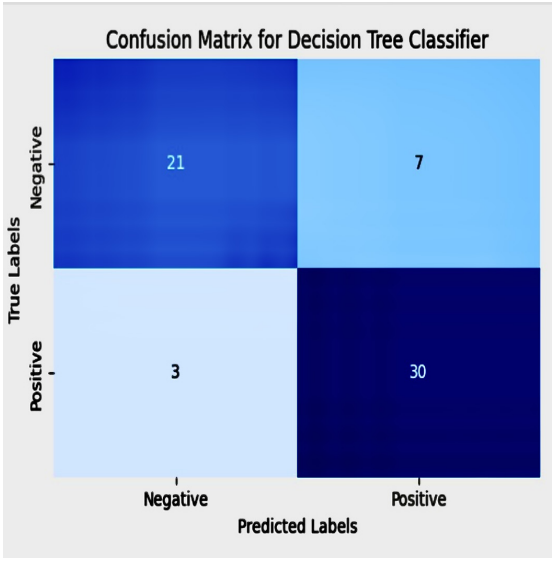

Fig. 13. Confusion Matrix for Decision Tree

at 0.88, reflecting the harmonious blend of precision and recall. In terms of overall accuracy, the model achieves an accuracy score of 0.869, which indicates its ability to make correct predictions across both classes. These performance metrics collectively provide a comprehensive assessment of the logistic regression model's effectiveness in classifying data points into the designated classes.

B. Decision Tree

The resulted accuracy for the decision tree model is 0.8360655737704918, which translates to approximately 83.61%. This accuracy score reflects the model's ability to make correct predictions across the entire dataset.

In other words, the decision tree model correctly classifies roughly 83.61% of the data points, demonstrating its overall effectiveness in performing the specified classification task.

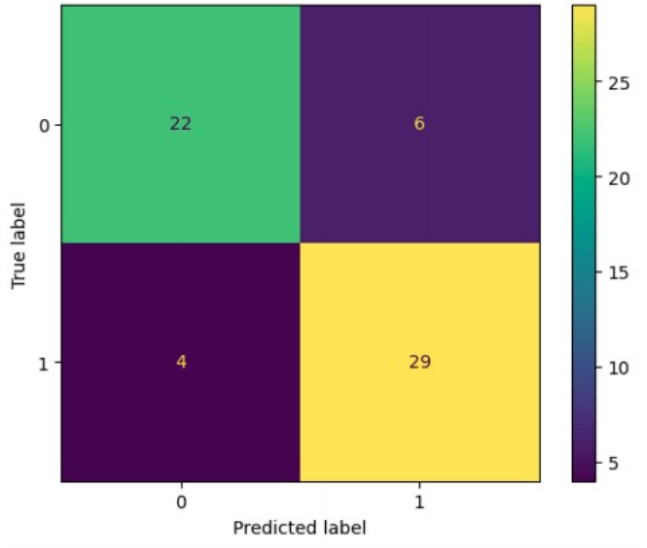

Fig. 14. Confusion Matrix for Random Forest

Accuracy is a fundamental evaluation metric that measures the proportion of correctly predicted instances out of the total dataset, making it an essential indicator of the model's performance.

C. RandomForest

The resulted accuracy for the Random Forest model is also 0.8360655737704918, which is approximately 83.61%. This means that the Random Forest model correctly predicts approximately 83.61% of the data points in the dataset. It's interesting to note that both the Decision Tree and Random Forest models have the same accuracy score, suggesting that they perform equally well in terms of making correct predictions for the given classification task. However, it's important to consider other evaluation metrics and potentially perform a more in-depth analysis to fully assess the strengths and weaknesses of each model and make an informed choice between them for the specific problem at hand as shown in figure 13.

D. Xgboost Cross Validation

The cross-validation results for the XGBoost model offer a comprehensive assessment of its performance in a binary classification task involving classes labeled as 0 and 1. For Class 0, the XGBoost model exhibits a precision of 0.85, indicating that when it predicts instances as belonging to Class 0, it is accurate approximately 85% of the time. The recall for Class 0 stands at 0.82, signifying the model's capability to correctly identify about 82% of the actual Class 0 instances. The corresponding F1-score for Class 0 is 0.84, representing a harmonious balance between precision and recall. Furthermore, there are 28 instances of Class 0 in the dataset.

Turning to Class 1, the XGBoost model demonstrates a precision of 0.85, implying that when it predicts instances as

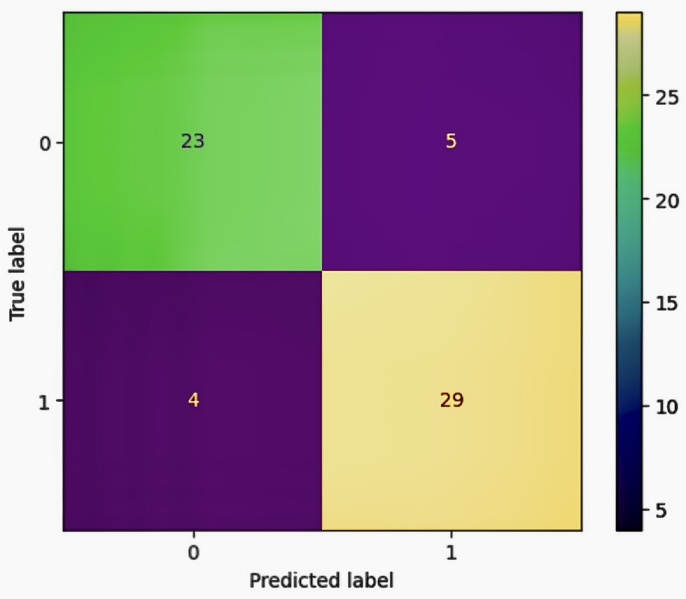

Fig. 15. Confusion Matrix for Xgboost Cross Validation

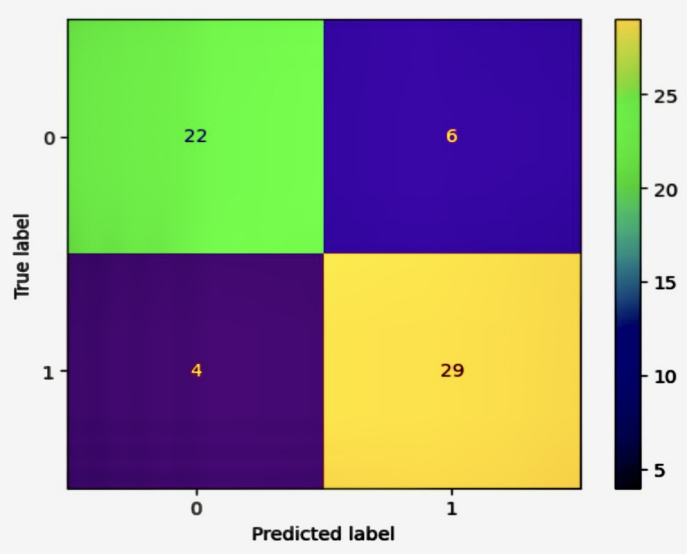

Fig. 16. Confusion Matrix for Ensemble Learning

belonging to Class 1, it is accurate approximately 85% of the time. The recall for Class 1 is 0.88, underscoring the model's robust ability to correctly identify approximately 88% of the actual Class 1 instances. The F1-score for Class 1 is 0.87, reflecting the harmonious blend of precision and recall. In terms of overall accuracy, the XGBoost model achieves an accuracy score of 0.85, which indicates its proficiency in making accurate predictions across both classes. These cross-validation metrics collectively offer a comprehensive evaluation of the XGBoost model's performance and its effectiveness in classifying data points into the specified classes as shown in figure 14.

E. Ensemble learning

XGBoost can be combined with other machine learning algorithms to create ensembles, resulting in improved model performance and generalization. The ensemble learning results provide a detailed assessment of the model's performance in a binary classification task, encompassing two classes labeled as 0 and 1. Regarding Class 0, the ensemble learning model demonstrates a precision of 0.85, indicating that when it predicts instances as belonging to Class 0, it is accurate approximately 85% of the time. However, the recall for Class 0 is slightly lower at 0.79, signifying that the model correctly identifies about 79% of the actual Class 0 instances. The corresponding F1-score for Class 0 is 0.81, representing a balanced measure of precision and recall. This evaluation is conducted on a dataset comprising 28 instances of Class 0.

Turning to Class 1, the ensemble learning model displays a precision of 0.83, implying that when it predicts instances as belonging to Class 1, it is accurate around 83% of the time. The recall for Class 1 is notably higher at 0.88, underscoring the model's robust ability to correctly identify approximately 88% of the actual Class 1 instances. The F1-score for Class 1 stands at 0.85, reflecting a harmonious blend of precision and recall. In terms of overall accuracy, the ensemble learning model achieves an accuracy score of 0.84, indicating its proficiency in making accurate predictions across both classes. These cross-validation metrics collectively offer a comprehensive evaluation of the ensemble learning model's performance, highlighting its effectiveness in classifying data points into the specified classes.

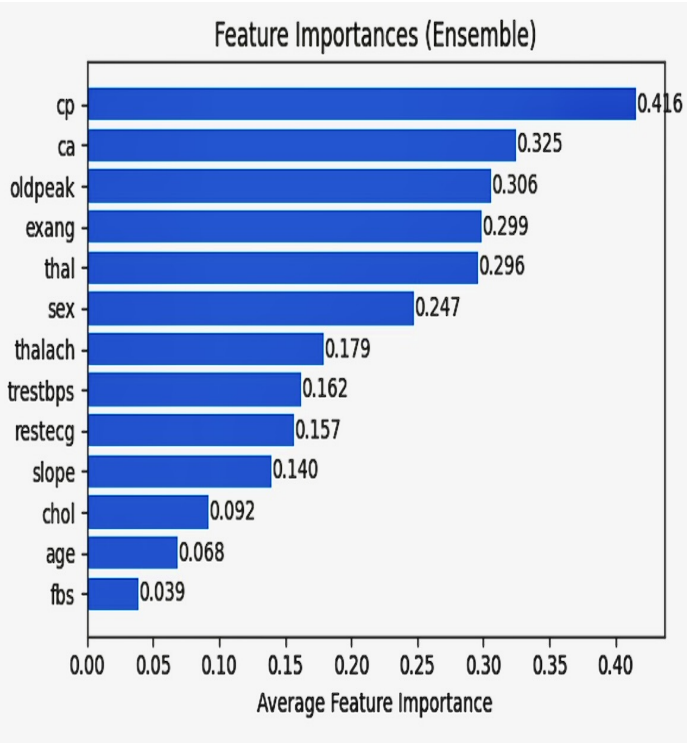

Fig. 17. Feature Importance

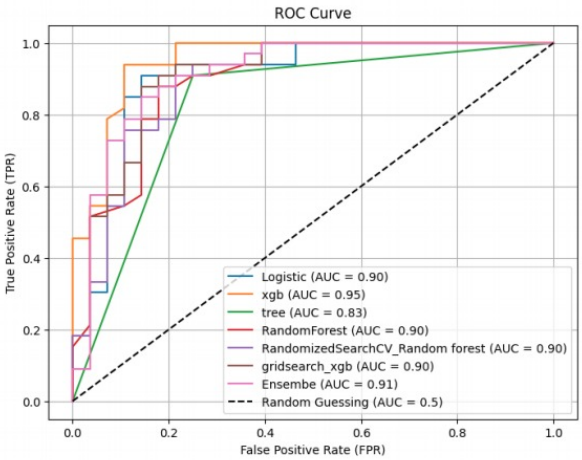

Fig. 18. ROC-AUC curve

F. Feature Importance

G. Metric Evaluation

The evaluation of the predictive models in this study encompassed a comprehensive set of metrics to assess their performance and effectiveness in predicting heart disease. These metrics include: **Accuracy**: Accuracy measures the overall correctness of the model's predictions, indicating the ratio of correctly predicted instances to the total number of instances in the dataset.

Precision: Precision quantifies the model's ability to make accurate positive predictions, representing the ratio of true positives to the sum of true positives and false positives.

Recall: Recall, also known as sensitivity or true positive rate, evaluates the model's capability to correctly identify all positive instances. It is calculated as the ratio of true positives to the sum of true positives and false negatives.

F1-Score: The F1-Score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance, particularly when dealing with imbalanced datasets.

H. Model Perfomance Comparison

According to the ROC-AUC curve (figure 15), the Logistic Regression model achieved a score of 0.90, which corresponds to 90%. This score indicates that the model exhibited strong discriminatory power, with a 90% probability of correctly distinguishing between positive and negative classes based on the area under the ROC curve. A score of 0.90 reflects a high level of accuracy in classifying instances, making Logistic Regression a robust choice for this binary classification task.

The XGBoost model outperformed the others with an impressive ROC-AUC score of 0.95, equivalent to 95%. This exceptional score demonstrates the model's superior ability to make highly accurate binary classification decisions, surpassing all other evaluated models. With a score close to 1.0, XGBoost showcases its remarkable discriminative capability, making it the top-performing model among those considered for this specific classification problem.

The Decision Tree model achieved a ROC-AUC score of 0.83, indicating that it successfully discriminated between positive and negative classes with an 83% accuracy rate. While this score reflects decent classification performance, it falls slightly behind the top-performing models.

The Random Forest model, along with RandomizedSearchCV all achieved a ROC-AUC score of 0.90, equivalent to 90%. These models demonstrated strong discriminatory power, accurately classifying instances with a 90% probability.

The Ensemble model achieved a ROC-AUC score of 0.91, representing a 91% accuracy rate in distinguishing between positive and negative classes. This score signifies the ensemble model's robust performance and its ability to make highly accurate binary classification decisions.

In summary, the ROC-AUC scores provide valuable insights into the discriminatory capabilities of each model, with XGBoost emerging as the top performer, closely followed by the Ensemble model and the Logistic Regression model. These scores serve as a critical metric for model evaluation and selection in binary classification tasks.

VII. Integration of AI Models into Streamlit Web Application

In this section, we provide a detailed report on the integration of AI models into the Streamlit web application for heart disease prediction. This integration aims to make predictive models accessible and user-friendly, allowing users to assess their cardiovascular risk easily and effectively.

A. Rationale for Using Streamlit

The choice of Streamlit as the framework for this web application is driven by several advantages that align with the project's goals:

Simplicity and Accessibility:Streamlit's simplicity and ease of use make it an ideal choice, even for developers without extensive web development experience. This accessibility allows for a faster development cycle.

Seamless Data Science Integration:Streamlit seamlessly integrates with popular data science libraries like Pandas, enabling the presentation of machine learning models and data analysis in a single application.

Real-Time Updates: Streamlit offers real-time updates, making it suitable for dynamic applications where predictions and data visualizations need to be generated and displayed instantly.

Interactivity: Streamlit provides interactive widgets that enable users to input data and explore predictions and visualizations. This interactivity enhances user engagement.

Rapid Prototyping and Deployment: Streamlit facilitates rapid prototyping, reducing development time and effort. It also simplifies deployment, ensuring that the application is accessible to users.

B. Architectural Design

Data pipeline: This component collects and preprocesses the data used to train and evaluate the machine learning

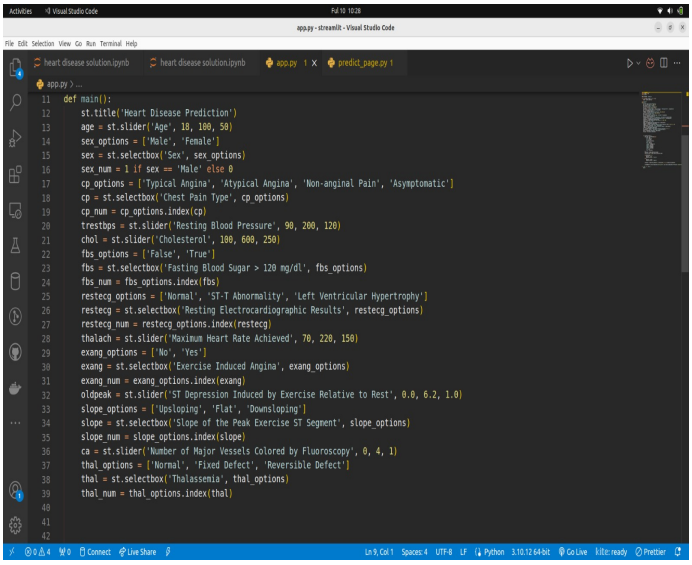

Fig. 19. Model Intergration

model. The data can be collected from a variety of sources, such as electronic health records, clinical trials, and wearable devices. The data pipeline may also include steps to clean and transform the data, as well as to split the data into training and testing sets.

Machine learning model: This component trains and evaluates a machine learning model to perform a specific task. The model can be trained using a variety of machine learning algorithms, such as logistic regression, support vector machines, and random forests.

Streamlit application: This component serves the machine learning model to users through a web application. The Streamlit application allows users to input their data and receive a prediction from the machine learning model.

C. Process of Integration

The technical integration of AI models into the Streamlit app involves the following steps:

Model Loading: Pre-trained machine learning models are loaded into the Streamlit application, ensuring that the models are ready for predictions.

User Input: Streamlit provides interactive widgets that collect user input, such as age, gender, blood pressure, cholesterol levels, and other relevant data.

Prediction: User input data is passed to the integrated models, which generate predictions regarding the likelihood of a major adverse cardiovascular and cerebrovascular event (MACCE).

Result Presentation: The predictions and insights are presented in real-time within the Streamlit app's interface. Users can view predictions, visualizations, and explanations.

D. Challenges and Solutions in Streamlit Integration

Model Compatibility: Ensuring that the models are compatible with the Streamlit environment required careful validation and testing. Compatibility issues were resolved through model adjustments and code optimization.

User Interaction: Ensuring a smooth and intuitive user experience required the design and implementation of interactive widgets and user interfaces. User feedback was considered to enhance usability.

E. Advantages of the Streamlit Web Application

The Streamlit web application offers numerous advantages: **User-Friendly Interface**: The application provides an intuitive interface for users, enabling easy data input, predictions, and data exploration. **Real-Time Predictions**: Predictions are generated in real-time, providing immediate feedback to users. **Interactivity**: Users can interact with the application through widgets, gaining insights into the factors influencing predictions. **Scalability**: The application is scalable, allowing for future enhancements, additional features, and updated datasets.

F. Description and Demo of the Streamlit AI App

Link provided: https://heart-disease-webapp.streamlit.app/

VIII. Validation and Comparison

In this section, we conducted a comprehensive comparative analysis of the results generated by the integrated machine learning models. We validated the performance of these models and compared them against previous works or established benchmarks in the field of heart disease prediction. Our aim was to provide insights into the effectiveness of the models and their predictive accuracy.

IX. Discussion

In this section, we delved into several critical aspects related to the robustness, generalizability, and real-world utility of our models:

Robustness to Different Data: We assessed the robustness of our models to different kinds of data. This involved examining how well the models performed when presented with data from diverse sources or patient populations. We discussed the models' ability to adapt and maintain predictive accuracy in various scenarios.

Generalizability: We reflected on the generalizability of our model results to other datasets or contexts. This was particularly important for Q1 journals, as they sought contributions that offered knowledge applicable beyond the immediate dataset. We discussed the potential for our models to be applied to different healthcare settings and populations, emphasizing the transferability of our findings.

User Study or Feedback: Given the development of our Streamlit application, we conducted a user study involving clinicians, stakeholders, and individuals who interacted with the application. We sought feedback to evaluate the real-world utility of our application and gathered suggestions for improvement. In this section, we presented the feedback received, highlighting the practical implications of our application in healthcare settings. We also discussed any valuable suggestions for enhancing user experience or the application's functionality.

X. Ethical Considerations

Ethical considerations were paramount in healthcare-related research. In this section, we address the ethical aspects of our study, including data privacy, informed consent, and responsible AI practices. We discussed how we ensured the protection of sensitive patient information, how we obtained necessary approvals, and the steps taken to minimize bias and discrimination in our models. Ethical considerations were crucial for the responsible deployment of AI in healthcare and were discussed comprehensively in this section.

XI. conclusion

In summary, this study has contributed significantly to the field of cardiovascular disease prediction through the development of a robust and accurate predictive model. However, it is imperative to acknowledge several key aspects for a more comprehensive understanding of the study's implications.

Limitations: While the model's performance is promising, we must recognize the limitations and constraints of this study. One crucial limitation pertains to the dataset's origin and representativeness, which may influence the model's generalizability to broader populations. Moreover, the study's reliance on retrospective data raises the possibility of unmeasured confounders or biases that could impact predictive accuracy. These limitations underscore the need for cautious interpretation and further investigation into the model's performance across diverse patient demographics and healthcare settings.

Practical Implications: The significance of this research extends to practical applications in healthcare. By providing accurate predictions for Major Adverse Cardiovascular and Cerebrovascular Events (MACCE), the model offers valuable support for healthcare professionals in risk assessment and early intervention. Patients, particularly those at higher risk of MACCE, stand to benefit from more timely and targeted medical attention, potentially reducing the burden of cardiovascular diseases on individuals and healthcare systems.

Future Work: The path forward involves several avenues for future research. Firstly, the model's generalizability should be rigorously tested across different datasets and populations to ascertain its broader applicability. Additionally, prospective studies and clinical trials are warranted to validate the model's effectiveness in real-world healthcare settings. Further enhancements in feature selection and data preprocessing techniques may improve the model's accuracy and robustness. Lastly, the incorporation of patient-specific variables and external factors could enhance the model's predictive power and clinical utility.

Generalizability and Stakeholder Impact: It is crucial to emphasize that while this study has made significant strides in cardiovascular disease prediction, the model's results should be considered within the context of the dataset used. Future research should aim to validate and generalize these findings to ensure their applicability across diverse healthcare scenarios. Ultimately, the beneficiaries of this research encompass a broad spectrum of stakeholders, including patients who stand to benefit from improved risk assessment, healthcare providers equipped with a valuable decision support tool, and healthcare institutions striving for more effective and efficient cardiovascular disease management. This collaborative effort holds the potential to make a substantial impact on public health by enhancing early detection and intervention in cardiovascular diseases.

XII. Acknowledgments

We extend our heartfelt gratitude to the entire research team for their unwavering commitment and dedication throughout this study. We also express our appreciation to the healthcare institutions and individuals who generously provided access to the valuable patient data used in this research. Their contributions were instrumental in the success of this project.

We acknowledge the support and guidance received from our mentors and advisors, whose expertise and insights greatly enriched our understanding and approach to cardiovascular disease prediction.

Additionally, we would like to thank the participants of our user study, clinicians, stakeholders, and users of the Streamlit application for their valuable feedback, which played a pivotal role in shaping the application's usability and functionality.

Finally, we are grateful to the broader scientific community for their ongoing research and contributions to the field of heart disease prediction, which provided valuable context and benchmarks for our study.

XIII. references

[1] Cardiovascular Diseases. Available online: https://www.who.int/health-topics/cardiovascular-diseases/ (accessed on 10 September 2022).

[2]Shah, S.; Shah, F.; Hussain, S.; Batool, S. Support Vector Machines-based Heart Disease Diagnosis using Feature Subset, Wrapping Selection and Extraction Methods. Comput. Electr. Eng. 2020, 84, 106628. [Google Scholar] [CrossRef]

[3]Che, C.; Zhang, P.; Zhu, M.; Qu, Y.; Jin, B. Constrained transformer network for ECG signal processing and arrhythmia classification. BMC Med. Inform. Decis. Mak. 2021, 21, 184. [Google Scholar] [CrossRef]

[4]Hoodbhoy, Z.; Jiwani, U.; Sattar, S.; Salam, R.; Hasan, B.; Das, J. Diagnostic Accuracy of Machine Learning Models to Identify Congenital Heart Disease: A Meta-Analysis. Front. Artif. Intell. 2021, 4, 197. [Google Scholar] [CrossRef]

[5]Wang, Z.; Chen, L.; Zhang, J.; Yin, Y.; Li, D. Multi-view ensemble learning with empirical kernel for heart failure mortality prediction. Int. J. Numer. Methods Biomed. Eng. 2020, 36, e3273. [Google Scholar] [CrossRef]

[6]Modepalli, K.; Gnaneswar, G.; Dinesh, R.; Sai, Y.R.; Suraj, R.S. Heart Disease Prediction using Hybrid machine Learning Model. In Proceedings of the 2021 6th International Conference on Inventive Computation Technologies

(ICICT), Coimbatore, India, 20–22 January 2021. [Google Scholar]
[7]Joo, G.; Song, Y.; Im, H.; Park, J. Clinical Implication of Machine Learning in Predicting the Occurrence of Cardiovascular Disease Using Big Data (Nationwide Cohort Data in Korea). IEEE Access 2020, 8, 157643–157653. [Google Scholar] [CrossRef]
[8]Li, J.; Haq, A.; Din, S.; Khan, J.; Khan, A.; Saboor, A. Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare. IEEE Access 2020, 8, 107562–107582. [Google Scholar] [CrossRef]
[9]Ali, F.; El-Sappagh, S.; Islam, S.M.R.; Kwak, D.; Ali, A.; Imran, M.; Kwak, K. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Inf. Fusion 2020, 63, 208–222. [Google Scholar] [CrossRef]
[10]Rahim, A.; Rasheed, Y.; Azam, F.; Anwar, M.; Rahim, M.; Muzaffar, A. An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases. IEEE Access 2021, 9, 106575–106588. [Google Scholar] [CrossRef]
[11]Ishaq, A.; Sadiq, S.; Umer, M.; Ullah, S.; Mirjalili, S.; Rupapara, V.; Nappi, M. Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques. IEEE Access 2021, 9, 39707–39716. [Google Scholar] [CrossRef]
[12]Khurana, P.; Sharma, S.; Goyal, A. Heart Disease Diagnosis: Performance Evaluation of Supervised Machine Learning and Feature Selection Techniques. In Proceedings of the 8th International Conference on Signal Processing and Integrated Networks, SPIN 2021, Matsue, Japan, 18–22 October 2021. [Google Scholar]
[13]Ashri, S.E.A.; El-Gayar, M.M.; El-Daydamony, E.M. HDPF: Heart Disease Prediction Framework Based on Hybrid Classifiers and Genetic Algorithm. IEEE Access 2021, 9, 146797–146809. [Google Scholar] [CrossRef]
[14]Bashir, S.; Almazroi, A.; Ashfaq, S.; Almazroi, A.; Khan, F. A Knowledge-Based Clinical Decision Support System Utilizing an Intelligent Ensemble Voting Scheme for Improved Cardiovascular Disease Prediction. IEEE Access 2021, 9, 130805–130822. [Google Scholar] [CrossRef]
[15]Odhiambo Omuya, E.; Onyango Okeyo, G.; Waema Kimwele, M. Feature Selection for Classification using Principal Component Analysis and Information Gain. J. Biomed. Inform. 2021, 174, 114765. [Google Scholar] [CrossRef]
[16]Le, T.; Lee, M.; Park, J.; Baik, S. Oversampling techniques for bankruptcy prediction: Novel features from a transaction dataset. Symmetry 2018, 10, 79. [Google Scholar] [CrossRef][Green Version]
[17]Vandewiele, G.; Dehaene, I.; Kovács, G.; Sterckx, L.; Janssens, O.; Ongenae, F.; Backere, F.D.; Turck, F.D.; Roelens, K.; Decruyenaere, J.; et al. Overly optimistic prediction results on imbalanced data: A case study of flaws and benefits when applying over-sampling. Artif. Intell. Med. 2021, 111, 101987. [Google Scholar] [CrossRef] [PubMed]
[18]Xu, Z.; Shen, D.; Nie, T.; Kou, Y. A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. J. Biomed. Inform. 2020, 107, 103465. [Google Scholar] [CrossRef] [PubMed]
[19]Budholiya, K.; Shrivastava, S.; Sharma, V. An optimized XGBoost based diagnostic system for effective prediction of heart disease. J. King Saud-Univ.–Comput. Inf. Sci. 2020, 34, 4514–4523. [Google Scholar] [CrossRef]
[20]Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016. [Google Scholar]
[21]Asadi, S.; Roshan, S.; Kattan, M.W. Random forest swarm optimization-based for heart diseases diagnosis. J. Biomed. Inform. 2021, 115, 103690. [Google Scholar] [CrossRef]
[22]Bansal, M.; Goyal, A.; Choudhary, A. A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. Decis. Anal. J. 2022, 3, 100071. [Google Scholar] [CrossRef]
[23]Ksiażek, W.; Gandor, M.; Pławiak, P. Comparison of various approaches to combine logistic regression with genetic algorithms in survival prediction of hepatocellular carcinoma. Comput. Biol. Med. 2021, 134, 104431. [Google Scholar] [CrossRef]
[24]Ghiasi, M.M.; Zendehboudi, S.; Mohsenipour, A. Decision tree-based diagnosis of coronary artery disease: CART model. Comput. Methods Prog. Biomed. 2020, 192, 105400. [Google Scholar] [CrossRef]
[25]Chen, S.; Webb, G.I.; Liu, L.; Ma, X. A novel selective naïve Bayes algorithm. Knowl.-Based Syst. 2020, 192, 105361. [Google Scholar] [CrossRef]