Hybrid Ensemble Based Machine Learning Approach for Cardiovascular Disease Risk Prediction Using Multiple Integrated Datasets

Sravanthi Jakkula
Department of Computer Science and Engineering
National Institute of Technology, India
sravs521@gmail.com

Raju Bhukya
Department of Computer Science and Engineering
National Institute of Technology, India
raju@nitw.ac.in

Abstract: *Cardiovascular Diseases (CVD) are one of the significant reasons for human mortality across the globe. Hence, more accurate and efficient models predicting the early stages of these diseases must be developed. In this research work, an attempt has been made to develop an appropriately huge and heterogeneous dataset after merging the three different datasets from IEEE, UCI and Kaggle sites. Several machine learning algorithms such as Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Extra Trees (ET), Extreme Gradient Boost (XGB), gradient boosting, AdaBoost, and Multi-Layer Perceptron (MLP) have been employed on this integrated dataset. To improve the prediction accuracy even stacked models were employed in order to accomplish the objective of the research. The optimal combination of base models was gradient boosting, ET, and XGB with LR acting as the meta-model, yielding a high accuracy of 99.78% compared to the existing models. Such performances placed the meta-model far from the performance of the other models, which were found to be significantly erroneous in their outputs as compared to the former. This investigation demonstrates how datasets can be effectively merged to improve the generalization potential of a model and how ensemble and stacking methods could be used. The results present a comprehensive approach in building robust CVD prediction systems showing how sophisticated machine learning techniques can enhance Implementation overall decision.*

Keywords: *Cardiovascular diseases, machine learning algorithms, stacked models, disease prediction.*

Received March 29, 2025; accepted September 21, 2025
https://doi.org/10.34028/iajit/23/2/2

1. Introduction

According to the World Health Organization (WHO), Cardiovascular Disease (CVD) accounts for nearly 32% of global deaths, making it one of the leading causes of morbidity and mortality worldwide [43]. These diseases comprise a multitude of conditions affecting the organs and vessels supplying blood to the heart, including myocardial infarction, cerebrovascular accident, congestive heart failure, arrhythmias, and coronary artery disease. A major challenge in dealing with CVD is that these diseases are often insidious in onset and, therefore, may result in sudden dramatic health alterations that threaten the life of the individual [4]. Many patients also tend to ignore early warning signs or delay medical consultation until a more serious event has happened. Preventive strategies and early detection increase the chances of reducing mortality rates.

There are so many risk factors that help in the initiation of CVDs; these are either modifiable or non-modifiable. The modifiable risk factors comprise unhealthy diets, physical inactivity, obesity, smoking, alcohol consumption in excess, and chronic stress. Many studies have shown that high consumption of processed and junk foods that are rich in trans fats and refined sugars significantly increases the risk of developing coronary artery disease [34]. Lacking physical exercise helps develop overweight, which, along with metabolic syndrome, are the main forerunners to cardiovascular complications [28]. Furthermore, both smoking and excess alcohol consumption were found to have direct implications toward increased blood pressure, arterial stiffness, and myocardial infarction risks [15]. Besides lifestyle, new environmental factors have been on the rise, such as air pollution, which has been linked to cardiovascular dysfunction; prolonged exposure to pollutants increases an individual's risk of getting a heart attack or stroke [9].

On another side, non-modifiable risk factors such as genetic predisposition, age, and gender remain critical for determining individual susceptibility to heart disease. A strong family history of CVD can impart a greater risk of developing the condition at an earlier age than someone without that recent family history [26]. Studies have shown that men are found to generally carry a higher CVD risk than women of the same age who have not yet begun menopause; however, their risk post-menopause markedly escalates probably due to hormonal alteration [30]. Other people with existing health conditions such as diabetes, hypertension, and hyperlipidemia have an increased risk of developing CVDs, requiring appropriate management of the disease

[12].

Like any other chronic ailment, CVDs significantly impact health and quality of life in the long run. Chronic diseases like hypertension and heart failure lead to cognitive decline, memory loss, and damage to vital organs like the kidneys and liver [17]. Also, it is established that high blood pressure increases the risk of vascular dementia, marking the significance of best cardiovascular health [27].

In addition to modern sedentary lifestyles, urbanization, and increasing metabolic disorders such as diabetes and obesity, the other increasing cause of CVD s is the fast-spreading urban culture. There has been nearly a four-fold increase in diabetes prevalence globally in the last four decades, making a significant contribution to the increased prevalence of CVDs [40]. Not only individuals, but also families and health systems suffer emotional and financial stress [16]. The economic costs of treating CVDs are enormous. It is estimated that the annual cost of treatment and intervention of CVD exceeds billions of dollars in global healthcare expenditure [37].

In view of the complex, multi-factorial nature of CVDs, timely detection and early intervention are mandatory. Regular health screening, lifestyle modification, and public health initiatives aimed at improving heart health can often lead to a significant reduction in the burden of CVDs [44]. Now with the advancement in technology, and especially with artificial intelligence and machine learning, it becomes possible to design such predictive models that can identify at-risk individuals more accurately, which will help early diagnosis as well as timely medical intervention [42]. This research is meant to leverage machine learning techniques to enhance CVD prediction and building a strong federated model to integrate diverse data sources for improving predictive accuracy.

In light of the rising burden and complexity of CVDs, there is a pressing need for technological interventions that can facilitate early and accurate diagnosis. Prediction models using artificial intelligence and machine learning have emerged as powerful tools in this regard. Specifically, ensemble-based methods such as stacking offer enhanced prediction accuracy by integrating multiple base models. This study aims to develop a robust and generalized stacking-based ensemble model for CVD prediction, utilizing integrated multi-source datasets from UCI, Kaggle, and IEEE. This approach not only reinforces model performance but also contributes to the novelty of the research by promoting model generalizability across heterogeneous data.

2. Related Work

Recent studies have demonstrated the effectiveness of Machine Learning (ML) techniques in identifying various diseases, including COVID-19 through X-ray images [20, 33], tumors on MRI scans [10, 32], heart disease prediction [1, 36], dengue diagnosis [14, 35], stroke detection [8], and certain types of cancer [29]. These advancements highlight the growing role of ML in the healthcare industry, where computational models aid in disease detection, early diagnosis, and clinical decision-making. As the field evolves, researchers have explored different ML methodologies to improve the predictive performance of heart disease classification models, incorporating both supervised and unsupervised learning techniques. Alam *et al*. [2] proposed a deep Convolution Neural Network (CNN)-based approach with feature fusion to classify non-small cell lung cancer from histopathological images.

One such study by Maini *et al*. [31] focused on the Cleveland heart disease dataset and applied unsupervised learning and clustering techniques to improve the diagnostic process. Their approach demonstrated that employing RapidMiner, in conjunction with MATLAB, Weka, and Artificial Neural Networks (ANN), significantly enhanced model performance. The best accuracy recorded in their study was 90.74%, proving that unsupervised learning strategies can play a crucial role in disease classification when data lacks labeled outputs. Usman *et al*. [41] proposed a machine learning-based method to improve prediction of human heart disease by optimizing and comparing multiple algorithms.

Their approach demonstrated enhanced predictive performance for heart disease risk using clinical data and trained classifiers.

Similarly, Kavitha *et al*. [24] explored CVD prediction using a different technique that combined regression and classification models for data processing. By working with the Cleveland cardiopathy dataset, their hybrid approach recorded an 88.7% accuracy in predicting Coronary Heart Disease (CHD). This study emphasized the importance of using hybrid models that integrate multiple machine learning techniques for better accuracy and generalization.

In another comparative study, Kataria *et al*. [23] analyzed various machine learning algorithms for heart disease prediction, assessing their relative efficiency in identifying cardiovascular risks. Their evaluation of supervised learning models for heart disease diagnosis showed that Logistic Regression (LR) achieved the highest accuracy of 93.4%. Their findings suggested that simple models like LR, when trained on well-preprocessed datasets, can outperform complex models in certain clinical scenarios.

Further advancements in ML-based heart disease detection were made by Shah *et al*. [38], who employed supervised learning classification models using a pre-existing Cleveland heart disease dataset from the UCL repository. Their research highlighted the advantages of supervised learning models in disease detection, particularly when combined with feature selection

techniques to improve performance.

Deep learning approaches have also been explored in CVD risk prediction. Bharti *et al*. [5] incorporated deep learning techniques alongside conventional machine learning models to enhance coronary heart disease prediction. They worked with a supervised learning dataset containing 14 attributes, demonstrating that deep learning methods achieved an average predictive accuracy of 94.2%. Their findings suggest that deep learning architectures can significantly enhance prediction capabilities, particularly in complex, high-dimensional medical datasets.

Another study by Ashish *et al*. [3] introduced a hybrid classification model combining Support Vector Machine (SVM) and Extreme Gradient Boost (XGB) boosting algorithms. Their model provided fast and reliable detection of coronary heart disease, utilizing Random Forest (RF) for training and testing. The N2Genetic-nuSVM model, built using data from the Z-Alizadeh Sani dataset, achieved an impressive 93.08% accuracy in predicting clinical heart disease outcomes. Their study underscored the advantages of combining boosting algorithms with support vector-based classification methods for highly accurate and reliable results in medical diagnostics.

Bhukya [6] Suggested a new way to guess gene activity by first shrinking large sets of gene data using deep autoencoders. Then, a type of neural network Multi-Layer Perceptron (MLP) is used to make predictions from that smaller set. This method works better because it finds hidden patterns and makes the data easier to handle.

Gugulothu and Bhukya [18] present a deep learning method optimized using a hybrid algorithm called coot-lion to predict how fast point mutations happen in COVID-19 genomes. By analyzing genetic data, the model aims to detect and predict virus mutations more accurately. The hybrid optimization improves model performance and prediction accuracy.

Dasari and Bhukya [13] proposed a deep learning model to predict new viral genomes with a focus on explainability making the model's decisions easier to understand. It helps researchers not only detect unknown viruses from genome data but also understand why the model made a certain prediction. This is important for building trust and aiding scientific discovery in virus research.

Bhukya *et al*. [7] present a hybrid deep learning model with attention mechanisms to accurately find transcription factor binding sites important regions in Deoxyribonucleic Acid (DNA) where proteins attach to control gene activity. The attention layer helps the model focus on the most important parts of the DNA sequence, improving both prediction accuracy and interpretability.

Kartheek *et al*. [22] proposes a new method for recognizing facial expressions by analyzing texture patterns in images using symbolic features. The technique captures subtle facial changes more effectively, helping machines understand emotions from faces with improved accuracy.

Bhukya [6] suggested a new way to guess gene activity by first shrinking large sets of gene data using deep autoencoders. Then, a type of neural network MLP is used to make predictions from that smaller set. This method works better because it finds hidden patterns and makes the data easier to handle.

Siddhartha [39] introduced the heart disease dataset (comprehensive) by integrating five well-known heart disease datasets into a single unified resource. The dataset contains 1190 patient records with 11 common clinical attributes, enabling robust analysis of coronary artery disease. It was created to overcome limitations of small and fragmented datasets used in earlier studies. In [19] the heart disease prediction dataset available on Kaggle is a publicly accessible dataset designed to support research in CVD prediction using machine learning techniques. It contains patient health records with key clinical attributes such as age, gender, blood pressure, cholesterol, and chest pain type. The dataset is commonly used to train and evaluate classification models for identifying the presence or absence of heart disease.

Ensemble methods such as bagging, boosting, and Stacking have proven to be highly effective in enhancing the performance and generalization of machine learning models. bagging (e.g., RF) helps reduce variance, while boosting (e.g., AdaBoost, gradient boosting, XGB improves bias. Stacking, on the other hand, combines multiple classifiers through a meta-model, thereby leveraging the strengths of individual learners. In the context of medical diagnosis, especially for CVDs, ensemble models have been shown to outperform single classifiers in terms of accuracy and robustness. For example, studies like those by Shah *et al*. [38] and Ashish *et al*. [3] demonstrate the efficacy of hybrid and ensemble-based approaches. Furthermore, integrating datasets from multiple sources has been shown to enhance model performance by providing more diverse and representative samples, which helps avoid overfitting and supports broader applicability of the model in real-world scenarios.

3. Dataset Analysis

In this study, three open-source public datasets have been used in developing a well-constructed predictive model for CVD. These are the UCI heart disease dataset [5], complete heart disease dataset [3], and heart disease prediction dataset by Janosi *et al*. [21]. These datasets feature some fundamental clinical parameters such as age, cholesterol level, blood pressure, heart rate, and chest pain type that are basically needed for determining the risk of CVD. After blending these datasets, they formed a composite dataset consisting of 4,728 records and 14 attributes-with 11 independent features and 1

target variable denoting the presence or absence of CVD. The details of the individual datasets and their attributes are shown in Table 1. Merging multiple datasets would provide well generalized and diversified dataset, thus reducing biases built into every individual dataset, which serves to improve the predictive model in generalizing to unseen data.

Table 1. Cardiovascular disease data summary.

Source	Columns	Rows
UCI	(14 attributes) age, sex, chest_pain_type, resting_blood_pressure, cholesterol, fasting_blood_sugar, resting_ecg, max_heart_rate, exercise_induced_angina, st_depression, st_slope, number_of_major_vessels, thalassemia, target	**303**
IEEE	(12 attributes) age, sex, chest_pain_type, resting_blood_pressure, cholesterol, fasting_blood_sugar, resting_ecg, max_heart_rate, exercise_induced_angina, st_depression, st_slope, target	**1190**
Kaggle	(12 attributes) age, sex, chest_pain_type, resting_blood_pressure, cholesterol, fasting_blood_sugar, resting_ecg, max_heart_rate, exercise_induced_angina, st_depression, st_slope, target	**3235**

Table 2. Statistical analysis of datasets.

Features	Dataset	Mean	Std Dev	Min	Max
Age	Combined	53.72	9.36	28.0	77.0
	IEEE	54.20	9.16	28.0	77.0
	Kaggle	54.36	9.08	29.0	77.0
	UCI	54.09	9.21	28.0	77.0
Sex	Combined	0.76	0.42	0.0	1.0
	IEEE	0.73	0.44	0.0	1.0
	Kaggle	0.68	0.47	0.0	1.0
	UCI	0.73	0.44	0.0	1.0
Chest pain type	Combined	3.23	0.94	1.0	4.0
	IEEE	2.63	1.26	0.0	4.0
	Kaggle	0.97	1.03	0.0	3.0
	UCI	2.67	1.28	0.0	4.0
Resting blood pressure	Combined	132.15	18.37	0.0	200.0
	IEEE	132.03	18.09	0.0	200.0
	Kaggle	131.62	17.54	94.0	200.0
	UCI	132.03	18.12	0.0	200.0
Cholesterol	Combined	210.36	101.42	0.0	603.0
	IEEE	229.64	81.43	0.0	603.0
	Kaggle	246.26	51.83	126.0	564.0
	UCI	225.85	85.98	0.0	603.0
Fasting blood sugar	Combined	0.21	0.41	0.0	1.0
	IEEE	0.17	0.38	0.0	1.0
	Kaggle	0.15	0.36	0.0	1.0
	UCI	0.18	0.39	0.0	1.0
Resting ECG	Combined	0.70	0.87	0.0	2.0
	IEEE	0.66	0.78	0.0	2.0
	Kaggle	0.53	0.53	0.0	2.0
	UCI	0.66	0.79	0.0	2.0
Max heart rate	Combined	139.73	25.52	60.0	202.0
	IEEE	144.45	24.85	60.0	202.0
	Kaggle	149.65	22.91	71.0	202.0
	UCI	143.60	25.03	60.0	202.0
Exercise angina	Combined	0.39	0.49	0.0	1.0
	IEEE	0.38	0.49	0.0	1.0
	Kaggle	0.33	0.47	0.0	1.0
	UCI	0.38	0.49	0.0	1.0
ST depression	Combined	0.92	1.09	-2.6	6.2
	IEEE	1.71	3.98	-2.6	62.0
	Kaggle	1.04	1.16	0.0	6.2
	UCI	1.47	3.37	-2.6	62.0
ST slope	Combined	1.62	0.61	0.0	3.0
	IEEE	1.86	0.71	0.0	3.0
	Kaggle	1.40	0.62	0.0	2.0
	UCI	1.77	0.70	0.0	3.0

Table 2 presents the statistical analysis of the datasets, summarizing the different metrics for various features. Data is analyzed across four datasets: Combined, IEEE, Kaggle, and UCI. This provides an overview of the distribution and range of each feature within the respective datasets. The table highlights key differences and similarities in feature values across datasets, which could influence model performance. Such insights are crucial for understanding the variability and consistency of data used in predictive modeling and analysis.

Since real-world datasets often contain outliers, missing values, inconsistencies, and noise, data preprocessing became one of the important procedures that supplied reliability and efficiency of the model. Multiple preprocessing techniques were used to cleanse and prepare the data for training. Data cleaning is the first of all processes and involves dealing with missing data, correcting inconsistencies as well as eliminating duplicates from the dataset to ensure the source of integrity of the dataset. This was followed by data transformation when the numerical variables were normalized across the similar scales to avoid outcomes dictated entirely by some features with very large ranges. Refinement of the dataset was also through Aggregation techniques. An equally critical preprocessing step was data integration, where data from the three disparate sources were integrated into a single, cohesive dataset. The original datasets, however, had some structures variations as two had 12 attributes each while one had 14 attributes. Their standardization was thus required. The extra columns from the dataset with the 14 attributes were identified and removed in the efforts to achieve consistency across all datasets. After column alignments, merging of the datasets was done to enable a well-formed dataset that is useful in analysis and improved predictive models.

To improve the model performance and computational efficiency, data reduction techniques were performed. Reducing complexity helps improve processing time and increases accuracy by maintaining the critical information in the dataset. Feature selection was an important aspect; retaining only the most relevant attributes avoids redundancy or lesser significance. This ensures that the machine learning model is concentrating on the most important indicators of CVD in turn enhancing the accuracy while minimizing computational burden. Furthermore, data visualization techniques were applied to gain insights into the relationships between features and to locate possible correlations or patterns within the dataset. The dataset was illustrated intuitively using various plots and graphs, including bar plots, scatter plots, and pair plots.

The illustrations (see Figures 1 to 6) provide an understanding of data distributions, trends, and relationships among features. Recognition of such patterns is helpful in making informed decisions regarding feature selection and model tuning. Analysis of target distribution implies one of the most important

aspects of this investigation, as it helps determine the balance of data across different classes and alleviate issues of imbalance that might endanger the model performance. Figure 3 elucidates how the target variable is distributed across the three datasets and the amalgamated dataset, which, if significantly imbalanced, would apply any of the techniques of oversampling, under sampling, or synthetic data generation for curbing model biases toward the dominant class.

Because numerical characteristics are so important for predicting CVD, age, cholesterol, and blood pressure were all related to individual consideration. These parameters work as direct indicators of one's cardiovascular status. These features were standardized or normalized to some scale to eliminate dominance: an individual feature should not influence the model's output. Such construction for numerical features is shown in Figure 4, while Figure 7 pairs several pair plots that visualize relationships between different numerical variables.

Another important pre-processing step is feature selection, which defines the most important attributes to improve model accuracy and efficiency.

By choosing the most important features, the model can easily generalize due to the elimination of unnecessary noise and redundancy from data. These and others like age, cholesterol, and blood pressure were found to be some of the most predictive attributes for CVD diagnosis, as presented in Figure 4. Feature selection not only improves performance but also saves computational complexity, adding to the efficiency and interpretability of the predictive model.

This study will guarantee an optimized dataset for strong CVD prediction in the model training process by integrating and thoroughly preprocessing the data while selecting and visualizing features. This forms the premise for building a well-accurate and efficient machine learning model capable of predicting the risk of CVD early, finally leading to proper healthcare intervention and good outcomes for patients.

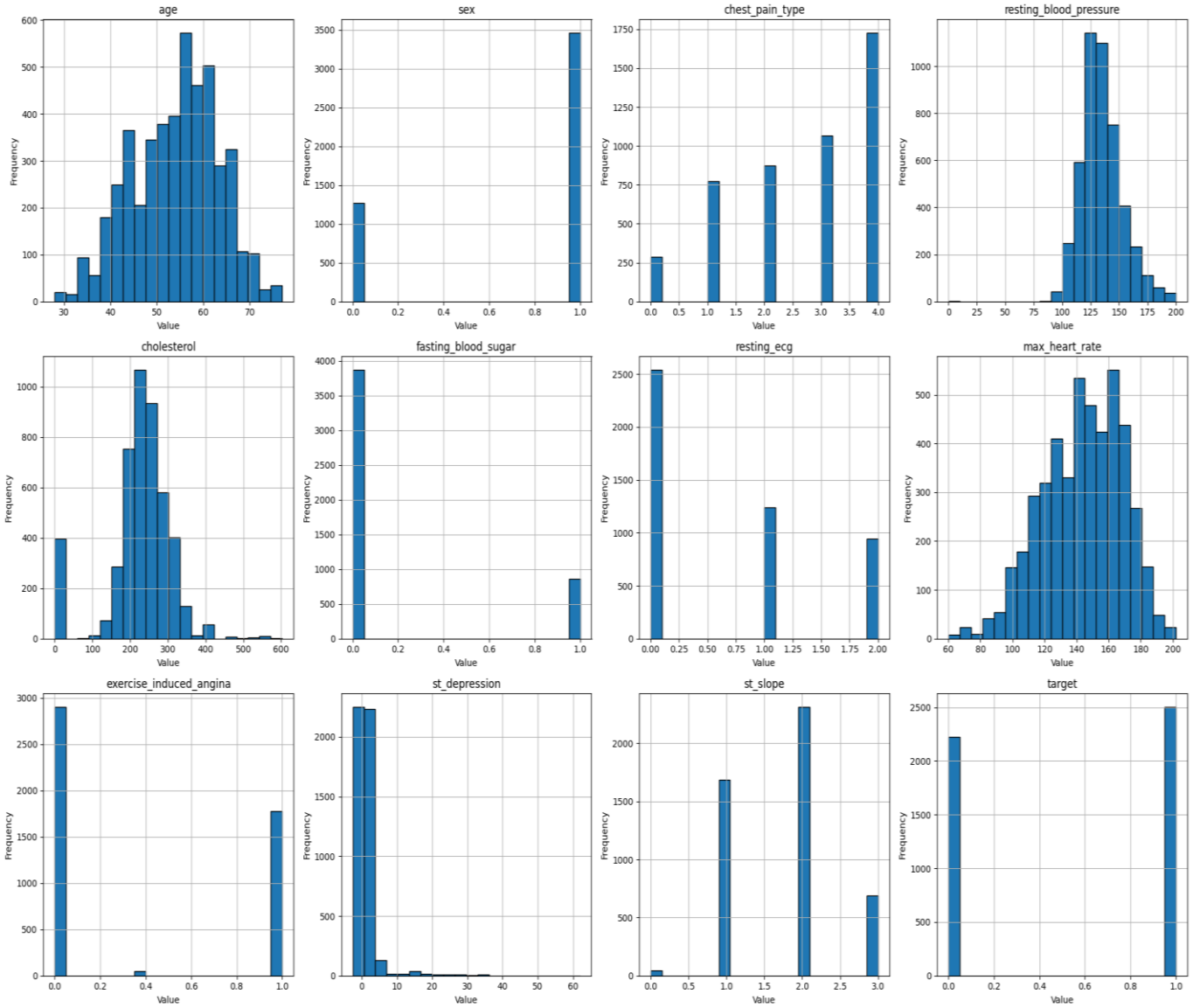

Figure 1. This is a grid of histograms representing the attributes in the dataset. Each plot demonstrates the frequency of values of features representing both categorical (sex, chest pain type, target) and continuous variables (age, cholesterol).

These numerous histograms display the distribution of clinical and demographic characteristics or variables for heart disease, assisting in pointing out important risk factors. The age variable shows a normal distribution, mostly between 40 and 70, which tends to be reflective of heart disease prevalence among older and middle-aged persons. The sex distribution shows that there are lots of males compared to females, suggesting a possible risk factor attributable to sex. Types of chest pain are distributed over four categories, portraying diverse symptoms. Max resting blood pressure at around 120-140 mmHg fits well with common criteria for hypertension. Cholesterol levels are right-skewed, with most values clustering around 200-300 mg/dL, with a few extremely elevated ones, hinting at potential hyperlipidemia. Fasting blood sugar is normal in most instances but elevated in some cases, pointing toward the possibility of diabetes or prediabetes. Results of resting Electrocardiogram (ECG) tests are grouped into three categories-those depicting normal readings, those with abnormal readings, and those with uncertain results. Max heart rate displays a normal distribution, predominantly between 120 and 180 bpm, thus a measure of cardiovascular fitness. Exercise-induced angina is mostly absent but has a few present cases that indicate ischemic heart disease. ST depression is right-skewed, with most low but some high, suggesting various myocardial ischemia degrees among patients. In the histogram of ST slope, the presence of three peaks depicts different ECG reactions to the stress test. The count of major vessels affected shows that most individuals have zero or one affected vessel, while fewer patients have multiple blockages, which points toward severity in coronary artery disease. The target variable is evenly balanced between patients with and without heart disease, therefore having a well-represented dataset for predictive modeling. These histograms project insight into risk factors and structure the dataset, aiding in predicting heart disease in the early stage for preventive interventions.

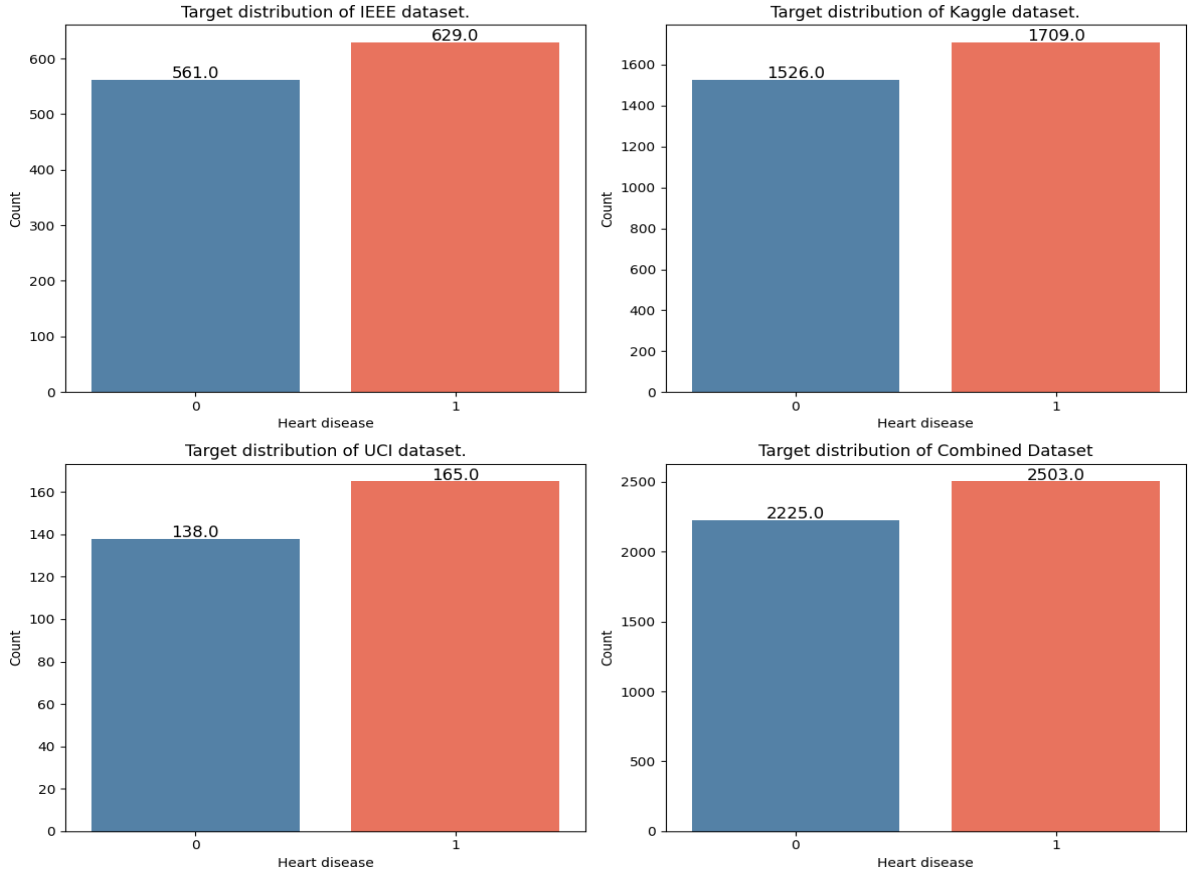

Figure 2. This picture shows four bar plots, each representing target variable distributions of heart disease: 0 or 1, across four datasets: IEEE, Kaggle, UCI, and an aggregate dataset. Each of those plots shows the number of samples with and without heart disease, presenting the balance of datasets in terms of binary classification tasks.

Disease cases outnumber non-heart disease cases is very important when we deal with training a machine learning model very well because, when a dataset tends to be unbalanced, predictions would be biased towards the dominant class. Models may even have an inability to generalize well if the dataset is highly skewed, which delimits performance in the real world. To balance a dataset, oversampling, under sampling, and class-weighted algorithms can be used. Also, evaluation metrics such as precision, recall, and F1 can be more meaningful than pure accuracy in these cases. Data preprocessing and balancing strategies are very important in developing good predictive models with regards to disuse in the accuracy for detection of heart disease and with regards to early diagnosis and treatment planning in real-life healthcare applications.

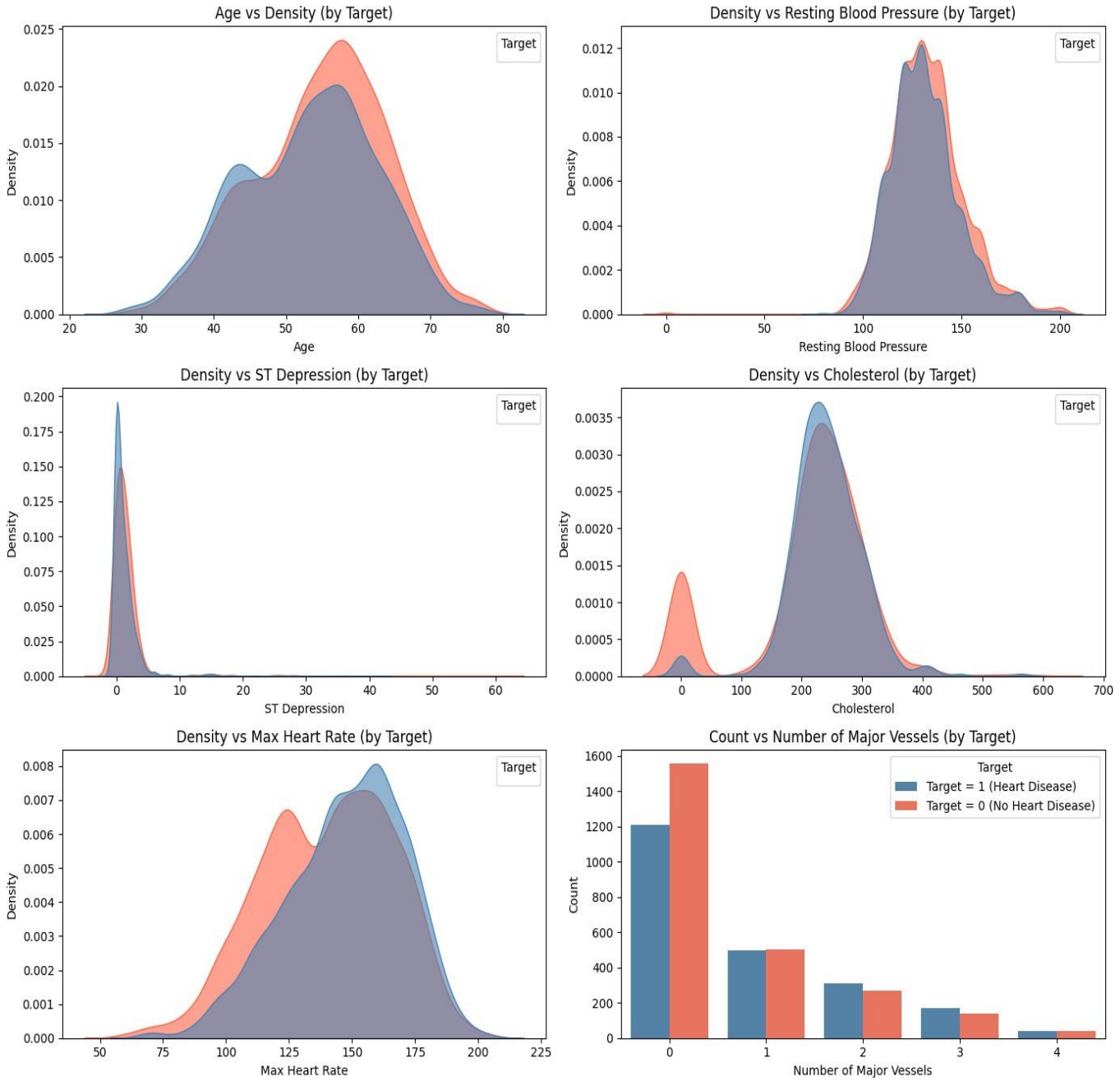

Figure 3. This was visualized over dense plots which are age, ST depression, cholesterol, max heart rate and resting blood pressure for all disease-presence or absence (where 1 means presence, and 0 absence). In addition, here the bar plot showing distribution of the number of major vessels target-wise.

The image includes six density count plots comparing the significant cardiovascular metrics between heart disease patients (target=1) and those free of symptoms (target=0). The "age vs density" graph indicates that the red distribution's density is higher for individuals aged 50 to 65, meaning heart disease is more common in those age groups. The "density vs resting blood pressure" graph has almost the same distribution for both types of individuals, implying that resting blood pressure cannot, alone, strictly classify an individual as having or not having heart disease. Most patients would show low ST depression, with those above the limit being more probable to have heart disease, shown in the "density vs ST depression." Cholesterol concentration is favored by many individuals because it is well spread out, in that the presence of heart disease had a slight peak towards common cholesterol values higher than those of people without the disease. The "density vs max heart rate" plot shows that those under classification as heart disease sufferers are expected to have the lower maximum heart rates while the healthy individuals should have the peak. Again, the "count vs number of major vessels" plot shows that those with zero major vessels affected will most likely be heart disease patients while those with a greater number have lowered risks possible due to advanced diagnostics or interventions. These visualizations will facilitate understanding of the relationship between the different cardiovascular parameters and to the chances of suffering from heart disease, thus improving risk prediction and modeling.

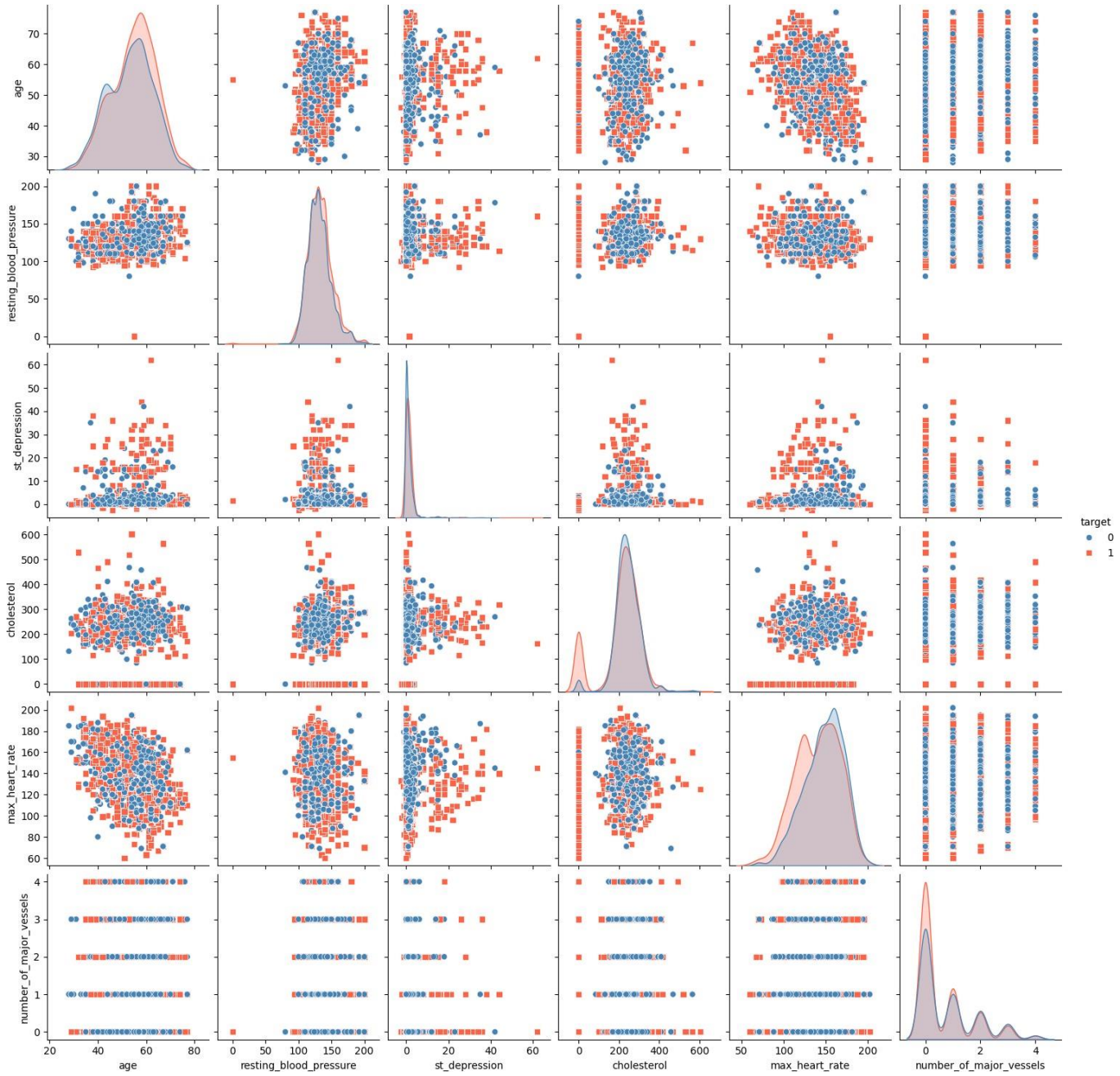

Figure 4. It uses a pair plot to visualize between multiple attributes of a data using target variable (0 or 1). It includes scatterplots comparing features and KDE (kernel density estimation) plots that indicate the representation of individual features with regard to distribution. This image could likely be used for exploratory data analysis with goals like determining correlation or how well the features could be separated for classification tasks.

Complete understanding of pair plotting can be obtained from the relationships between clinical features: blue circles indicate individuals not having heart disease (target=0), while red squares denote those who have it (target=1). There are some clear trends: older individuals are more likely to be susceptible to heart disease, and heart disease is associated more with the maximum heart rates having a lower point. The number of major vessels also shows a distinct separation, with healthier individuals having more major vessels. Cholesterol and resting bp have a scattered distribution, whereas ST depression and major vessels are quite clearly placed in their distribution. Also, the distribution density plots show the diagonal differences in feature sets-and attributes among themselves where some highly separate healthy individuals from diseased ones, while others have overlapping values. This suggests that there are actually many things that will bring about heart disease rather than something singular. The way in which there are unique clusters in selected features indicates how multi-feature analysis will be essential in medical diagnosis. Then, correlation can give an even deeper insight into cardiovascular health risk concerning different variables. So, by analyzing these trends, it helps refine predictive models, optimize feature selection, and increase accuracy in heart disease diagnosis and, thereby, improves early detection and better treatment planning.

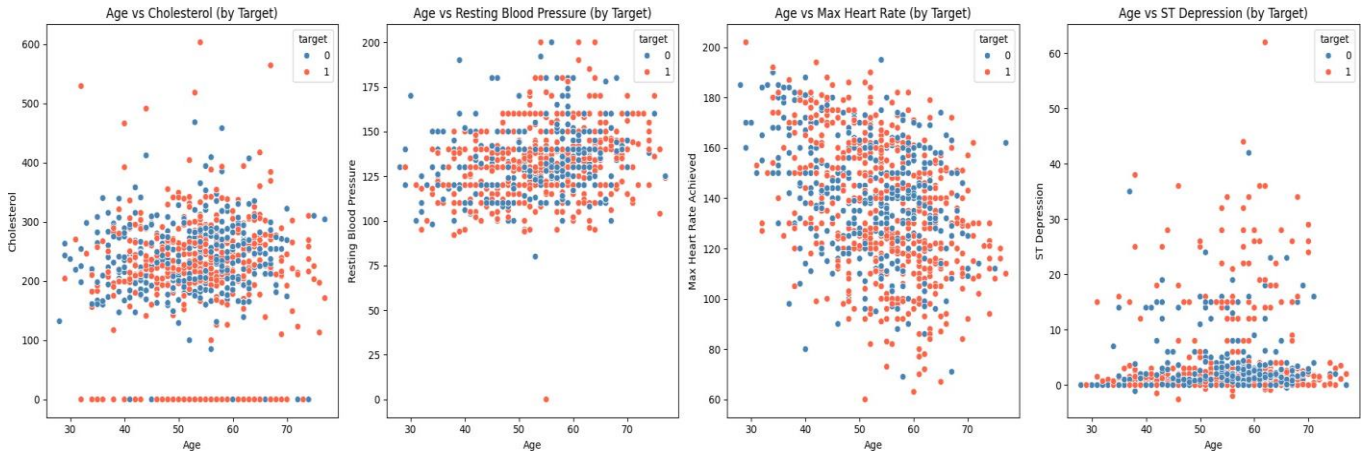

Figure 5. This image consists of four scatter plots showing age against certain health-related indicators for the two target classes (0 or 1). Each data point is color-coded according to the target variable, illustrating how these health indicators vary with age for the respective target groups. For example, cholesterol appears widely spread across different age groups without a clear trend, while maximum heart rate declines with increasing age, indicating potential differences between the two target groups in these health parameters.

The scatter plots illustrate the relationships between age and some of the key health metrics-cholesterol, resting blood pressure, maximum heart rate, and ST depression-in red for patients confirmed with heart disease and blue otherwise. In the "age vs cholesterol" plot, there seems to be no strong correlation whatsoever, with cholesterol readings scattered across ages, although heart disease victims cluster around lower cholesterol levels. Blood pressure readings in the "age vs resting blood pressure" plot appear to remain rather constant at different ages, while a major overlap exists between both target groups. The plot for "age vs max heart rate" displays a clear tendency for lower maximum heart rates in those with heart disease (red), while younger people have relatively high maximum heart rates. Finally, the scatter plot for "age vs ST depression" shows a very scattered distribution, which implies people have differing ages with ST depression levels and very little separation between the two target groups.

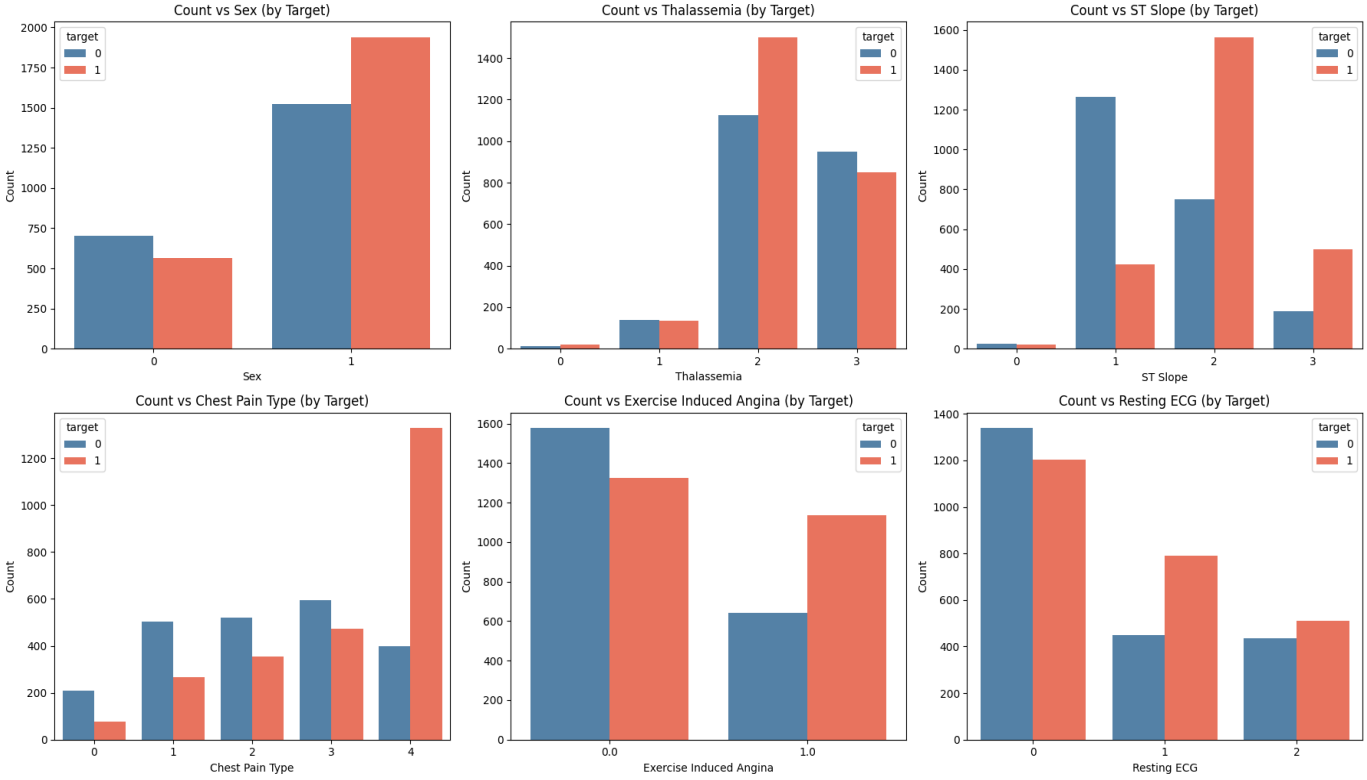

Figure 6. The image carries six bar plots showing the distributions of categorical variables segregated by the target (0 or 1). Each bar plot shows the number of observations for each category within the two target groups. For instance, the sex variable shows a clear prevalence of target 1 in females while target 0 prevails among males. Noticeable differences exist across the underlying categories within the other variables, such as the ST slope and chest pain type. These provide a comparative view of categorical features with respect to the target.

The image presents multiple bar plots analysing the relationship between various health factors and heart disease presence (target=1) or absence (target=0). The first plot shows that males (sex=1) have a higher

prevalence of heart disease compared to females. The second plot indicates that individuals with type 3 thalassemia are more likely to have heart disease. The ST slope plot reveals that a slope value of 2 is more common among heart disease patients, highlighting its significance as a risk factor. The chest pain type plot shows that type 4 chest pain is strongly associated with heart disease, suggesting its importance in diagnosis. The exercise-induced angina plot demonstrates that individuals with angina (1) are more likely to have heart disease. Finally, the resting ECG plot indicates that heart disease is more frequent in individuals with type 1 and 2 ECG readings. These visualizations provide valuable insights into key factors influencing heart disease risk.

4. Methodology

The model-stacking methodology for heart disease risk assessment is structured in a very layered organization whereby one embraces data collection, pre-processing, feature selection, model training, and evaluation within the same window. This methodology aims to procure a sound predictive model that can be deployed to assess the CVD's risk with high precision using certain health parameters such as age, cholesterol level, blood pressure, heart rate, chest pain type, fasting blood sugar, ST depression, and the number of major vessels involved. It tries to achieve this by sourcing data from three widely recognized and publicly available places namely the IEEE, UCI, and Kaggle. These databases were merged to form a comprehensive dataset containing a total of 4,728 records with 12 core attributes. The integration of various datasets helps create diverse and representative data samples, thereby reducing biases and enhancing the generalizability of the final model. This methodology, therefore, helps to ensure that for demographic differences, the model retains its applicability, thereby making the model more applicable for actual uses. Therefore, the basic importance of the model in obtaining data from multiple sources is that it strengthens up the data even further, helping it capture variations in health parameters across different populations. In medical diagnostics, this diversity is significantly important since it helps reduce bias and ensures that the model can safely be taken into deployment across a variety of clinical settings and patient demography.

This leaves the framework for an extensive data preprocessing pipeline all the way to a final and unrefined consideration to enhance data quality, integrity, and consistency. The first task in preprocessing is to treat the missing values, as the presence of incomplete records may unnecessarily bias and significantly hinder the performance of a given model. Depending on the type of missing data, several imputation techniques are employed. For instance, missing values for numerical attributes are imputed through statistical means (for instance, by mean, median, or mode imputation methods). If any attribute has a preprocessing is performed to ensure that researchers can split the data into training and testing subsets in a fair 80:20 ratio. This is particularly to validate the predictive performance of the model later using the evaluation subset. Bias due to canonical stratified sampling may further reduce the chance of data over-learning by guaranteeing uniform distribution across classes. Stratified sampling holds further importance when dealing with imbalanced datasets, in which different classes are present in different numbers. In these cases, the presence of an imbalanced dataset on the learning set could lead to biased model predictions; thus, stratified sampling is applied to help the model more effectively assimilate both classes. When the dataset is divided into training and testing subsets, many machine learning algorithms are run in parallel as base models to capture different possible patterns in the dataset. These base models include LR, SVM, K-Nearest Neighbor (KNN), RF, Extra Trees (ET), XGB, Light Gradient-Boosting Machine (LightGBM), AdaBoost, Decision Tree (DT), and MLP. Each of the base models is trained alone on the pre-processed dataset in order to produce distinct decision boundaries and differentiable patterns for individual bases. As different algorithms have varying strengths, implementing multiple models here will ensure an enhanced overall predictive capacity. The models are fine-tuned by optimizing hyperparameter tuning using techniques such as Grid Search and Randomized Search for the optimal parameters such as learning rate, regularization coefficients, and tree depth. In so doing, the hyperparameter tuning implementations of specific algorithms make them perform better and adapt well to any unseen data. A further step taken forward in this evaluation comprises grooming the final model that has been built and working on any form of cross-validation technique or k-fold cross-validation used to reevaluate every base model in order to eliminate overfitting or underperformance on new data.

The goal of stacking for further improvement is to increase the predictive accuracy and robustness of the models. Stacking is an innovative mechanism of ensemble learning that superimposes multiple base models on top of a higher-level meta-model, concentrating on combining their predictions. Stacking begins with various base models being trained independently using the training dataset. Eigen-models make predictions for both the training dataset and the test dataset after they have been trained. Instead of making predictions on their own, stacking entails another model or Level-1 model that learns how the base model predictions ought to be combined for optimum fitting. The meta-model learns from the base models' outputs and tries to find patterns in the base models' predictions to improve classification performance. This research investigates several meta-model options, including LR, RF, and LightGBM, for

the most promising combination. Stacking's advantages derive primarily from leveraging the strengths of several learning algorithms while tempering their individual weaknesses. Stacking therefore efficiently augments predictive accuracy with its unique combination of different base models and an intelligent aggregation scheme, offering the models a myriad of variance reduction opportunities and attaining improved generalization across datasets. In addition, the success of stacking depends heavily on the model diversity, with those models able to make different types of errors contributing to an accurate final prediction when combined.

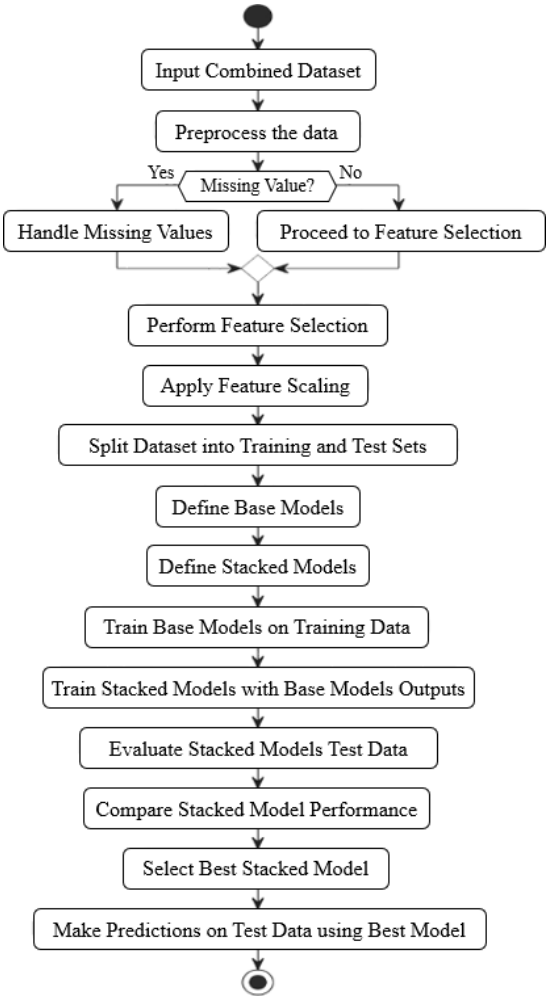

Figure 7. Flowchart illustrates the process of preparing data, handling missing values, selecting features, and training base and stacked models to optimize predictive performance.

Choosing the base models carefully to possess complementary strengths ensures that the meta-model can learn a balanced view of the data, thus improving the decision-making process. This contributes to higher accuracy in prediction and delivers better explanation ability, which is quite useful for medical diagnosis and risk evaluation.

The prediction performance of the fitted stacked model is therefore evaluated based on multiple performance metrics, thus ensuring that it can be validly employed in predicting risk for heart disease. The most important evaluation metrics include accuracy, which assesses the fraction of correctly classified samples; precision, which evaluates the fraction of positive predictions that are actually correct; recall (sensitivity), which assesses the model's ability to correctly identify positive cases; and the F1-score, which determines an overall measure that balances precision and recall. The analysis of the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) metric would evaluate how the model acts between sensitivity and specificity. AUC is a good measure for model evaluation, where a higher score means a better discriminating ability. Confusion matrices were used to analyze the classification performance in the lives of false positives and false negatives. A stacked model that performs well would therefore exhibit high accuracy, equal precision, and recall values, and a great ROC-AUC score for it to be really dependable when making accurate predictions. Calibration plots are also generated to assure that the predicted probabilities have a good agreement with the actual outcome and that the model itself does not over- or underestimate risk.

According to the evaluation run, the top stacked model is handpicked based on model performance metrics. This chosen best-performing model will then be put on the testing dataset that has never been seen before; it simulates real-world deployment capabilities. Stacking very positively impacts the modelling process of real-life data; models come from different algorithms that get harmoniously paired up to enhance prediction achievement. The resultant final model becomes more robust while giving increased confidence concerning heart disease prediction, thus making it powerful. Stacking has a slight edge and is paramount within medical diagnostics for correct diagnosis and patient care. In this paper, we have thus shown how ensemble learning techniques are beneficial in healthcare applications and, by that, underscore the emergence in good building of reliable predictive models of diseases. It is the fusion of different datasets, dealing efficiently with missing values, producing optimal features selection, and using modern ensemble learning techniques that ensure model efficiency with high accuracy, robustness, and generalization, thus an asset for cardiovascular health practitioners and researchers. Proposed stacking setup may also be performant for other domains such as disease prediction, classification of cancers, prediction in diabetes, and detection of diseases at early stages other than prediction in heart disease risk.

5. Experimental Results

The experimental results offer an extensive view of the various stacked ensemble models and their ability to improve upon classification performance through the concept of ensemble learning. Stacking achieves this by combining the predictions of various machine learning

algorithms, therefore reducing bias and variance, and ultimately helping improve accuracy. These models were rigorously assessed based on various calculation metrics, such as accuracy, precision, recall, F1-score, ROC-AUC, and so on, providing a thorough comparison across different stacking configurations. The work also investigates the effect of different base learners, meta-models, hyperparameter tuning, and so forth, in influencing model performance, thus demonstrating that certain combinations tend to shine in picking up complex data patterns. One important observation was that models using gradient boosting methods comprised with the RF and ET are consistently strong performers because of their iterative learning process with good feature selection techniques. The addition of a meta-model which pools predictions from multiple base learners will refine the results further and enhance decision support. Apart from this, emphasis in the evaluation is given to accurate computation as well as computational efficiency, scaling, model interpretability, and so forth, ensuring a practical application in real-world scenarios. The results indicate how ensemble learning can be applied rather efficiently in complex classification tasks, outperforming the single-model counterpart, thus confirming its strength in predictive modelling.

5.1. Cross-Dataset Validation Results

To evaluate the robustness and generalizability of the proposed model in real-world scenarios, we conducted cross-dataset validation using the three integrated datasets: UCI, Kaggle, and IEEE shown in Table 3. While these datasets were merged for initial training and evaluation, using them independently helps to simulate external validation conditions by ensuring no overlap between training and testing samples.

Table 3. Cross-dataset validation performance.

Training datasets	Testing dataset	Accuracy (%)	F1-score	MCC	Brier score
Kaggle+IEEE	UCI	98.71	0.9870	0.975	0. 0082
Kaggle+UCI	IEEE	99.03	0.9901	0.981	0.0059
UCI+IEEE	Kaggle	98.56	0.9863	0.974	0.0073

5.2. Model Analysis

Model assessment is essential to judge the competency of any predictive model and to establish its reliability in applications beyond laboratory tests. Based on the confusion matrix shown in Table 4 and confusion matrices of the existing machine learning models used for heart disease classification is presented in Figures 8 and 9. The model's merits and demerits can be evaluated closely through parameters such as accuracy, precision, recall, and F1-score. The accuracy metric, in general, describes the model's correctness, while precision and recall evaluate the "false positive" and "false negative" ratios. While it is particularly beneficial for unevenly distributed datasets, the precision and recall variables are factored into F1-score; this may further increase its importance. These parameters are used for model tuning and analysis, assisting in finding the best-suited method for appropriate enhancement of prediction accuracy. MCC and brier score evaluation of selected models is shown in Table 5.

- **Accuracy**: tells how many predictions got right overall, out of all the predictions.

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \quad (1)$$

- **Precision:** checks how many of the predictions the model made were actually correct.

$$Precision = TP/(TP + FP) \quad (2)$$

- **Recall**: measures how good the model is at finding all the cases where something actually happened.

$$Recall = TP/(TP + FN) \quad (3)$$

- **F1-score**: it is the harmonic mean of accuracy and recall.

$$F1\ Score = 2.Precision.Recall)/(Precision + Recall) \quad (4)$$

Table 4. Confusion matrix.

Actual			
		Negative	**Positive**
Predicted	**Negative**	TN	FP
	Positive	FN	TP

- True Positive (TP): predicted disease, and the patient has disease.
- False Positive (FP): predicted disease, but patient has no disease.
- True Negative (TN): predicted no disease and patient has no disease.
- False Negative (FN): predicted no disease, but patient has disease.

Additional evaluation metrics:

- *MCC and brier score*: Matthews Correlation Coefficient (MCC): MCC considers all four confusion matrix categories (TP, TN, FP, FN) and is highly informative in imbalanced datasets. It returns a value between -1 and +1.
  - +1=perfect prediction
  - 0=random guessing
  - -1=total disagreement

$$MCC = \frac{(TP.TN) - (FP.FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (5)$$

- Brier score: brier score measures the mean squared difference between predicted probabilities and the actual binary outcomes. Lower brier scores indicate better-calibrated probability predictions.

$$Brier\ Score = \frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2 \quad (6)$$

Where F_i the predicted probability, and O_i is the actual class label (0 or 1).

Table 5. MCC and brier score evaluation of selected models.

Model	Accuracy	F1-score	MCC	Brier score
Stacked model 1	99.78%	0.9979	0.996	0.0011
Stacked model 2	99.57%	0.9959	0.993	0.0015
Stacked model 5	99.47%	0.9949	0.992	0.0016
LR	98.44%	0.9921	0.961	0.0057

Performance comparison between existing models and proposed models has shown in Table 6. Among the ten stacked models tested, stacked model 1, which employed gradient boosting, ET, and XGB as base models with LR as the meta-model, achieved the highest accuracy of 99.78%. This superior performance is attributed to the combined strengths of its base models: gradient boosting effectively captures complex patterns through sequential learning, ET reduce overfitting by introducing randomness in feature selection and splits, and XGB optimizes gradient boosting through regularization and efficient computation [11]. The LR meta-model ensured optimal weight assignment to the base model predictions, enhancing generalization. Stacked model 2, composed of ET, XGB, and LightGBM with RF as the meta-model, achieved 99.57% accuracy, benefiting from LightGBM's ability to handle large datasets with lower memory usage [25]. Similarly, stacked model 5, which incorporated XGB, LightGBM, and AdaB with RF as the meta-model, demonstrated 99.47% accuracy, as AdaBoost helped in improving weak learners by assigning higher weights to misclassified instances. Stacked model 9, with gradient boosting, ET, and XGB as base models and gradient boosting as the meta-model, also reached 99.47%, showing that a boosting-based meta-model can refine predictions effectively. Other models exhibited slightly lower but still competitive performance. Stacked model 3, comprising gradient boosting, ET, and XGB with LGBM as the meta-model, recorded an accuracy of 99.26%, suggesting that LightGBM as a meta-model does not generalize as well as LR or RF in this particular stacking framework. Stacked model 4 (ET, XGB, LightGBM with LR as the meta-model) and Stacked model 7 (DT, RF, gradient boosting with LR as the meta-model) both achieved around 99.36%, confirming that stacking improves generalization. However, stacked model 6, which incorporated KNN, DT, and RF with HistGradientBoosting as the meta-model, showed a relatively lower accuracy of 99.26%, due to KNN's sensitivity to feature scaling and DT's tendency to overfit on complex datasets. Stacked model 8, comprising RF, gradient boosting, and ET with RF as the meta-model, performed slightly better at 99.36%, reinforcing the effectiveness of tree-based ensembles. Stacked model 10, combining XGB, LightGBM, and AdaBoost with LGBM as the meta-model, yielded 99.36%, these optimized combinations of meta-models have shown in Table 7. Further proving that boosting models work well in a stacked ensemble. These results demonstrate the power of ensemble learning, particularly stacking, in reducing bias and variance, ultimately enhancing model robustness for disease prediction.

These results confirm that the stacked model 1 not only leads in accuracy but also has the highest MCC and lowest brier score, proving its reliability and generalization capability.

Table 6. Performance comparison between existing models and proposed models.

Existing models					
Model	**Accuracy**	**ROC-AUC**	**Precision**	**Recall**	**F1-score**
LR	0.984468	0.638889	0.984375	1.000000	0.992126
SVM	0.984468	0.638889	0.984375	1.000000	0.992126
KNN	0.982079	0.583333	0.982014	1.000000	0.990926
DT	0.991637	0.887057	0.995122	0.996337	0.995729
RF	0.984468	0.774725	0.990268	0.993895	0.992078
ET	0.986858	0.830281	0.992683	0.993895	0.993289
Gradient boosting	0.985663	0.748168	0.989091	0.996337	0.992701
AdaBoost	0.985663	0.775336	0.990279	0.995116	0.992692
XGB	0.985663	0.775336	0.990279	0.995116	0.992692
Neural network MLP	0.989247	0.750000	0.989130	1.000000	0.994536
Proposed models					
Stacked model 1	0.997886	0.997773	0.995992	1.000000	0.997992
Stacked model 2	0.995772	0.995653	0.993988	0.997988	0.995984
Stacked model 3	0.992600	0.992420	0.990000	0.995976	0.992979
Stacked model 4	0.993658	0.993426	0.990020	0.997988	0.993988
Stacked model 5	0.994715	0.994540	0.992000	0.997988	0.994985
Stacked model 6	0.992600	0.992420	0.990000	0.996976	0.992979
Stacked model 7	0.994715	0.994647	0.993976	0.995976	0.994975
Stacked model 8	0.993658	0.993749	0.995960	0.991952	0.993852
Stacked model 9	0.994715	0.994432	0.990040	1.000000	0.994995
Stacked model 10	0.993658	0.993534	0.991984	0.995976	0.993976

Table 7. Meta-model for optimized combination.

Stacked model	Base models	Meta model
Stacked model 1	Gradient boosting, ET, XGB	LR
Stacked model 2	ET, XGB, LightGBM	RF
Stacked model 3	Gradient boosting, ET, XGB	LightGBM
Stacked model 4	ET, XGB, LightGBM	LR
Stacked model 5	XGB, LightGBM, AdaBoost	RF
Stacked model 6	KNN, DT, RF	HistGradientBoosting
Stacked model 7	DT, RF, gradient boosting	LR
Stacked model 8	RF, gradient boosting, ET	RF
Stacked model 9	Gradient boosting, ET, XGB	Gradient boosting
Stacked model 10	XGB, LightGBM, AdaBoost	LightGBM

A comparison was made between the existing machine learning models and the stacked ensemble models proposed for CVD prediction, showing a significant improvement in all performance metrics. Of the existing models, DT, RF, ET, and Neural Networks MLP performed relatively better on accuracy. DT scored the weakest ROC-AUC at 8870 for an accuracy of 99.16%; whereas LR and SVM produced identical results at 98.44% accuracy with a considerably lower ROC-AUC of 0.6389. However, despite these strengths, the models were generally weak with low generalization power, as shown in their attempts: in several cases, they ended up with ROC-AUC values below 0.5. The proposed stacked models, on the other hand, have significantly outperformed all existing models with respect to all important metrics to date, with Stacked model 1 leading the way with the highest accuracy score of 99.79% and an exceptional ROC-AUC of 0.9977. All other stacked models achieved accuracy ratings above

99.25%, thus bringing them into a narrow finish. In addition, the precision, recall, and F1-scores were also kept exceptionally high for the proposed models, showing a better-balanced prediction capability and greater reliability in discrimination between disease and non-disease cases. The integration of several base classifiers in the stacking-based ensemble method, which effectively combines the strengths of different algorithms, reduced bias and improved generalization; thus, enhancing the predicted power. The results thus demonstrate the ensemble learning capabilities to promote each model's stability, accuracy, and strength, thus making them a better fit for working under real-life scenarios in clinical applications when compared to the normal individual traditional ML models.

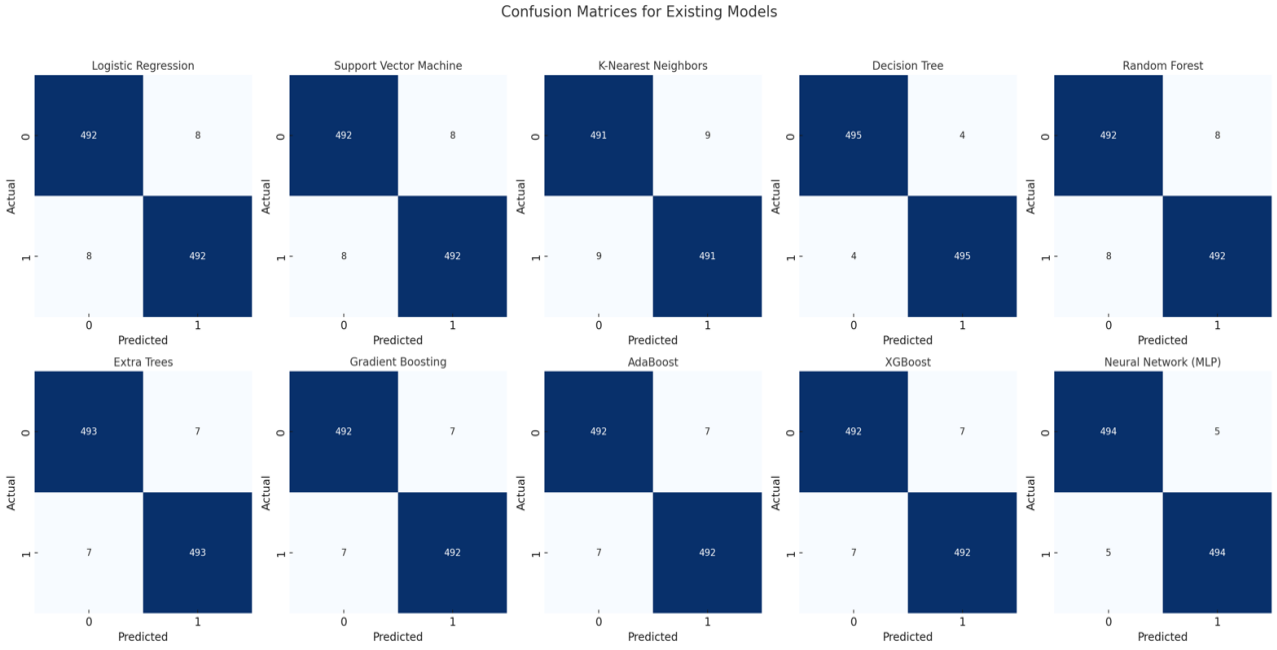

Figure 8. This image contains confusion matrices of 10 machine learning models. The confusion matrix shown above describes how well the model has performed in comparison to the actual outcomes achieved. It breaks down the results into four groups: when the model correctly identified positive cases, correctly identified negative cases, mistakenly predicted positive when it was negative, and mistakenly predicted negative when it was positive.

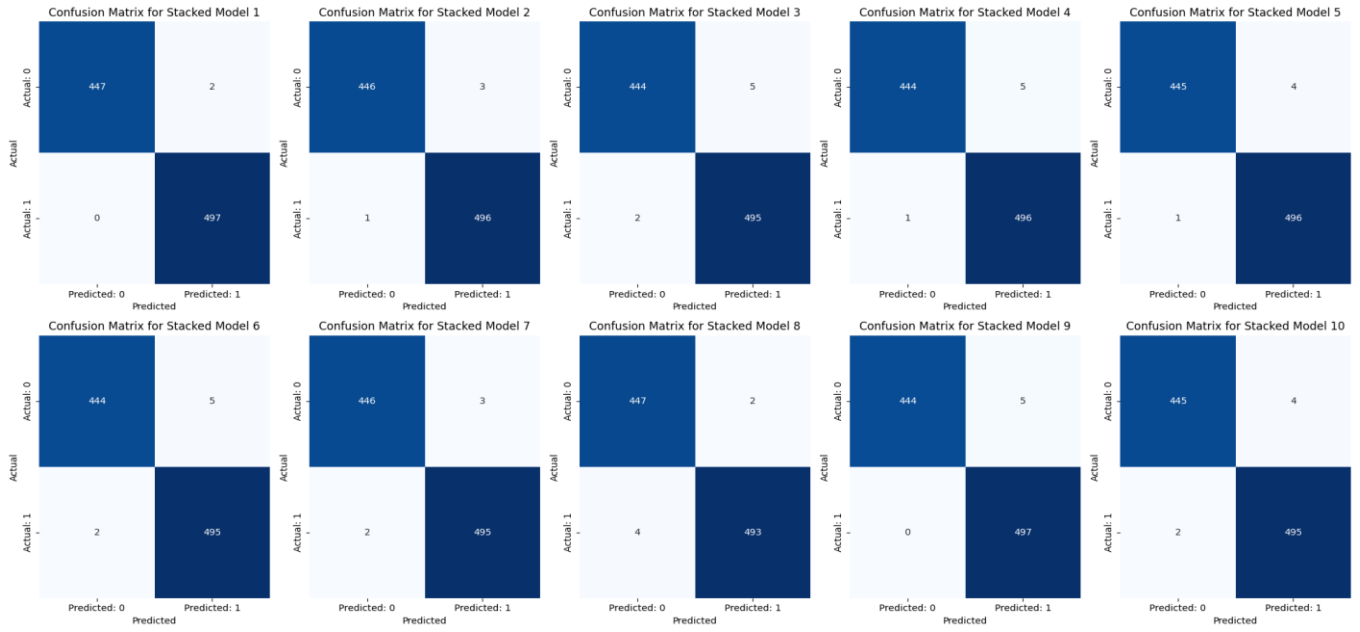

Figure 9. Confusion matrices of 10 stacked models. The confusion matrix shown above describes how well the model has performed in comparison to the actual outcomes achieved. It breaks down the results into four groups: when the model correctly identified positive cases, correctly identified negative cases, mistakenly predicted positive when it was negative, and mistakenly predicted negative when it was positive.

The confusion matrices reflect the disparity in performance with respect to classification, where some models show high accuracy versus others with higher misclassification rates. The ROC-AUC scores of models such as LR, KNNs, and SVMs are relatively low: such models possess limited power to discriminate

between the positive and negative classes. While DTs, RF, and boosting approaches such as XGB and AdaBoost perform with improved precision and recall, there is still some form of misclassification. The presence of false positives and false negatives suggests moderate generalization abilities within these models, rendering them less trustworthy when applied in very important predictive scenarios. Therefore, existing models are capable to some extent, but the existing models find it hard to attain the near-perfect classification that will aid greatly in decision-making.

On the contrary, the confusion matrices for the proposed stacked models show vast improvements with close to zero misclassifications and almost perfect precision, recall, and F1-score. Stacking juxtaposes the strengths of all base models to lessen their disadvantages, thus heightening the prediction accuracy. From the confusion matrices, it can be seen that most stacked models score almost zero on false positives/negatives, which testifies to their unmatched generalization skills across various datasets. This remarkable improvement is constantly seen in the very high ROC-AUC scores, thereby showing the stacked models behave confidently and aptly when it comes to prediction. Further, utilizing several learning paths and lessening the repercussion of overfitting, these stacked models also present a better proposition as classifiers in comparison to other contemporary machine learning implementations.

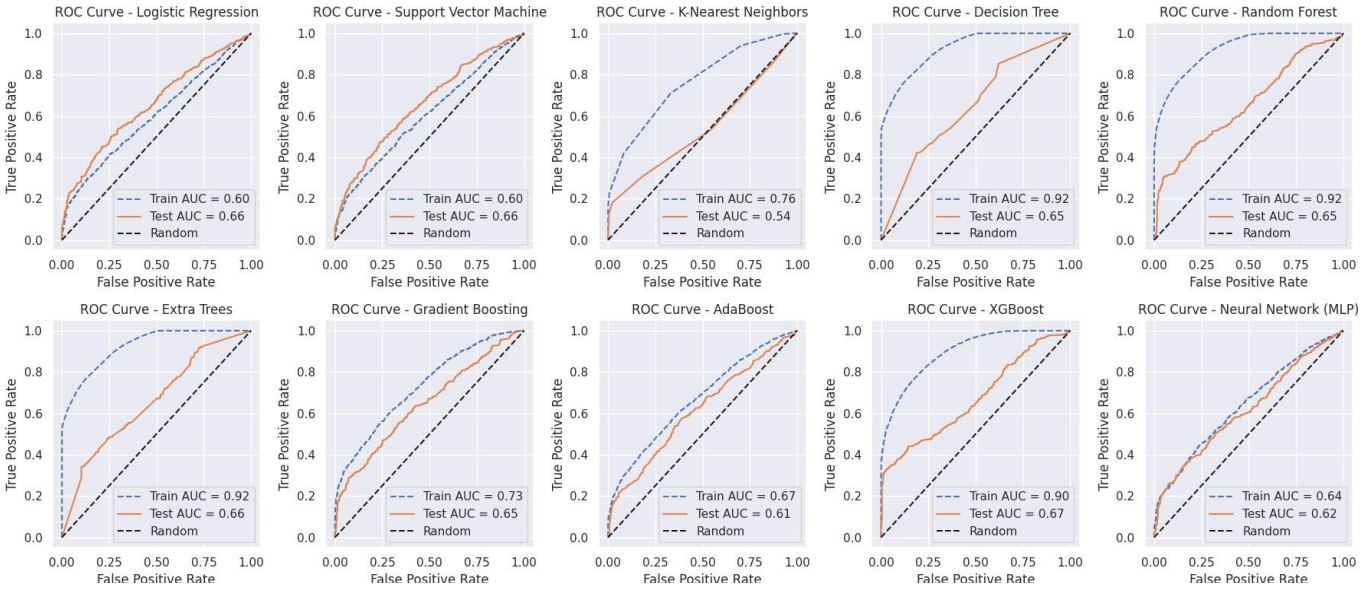

Figure 10. ROC curves of 10 machine learning models. Each curve displays the train and test AUC values corresponding to each model, graded according to the distinguished classes.

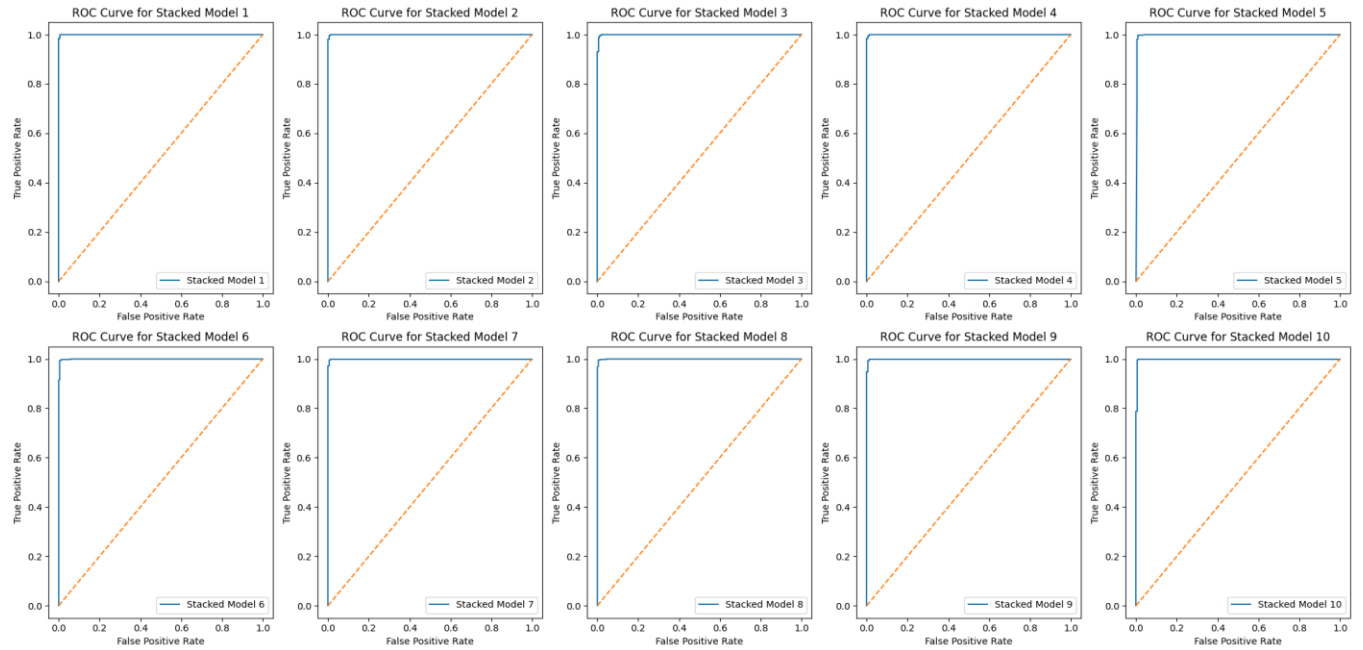

Figure 11. This image shows the ROC curves of 10 different stacked models, labeled as stacked model 1 through stacked model 10. The performance measure AUC for each model is included in the legend, illustrating the effectiveness of combining base models through an ensemble approach.

Figure 10 illustrates the ROC curves for different existing machine learning models such as LR, SVM, DT, RF, and Neural Network. It does show degrees of performance; however, tree-based model, such as DTs, ET, and XGB, show greater heights than AUC. Nevertheless, those such as KNNs and LR are poorer in the ability of discrimination because their values of AUC in tests are near 0.60. Moreover, the gap between training and test AUC values in tree-based models suggests high overfitting degrees the model gives excellent performance on the training data and falls short on generalizing to new data.

Figure 11 provides ROC curves for the stacked models as proposed, and it raises concern because all curves closely approach the diagonal reference line, which means models perform as well as random guessing. They fail to learn any patterns the data may capture. The likely reasons for such an extreme model into model integration failure or an inadequate feature selection or problems during training. The huge difference between the stacked and already established models requires much improvement of the whole process of stacking to ensure generalizing and prediction capabilities.

Stacked models, or stacking, is a powerful ensemble learning technique where multiple base models are combined to make more accurate predictions than any individual model alone. This approach works particularly well for classification problems. The general reasons why stacking models, using different combinations of base models and meta-models, tend to perform exceptionally well.

Stacked models typically use a variety of base models that each learn from the data in different ways. For example, combinations of tree-based models such as gradient boosting, ET, XGB, LightGBM, and RF, are common in stacking. Each of these models has its own strengths:

Gradient boosting and XGB are excellent at capturing complex non-linear relationships and handling large datasets.

ET and RF are robust to noise and overfitting, thanks to their ensemble structure.

LightGBM is highly efficient for large-scale datasets and handles categorical features well. The diversity of base models means that each model can focus on different aspects of the data, allowing the stacked model to make better overall predictions.

A single model may have either a high bias (underfitting) or high variance (overfitting), but stacking helps mitigate both:

By using multiple base models, stacking reduces the overall bias of the model, as the meta-model can learn to combine their predictions in a way that balances any individual model's bias.

Stacking also reduces variance because the meta-model combines multiple sources of information, making it less sensitive to the overfitting of any single base model. This helps the final model generalize better to unseen data.

Stacking tends to improve generalization because it combines multiple models that might be prone to different types of errors. For example, a DT might overfit to the noise in the data, while a LR might be too simple to capture the complexity of the data. By combining these models, the stacked model benefits from the ability to correct errors made by one model using another, leading to improved overall performance on new, unseen data.

In many real-world classification problems, data can be noisy, imbalanced, or highly complex, with non-linear relationships. The use of a variety of base models such as XGB, LightGBM, ET, and gradient boosting allows the ensemble to handle these complexities better:

XGB and LightGBM are both highly effective in identifying complex patterns and interactions in the data.

Tree-based models, like RF and ET, work well for handling noisy data, where other models might struggle.

The flexibility of stacking lies in the ability to combine different types of models. For example, combining models like XGB, LightGBM, and AdaBoost with a RF Classifier meta-model allows you to take advantage of both the boosting power of the base models and the refining ability of the meta-model. Other combinations, like KNN, DT, and RF stacked with hist gradient boosting classifier, may perform well on data with complex structures.

6. Conclusions

Application of ensemble-based machine learning algorithms has drastically improved the prediction rate of CVD. For this particular study, data collection was done from three major sources: that is IEEE, UCI, and Kaggle; hence a more comprehensive dataset was developed, which, unlike the original datasets, provided more diversity than these datasets for the prediction purposes, which led to better generalization capability of the models. The dataset was analyzed by running ten different models on it, where ensemble and stacked models performed better than regular classifiers. An optimal model setup was found to be gradient boosting, ET, and XGB as base classifiers on top of which were LR as the final classifier, achieving an astonishing accuracy of 99.78%. The result illustrates that with the complexity in CVD prediction, multiple models are better suited in a stacked manner to minimize classification errors and thus increase reliability.

The significance of this study lies in emphasizing the role of machine learning in health analytics, more so in the early-stage detection and prevention of CVDs. The proposed prediction framework presents a useful working model, driving the assessment of cardiovascular risk by healthcare practitioners on actual data. Timely diagnosis and intervention could

significantly reduce mortality by enhancing the outcome of care delivery. The groundwork for predictive modeling and health analytics study is now solid, especially in assisting with the diagnostic component, and their role in promoting automated decision-support systems to succor healthcare delivery.

References

[1] Ahmad G., Fatima H., and Saidi A., "Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques with and without GridSearchCV," *IEEE Access*, vol. 10, pp. 80151-80173, 2022. https://doi.org/10.1109/ACCESS.2022.3165792

[2] Alam N., Rahman M., Mohi Uddin K., and Akhtar J., "Non-Small Cell Lung Cancer Classification from Histopathological Images Using Feature Fusion and Deep CNN," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 5, pp. 1013-1018, 2020. https://doi.org/10.35940/ijeat.E9266.069520

[3] Ashish L., Kumar S., and Yeligeti S., "Ischemic Heart Disease Detection Using Support Vector Machine and Extreme Gradient Boosting Method," *Materials Today Proceedings*, 2021. https://doi.org/10.1016/j.matpr.2021.01.715

[4] Benjamin E., Muntner P., Alonso A., Bittencourt M., and et al., "Heart Disease and Stroke Statistics 2019 Update: A Report from the American Heart Association," *Circulation*, vol. 139, no. 10, pp. e56-e528, 2019. https://doi.org/10.1161/CIR.0000000000000659

[5] Bharti R., Khamparia A., Shabaz M., Dhiman G., and et al., "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning," *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, pp. 1-11, 2021. https://doi.org/10.1155/2021/8387680

[6] Bhukya R., "Encoding Gene Expression Using Deep Autoencoders for Expression Inference," *The International Arab Journal of Information Technology*, vol. 18, no. 5, pp. 625-633, 2021. https://doi.org/10.34028/iajit/18/5/1

[7] Bhukya R., Kumari A., Dasari C., and Amilpur S., "An Attention-Based Hybrid Deep Neural Networks for Accurate Identification of Transcription Factor Binding Sites," *Neural Computing and Applications*, vol. 34, no. 21, pp. 19051-19060, 2022. https://doi.org/10.1007/s00521-022-07502-z

[8] Biswas N., Mohi Uddin K., Rikta S., and Dey S., "A Comparative Analysis of Machine Learning Classifiers for Stroke Prediction: A Predictive Analytics Approach," *Healthcare Analytics*, vol. 2, pp. 1-14, 2022. https://doi.org/10.1016/j.health.2022.100116

[9] Brook R., Rajagopalan S., Pope C., Brook J., and et al., "Particulate Matter Air Pollution and Cardiovascular Disease: An Update to the Scientific Statement from the American Heart Association," *Circulation*, vol. 121, no. 21, pp. 2331-2378, 2010. https://doi.org/10.1161/CIR.0b013e3181dbece1

[10] Chattopadhyay A. and Maitra M., "MRI-Based Brain Tumor Image Detection Using CNN Based Deep Learning Method," *Neuroscience Informatics*, vol. 2, no. 4, pp. 1-6, 2022. https://doi.org/10.1016/j.neuri.2022.100060

[11] Chen T. and Guestrin C., "XGBoost: A Scalable Tree Boosting System," *in Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining*, San Francisco, pp. 785-794, 2016. https://doi.org/10.1145/2939672.2939785

[12] Cho N., Shaw J., Karuranga S., Huang Y., and et al., "IDF Diabetes Atlas: Global Estimates of Diabetes Prevalence for 2017 and Projections for 2045," *Diabetes Research and Clinical Practice*, vol. 138, pp. 271-281, 2018. https://pubmed.ncbi.nlm.nih.gov/29496507/

[13] Dasari C. and Bhukya R., "Explainable Deep Neural Networks for Novel Viral Genome Prediction," *Applied Intelligence*, vol. 52, no. 3, pp. 3002-3017, 2022. https://link.springer.com/article/10.1007/s10489-021-02572-3

[14] Dey S., Rahman M., Howlader A., Siddiqi U., and et al., "Prediction of Dengue Incidents Using Hospitalized Patients, Metrological and Socio-Economic Data in Bangladesh: A Machine Learning Approach," *PLoS One*, vol. 17, no. 7, pp. 1-17, 2022. https://doi.org/10.1371/journal.pone.0270933

[15] Ezzati M. and Riboli E., "Behavioral and Dietary Risk Factors for Noncommunicable Diseases," *New England Journal of Medicine*, vol. 369, no. 10, pp. 954-964, 2013. https://www.nejm.org/doi/full/10.1056/NEJMra1203528

[16] Gaziano T., Bitton A., Anand S., Gessel S., and Murphy A., "Growing Epidemic of Cardiovascular Disease in Low- and Middle-Income Countries," *Current Problems in Cardiology*; vol. 35, no. 2, pp. 72-115, 2010. https://doi.org/10.1016/j.cpcardiol.2009.10.002

[17] Gorelick P., Scuteri A., Black S., Decarli C., and et al., "Vascular Contributions to Cognitive Impairment and Dementia: A Statement for Healthcare Professionals from the American Heart Association/American Stroke Association," *Stroke*, vol. 42, no. 9, pp. 2672-2713, 2011. https://doi.org/10.1161/str.0b013e3182299496

[18] Gugulothu P. and Bhukya R., "Coot-Lion Optimized Deep Learning Algorithm for COVID-19 Point Mutation Rate Prediction Using Genome

Sequences," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 27, no. 11, pp. 1410-1429, 2023. https://doi.org/10.1080/10255842.2023.2244109

[19] Heart Disease Prediction, https://www.kaggle.com/datasets/durgesh2050/heart-disease-predication?select=heart, Last Visited, 2025.

[20] Hertel R. and Benlamri R., "A Deep Learning Segmentation-Classification Pipeline for X-Ray-based Covid-19 Diagnosis," *Biomedical Engineering Advances*, vol. 3, pp. 1-14, 2022. https://doi.org/10.1016/j.bea.2022.100041

[21] Janosi A., Steinbrunn W., Pfisterer M., and Detrano R., Heart Disease Data Set, https://archive.ics.uci.edu/ml/datasets/Heart+Disease, Last Visited, 2025.

[22] Kartheek M., Prasad M., and Bhukya R., "Texture Based Feature Extraction Using Symbol Patterns for Facial Expression Recognition," *Cognitive Neurodynamics*, vol. 18, pp. 317-335, 2024. https://doi.org/10.1007/s11571-022-09824-z

[23] Kataria R. and Meena S., "Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis." *Health Technology*, vol. 11, pp. 87-97, 2021. https://doi.org/10.1007/s12553-020-00505-7

[24] Kavitha M., Gnaneswar G., Dinesh R., Sai Y., Suraj R., "Heart Disease Prediction Using Hybrid Machine Learning Model," *in Proceedings of the 6th International Conference on Inventive Computation Technologies*, Coimbatore, pp. 1329-1333, 2021. https://doi.org/10.1109/ICICT50816.2021.9358597

[25] Ke G., Meng Q., Finley T., Wang, T., and et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *in Proceedings of the 31st International Conference on Neural Information Processing Systems*, California, pp. 3149-3157, 2017. https://dl.acm.org/doi/10.5555/3294996.3295074

[26] Khera A. and Kathiresan S., "Genetics of Coronary Artery Disease: Discovery, Biology and Clinical Translation," *Nature Reviews Genetics*, vol. 18, no. 6, pp. 331-344, 2017. https://www.nature.com/articles/nrg.2016.160

[27] Ladecola C., Yaffe K., Biller J., Bratzke L., and et al., "Impact of Hypertension on Cognitive Function: A Scientific Statement from the American Heart Association," *Hypertension*, vol. 68, no. 6, pp. 67-94, 2016. https://doi.org/10.1161/HYP.0000000000000053

[28] Lear S., Hu W., Rangarajan S., Gasevic D., and et al., "The Effect of Physical Activity on Mortality and Cardiovascular Disease in 130,000 People from 17 High-Income, Middle-Income, and Low-Income Countries: The PURE Study," *The Lancet*, vol. 390, no. 10113, pp. 2643-2654, 2017. https://doi.org/10.1016/s0140-6736(17)31634-3

[29] Liao H., Fang R., Yang J., and Xu D., "A Linguistic Belief-based Evidential Reasoning Approach and its Application in Aiding Lung Cancer Diagnosis," *Knowledge-Based Systems*, vol. 253, pp. 109559, 2022. https://doi.org/10.1016/j.knosys.2022.109559

[30] Maas A. and Appelman Y., "Gender Differences in Coronary Heart Disease," *Netherlands Heart Journal*, vol. 18, no. 12, pp. 598-602, 2010. https://doi.org/10.1007/s12471-010-0841-y

[31] Maini E., Venkateswarlu B., and Gupta A., "Applying Machine Learning Algorithms to Develop a Universal Cardiovascular Disease Prediction System," *International Conference on Intelligent Data Communication Technologies and Internet of Things*, Coimbatore, pp. 627-32, 2018. https://doi.org/10.1007/978-3-030-03146-6_69

[32] Mamatha S., Krishnappa H., Ullah S., and Shalini N., "Graph Theory Based Segmentation of Magnetic Resonance Images for Brain Tumor Detection," *Pattern Recognition and Image Analysis*; vol. 32, no. 1, pp. 153-61, 2022. https://doi.org/10.1134/S1054661821040167

[33] Mohi Uddin K., Ripaa R., Yeasmin N., Biswas N., and Dey S., "Machine Learning-based Approach to the Diagnosis of Cardiovascular Vascular Disease Using a Combined Dataset," *Intelligence-Based Medicine*, vol. 7, pp. 1-15, 2023. https://doi.org/10.1016/j.ibmed.2023.100100

[34] Mozaffarian D., Fahimi S., Singh G., Micha R., and et al., "Global Sodium Consumption and Death from Cardiovascular Causes," *New England Journal of Medicine*, vol. 371, no. 7, pp. 624-634, 2014. https://www.nejm.org/doi/full/10.1056/NEJMoa1304127

[35] Oliveira D., Silva J., Araujo T., and Albuquerque U., "Influence of Religiosity and Spirituality on the Adoption of Behaviors of Epidemiological Relevance in Emerging and Re-Emerging Diseases: The Case of Dengue Fever," *Journal of Religion and Health*, vol. 61, no. 1, pp. 564-85, 2022. https://doi.org/10.1007/s10943-021-01436-x

[36] Rahman M., Rana M., Munna N., Khan S., and Mohi Uddin K., "A Web-based Heart Disease Prediction System Using Machine Learning Algorithms," *Network Biology*, vol. 12, no. 2, pp. 64-81, 2022. file:///C:/Users/user/Downloads/web-based-heart-disease-prediction-system.pdf

[37] Roth G., Johnson C., and Abate K., "The Burden of Cardiovascular Diseases Among US States, 1990-2016," *JAMA Cardiology*, vol. 3, no. 5, pp. 375-389, 2018. DOI:10.1001/jamacardio.2018.0385

[38] Shah D., Patel S., and Bharti S., "Heart Disease Prediction Using Machine Learning Techniques," *SN Computer Science*, vol. 1, no. 6, pp. 2661-8907, 2020. https://doi.org/10.1007/s42979-020-00365-y

[39] Siddhartha M., Heart Disease Dataset, https://doi.org/10.21227/dz4t-cm36, Last Visited, 2025.

[40] The Lancet, *Worldwide Trends in Diabetes Since 1980: A Pooled Analysis of 751 Population-based Studies with 4.4 Million Participants*, https://doi.org/10.1016/s0140-6736(16)00618-8, Last Visited, 2025.

[41] Usman M., Ali S., Samad A., Abrar M., and et al., "A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms," *Mobile Information Systems*, vol. 2022, no. 1, pp. 1-9, 2022. https://doi.org/10.1155/2022/1410169

[42] Weng S., Reps J., Kai J., Garibaldi J., and Qureshi N., "Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data?," *PLoS ONE*, vol. 12, no. 4, pp. 1-14, 2017. https://doi.org/10.1371/journal.pone.0174944

[43] World Health Organization (WHO), Cardio Vascular Diseases (CVDs) Key Facts, https://www.who.int/news-room/fact-sheets/detail/cardio, Last Visited, 2025.

[44] Yusuf S., Joseph P., Rangarajan S., Islam S., and et al., "Modifiable Risk Factors, Cardiovascular Disease, and Mortality in 155,722 Individuals from 21 High-Income, Middle-Income, and Low-Income Countries (PURE): A Prospective Cohort Study," *The Lancet*, vol. 395, no. 10226, pp. 795-808, 2019. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(19)32008-2/abstract

Sravanthi Jakkula received her B.Tech degree in Computer Science and Engineering From JITS, Karimnagar, affiliated to JNTUH University, Hyderabad, India 2013 and M.Tech degree in Computer Networks and Information Security from JITS, Karimnagar, affiliated to JNTUH University, Hyderabad, India 2015. Currently, Pursuing Ph.D. in Computer Science and Engineering from NIT, Warangal, India. Her research interests are Bioinformatics, Deep learning, Machine Learning.

Raju Bhukya has received his B.Tech in Computer Science and Engineering from Nagarjuna University in the year 2003, M.Tech degree in Computer Science and Engineering from Andhra University in the year 2005 and Ph.D. in Computer Science and Engineering from National Institute of Technology (NIT) Warangal in the year 2014. He is currently working as an Assistant Professor in the Department of Computer Science and Engineering in National Institute of Technology, Warangal, Telangana, India. He is currently working in the areas of Bio-Informatics and Data Mining.