Predicting Heart Failure Risk in Clinical Data: An Ensemble Learning Approach with SHAP Interpretability

Omri Newman

Student ID: 806646

Reichman University
Efi Arazi School of Computer Science
Master of Science in Machine Learning and Data Science

In collaboration with Nadav Loebl, Head of AI at Beilinson Innovation
March 2025

Contents

1 Introduction

Heart disease remains one of the leading causes of mortality around the world, affecting millions of people each year. Heart failure — a common outcome of various cardiovascular diseases — occurs when the heart is unable to pump enough blood to meet the needs of the body. Early identification of patients at risk of heart failure is crucial for timely intervention, but predicting heart failure remains a significant challenge in clinical practice today.

This project aims to address this challenge by developing a machine learning model that accurately predicts heart failure risk using clinical data. The primary goals are to achieve high predictive performance and ensure model interpretability. An ensemble method is utilized, effectively combining machine learning models to improve robustness and generalization.

Ensemble methods are known for their ability to enhance the performance of individual models by leveraging their strengths and mitigating their weaknesses. In this project, two ensemble strategies are explored: a weighted ensemble and a blended ensemble. The weighted ensemble aggregates predictions based on model performance, while the blended ensemble uses a meta-model to combine base model predictions. Additionally, SHAP (SHapley Additive exPlanations) values are utilized to provide interpretability, enabling clinicians to understand the importance of different features in the prediction process.

By combining robust ensemble techniques with interpretability, this project aims to contribute to the development of practical, data-driven tools for heart failure risk assessment.

2 Background and Problem Statement

Heart failure is a specific condition within the broader category of cardiovascular diseases (CVD), which encompasses a range of disorders affecting the heart and blood vessels. While coronary artery disease (CAD) refers to the narrowing of coronary arteries due to plaque buildup, heart failure represents the heart's inability to pump blood efficiently. These conditions are often interrelated, as CAD can lead to heart failure over time.

Cardiovascular diseases, including heart failure, are a major global health concern. According to the World Health Organization, CVDs are responsible for nearly 18 million deaths annually, accounting for approximately 32% of all global deaths [1]. Among these, heart failure contributes significantly, leading to frequent hospitalizations and reduced quality of life for patients. The high prevalence of heart failure highlights the need for improved risk prediction models that can aid in early detection and management.

Known risk factors for CVD and heart failure include age, hypertension, diabetes, high cholesterol, smoking status, obesity, and sedentary lifestyle. Traditional clinical methods rely on these factors to assess a patient's risk, often using clinical scoring systems such as the Framingham Risk Score and ASCVD Risk Calculator. These models provide population-based risk estimates but do not fully account for complex interactions between risk factors or individual variations in genetics and lifestyle. In addition to scoring systems, imaging techniques such as echocardiography, cardiac MRI, coronary CT angiography, and myocardial perfusion imaging play a crucial role in diagnosing structural and functional heart abnormalities. However, these imaging methods can be resource-intensive, requiring specialized equipment and trained personnel, and their interpretation is subject to clinician opinion.

While these traditional approaches are widely used in clinical practice, they have limitations in providing holistic, data-driven risk assessments. Machine learning models offer an alternative by leveraging large datasets to uncover hidden patterns and complex interactions that may not be captured by predefined risk scores. By integrating multiple sources of clinical data these models have the potential to improve predictive accuracy, ultimately aiding in early intervention for patients at risk of heart failure.

This project seeks to bridge this gap by leveraging machine learning techniques to develop a more accurate and scalable risk prediction tool.

3 Objectives

The primary objective of this project is to develop an accurate and interpretable machine learning model to predict the risk of heart failure using clinical data. Accurate prediction is essential to enable clinicians to identify high-risk patients early and develop effective treatment plans, saving lives and reducing healthcare costs in the process.

To achieve this goal, the project focuses on utilizing ensemble learning techniques to enhance predictive performance. Specifically, two ensemble methods are implemented: a weighted ensemble and a blended ensemble. These methods are designed to leverage the strengths of multiple base models while mitigating their individual weaknesses, resulting in more robust predictions.

Model interpretability is a core objective — understanding how a model arrives at its predictions is critical for building trust with clinicians and ensuring that the model's recommendations can be effectively integrated into clinical decision-making. SHAP values are used to provide detailed insights into feature importance, allowing clinicians to understand the role of various clinical factors in the model's predictions.

4 Related Works

The development of machine learning models for heart disease prediction has been an active area of research. Several studies have explored different techniques, frameworks, and methodologies aimed at improving prediction accuracy and real-time monitoring of patients.

Nashif, Raihan, and Imam (2018) proposed a cloud-based heart disease prediction system designed to detect heart disease using machine learning techniques and continuously monitor patients in real-time [2]. The study involved a comparative analysis of several machine learning algorithms on two open-source databases, with the support vector machine algorithm achieving the highest accuracy of 97.53%. The system incorporated a real-time patient monitoring framework using Arduino, which collected vital signs such as body temperature, blood pressure, humidity, and heartbeat. Data was transmitted to a central server every 10 seconds, enabling doctors to monitor patients remotely and receive instant notifications if any parameter exceeded a critical threshold. This integrated approach aimed to improve early detection and enhance patient care through continuous monitoring.

Liu, Wang, and colleagues (2017) developed a hybrid classification system for heart disease diagnosis based on the ReliefF and Rough Set (RFRS) method [3]. Their proposed system includes two components: a feature selection subsystem and a classification subsystem. The feature selection subsystem involves data discretization, feature extraction using the ReliefF algorithm, and feature reduction through a heuristic Rough Set reduction algorithm. The classification subsystem employs an ensemble classifier built on the C4.5 algorithm. Experiments conducted on the Statlog (Heart) dataset from the UCI database yielded a maximum classification accuracy of 92.59%. The study demonstrates that the hybrid system outperformed previously reported classification techniques, highlighting the potential of combining advanced feature selection methods with ensemble classifiers for heart disease diagnosis.

Chaki, Das, and Zaber (2015) conducted a comparative study of three discrete classification methods for heart disease prediction using the Cleveland Clinic Foundation Heart Disease Data Set from the UCI Machine Learning Repository [4]. The classifiers evaluated were C4.5 decision tree, naive Bayes, and support vector machines (SVM). Their findings revealed that SVM outperformed both naive Bayes and C4.5, achieving the highest accuracy. This study highlighted the importance of selecting appropriate classification techniques for heart disease diagnosis based on dataset characteristics.

Elwahsh, El-Shafeiy, and Tawfeek (2021) proposed a smart healthcare framework (SHDML) for real-time heart disease detection using deep and machine learning techniques [5]. The SHDML framework consists of two stages: a real-time monitoring system and a decision support system for heart disease prediction. The monitoring system employs an ATmega32 microcontroller with pulse rate sensors to measure heart rate per minute, broadcasting the collected data to a Firebase Cloud database every 20 seconds. The smart application displays the sensor data and provides notifications when critical thresholds are exceeded. The second stage involves using deep and machine learning models to predict and diagnose heart diseases in real-time. The models were trained and tested on widely used open-source datasets, achieving a high accuracy of 99%, sensitivity of 94%, specificity of 85%, and an F1-score of 87%.

5 Discussion of Data

The dataset used in this project, titled "Heart Failure Prediction Dataset", was obtained from Kaggle [6]. It combines data from five different datasets and consists of 918 samples, each representing a unique patient. The original datasets include four unique countries which combined share 11 heart disease features. These features contain a mixture of demographic, clinical, and exercise-related information about each patient, all of which are potential predictors of heart failure risk. Features include age,

sex, resting blood pressure, serum cholesterol levels, maximum heart rate achieved, and the presence of exercise-induced angina. Additionally, the dataset provides electrocardiogram (ECG) results, old peak ST depression, and the slope of the peak exercise ST segment. The target variable is a binary indicator representing the presence or absence of heart disease.

One notable challenge in working with this dataset is the absence of certain features commonly associated with heart disease, such as body mass index (BMI) and smoking status. These missing features could potentially limit the model's ability to capture important risk factors. Furthermore, combining patient data from different countries with varying lifestyle habits can introduce variability in feature distributions, potentially affecting model performance.

5.1 Exploratory Data Analysis

To investigate the relationship between various clinical features and heart disease, we analyze the distribution of features stratified by the presence or absence of heart disease. The kernel density estimation plots for continuous variables and bar plots for categorical variables illustrate relevant differences between the two groups. Out of the 918 patients in the dataset, 508 (55%) are diagnosed with heart disease, while 410 (45%) do not have the condition.

Patients with heart disease tend to have slightly higher cholesterol levels on average compared to those without, but their distributions are very similar (Figure 3). Resting blood pressure (RestingBP) shows similar distributions between patients with and without heart disease as well (Figure 4). Maximum heart rate achieved (MaxHR) reveals a clearer separation, as heart disease patients tend to have a lower MaxHR compared to non-diseased individuals, suggesting it is a potential indicator of heart failure risk (Figure 5). The Oldpeak variable, which measures ST depression during exercise, shows heart disease patients have a significantly higher Oldpeak value on average than those without (Figure 6).

Exercise-induced angina is considerably more prevalent among heart disease patients, emphasizing its role as a clinical risk factor (Figure 7). Chest pain type (TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic) also varies significantly, with ASY chest pain being the most common among individuals with heart disease (Figure 8).

Fasting blood sugar (FastingBS) indicates whether a patient's fasting blood glucose level exceeds 120 mg/dL, with 1 representing elevated levels and 0 otherwise. Lower fasting blood sugar levels are more common in both groups, but they appear slightly more frequently among individuals without heart disease. In contrast, higher fasting blood sugar levels occur more frequently in heart disease patients compared to those without the condition (Figure 9).

Resting electrocardiogram (RestingECG) measures the heart's electrical activity at rest and is categorized into three classes: Normal, ST, and LVH. A Normal reading indicates no significant abnormalities, ST represents ST-T wave abnormalities such as T wave inversions or ST segment elevation/depression greater than 0.05 mV, and LVH indicates probable or definite left ventricular hypertrophy based on Estes' criteria. The distribution of RestingECG results show that while normal readings are most common in both groups, they appear slightly more frequently in individuals with heart disease. ST-T wave abnormalities are notably more prevalent among heart disease patients compared to those without the condition. Similarly, LVH is more frequently observed in heart disease patients, reinforcing this clinical features relevance in diagnosing cardiovascular issues.(Figure 10).

The dataset exhibits a skewed sex distribution, with 79% male and 21% female patients. Among males, 63% have heart disease, whereas among females, 26% are diagnosed with the condition (Figure 11). ST Slope represents the slope of the peak exercise ST segment, and its distribution serves as a particularly strong distinguishing factor. A normal "Up" sloping ST segment is more common among non-diseased individuals, while a "Flat" ST slope overwhelmingly corresponds to heart disease cases. The less frequent "Down" slope is also more prevalent among individuals with heart disease, further supporting this features diagnostic relevance (Figure 12).

The dataset consists of patients aged between 28 and 77 years, with a median age of 54. The interquartile range spans from 47 to 60 years, indicating that most patients fall within this middle-aged group. To analyze trends in heart disease prevalence, patients were categorized into three age groups: those under 40 years, those between 40 and 60 years, and those 60 years and above. A clear trend emerges across these groups, with heart disease prevalence increasing with age, reinforcing the well-established link between aging and cardiovascular risk. The age distribution (Figure 2) reinforces this trend, showing that patients diagnosed with heart disease tend to be slightly older on average (56) than those without the condition (50). Notably, while older individuals exhibit the highest prevalence (73.1% in those over 60), a substantial proportion of younger patients (32.5%) are also diagnosed with heart disease. This

highlights the importance of considering additional clinical indicators beyond age alone when assessing cardiovascular risk.

Age Group Summary

Age Group	Number of Patients	Heart Disease	No Heart Disease
Less than 40	80 (9%)	26 (32%)	54 (68%)
40-60	585 (64%)	297 (51%)	288 (49%)
60 and above	253 (28%)	185 (73%)	68 (27%)
Total	918 (100%)	508 (100%)	410 (100%)

Table 1: Heart Disease Prevalence by Age Group

5.2 Cardiovascular Score

While age is a well-established risk factor for heart failure, it alone is insufficient for determining risk. Younger patients may be at elevated risk due to other clinical indicators, while older patients may remain in good cardiovascular health. To more effectively distinguish between patients a CardiovascularScore was developed to make this distinction. The score is an interpretive metric designed to consolidate cardiovascular risk factors into a single explainable score. Its aim is:

- To aid clinicians in identifying high-risk patients by providing a structured, intuitive metric.
- To improve visualization and interpretability of patient data, ensuring a holistic assesment of clinical factors on heart failure risk.

The CardiovascularScore is a composite metric designed to enhance interpretability by aggregating features—RestingBP, Cholesterol, MaxHR, and Oldpeak—with weights assigned based on clinical relevance. This score was normalized to a range of 0–100, allowing for stratification of patients into categories such as Poor Health, Mid Health, and Healthy. The CardiovascularScore's features also appear prominently in SHAP analyses, validating its relevance and demonstrating its utility as an explanatory tool.

5.3 Unsupervised Learning Methods

Clustering techniques were used as part of the initial exploratory data analysis to assess whether natural groupings exist within the dataset and to evaluate whether such clusters correspond to different levels of heart failure risk. The primary goal was to uncover patterns that could potentially inform feature engineering and model development. Several clustering methods were applied, including K-Means, hierarchical clustering, DBSCAN, and K-Medoids. Additionally, dimensionality reduction techniques such as Principal Component Analysis (PCA), t-SNE, and UMAP were used to visualize the structure of data in lower-dimensional space.

The results from K-Means clustering indicate that while some separation exists between patient groups, there is no clear partitioning that aligns perfectly with heart disease labels. A pair plot of clinical features colored by K-Means clusters shows that while certain features contribute to differentiation of cluster labels, significant overlap remains (Figure 13). Hierarchical clustering provides a more granular view of data groupings, but the resulting clusters made it difficult to define distinct patient subgroups (Figure 14). DBSCAN, which identifies clusters of varying densities, struggled to find well-defined groups in these data. This was confirmed by visualizing its cluster assignments using scatter plots of clinical features, which reveals no strong cluster boundaries (Figure 15). DBSCAN tends to perform best when there are high-density clusters separated by low-density regions. This suggests that the heart failure dataset lacks well-defined density-based separations, leading DBSCAN to classify many samples as noise or merge them into a single dominant cluster. The absence of distinct density variations indicate that heart failure risk does not naturally segment into clear subgroups.

The dimensionality reduction techniques—PCA, t-SNE, and UMAP—offered insights into the structure of the dataset, highlighting areas where patients with and without heart disease overlapped (Figures 16, 17, 18). Interestingly, the structural similarities observed across t-SNE and UMAP were also reflected in PCA, suggesting a degree of organization within the dataset. The t-SNE and UMAP plots

both exhibit a distinct high-risk cluster in the upper left, suggesting a subgroup of patients with different clinical profiles. These patterns suggest that while unsupervised learning methods struggle to define definitive clusters, they may still capture meaningful insights on these data.

Despite these findings, the clustering analysis ultimately did not influence the modeling approach. The overlap between heart failure and non-heart failure cases across clusters suggests that additional domain knowledge or feature engineering is required to extract distinct subgroups. Moreover, the results highlighted the difficulty of defining risk groups purely from unsupervised clustering, reinforcing the importance of supervised learning to better capture underlying patterns in patient data.

6 Methodology

6.1 Modeling Approach

The primary objective of this project is to develop an accurate and interpretable predictive model for heart failure risk using clinical data. To achieve this, an ensemble learning approach is employed, leveraging multiple machine learning base models to improve robustness and generalization. Two ensemble strategies are implemented: a weighted ensemble and a blended ensemble. The weighted ensemble assigns dynamic weights to base models based on their misclassified samples, adjusting their influence in the final prediction. In contrast, the blended ensemble utilizes a logistic regression meta-model, learning the optimal combination of base model predictions to enhance overall performance.

The dataset comprises a mix of numerical and categorical features, categorical variables are one-hot encoded while numerical variables are standardized to ensure consistent scaling across models. Data is split into training (75%), validation (15%), and testing (10%) sets, maintaining sufficient samples for model training, hyperparameter tuning, and independent evaluation.

The ensemble models are built using three base classifiers: Random Forest, XGBoost, and Multi-Layer Perceptron (MLP). Random Forest is selected for its robustness in handling both numerical and categorical data while mitigating overfitting. XGBoost — which is an ensemble learning method on its own — is included due to its efficiency and strong performance in tabular datasets. MLP, a type of feedforward neural network, is incorporated to capture complex, non-linear patterns within the data.

To ensure the stability and reliability of the ensemble models, random state analysis is conducted by evaluating model performance across five random seeds. Since model training and data splitting can be sensitive to initialization conditions, running the entire pipeline across different random states helps assess the consistency of results. Performance metrics, SHAP values, and relevant plots are aggregated across these random states to obtain median values, ultimately mitigating the risk of performance variations due to a single train-validation-test split.

6.2 Pipeline Overview

The development of the heart failure risk prediction pipeline was designed to ensure robust learning, effective interpretability, and unbiased evaluation. This pipeline consists of three phases: training, validation, and testing, each serving a distinct role while maintaining strict separation to prevent data leakage.

In the training phase, base models are utilized to identify patterns associated with heart disease and heart failure risk. Grid search hyperparameter tuning with five-fold cross-validation is applied to optimize model performance. This phase focuses on learning meaningful relationships in the data and building a strong predictive foundation.

The validation phase plays a critical role in refining model selection, it provides an unseen dataset to evaluate how models generalize beyond their training data. During this phase, base model misclassified samples are recorded for use in the weighted ensemble, and base model predictions are stored for use in the blended ensemble. The validation phase ensures that these ensemble strategies leverage the strengths of individual base models while mitigating their weaknesses.

The test phase is used exclusively for final evaluation, applying the base and ensemble models to an entirely unseen dataset. Performance metrics—including **Accuracy**, **Precision**, **Recall**, **F1-score**, **ROC AUC** and **PR AUC**—are recorded to assess real-world generalization. Since the test set is never used during training or validation, this phase provides the most unbiased estimate of model performance.

Model interpretability is ensured with the integration of SHAP analysis and feature importance. SHAP values provide insight into feature importance and help explain why a model makes a particular prediction, they are computed for each base model to understand which clinical features influence their

decisions. In the weighted ensemble, these values are aggregated from base models using a weighted average, ensuring it reflects the ensemble's final decision. In the blended ensemble, SHAP values are computed on the logistic regression meta-model to assess how each base model's predictions contribute to the final outcome. To improve efficiency, SHAP caching is implemented to store precomputed values across random states, reducing computational overhead. By integrating SHAP analysis into the pipeline, the model enhances interpretability, making it easier to understand model behavior and ensuring that predictions align with clinical intuition.

This structured machine learning pipeline ensures that model development follows a rigorous and reproducible framework. The integration of ensemble learning and SHAP-based interpretability allows the model to balance predictive accuracy with transparency, making it more suitable for real-world clinical deployment.

6.3 Weighted Ensemble

The weighted ensemble refines predictions by assigning weights based on the relationship between new test samples and previously misclassified validation samples. This approach leverages model-specific errors to determine how much influence each base model should have when making predictions for new data.

6.3.1 Identifying Misclassified Samples

Each base model's misclassified samples from the validation set are recorded and stored. Then the median feature values of these samples are computed to create a representative "difficult-to-classify" reference point for each base model. Comparing against this median misclassified sample assesses how difficult a new test sample may be to classify:

For each base model m consider the set of misclassified validation samples, where each sample is represented as a feature vector. Let N_m be the number of misclassified samples for model m, and define the set of misclassified samples as:

$$X_{\text{misclassified},m} = \{\mathbf{x}_1^{(m)}, \mathbf{x}_2^{(m)}, \ldots, \mathbf{x}_{N_m}^{(m)}\}$$

where each sample i is a p-dimensional feature vector:

$$\mathbf{x}_i^{(m)} = (x_{i1}^{(m)}, x_{i2}^{(m)}, \ldots, x_{ip}^{(m)})$$

with $x_{ij}^{(m)}$ representing the value of feature j for sample i misclassified by model m.

To obtain a single representative vector for model m, we compute the median feature value across all N_m misclassified samples for each feature j:

$$\tilde{x}_j^{(m)} = \text{median}\left(\{x_{1j}^{(m)}, x_{2j}^{(m)}, \ldots, x_{N_m j}^{(m)}\}\right)$$

for all features $j = 1, 2, \ldots, p$.
Thus, the median misclassified sample for model m is:

$$\tilde{\mathbf{x}}^{(m)} = (\tilde{x}_1^{(m)}, \tilde{x}_2^{(m)}, \ldots, \tilde{x}_p^{(m)})$$

This vector serves as a representative "difficult-to-classify" reference point for model m.

6.3.2 Calculating Weighted Distances

To quantify the similarity between a new test sample and each model's misclassified samples, a Euclidean distance metric is used, weighted by feature importance scores from XGBoost:

Let the test sample be represented as a p-dimensional feature vector:

$$\mathbf{x}_{\text{test}} = (x_{\text{test},1}, x_{\text{test},2}, \ldots, x_{\text{test},p})$$

For each model m, the median feature values of its misclassified samples serve as a representative "difficult-to-classify" sample, denoted as:

$$\tilde{\mathbf{x}}^{(m)} = (\tilde{x}_1^{(m)}, \tilde{x}_2^{(m)}, \ldots, \tilde{x}_p^{(m)})$$

where $\tilde{x}_j^{(m)}$ is the median value of feature j across the misclassified samples of model m.

The feature-weighted Euclidean distance between the test sample and the median misclassified sample is computed using XGBoost-derived feature importance scores. Let α_j be the importance weight for feature j, then the weighted Euclidean distance is:

$$d_m = \sqrt{\sum_{j=1}^{p} \alpha_j (x_{\text{test},j} - \tilde{x}_j^{(m)})^2}$$

this weighted Euclidean distance ensures that influential features contribute more significantly to the distance. Larger α_j values indicate features with greater predictive power, amplifying their contribution in the distance metric. Thus, d_m provides a feature-aware similarity measure between the test sample and the difficult-to-classify region of each model.

6.3.3 Assigning Sample-Specific Weights

The weighted distances are used to determine each base model's contribution to the final ensemble prediction. To achieve this, the distances are normalized to create a proportional weighting system that sums to one. Models whose misclassified validation samples are closer to the test sample receive lower weights, reducing their influence on the final prediction. Conversely, when the test sample is farther from the misclassified samples, it suggests the model generalizes better to this instance, warranting a higher weight. This weighting scheme penalizes base models when the test sample closely resembles their representative misclassified sample, as these models have demonstrated difficulty in handling similar cases. Conversely, it rewards base models when the test sample significantly differs from their representative misclassified sample, reflecting higher confidence in the ability to make accurate predictions for this test case.

Given the weighted Euclidean distance d_m for each model m, the final weight assignment is computed as:

$$w_m = \frac{d_m}{\sum_{k=1}^{M} d_k}$$

where:

- w_m is the normalized weight assigned to model m.
- d_m is the weighted Euclidean distance between the test sample and the median misclassified sample of model m.
- d_k represents the weighted Euclidean distance for any model k in the ensemble.
- M is the total number of models in the ensemble.

6.3.4 Final Predictions

Once the weights are assigned, the final ensemble prediction is computed as a weighted sum of probability outputs from each base model. Each base model's predicted probability is multiplied by its assigned weight, and these weighted probabilities are summed to form the final ensemble prediction. A threshold of 0.5 is then applied to classify the final prediction as either positive or negative for heart disease.

The final ensemble prediction $\hat{y}$ is computed as the weighted sum of individual model predictions:

$$\hat{y} = \sum_{m=1}^{M} w_m \hat{y}_m \tag{1}$$

where:

- $\hat{y}$ is the final ensemble prediction.
- w_m is the normalized weight for model m.
- $\hat{y}_m$ is the predicted probability from model m.

For classification:

$$\hat{y} = \begin{cases} 1, & \text{if } \hat{y} > 0.5 \\ 0, & \text{otherwise} \end{cases}$$

This weighting scheme effectively balances the contributions of different models based on their reliability, leveraging their individual strengths while reducing dependence on models that have historically struggled with similar instances.

6.3.5 Key Benefits

The weighted ensemble dynamically adjusts model contributions per sample, improving adaptability to different data distributions. By leveraging validation misclassifications, the ensemble accounts for model-specific weaknesses. This sample-specific weighting refines model predictions, offering flexibility that can improve precision in decision-making. The approach remains interpretable and responsive, making it a valuable method for optimizing ensemble performance.

6.4 Blended Ensemble

The blended ensemble approach enhances predictive performance by leveraging a meta-learning strategy that combines the outputs of multiple base models. Unlike the weighted ensemble, which assigns sample-specific weights based on misclassified validation samples, the blended ensemble utilizes a logistic regression meta-model to learn the optimal way to integrate base model predictions. This method capitalizes on the complementary strengths of the base classifiers, refining the final prediction through an additional layer of learning.

6.4.1 Generating Meta-Features

Once the base models are trained, rather than directly using their predictions on the validation set for classification, their probability estimates of heart disease serve as input features for a meta-model.

For each sample i in the validation set, each base model m outputs a probability score representing its confidence in predicting the presence of heart disease. These probabilities form a new meta-feature matrix $\mathbf{P}$:

$$\mathbf{P} = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,M} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N,1} & p_{N,2} & \cdots & p_{N,M} \end{bmatrix}$$

where:

- $p_{i,m}$ is the probability score assigned by base model m for sample i.
- N is the total number of validation samples.
- M is the total number of base models.

Rather than relying on the original clinical features, the meta-features represent the confidence levels of each base model in predicting heart failure. These meta-features are then used to train the logistic regression meta-model, which learns to optimally combine base model outputs.

6.4.2 Training the Meta-Model

The logistic regression meta-model is trained using the meta-feature matrix $\mathbf{P}$, where each row corresponds to a validation sample and each column represents the probability scores predicted by a base model. In this step the meta-model learns to optimally combine the base model outputs using the validation set's true labels as the target variable.

During training, the logistic regression model assigns a learned weight β_m to each base model's predictions and includes an intercept term β_0 to account for systematic bias. The probability of heart disease for validation sample i is given by:

$$p_{\text{final},i} = \sigma\left(\beta_0 + \sum_{m=1}^{M} \beta_m p_{i,m}\right)$$

where:

- $p_{\text{final},i}$ is the final probability prediction for sample i.
- $p_{i,m}$ is the probability score from base model m for sample i.
- $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid activation function.
- β_0 is the learned intercept term, which captures the overall baseline probability when all model predictions are zero.
- β_m is the learned coefficient for base model m, determining its relative contribution to the final prediction.

The meta-model is trained by minimizing the binary cross-entropy loss, which quantifies the difference between the predicted probabilities and the true labels of the validation set. The objective function is:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} [y_i \log p_{\text{final},i} + (1 - y_i) \log(1 - p_{\text{final},i})]$$

where:

- N is the total number of validation samples.
- y_i is the true label for sample i (1 for heart disease, 0 otherwise).
- $p_{\text{final},i}$ is the predicted probability from the logistic regression meta-model.

The optimization process adjusts the coefficients β_m and intercept β_0 to minimize this loss function, ensuring the blended ensemble generates probability estimates that align closely with the true class labels.

Since logistic regression provides a probabilistic output, the learned coefficients β_m indicate the relative importance of each base model's predictions. If a base model consistently provides valuable predictive information, its corresponding coefficient will have a larger magnitude. Conversely, if a base model is unreliable, its coefficient will be close to zero.

The logistic regression meta-model undergoes hyperparameter tuning using `GridSearchCV`. This process optimizes parameters such as regularization strength and solver choice, ensuring an optimal balance between model complexity and predictive performance.

6.4.3 Final Predictions

Once trained, the blended ensemble is evaluated on the test set. For each test sample i, the base models produce probability scores $p_{i,m}$, which are then used as inputs to the trained logistic regression meta-model. The meta-model applies the learned coefficients to compute a final probability prediction:

$$p_{\text{final},i} = \sigma\left(\beta_0 + \sum_{m=1}^{M} \beta_m p_{i,m}\right) \tag{2}$$

where:

- $p_{\text{final},i}$ is the final probability estimate for test sample i.

- $p_{i,m}$ is the probability output from base model m for sample i.
- $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid activation function.
- β_0 is the learned intercept term, which corrects for systematic biases in base model predictions.
- β_m is the learned coefficient for base model m, determining its relative contribution to the final prediction.

This process mirrors the training phase, where the logistic regression model combines base model outputs using learned coefficients. However, in this stage, no additional learning occurs—the meta-model simply applies the previously optimized weights to generate final predictions. The intercept term β_0 allows the blended model to adjust the probability scale, compensating for potential biases in the base model outputs.

The predicted probability $p_{\text{final},i}$ is then converted into a binary classification decision using a fixed threshold of 0.5:

$$\hat{y}_i = \begin{cases} 1, & \text{if } p_{\text{final},i} \geq 0.5 \\ 0, & \text{otherwise} \end{cases}$$

where $\hat{y}_i$ is the final predicted label for test sample i.

Unlike the weighted ensemble, which dynamically assigns sample-specific weights based on feature importance and model-specific distances, the blended ensemble learns global static weights β_m and an intercept β_0 that remain fixed for all test samples.

6.4.4 Key Benefits

The blended ensemble optimally combines base model predictions without relying on fixed heuristics, improving adaptability across datasets — it provides a structured approach to model combination while maintaining interpretability. By learning from validation performance, it enhances generalization and ensures well-calibrated probability estimates for reliable decision-making. The approach remains transparent and computationally efficient, making it a strong choice for heart failure risk assessment.

6.5 Model Interpretability

SHAP values are utilized to interpret the decision-making process of both the base models and ensemble models. By quantifying the contribution of each input feature to a model's prediction, SHAP provides transparency into how clinical variables influence heart failure risk predictions. This interpretability is essential for ensuring the models are explainable in a clinical context.

For the **weighted ensemble**, SHAP values are computed directly on the original input features. Since this model aggregates predictions from the base models while maintaining a direct relationship with the input data, the values provide feature-level explanations. The SHAP values from the base models are averaged and scaled by their assigned weights, ensuring that models with greater influence on the final prediction contribute more significantly to the interpretability analysis. This method allows for a clear attribution of feature importance.

For the **blended ensemble**, the interpretability process is more complex. The logistic regression meta-model does not directly use the original input features but instead relies on the probability outputs from the base models as meta-features. Consequently, SHAP values computed for the blended ensemble explain the importance of these meta-features rather than the original clinical variables. This means the blended ensemble's values primarily indicate which base models had the most impact on the final prediction rather than why a specific feature was important. To trace back to individual feature contributions, separate analyses are conducted for each base model. This additional step ensures that feature importance can still be extracted, even though the blended ensemble itself does not operate directly on the original features.

Since SHAP computation can be computationally expensive, particularly when evaluating multiple random states, a caching mechanism is implemented. This caching system stores SHAP values for each model and reuses them within the same random state, reducing computation and improving efficiency. By integrating SHAP analysis into the ensemble methodology, the models achieve a balance between predictive performance and interpretability. This ensures the model's outputs can be understood and validated within a clinical decision-making framework.

7 Evaluation and Results

To assess the effectiveness of the ensemble models, performance metrics are computed and aggregated across five random states. This ensures the evaluation is not biased by a single train-test split and provides a more robust measure of generalization. The weighted ensemble and blended ensemble are compared against the base models to determine whether the ensemble strategies offer improvements in predictive performance. The evaluation focuses on standard classification metrics, including accuracy, precision, recall, F1-score, ROC AUC (Receiving operating characteristic Area Under the Curve) and PR AUC (Precision-Recall Area Under the Curve), each of which provides different insights into the model's effectiveness.

Accuracy is commonly used as a general metric for model performance and provides an overall measure of correctness. It is calculated as the proportion of correctly classified cases out of the total number of patients in the dataset. However, in healthcare applications, precision and recall provide more valuable insights.

Precision measures how often the model's positive predictions are correct—it is the proportion of true positives among all predicted positives. In a clinical setting, high precision is important to avoid false positives, which could lead to unnecessary tests or treatments. However, focusing too much on precision may reduce recall, meaning some high-risk patients could be missed.

Recall, or sensitivity, measures how well the model identifies actual positive cases—it is the proportion of true positives among all actual positives. High recall is crucial in healthcare to minimize false negatives, ensuring high-risk patients receive timely care. However, prioritizing recall without considering precision can lead to excessive false positives, straining resources and potentially causing alarm fatigue among healthcare providers. Balancing these metrics is key to an effective diagnostic model.

One approach to balancing precision and recall is to optimize **F1-score**, which is the harmonic mean of the two metrics. It is particularly useful when both false positives and false negatives carry significant consequences. A high F1-score suggests that the model maintains a good balance between identifying high risk heart failure cases and minimizing incorrect classifications.

ROC AUC evaluates the model's ability to discriminate between positive and negative cases across varying classification thresholds. A higher AUC indicates that the model is effective at distinguishing patients with heart failure from those without.

Finally, **PR AUC** provides insight into model performance in scenarios with imbalanced datasets and focuses on the trade-off between Precision (the proportion of correctly predicted positive cases) and Recall (the proportion of actual positive cases correctly identified). A high PR AUC indicates that the model effectively identifies heart failure cases while minimizing false positives.

7.1 Aggregating Performance Across Random States

To ensure results are not influenced by a specific train-test split, all performance metrics are averaged across five different random states. For each state, models are trained, validated, and tested independently, and the final reported metrics reflect the median performance across all states — this offers a broader understanding of how the models generalize. The results highlight the advantages of ensemble learning over individual base models, where both ensemble methods demonstrate improvements over the base models.

The **blended ensemble** achieved the highest median precision (0.8868) and median recall (0.9444), suggesting it is effective at identifying patients at risk of heart failure and also minimizing false positives. The **weighted ensemble** maintained a strong balance between median precision (0.8846) and median recall (0.9259), leading to a competitive median F1-score (0.9091).

Median Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC	PR AUC
MLP	0.8804	0.8679	0.9259	0.9000	0.9376	0.9512
XGBoost	0.8804	0.8704	0.9000	0.9000	0.9433	0.9423
Random Forest	0.8913	0.8824	0.9074	0.9091	0.9459	0.9515
Blended Ensemble	0.8804	0.8868	0.9444	0.9043	0.9518	0.9601
Weighted Ensemble	0.8913	0.8846	0.9259	0.9091	0.9498	0.9576

Table 2: Aggregated performance metrics for each model, evaluated across five random states.

7.2 Confusion Matrices

To better understand the trade-offs between precision and recall, confusion matrices were analyzed. These matrices provide an intuitive representation of how models classify patients into positive and negative cases (Figure 20). Each confusion matrix represents the median aggregated values across five random states, ensuring the results are not skewed by a single random initialization. By computing the median for each of the four confusion matrix entries (true positives, false positives, true negatives, and false negatives), we obtain a more stable representation of model performance.

7.3 ROC Curves

Receiver Operating Characteristic curves provide another perspective on model performance, illustrating the trade-off between recall (sensitivity) and specificity. The area under the ROC curve is a measure of a classifier's ability to distinguish between positive and negative cases (Figure 21). Since model performance varies across different random states, ROC AUC values were aggregated using the median across five random states to provide a more stable comparison.

The **blended ensemble** achieved the highest median ROC AUC score (0.9518), indicating superior discriminatory power, while the **weighted ensemble** achieved the second highest ROC AUC (0.9498). Notably, both ensemble methods outperformed the base models in this metric.

7.4 Precision-Recall Curves

Precision-Recall curves provide insight into model performance, particularly in scenarios with class imbalance, by illustrating the trade-off between precision and recall. The area under the PR curve serves as a measure of a classifier's ability to maintain high precision while capturing as many positive cases as possible (Figure 22). PR AUC values were aggregated using the median across five random states to ensure a stable evaluation.

The **blended ensemble** achieved the highest median PR AUC score (0.9601), highlighting a strong ability to maintain precision while identifying cases of heart failure risk. The **weighted ensemble** followed closely with a PR AUC of 0.9576, reinforcing the effectiveness of ensemble methods in balancing precision and recall. Notably, both ensemble models outperformed the base models in this metric, further demonstrating their robustness in classification performance.

7.5 SHAP Analysis and Feature Importance

SHAP values were computed and aggregated across five random states for each model. Their summary plots provide insights into which features contributed most to model predictions, along with the corresponding magnitude of influence. Each plot depicts the median absolute value of SHAP feature importance, which quantifies the average magnitude of impact each feature has on the model's decision.

Normally, SHAP values are bidirectional, meaning they indicate whether a feature increases or decreases the model's predicted probability of heart disease. However, since SHAP values were aggregated across multiple random states, taking the absolute value prevents cancellation when averaging. Without this adjustment, features that strongly influence predictions in opposite directions across different states might appear to have little importance overall. By using the median absolute SHAP values, the analysis ensures that features are ranked based on their overall contribution to model decisions, regardless of the direction of their effect. Examining the model's SHAP plots, several features consistently emerged as influential, each with relevant clinical impact.

- **XGBoost:** The top-ranked features include `FastingBS`, `ChestPainType`, and `MaxHR`(Figure 23). `FastingBS` is a well-known biomarker for cardiovascular risk, studies have shown that elevated fasting blood sugar levels are strongly associated with coronary artery disease and an increased likelihood of heart attack [7]. Similarly, `MaxHR` is a vital measure of cardiovascular risk, with research indicating that a lower maximum heart rate during stress testing correlates with a higher risk of heart disease [8]. Additionally, `ChestPainType`, is critical for diagnosing angina and other heart-related conditions [9].
- **Random Forest:** The most influential features are `Cholesterol`, `ChestPainType`, and `FastingBS`(Figure 24). `Cholesterol` is a major determinant of atherosclerosis progression, with elevated LDL cholesterol levels being a primary risk factor for coronary heart disease [10].

- **MLP:** The most impactful features in MLP include `Cholesterol`, `MaxHR` and `RestingECG` (Figure 25). `RestingECG` abnormalities can indicate underlying cardiac conditions such as left ventricular hypertrophy or atrial fibrillation, both of which are associated with an increased risk of cardiovascular disease [11].
- **Weighted Ensemble:** This model assigns high importance to `FastingBS`, `MaxHR`, and `ChestPainType`, consistent with top predictors identified by the base models. The inclusion of multiple high-ranking features from different base models indicates the weighted ensemble successfully integrates base predictions and maintains interpretability (Figure 26).
- **Blended Ensemble:** Unlike the other models, the blended ensemble SHAP values do not directly assign importance to clinical features. Instead they highlight probability outputs of the base models and their respective contributions to the final prediction. The most important model is **MLP** followed by **XGBoost** and lastly **Random Forest**. Important clinical features used in the blended ensemble can be inferred from the individual base model SHAP features above (Figure 27).

None of the base models share the same top SHAP features, indicating that each model considers new information when making a prediction. However, there is still a significant overlap in the top features defined by SHAP importance, reinforcing their clinical significance in assessing risk of heart failure.

SHAP Feature Importance

Rank	XGBoost	Random Forest	MLP	Weighted Ensemble
1	FastingBS	Cholesterol	Cholesterol	FastingBS
2	ChestPainType	ChestPainType	MaxHR	MaxHR
3	MaxHR	FastingBS	RestingECG	ChestPainType

Table 3: Top three most important features based on SHAP values for each model.

The integration of SHAP analysis into predictive modeling provides interpretability for clinical decision-making. By identifying which features drive model predictions, healthcare practitioners can assess whether a model's reasoning aligns with established medical knowledge, ensuring that patients at higher risk receive timely intervention.

The transparency provided by SHAP values addresses one of the main barriers to adopting AI in healthcare: the black-box nature of machine learning models. By providing clear explanations for each prediction, SHAP values demystify the model's decision-making process, making it easier for clinicians to trust and rely on its recommendations. These SHAP-driven insights not only validate existing clinical knowledge but also offer actionable, patient-specific information that can assist in risk assessment and personalized treatment planning. By integrating these insights into clinical workflows, these model have the potential to significantly enhance detection and management of heart failure.

8 Experiments

To successfully assess the influence of random state variations on model performance, several experiments were conducted:

8.1 Random State Analysis

Given the variability in model performance due to random initialization and data splitting, multiple random states were tested and averaged. Training 1000 models across 200 random states and aggregating their results provides a robust estimate of performance.

The **blended ensemble** achieves the highest median recall, F1-score, and ROC AUC, but at the expense of a slightly lower precision and PR ROC. The **weighted ensemble** appears as a top performer across all metrics as well, most notably achieving the highest median precision and PR AUC scores. Random Forest and MLP also deliver highly competitive results, even beating out or matching the ensembles in some metrics. These results highlight the inherent strength of the base models for this task and their performance suggests that simpler, well-tuned models can still provide reliable predictions even in the presence of more complex ensembling strategies.

Overall, this analysis demonstrates that predictive performance remains strong across different random states. The ability of ensemble methods to integrate diverse predictive signals contributes to their

robustness, making them a valuable tool for clinical decision-making. Most importantly, these results are promising because they suggest that recall —essential for minimizing false negatives in clinical applications — can be significantly improved with the blended ensemble method while maintaining a minimal trade-off in precision. This represents a valuable enhancement in predictive performance, particularly for medical risk assessment settings where early detection is critical.

Median Performance Metrics Across All States

Model	Precision	Recall	F1-Score	ROC AUC	PR AUC
MLP	0.8644	0.8958	0.8776	0.9328	0.9393
XGBoost	0.8571	0.8889	0.8762	0.9243	0.9302
Random Forest	0.8621	0.9020	0.8829	0.9250	0.9345
Blended Ensemble	0.8571	0.9167	0.8868	0.9333	0.9381
Weighted Ensemble	0.8654	0.9020	0.8846	0.9313	0.9401

Table 4: Aggregated performance metrics for each model, evaluated across 200 random states.

8.2 Ranking Random States

To further evaluate effectiveness of the ensemble methods compared to their underlying base models, a ranking analysis was conducted. The objective was to determine how frequently each model appears as a top performer across different evaluation criteria. A total of 1,000 models were trained across 200 random states and each model's performance was recorded based on Precision, Recall, F1-score, ROC AUC, and PR AUC.

To analyze which models consistently perform best, all 1,000 models were ranked separately for each metric, and the top 50 models were selected. The frequency of each model's appearance in these top 50 rankings was then calculated, providing insight into the proportion of times each model ranked among the best-performing models for a given performance metric.

The heatmap in Figure 1 summarizes these rankings, visually representing how often each model appeared in the top 50 across different performance metrics. Darker colors indicate a higher ranking frequency. The results demonstrate that the ensemble methods consistently outperform the base models in most performance metrics. The Weighted Ensemble dominates in Precision, ROC AUC, and PR AUC, suggesting it maintains strong confidence in its positive predictions. In contrast, the Blended Ensemble ranks highest in Recall, indicating its ability to capture more positive cases, though at the cost of some precision.

While base models still rank among the top performers, they appear less frequently than the ensembles, reinforcing the benefits of ensembling. The heatmap highlights the trade-offs between the two ensemble approaches: the Blended Ensemble prioritizes Recall, making it more suitable for identifying high-risk patients, whereas the Weighted Ensemble excels in Precision and AUC-based metrics, effectively reducing false positives while maintaining strong overall classification performance.

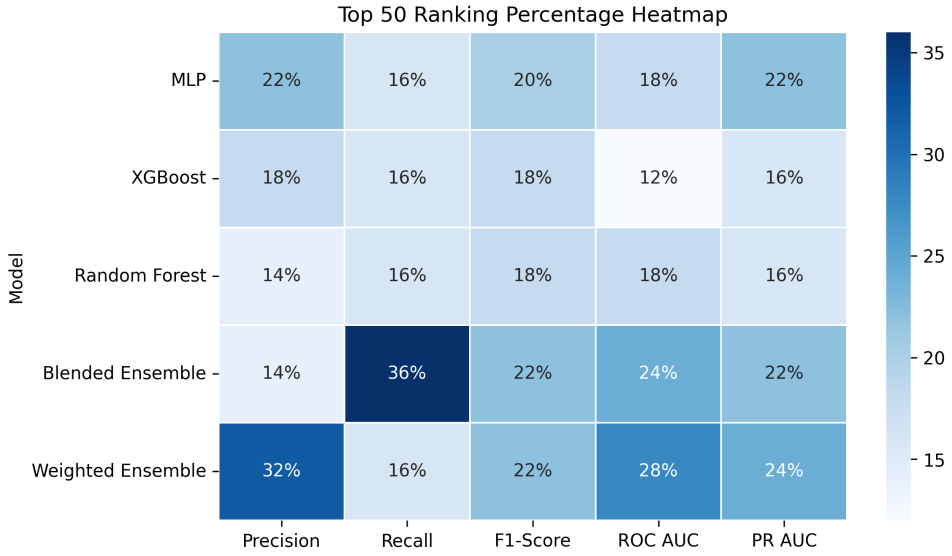

Figure 1: Top Ranked Models.

8.3 Hypothesis Testing

Aggregating results across all random states and ranking models based on their performance metrics provides strong empirical insights showcasing ensemble superiority. However, these analyses do not establish whether the observed improvements are statistically significant. A formal statistical test is necessary to validate whether the ensembles consistently provide a meaningful advantage over the base models.

To address this, we apply the **Wilcoxon Signed-Rank Test**, a non-parametric test suitable for paired comparisons when data distributions are unknown. This test evaluates whether the performance gains observed with ensemble methods are systematic and statistically significant across different random states. For each performance metric, we define the following hypotheses:

- Null Hypothesis (H_0): The ensemble model performs the same or worse than the base model.
- Alternative Hypothesis (H_1): The ensemble model significantly outperforms the base model.

The is conducted as a one-sided test with $\alpha = 0.05$:

- If $p < 0.05$, we reject H_0 and conclude that the ensemble model significantly outperforms the base model.
- If $p \geq 0.05$, we fail to reject H_0, indicating no strong evidence of superiority.

The results of the Wilcoxon Signed-Rank Test provide strong statistical evidence supporting the effectiveness of the ensemble methods over base models. The **weighted ensemble** significantly outperforms XGBoost and Random Forest in precision, indicating its superior ability to make confident positive predictions. It also demonstrates significant improvements in recall over MLP and XGBoost, suggesting its ability to capture positive cases. Furthermore, the weighted ensemble exhibits significant gains in F1-score across all base models and outperforms XGBoost in ROC AUC and PR AUC, reinforcing its overall superior discriminatory power.

The **blended ensemble** shows significant improvement in recall, F1-score, ROC AUC, and PR AUC across all base models but did not achieve statistical significance in precision. This aligns with the previously observed trends, where the blended ensemble consistently prioritized recall, making it particularly effective for identifying high-risk patients.

Hypothesis Testing Results

Metric	Weighted Ensemble Beats	Blended Ensemble Beats
Precision	XGBoost, Random Forest	None
Recall	MLP, XGBoost	MLP, XGBoost, Random Forest
F1-Score	MLP, XGBoost, Random Forest	MLP, XGBoost
ROC AUC	XGBoost	MLP, XGBoost, Random Forest
PR AUC	XGBoost, Random Forest	XGBoost, Random Forest

Table 5: Wilcoxon Signed-Rank Test results ($p < 0.05$ indicates significant improvement).

8.4 Best Model Recommendations

These experimental results consistently highlight the advantage of ensemble learning for heart failure risk prediction. The analysis across multiple random states demonstrates that ensemble methods outperform base models in key performance metrics. The ranking analysis further reinforces these findings, showing that blended and weighted ensembles consistently appear among the top-performing models across different evaluation criteria. Finally, the Wilcoxon Signed-Rank Test confirms that these performance differences are statistically significant, particularly in recall, F1-score, and AUC-based metrics.

Given the clinical context of heart failure prediction, where identifying high-risk patients is a priority, model selection should be guided by the need to maximize recall while maintaining strong overall performance. The blended ensemble is the preferred choice for high-risk patient identification, as it consistently achieves the highest recall, ensuring more patients at risk of heart failure are correctly identified. This is critical in medical applications where false negatives can lead to severe health consequences due to missed diagnoses. For balanced classification performance and confident positive predictions, the weighted ensemble is the most reliable option. It provides strong recall while maintaining high precision, reducing false positives without sacrificing sensitivity. This balance is valuable for optimizing both early detection and efficient resource allocation.

While base models like Random Forest and MLP demonstrate strong individual performance, the ensembles provide an **improvement** by effectively integrating base model predictive signals. Given their robust performance across multiple random states and the statistical significance of their improvements, ensemble methods should be the preferred approach for heart failure risk prediction in real-world applications, particularly when prioritizing recall to ensure that at-risk patients receive timely intervention.

9 Discussion and Future Work

9.1 Ensemble Performance and Decision-Making Tradeoffs

The findings from this study reinforce the effectiveness of ensemble learning for heart failure risk prediction, demonstrating that blended and weighted ensemble models outperform individual base models. It is important to recognize that ensemble methods have an inherent performance ceiling. Since they are ultimately built upon base models, their maximum potential is somewhat constrained by the individual models' strengths and weaknesses. This connection is particularly evident in the weighted ensemble, where the weighting scheme is explicitly influenced by the feature importances of XGBoost. The weighted ensemble determines sample importance based on the Euclidean distance to the median of misclassified samples in the validation set, with distances being scaled by XGBoost's feature importance scores. This approach ensures that features deemed more influential by XGBoost contribute more to the weighting scheme, thereby guiding the ensemble's decision-making. As a result, the SHAP-based feature importance of the weighted ensemble closely resembles that of XGBoost, reflecting the direct influence of XGBoost's internal rankings on the ensemble's weighting mechanism. This relationship suggests that while the weighted ensemble benefits from combining multiple models, its decision-making process is still heavily guided by the most dominant base model. Future work could explore alternative feature importance weighting strategies that incorporate a more balanced aggregation of feature influence across all base models, potentially improving model robustness.

One limitation of the current approach is the use of a fixed classification threshold of 0.5 to determine whether a patient is at risk of heart failure. In a real-world clinical setting, adjusting this threshold based on specific risk tolerance levels could improve decision-making. A lower threshold would increase

recall, capturing more high-risk patients but potentially leading to more false positives. Conversely, a higher threshold would prioritize precision, reducing false positives but at the risk of missing some high-risk cases. Future experiments should explore the effects of a dynamic thresholding mechanism to better align model predictions with clinical decision-making. Similarly, while the F1-score is used as a primary evaluation metric, it assumes equal weighting between precision and recall. However, in medical applications, recall is often more critical than precision, as missing a high-risk patient can have severe consequences. The F2-score, which places greater emphasis on recall, could be a more appropriate metric for this task.

Unlike dynamic thresholding, which modifies the decision boundary itself, changing the evaluation metric does not affect model predictions but rather how performance is assessed. These two approaches are independent but could also be applied together, using a recall-focused metric like the F2-score while simultaneously tuning the classification threshold to optimize for a desired balance with precision. Applying these refinements would provide a more clinically relevant evaluation of the models and help determine whether the performance gains from ensemble methods are more pronounced when recall is prioritized ultimately validating their usefulness in heart failure risk prediction.

9.2 Ensemble Comparison

While both ensemble methods aim to improve predictive performance by combining multiple base models, they differ in their approach to model combination.

The blended ensemble leverages logistic regression to learn the optimal way to integrate base model predictions, whereas the weighted ensemble applies a weighting scheme based on sample misclassification patterns. This makes the blended ensemble more adaptable, as it lets machine learning determine the best combination of base model outputs rather than relying on manually designed heuristics.

The blended ensemble learns **global** weights that remain fixed for all test samples. In contrast, the weighted ensemble assigns **sample-specific** weights, dynamically adjusting predictions based on individual samples. This makes the weighted ensemble more flexible but potentially more sensitive to noise in the feature space. Because the blended ensemble learns a single set of coefficients for all base models, it can naturally adjust for redundant information when they are highly correlated. In contrast, the weighted ensemble may not explicitly account for such dependencies, potentially leading to over-reliance on certain models if their weights are not carefully tuned.

A notable difference between the two ensemble methods is their impact on recall and precision. The blended ensemble generally achieves higher recall, meaning it is more effective at identifying heart disease cases, even at the cost of increased false positives. This occurs because logistic regression learns a global weighting scheme that prioritizes capturing positive cases. The weighted ensemble typically results in higher precision, as it dynamically adjusts predictions for each sample, often leading to a more conservative decision boundary that reduces false positives but may miss some true positives. This difference arises due to the global vs. sample-specific weighting mechanisms, where the blended model applies fixed weights across all samples, while the weighted model adapts per instance.

The weighted ensemble, by dynamically adjusting sample weights, can be more prone to overfitting, especially if it disproportionately emphasizes outlier predictions. The blended ensemble, by using logistic regression on validation set predictions, mitigates this risk by enforcing a structured learning process that generalizes better across datasets.

9.3 Variability in Model Performance Across Data Splits

One main observation from this study is the variability in model performance across different random states, particularly in the extent to which the ensemble models outperform the base models. In an ideal scenario, the ensembles would consistently provide superior performance regardless of the specific train/test split. However, certain random states yield noticeably stronger results, raising the question of whether some data splits are inherently more conducive to to high performance than others. Future work could investigate which characteristics define high-performing random states and whether these insights can be used to inform a more structured data-splitting strategy. Rather than relying solely on random sampling, an alternative approach could involve stratified sampling or clustering-based splits to ensure that critical variables are more evenly distributed. By reverse-engineering the patterns observed in high-performing random states, it may be possible to develop a more consistent and generalizable sampling methodology, which in turn would feed a more robust model.

9.4 Causal Inference and Model Explainability

While SHAP values offer valuable insights into which features contribute most to model predictions, they do not establish causality. The current analysis highlights predictors such as fasting blood sugar, cholesterol, and chest pain type, but it does not confirm whether these factors are causal determinants of heart failure risk or merely correlated with it. Future research could incorporate causal inference techniques, such as propensity score matching or structural causal models, to assess the true causal impact of these features. Additionally, while SHAP values provide transparency, further work should explore complementary explainability methods. For instance, integrating an interpretable clustering approach such as explainable K-means could offer clinicians actionable insights into patient subgroups, helping bridge the gap between black-box machine learning and real-world medical decision-making.

9.5 Computational and Implementation Improvements

While the ensemble models demonstrate strong performance, there are several areas where methodological refinements could further enhance runtime efficiency and overall usability. One such area involves optimizing the code structure, for example using parallelization to leverage all CPU cores to train models simultaneously, especially for grid search. Doing so would shorten the time spent on model training and hyperparameter tuning. Moreover, the SHAP calculations prove to be a computational bottleneck. The original implementation uses all available samples to approximate SHAP values, and doing so significantly increases computation time. A more efficient approach involves summarizing the background data using a representative samples rather than the whole dataset. This can be achieved through random sampling, or clustering-based sampling to reduce computation time while preserving interpretability.

To improve usability, the code should be made more adaptable to different datasets by addressing feature selection, handling missing data, and improving modularity. Current implementation relies on hard coded feature names, which limits flexibility. A more adaptive feature selection process should be implemented to automatically adjust based on data availability. Another consideration is handling missing values, as real-world medical datasets often contain incomplete data. Adding an imputation strategy such as mean imputation or more sophisticated techniques like k-nearest neighbors would improve robustness to other datasets. Lastly, breaking down large functions in the source code into smaller, modular helper functions would enhance maintainability and readability.

10 Conclusion

This project presents a machine learning framework for predicting heart failure risk using ensemble methods, with a focus on accuracy and interpretability. By combining multiple base models through weighted and blended ensembles, these approaches demonstrate significant improvements over individual models, highlighting the advantages of ensemble learning in medical risk prediction. The use of SHAP values ensures that models remain interpretable, a critical requirement for clinical adoption.

The experiments confirm that ensemble models consistently outperform base models, with both ensemble strategies offering distinct advantages depending on the clinical context. The weighted ensemble's higher precision makes it suitable for minimizing unnecessary interventions, while the blended ensemble's superior recall ensures that high-risk patients are less likely to be missed. These findings highlight the importance of selecting machine learning models based on specific healthcare priorities.

Future refinements in data handling, computational efficiency, and causal interpretability are necessary to ensure the models are both clinically useful and scalable. Addressing these areas will help bridge the gap between machine learning-driven risk prediction and real-world medical decisions. This work contributes to the growing body of research on machine learning for healthcare by providing a robust and interpretable approach to heart failure risk prediction.

GitHub Repository: https://github.com/omrinewman/MScFinal/tree/main

11 Plots and Graphs

11.1 Exploratory Data Analysis

11.1.1 Feature Distributions

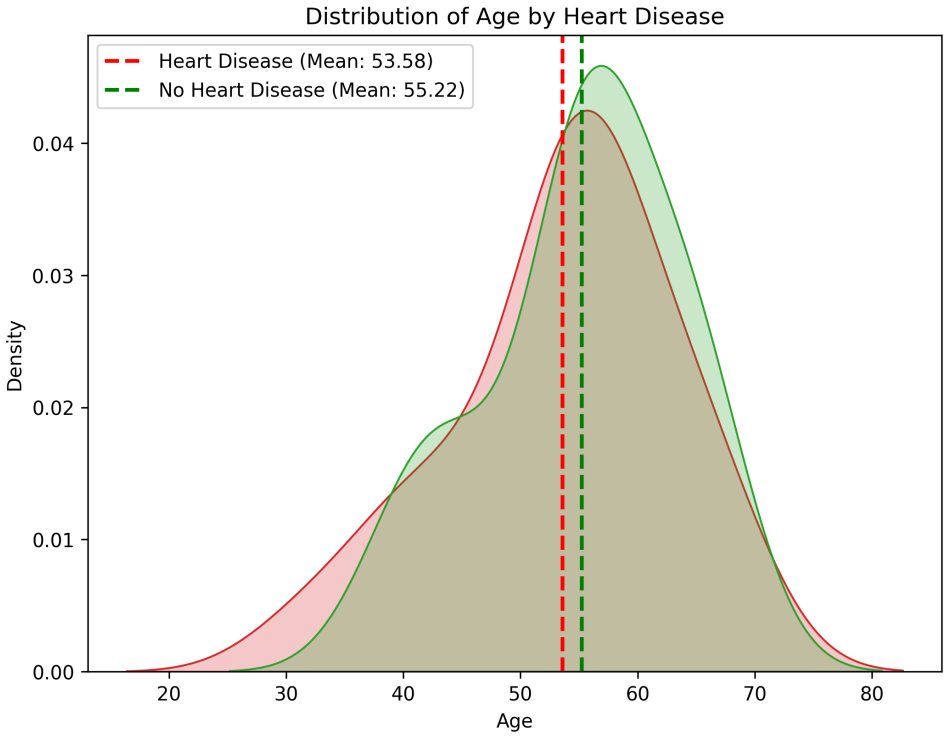

Figure 2: Distribution of Age by Heart Disease.

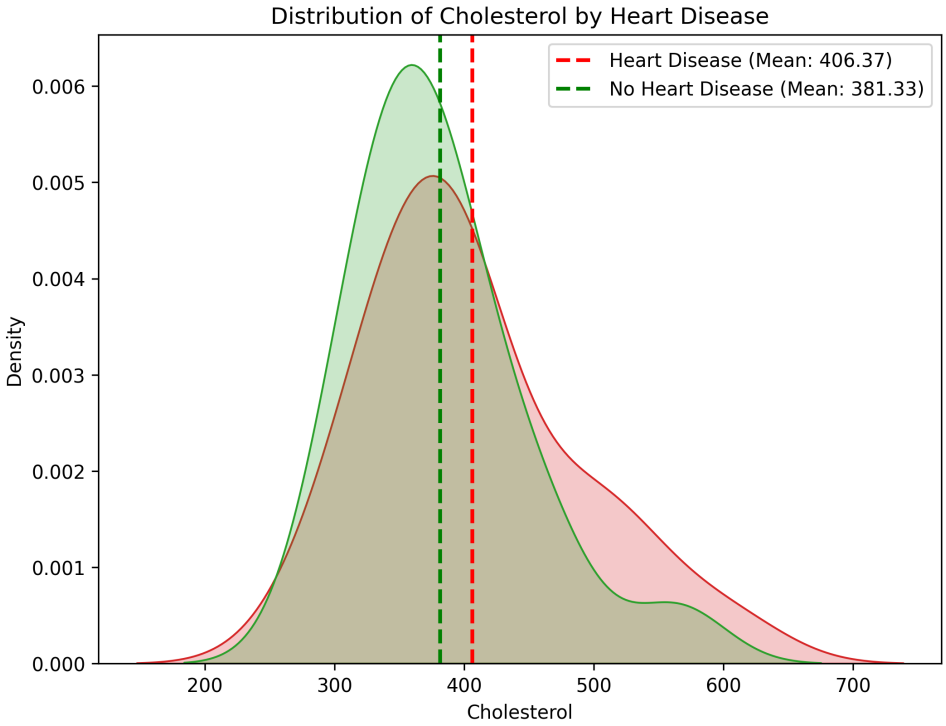

Figure 3: Distribution of Cholesterol by Heart Disease.

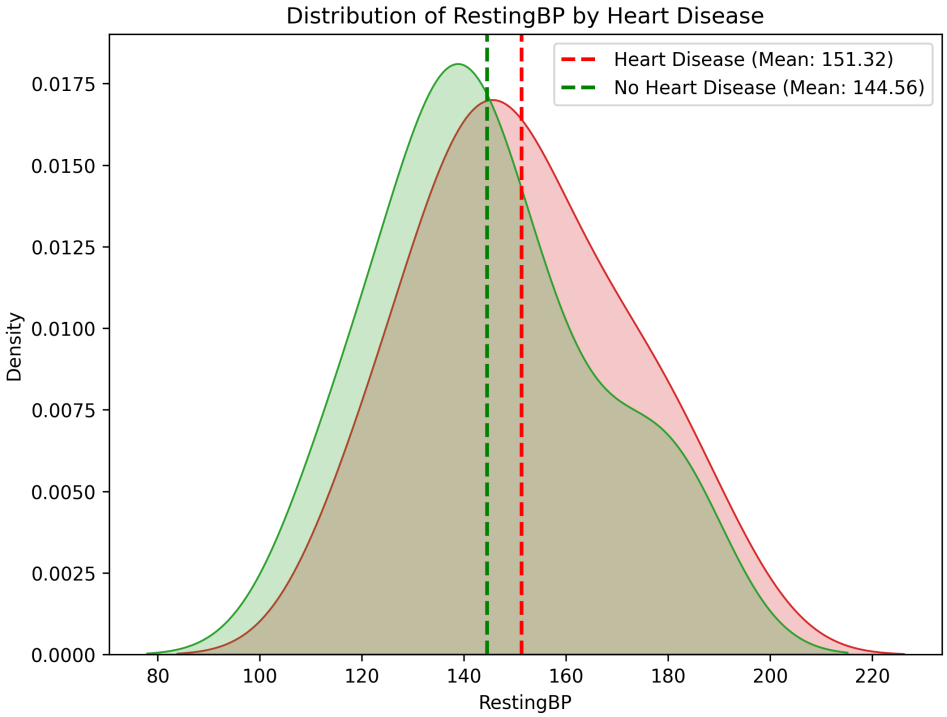

Figure 4: Distribution of RestingBP by Heart Disease.

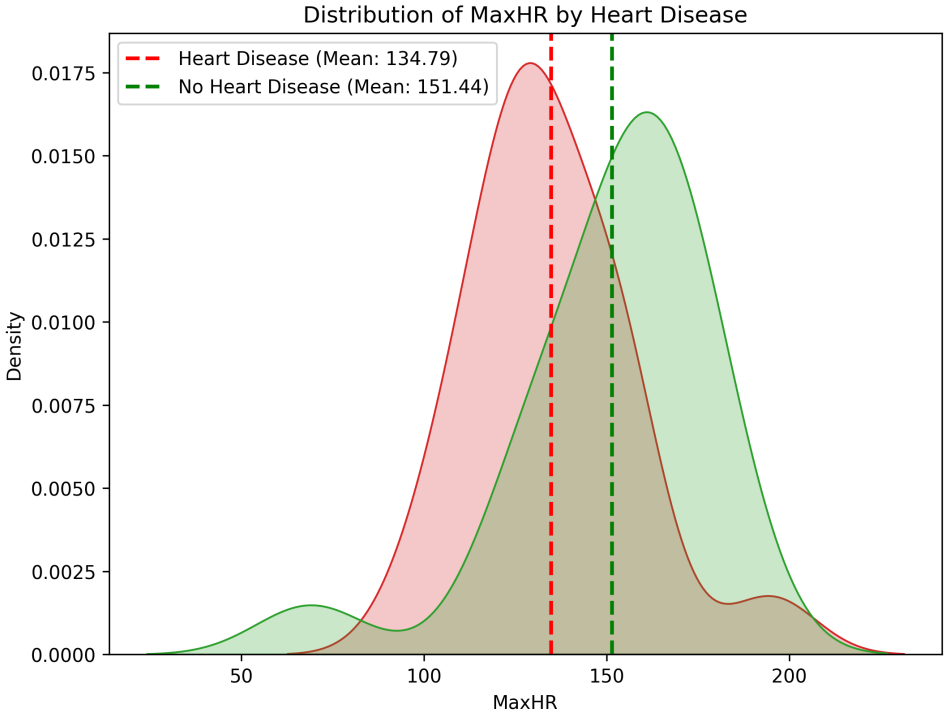

Figure 5: Distribution of MaxHR by Heart Disease.

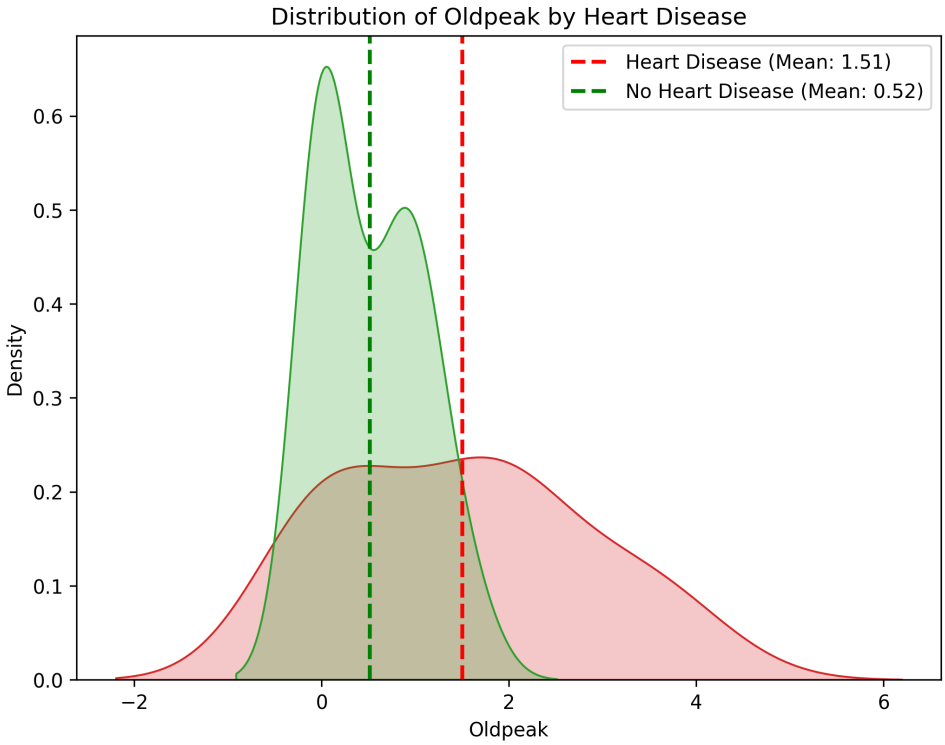

Figure 6: Distribution of Oldpeak by Heart Disease.

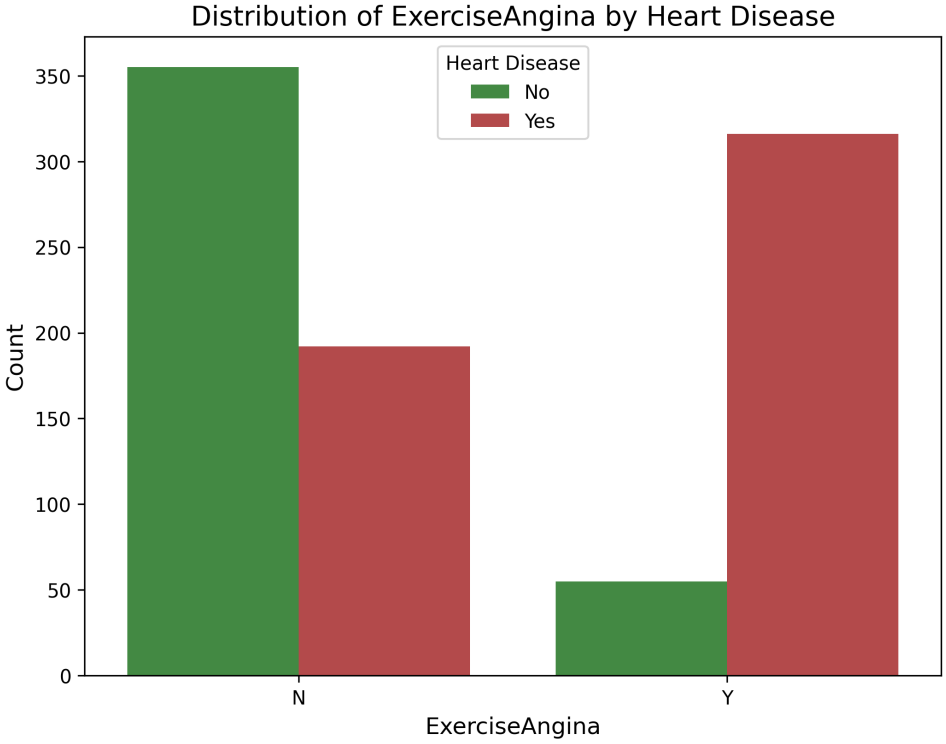

Figure 7: Distribution of Exercise Angina by Heart Disease.

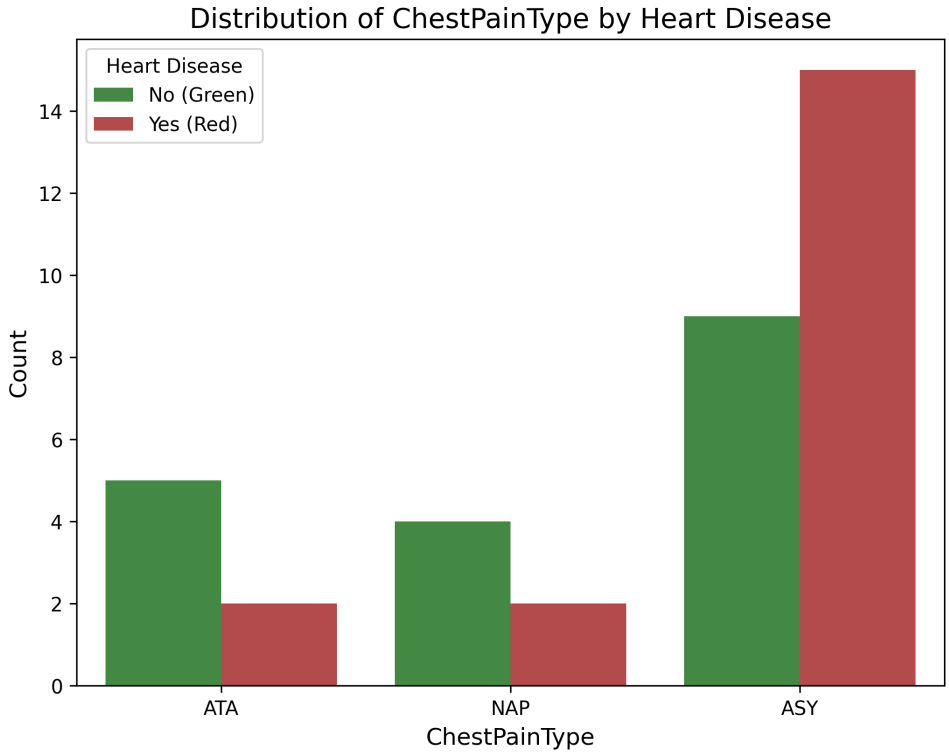

Figure 8: Distribution of Chest Pain Type by Heart Disease.

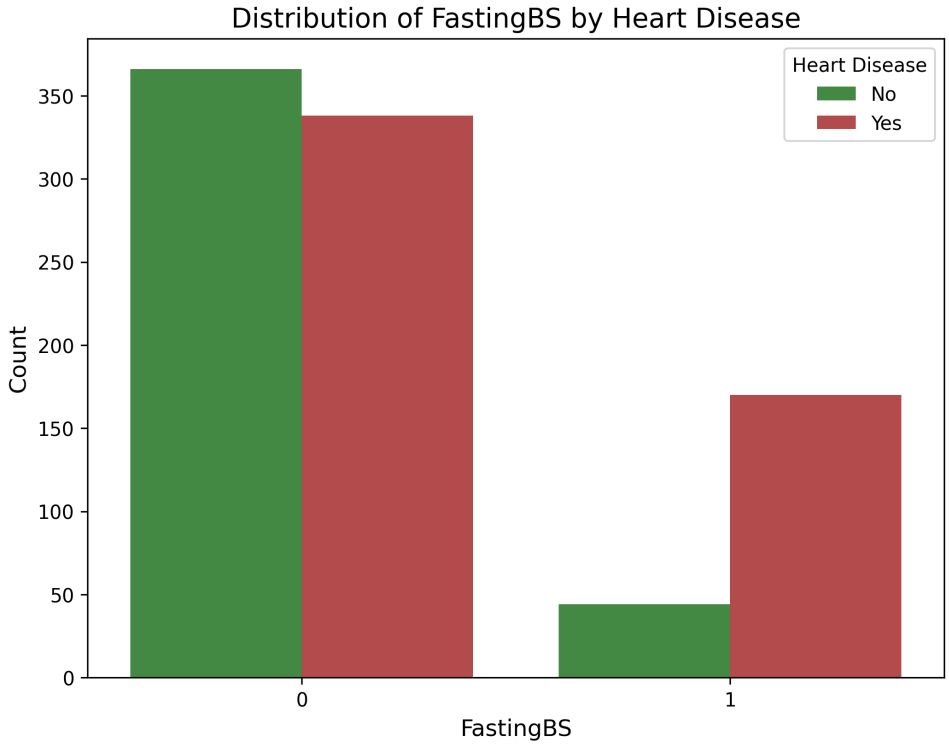

Figure 9: Distribution of FastingBS by Heart Disease.

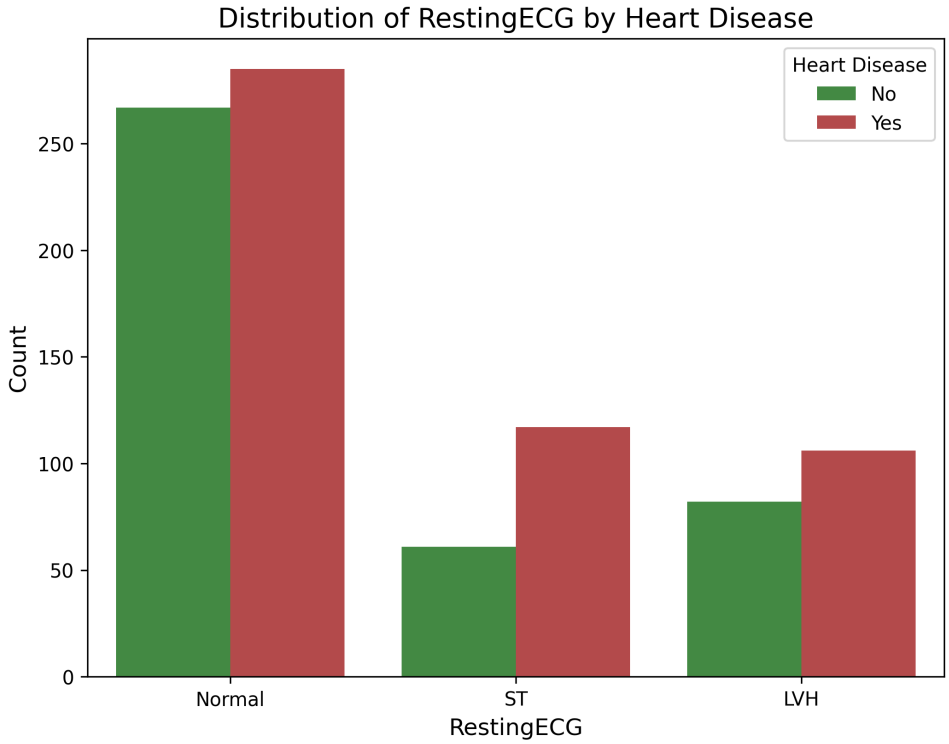

Figure 10: Distribution of RestingECG by Heart Disease.

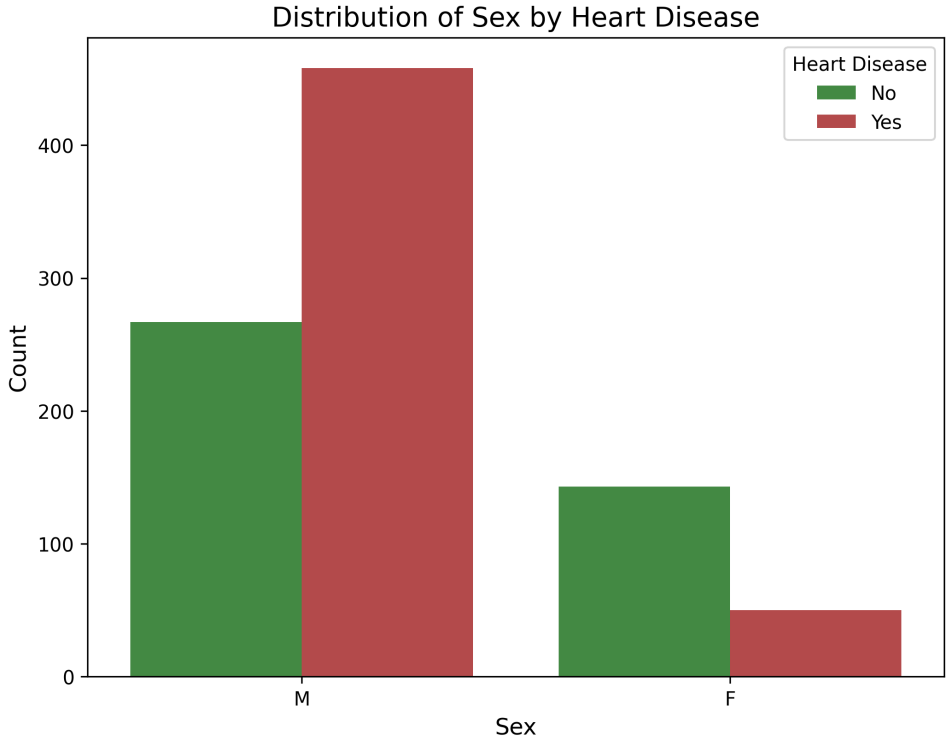

Figure 11: Distribution of Patient Sex by Heart Disease.

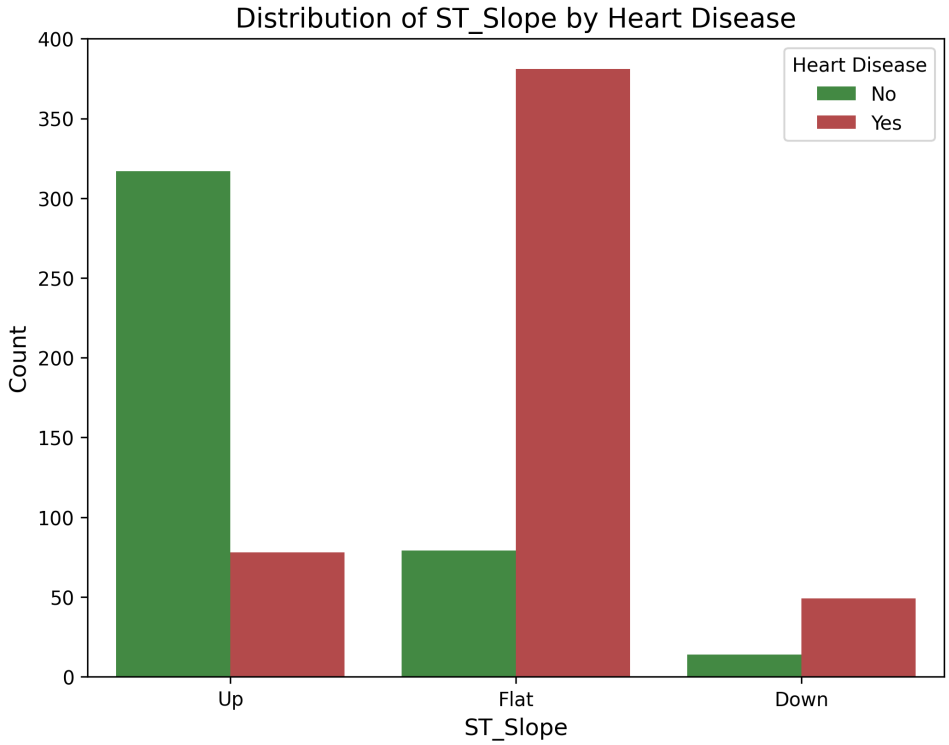

Figure 12: Distribution of ST Slope by Heart Disease.

11.1.2 Clustering & Dimensionality Reduction

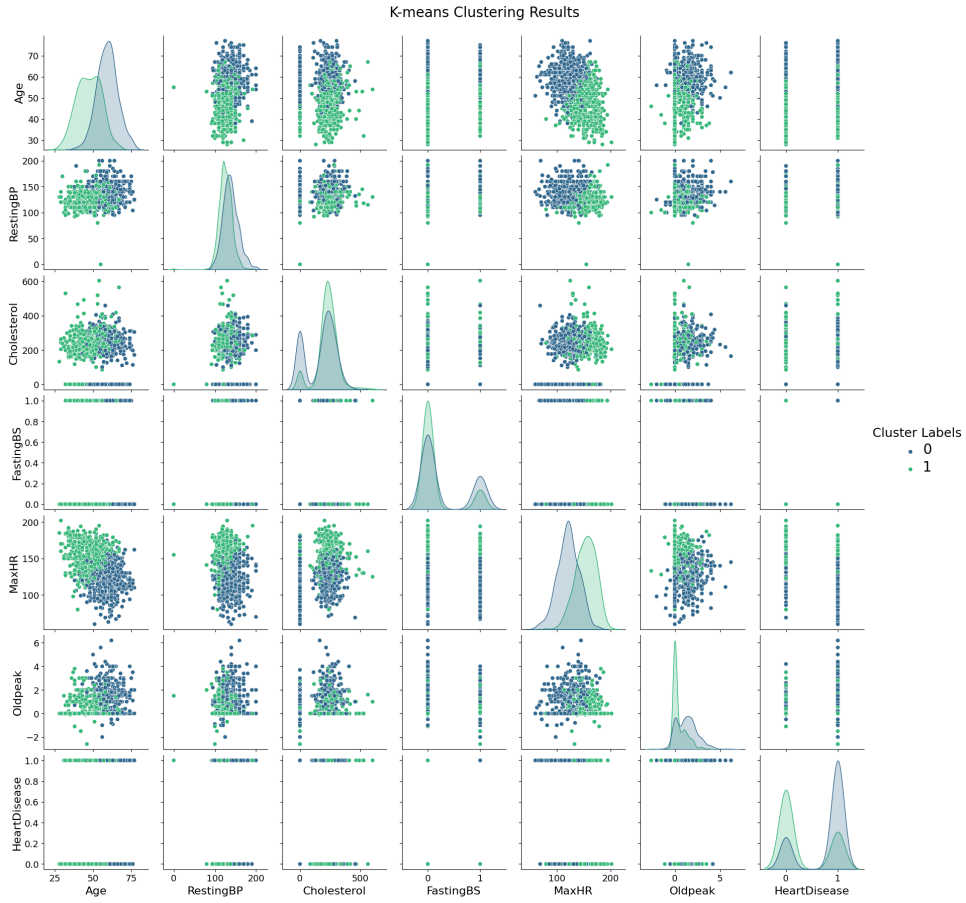

Figure 13: K-means clustering visualization of the dataset.

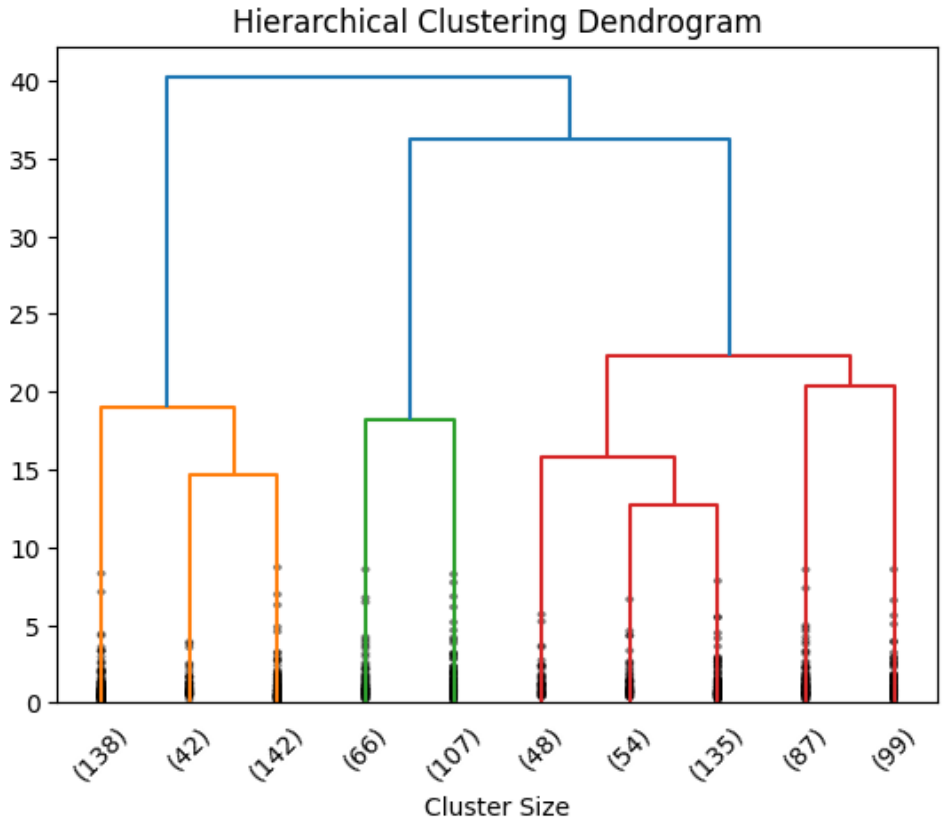

Figure 14: Hierarchical clustering visualization of the dataset.

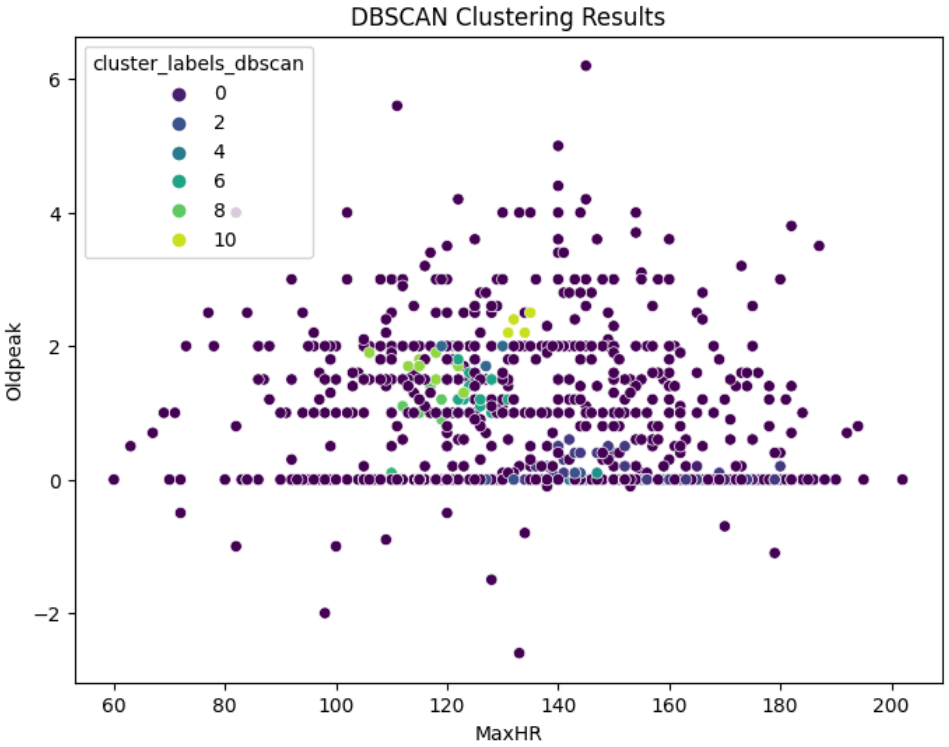

Figure 15: DBSCAN visualization using Oldpeak and MaxHR.

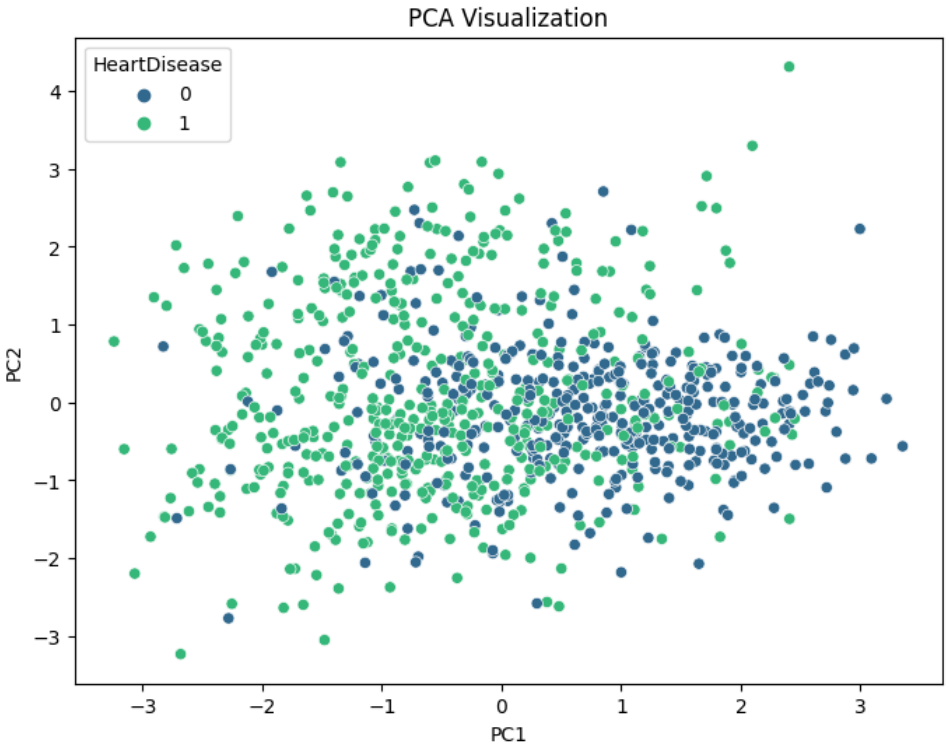

Figure 16: PCA visualization of the dataset.

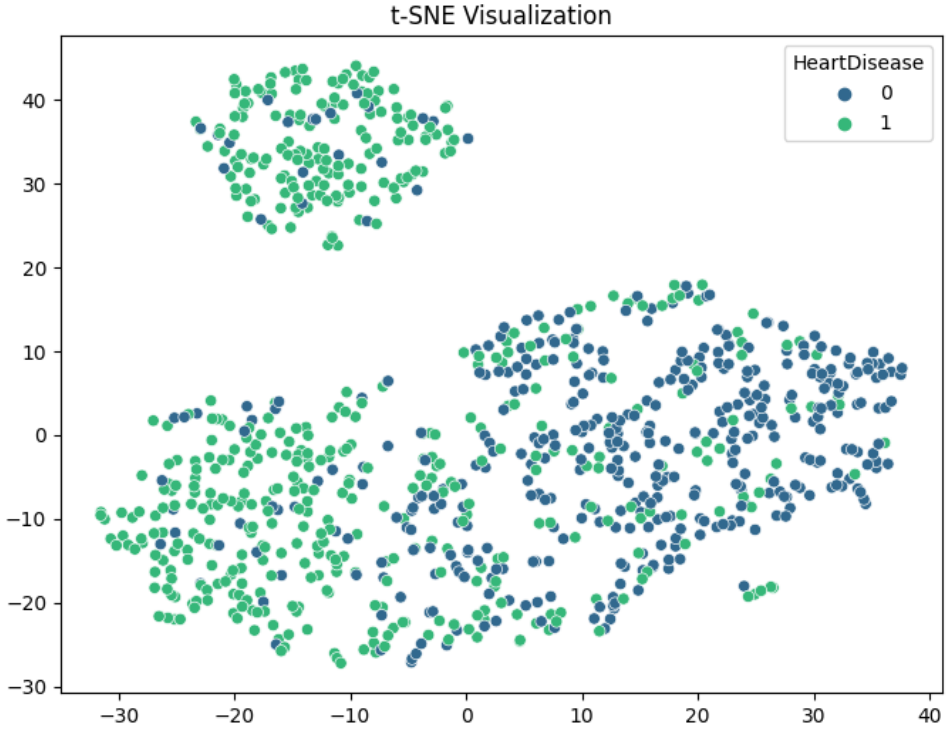

Figure 17: t-SNE visualization of the dataset.

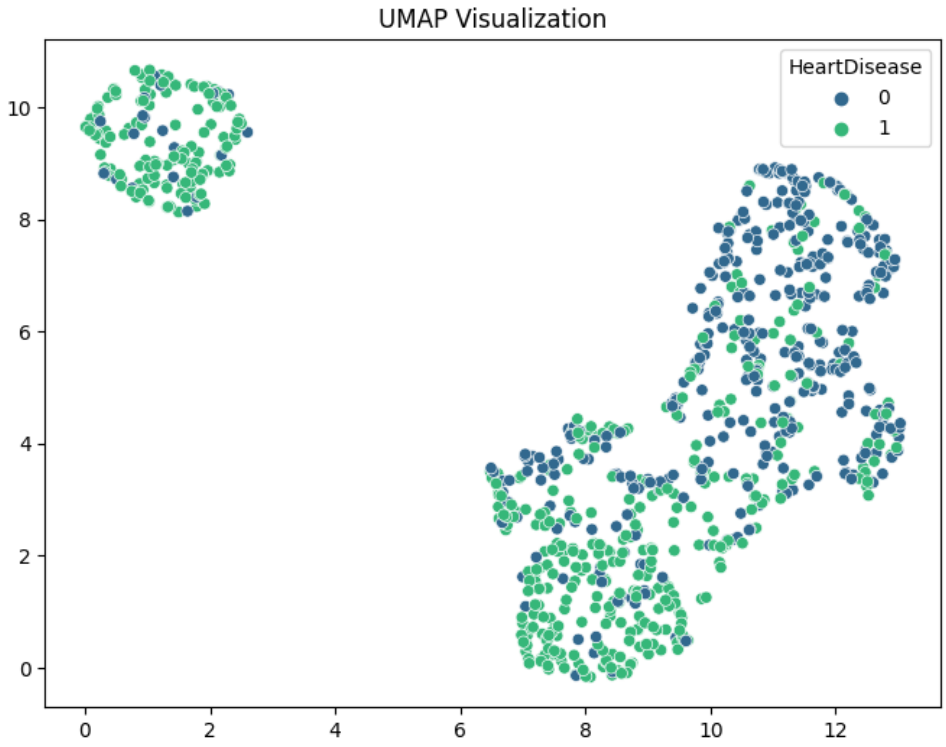

Figure 18: UMAP visualization of the dataset.

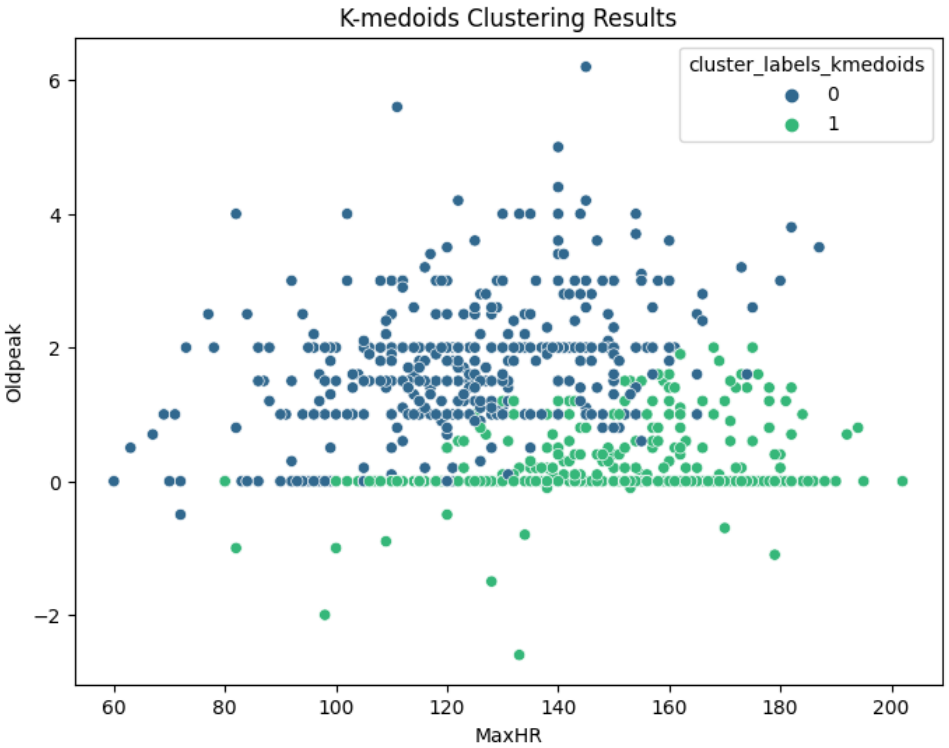

Figure 19: K-medoids clustering visualization using Oldpeak and MaxHR.

11.2 Model Performance

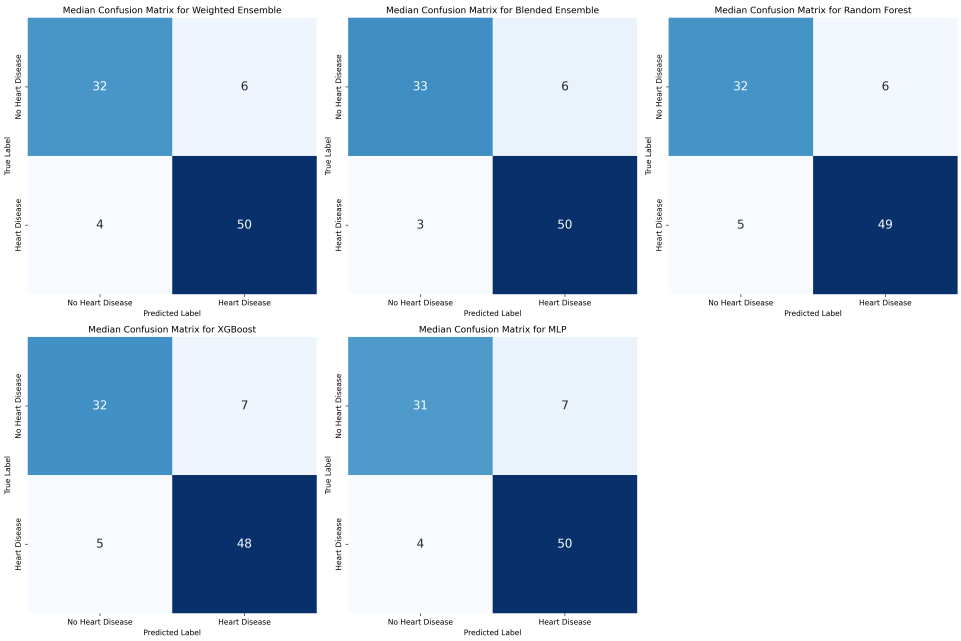

Figure 20: Median confusion matrices across five random states.

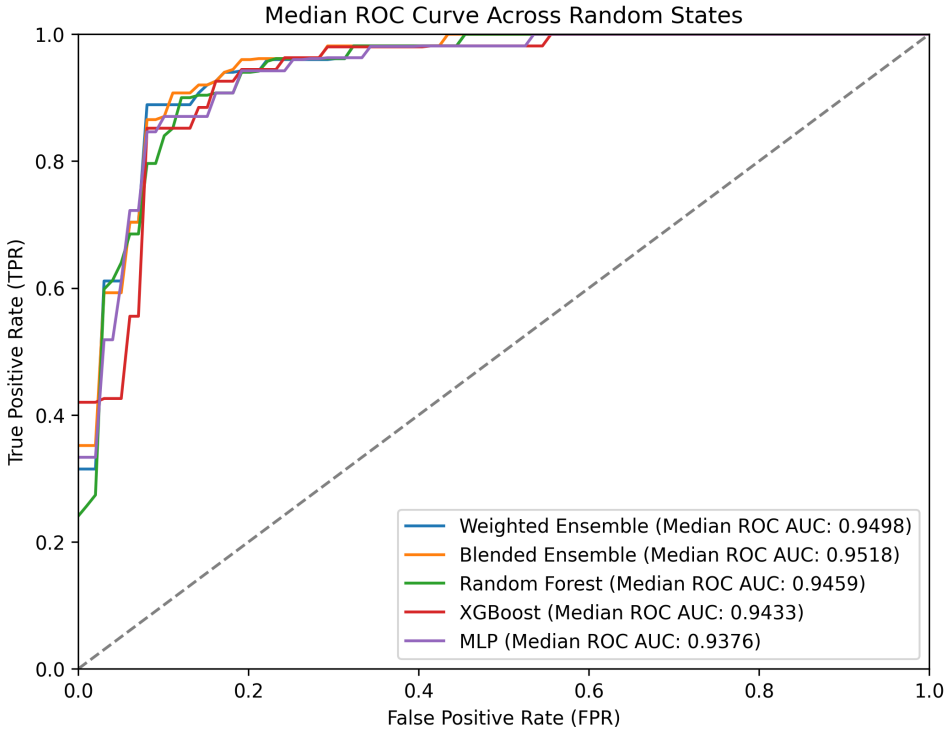

Figure 21: Median ROC curves across five random states.

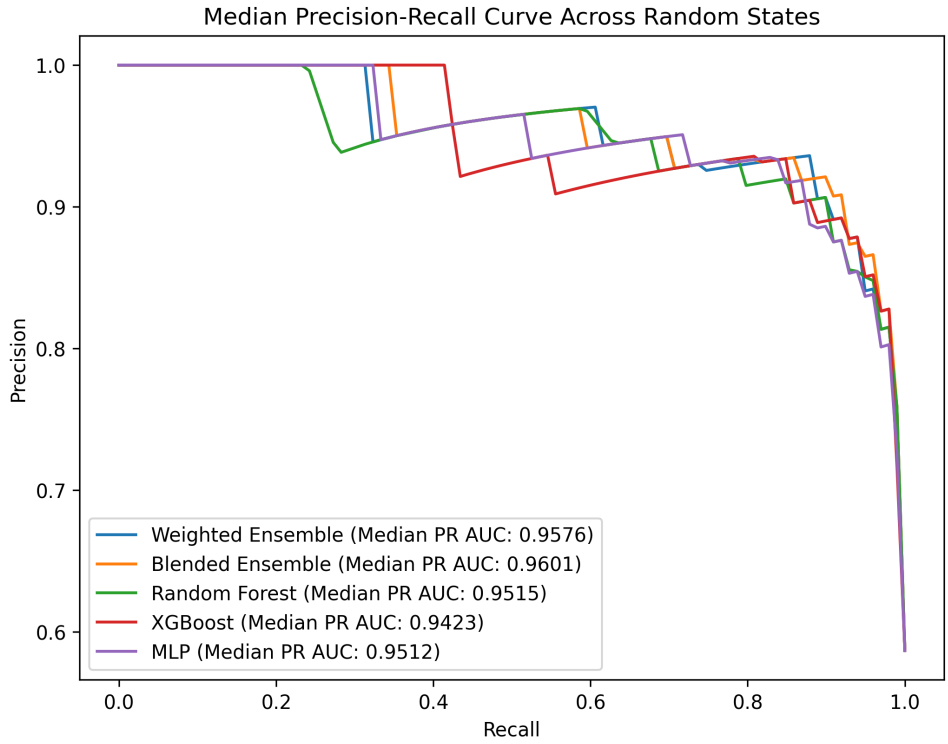

Figure 22: Median Precision-Recall curves across five random states.

11.3 SHAP Feature Importance

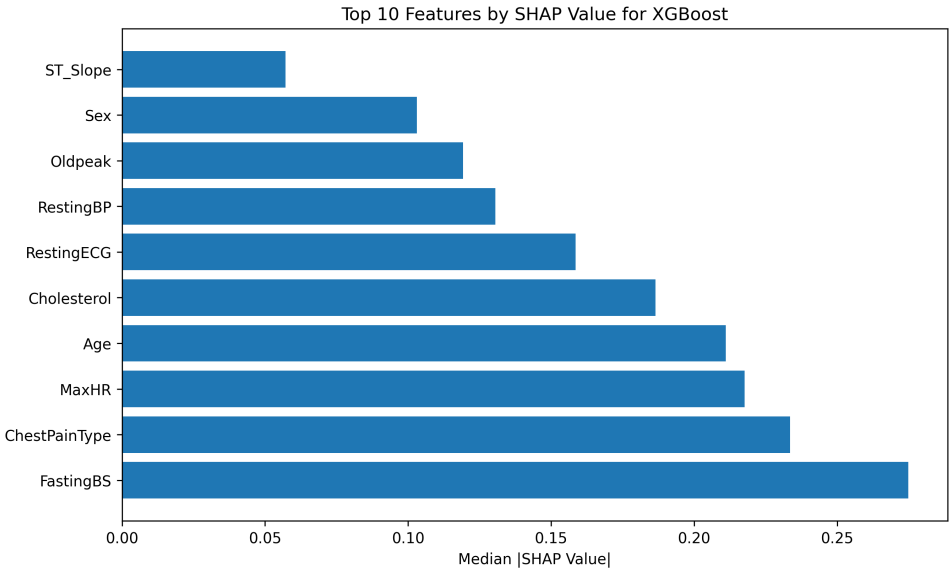

Figure 23: Median XGBoost SHAP values across five random states.

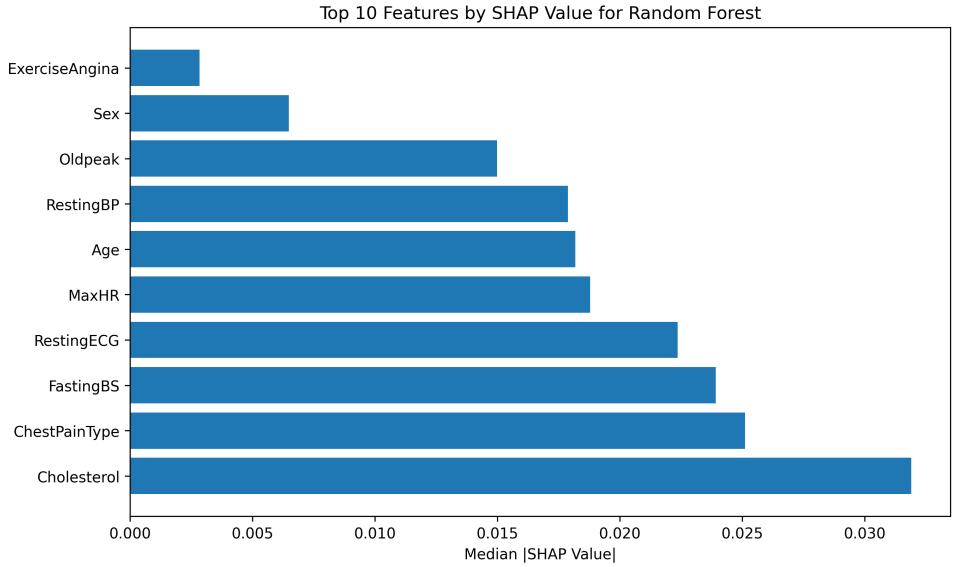

Figure 24: Median Random Forest SHAP values across five random states.

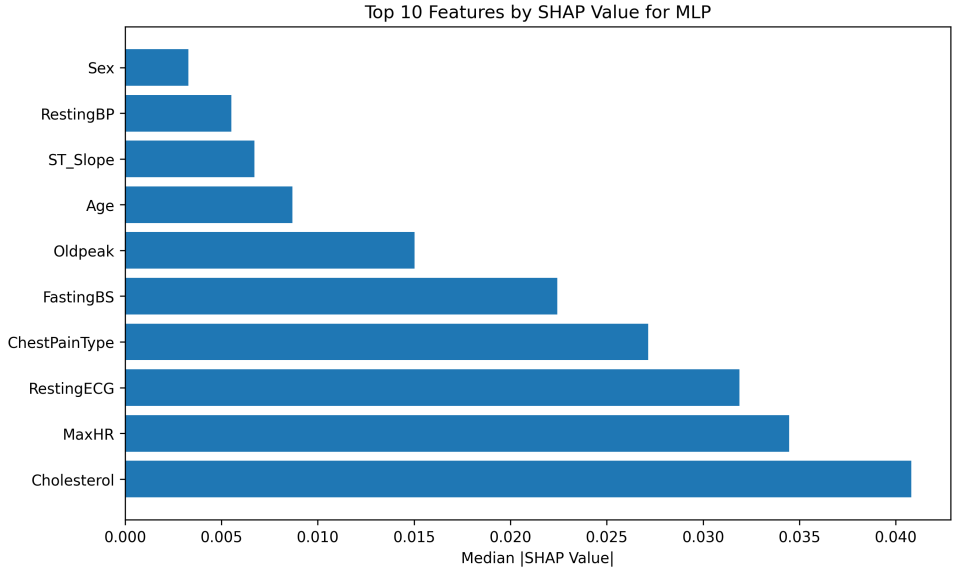

Figure 25: Median MLP SHAP values across five random states.

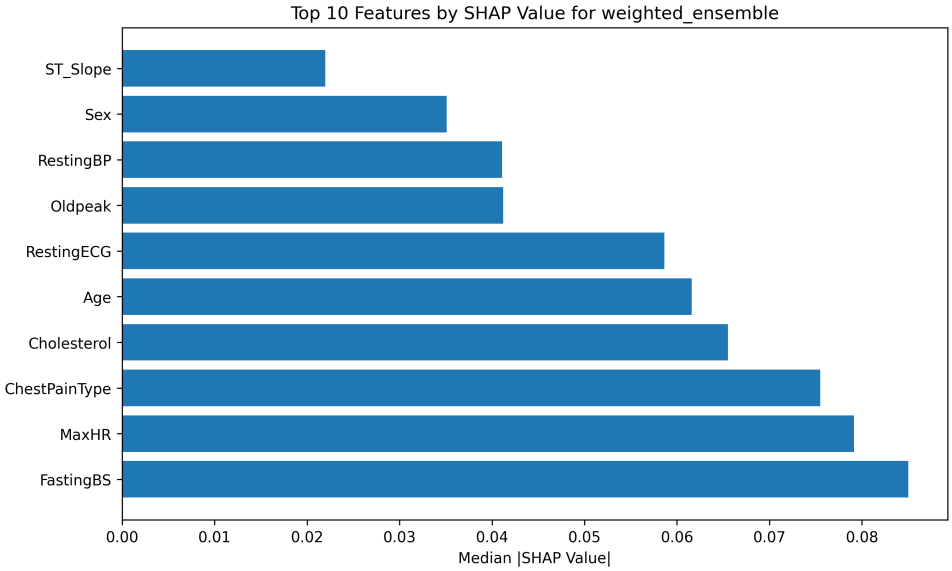

Figure 26: Median Weighted Ensemble SHAP values across five random states.

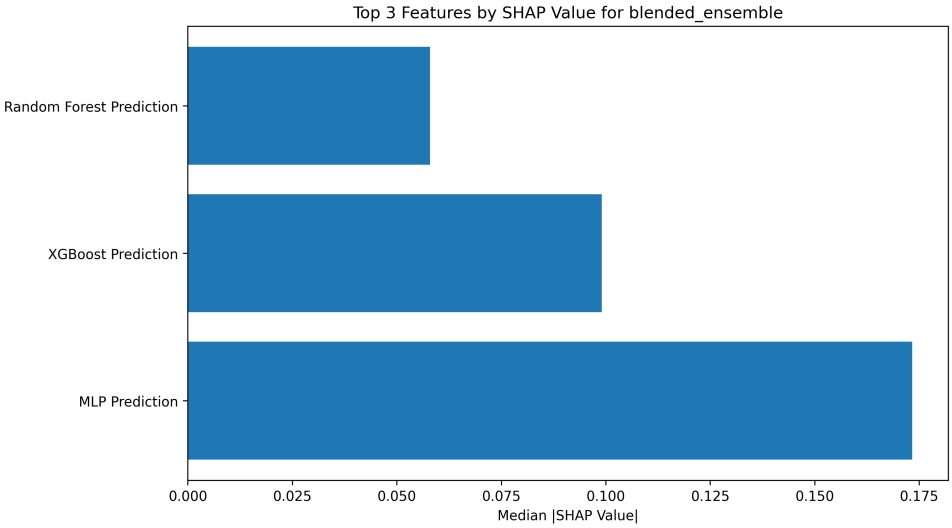

Figure 27: Median Blended Ensemble SHAP values across five random states.

References

[1] W. H. O. (WHO), "Cardiovascular diseases (cvds)," 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] S. M. Nashif and A. Raihan, "Heart disease detection by using machine learning algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 8, pp. 228–234, 2018. [Online]. Available: https://www.semanticscholar.org/paper/Heart-Disease-Detection-by-Using-Machine-Learning-a-Nashif-Raihan/69cc23011f417e6b20cfa631bafd7456521fa2f2

[3] Z. Liu and Y. Wang, "A hybrid classification system for heart disease on imbalanced data," *J. Healthc. Inform. Res.*, vol. 3, no. 4, pp. 412–425, 2019. [Online]. Available: https://www.semanticscholar.org/paper/A-Hybrid-Classification-System-for-Heart-Disease-on-Liu-Wang/63c554b10fe3fa8733b84de0467296ddaf406032

[4] R. Chaki and R. Das, "A comparison of three discrete methods for classification of heart disease data," *Biomed. Signal Process. Control*, vol. 68, p. 102600, 2021. [Online]. Available: https://www.semanticscholar.org/paper/A-comparison-of-three-discrete-methods-for-of-heart-Chaki-Das/c9478126d76c6f6481761c898eb432be3ade9ee5

[5] T. Elwahsh and M. El-Shafeiy, "A new smart healthcare framework for real-time prediction of heart disease," *J. Biomed. Inform.*, vol. 133, p. 104600, 2023. [Online]. Available: https://www.semanticscholar.org/paper/A-new-smart-healthcare-framework-for-real-time-on-Elwahsh-El-Shafeiy/641bfa6caceb4f352325869676e6631cbbf0e5d1

[6] F. Soriano, "Heart failure prediction dataset," 2022. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

[7] H. G. Stratmann and M. F. Wilson, "Fasting glucose level and the risk of incident atherosclerotic cardiovascular diseases," *Am. Heart J.*, 1993. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/23404299/

[8] M. S. Lauer and G. S. Francis, "Influence of the maximum heart rate attained during exercise testing on subsequent heart rate recovery," *Am. Heart J.*, 2010. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/19615487/

[9] D. C. Goff, D. M. Lloyd-Jones, and G. Bennett, "2013 acc/aha guideline on the assessment of cardiovascular risk," *Circulation*, vol. 129, no. 25, 2013. [Online]. Available: https://doi.org/10.1161/01.cir.0000437741.48606.98

[10] S. M. Grundy, "Cholesterol and coronary heart disease: A new era," *JAMA*, vol. 256, no. 20, pp. 2849–2858, 11 1986. [Online]. Available: https://doi.org/10.1001/jama.1986.03380200087027

[11] P. M. Rautaharju, B. Surawicz, and L. S. Gettes, "Aha/accf/hrs recommendations for the standardization and interpretation of the electrocardiogram," *JACC*, vol. 53, no. 11, pp. 982–991, 2009. [Online]. Available: https://www.jacc.org/doi/abs/10.1016/j.jacc.2008.12.014

[12] L. National Heart and B. I. (NHLBI), "Learn the difference: Heart attack, cardiac arrest, and heart failure," 2020. [Online]. Available: https://www.nhlbi.nih.gov/sites/default/files/publications/FactSheetKnowDiffDesign2020V4a.pdf

[13] N. H. S. (NHS), "Heart failure information," 2024. [Online]. Available: https://www.nhs.uk/

[14] P. K. Whelton and R. M. Carey, "Clinical practice guidelines for hypertension management," *Hypertension*, vol. 71, no. 6, pp. 1269–1324, 2018. [Online]. Available: https://doi.org/10.1161/HYP.0000000000000065

[15] T. H. Marwick, "Prediction of mortality in patients without angina: Use of an exercise score and exercise echocardiography," *Eur. Heart J.*, vol. 24, no. 13, pp. 1223–1230, 2003. [Online]. Available: https://doi.org/10.1016/S0195-668X(03)00192-1

[16] P. M. Okin and R. B. Devereux, "Heart rate adjustment of st segment depression for improved detection of coronary artery disease," *N. Engl. J. Med.*, vol. 79, no. 2, pp. 1126–1135, 1989. [Online]. Available: https://doi.org/10.1161/01.CIR.79.2.245